Conjunctive standards in OSCEs: the why and the how of number of stations passed criteria

Abstract

Introduction

Many institutions require candidates to achieve a minimum number of OSCE stations passed (MNSP) in addition to the aggregate pass mark. The stated rationale is usually that this conjunctive standard prevents excessive degrees of compensation across an assessment. However, there is a lack of consideration and discussion of this practice in the medical education literature.

Methods

We consider the motivations for the adoption of the MNSP from the assessment designer perspective, outlining potential concerns about the complexity of what the OSCE is trying to achieve, particularly around the blueprinting process and the limitations of scoring instruments. We also introduce four potential methods for setting an examinee-centred MNSP standard, and highlight briefly the theoretical advantages and disadvantages of these approaches.

Discussion and conclusion

There are psychometric arguments for and against the limiting of compensation in OSCEs, but it is clear that many stakeholders value the application of an MNSP standard. This paper adds to the limited literature on this important topic and notes that current MNSP practices are often problematic in high stakes settings. More empirical work is needed to develop understanding of the impact on pass/fail decision-making of the proposed standard setting methods developed in this paper.

Key words

OSCE; assessment; standard setting.

Introduction

Across a range of performance assessment settings, especially in OSCEs, institutions and regulators often require their candidates to pass a minimum number of stations in addition to achieving the aggregate pass mark (Ben-David 2000; Harden et al. 2015, p. 140; General Medical Council 2019). The stated rationale behind such an additional passing requirement, often referred to as a conjunctive standard (Haladyna & Hess 1999; Ben-David 2000; McKinley & Norcini 2014), is that it prevents excessive degrees of compensation across an assessment. In other words, it serves to prevent candidates passing the OSCE who do well on a few stations, gaining many marks, but who perform poorly on most others. Without such a standard, it is possible, hypothetically at least, to pass the exam having failed the majority of stations. In the medical education literature, however, we find there is very little written about this conjunctive standard setting approach, with a lack of underpinning theoretical argument or justification, and an absence of detail of how one might actually set such a standard in a defensible way.

In this paper, we consider these two questions:

1. Why have conjunctive standards in OSCEs become an accepted norm in many assessment contexts?

   In discussing this question, we suggest an alternative conception to that of limiting compensation in student performance; instead reframing the use of conjunctive standards as a way for assessment writers/designers to compensate for their understandable concern as to the limitations to the quality of decision-making that their OSCEs can afford.

2. How can we set a defensible minimum-number-of-stations-passed (MNSP) standard?

   We introduce and discuss a range of methods for the setting of a *post hoc* (examinee-centred) standard based on a minimum number of stations passed. These are, all to varying degrees, criterion-referenced, thereby ensuring the conjunctive standard adjusts based on the difficulty of the examination rather than being fixed *a priori* as is common practice.

We begin this paper with a brief discussion of why OSCEs have become the accepted method for assessing performance summatively in medical education, and then move on to the consider the issue of compensation and the justification of the need for a conjunctive MNSP standard. An overview of different methods for setting such a standard follows, and the paper concludes with some final thoughts on this under-researched area of medical education assessment.

Summative assessment of performance using the OSCE

Medical schools, in the UK under the inspection of the General Medical Council, must guarantee that their graduate doctors will model the professional attitudes, skills and behaviours that define a doctor (General Medical Council 2017; General Medical Council 2018). Consequently, assessments of individuals' practices are directed efforts, driven by the requirement for assurances to higher authorities about competence in practice that ensures safe patient care. This is also true in many other countries and jurisdictions.

Assessments of ‘performance’ can be considered the *showing* of the “capacity of an individual to successfully handle (according to certain formal or informal criteria, set by oneself or by somebody else) certain situations or complete a certain task or job” (Miller 1990; Eilström & Kock 2008; Boursicot et al. 2011), and comes in several guises. These include workplace based assessments (Norcini & Burch 2007), and various formats of summative examinations. However, the latter, the practice of ‘objectively judging’ the complexities of clinical practice for the purposes of admittance, progression or qualification, is dominated by the OSCE (Norman 2002). Conceived over forty years ago (Harden et al. 1975), OSCEs are widely used in many healthcare professions, and have become synonymous with decision-making about performance in ‘high-stakes’ situations (Miller 1990; Norman 2002; van der Vleuten & Schuwirth 2005; Boursicot et al. 2010; Pell et al. 2010; Khan, Ramachandran, et al. 2013; Khan, Gaunt, et al. 2013).

The OSCE is designed to assess standards of performance in a simulated environment across a series of ‘constructed realities’ known as stations (Harden et al. 2015, chap. 1). However, clinical performance is not a single tangible ‘position’, but consists of a collection of traits deemed desirable by a community of practice and wider society, and so an absolute measurement is always going to be challenging to achieve (Wilkinson et al. 2003; Newble 2004; White et al. 2008; Gormley et al. 2016). Generally, performance in an OSCE station is deemed a success when scoring sufficiently highly via a checklist or domain score. Overall success is generally based on station-level marks aggregated to a total for the exam. The method for deciding the ‘sufficient standard’ varies across institutions (Pell et al. 2010; McKinley & Norcini 2014) but, regardless of the chosen approach, the intended inference remains the same: those who are successful can be defended as having met the minimum acceptable standard.

OSCEs are purported to be a valid and reliable method for assessing whether candidates should be allowed to progress (van der Vleuten & Schuwirth 2005; Boursicot et al. 2010; Pell et al. 2010). However, the nature of OSCE decision-making often allows individuals to pass without necessarily having shown competent practice in *all* aspects of the assessment. Irrespective of the standard setting methodology employed, in a fully compensatory approach, it is acknowledged that candidates are able to compensate areas of deficient practice with an excellent performance in other domains, and still have a successful test outcome (Ben-David 2000; Cizek & Bunch 2007, chap. 2; McKinley & Norcini 2014). However, success in the exam will necessarily be interpreted by many stakeholders as a hallmark of competence in those individuals who 'make the grade'. Given the risk to patient safety of error, it is understandable, therefore, that assessment designers, who are being asked to provide the required assurances about their graduates/candidates, may seek additional assurances as to the quality of those candidates who pass the exam. One way to achieve this is in the form of a conjunctive MNSP standard, in an attempt to circumvent or alleviate any perceived weakness of the test and associated outcomes upon which all stakeholders are so reliant. We develop these arguments in the next section.

OSCE limitations as a motivator for conjunctive standards

As already stated, a conjunctive standard is an additional criterion (to the cut-score) that the candidate must also achieve (Cizek & Bunch 2007, chap. 2). Examples in OSCEs include having to also pass a particular station in order to pass the exam (often called informally a 'killer station' (Schuwirth & van der Vleuten 2006), although this is hard to defend psychometrically since a single station-level decision lacks reliability (Ben-David 2000; McKinley & Norcini 2014). Another example is the addition to the overall cut-score of a standard error of measurement (Hays et al. 2008), based on the idea that this will minimise false positive decisions; that is candidates passing because of measurement error in their favour. In some settings, proficiency has to be demonstrated across a number of domains simultaneously, such as communication and clinical skills. This is the case for United States Medical Licensing Examination (USMLE) Step 2 Clinical Skills examination, where candidates are scored in three separate domains, and must pass each of these in a single administration (Federation of State Medical Boards & National Board of Medical Examiners 2020).

We focus in this paper on the requirement to also pass a minimum number of stations in the examination (McKinley & Norcini 2014). Many institutions and regulators have such a conjunctive standard in their performance assessments. In a rare paper explicitly focussed explicitly on this issue, Clauser and colleagues (1996) find that expert judges are strongly of the opinion that fully compensatory models, where no MNSP standard is applied, are not viewed as appropriate in high stakes clinical settings. However, one could argue that the common practice of pre-determining a fixed MNSP across test administrations (e.g. '*in addition to achieving the aggregate passing score, candidates must also pass 11 out of 16 stations*') lacks defensibility in a world where criterion-based assessment standards are expected to be employed in such a high-stakes setting (Cizek & Bunch 2007, chap. 1).

Given the tension between the apparent lack of defensibility, and yet the desire to satisfy concerns over allowing excessive compensation, the question arises as to why incorporating an MNSP requirement in OSCEs has become an accepted norm in many contexts. In the following sub-sections, we focus respectively on the complexity of the OSCE, the problematic nature of the blueprinting process, and the difficulty of capturing appropriate and sufficient scoring information on candidate performance.

The complexity of the OSCE

Clinical practice is a multi-dimensional, highly complex entity, and designing a performance assessment which robustly tests all domains of practice is therefore very difficult (Khan, Ramachandran, et al. 2013; Khan, Gaunt, et al. 2013). It is acknowledged that despite being used as a high-stakes performance assessment, it can be a challenge to measure successfully in the OSCE crucial capabilities of clinical practice, such as professional values and behaviours, team working and leadership (Khan, Ramachandran, et al. 2013; General Medical Council 2017; General Medical Council 2018). To do this might require innovation and creativity to go beyond assessing the standard clinical tasks of diagnosis and treatment.

In the assessment of students in the senior years of undergraduate medicine, there is a desire to include highly integrated stations in summative OSCEs. This is seen as enhancing validity of the assessment by making it closer to 'real' clinical practice than more junior OSCEs where stations might be focussed on more narrowly defined sub-tasks (Daniels & Pugh 2018). However, in integrated stations, the individual 'constructs' of competence (known as a '*competency'* (Greenaway 2013)) cannot easily be separated, or may account for only a small proportion of requirements of the assessed task overall. Frequently, this is deliberate by design, reflecting the desire of the assessors to differentiate the truly competent practitioners from the 'lesser developed' individuals who are still reliant on explicit cues for clarity on appropriate action (Dreyfus et al. 1988; Ericsson 2004; Carraccio et al. 2008). Whilst the integrated approach to summative OSCE design clearly has its merits, it also means that specific gaps in competence (or proficiencies) may not easily be identified, either by those responsible for the OSCE design or by assessors during scoring. This problem is further compounded by the common lack of checklist/domain score item-level data analysis during production of OSCE results – understandable given the complexity of potential *post hoc* analysis that is recommended in the limited time available before results are ratified and released (Pell et al. 2010; Pell et al. 2015). Assessment designers and other faculty are aware that these challenges could lead to a situation in which false reassurance or overly optimistic conclusions on candidate competency across substantive domains might be drawn. How then can such problems be best avoided or ameliorated?

Blueprinting issues

Whilst doing their best to maximise the robustness of their assessments, OSCE developers have to acknowledge that the complexity of the domain being tested is, to a degree, always beyond the capabilities of the instrument being used. Despite the high stakes nature of the OSCE, the constraints of time and resources, and the limits of psychometric measurement (Schauber et al. 2018), imply that no perfect OSCE can possibly exist. In order to ensure that a pragmatic approach is sufficiently convincing to those looking for guarantees about competence of practitioners, OSCE developers need to confirm that the requisite domains of performance have been tested, and that this has been done as reliably and validly as possible. One way to help offer such assurances is with a well-developed examination blueprinting process (Coderre et al. 2009; Raymond & Grande 2019). This is a method of formally determining the content of any examination and ensuring that this is congruent with outcomes and learning experiences (Coderre et al. 2009; Khan, Ramachandran, et al. 2013). Undertaken prospectively in the preparation and planning stage of the OSCE development cycle (Khan, Gaunt, et al. 2013), blueprinting a summative OSCE aims to ensure an appropriate spread of sampled capabilities are to be assessed. Blueprinting is therefore essential as it assists in the constructive alignment of the task(s) and contributes to evidential support of for the OSCE in terms of validity and authenticity (Messick 1995; Biggs 1996; Downing 2003).

As usually carried out, blueprinting 'macroscopically' evidences included content, meaning that curriculum areas are broadly tabulated against specific stations/sub-tasks in the assessment. However, this process necessarily lacks fine detail. Once the blueprint is produced, the process is often viewed as complete – often before stations are formally chosen or, perhaps, even written. However, we argue that this process is often problematic because it assumes that blueprinting will automatically lead to the creation of valid and appropriate content. Under ideal conditions, once OSCE stations have been authored, practised and amended, the blueprinting process would be repeated. This would provide further confirmation that finalised station content and item scoring instruments properly align with the identified key learning outcomes, and would also allow the identification of other important capabilities or outcomes which have (inadvertently or fortuitously) been included in the test. However, the reality is often that, due to constraints of time and resources, the blueprinting of performance assessments is usually only undertaken in the planning process, and is not revisited after the content has been written and completely assembled. In short, in contexts where such revisiting does not take place, there remains potential for doubt about whether the OSCE truly tests what it was intended to assess at the initial design stage.

Considering the limitations of the blueprinting process in the context of MNSP standards, the inclusion of an additional 'safety net' therefore holds natural appeal. It is attractive to assessment designers because it appears to offer further assurances that, in a successful candidate, a broad range of key capabilities (learning outcomes) have been tested and passed a sufficient number of times, to a 'critical mass'. Hence, the final judgements made of candidates are then thought to be reliable and valid, with the inclusion of an additional degree of 'quality assurance' offered by the designers about their test and its associated decisions. This helps assuage concerns over a potential lack of tangible evidence as to the genuine complexity of content included, the weightings of various components in the assessment, or indeed detailed information of how candidates truly actually perform (e.g. at the item-level).

Scoring issues

Given the complexity of assessing medical practice, it is unsurprising that criticism of OSCEs is widespread for a number of reasons. With regards to the context in which the performance is being assessed, concerns about constructed station checklists being unable to capture the nuances and complexity of clinical practice seem reasonable; as are questions about the artificial (and frequently short) amounts of time that are allocated for demonstration of performance (Hodges et al. 1999; Hodges & McIlroy 2003; Hodges 2013) . Concerns may also be raised about the constraints that simulating reality place upon what content can authentically be assessed in the test (Govaerts et al. 2007; Gormley et al. 2020).

Dichotomous checklist items may be considered to binarily capture the 'mechanics' of clinical practice without recording the more holistic overview of how the performance 'looked' or 'felt' to the examiner and/or simulated patients (Hodges et al. 1999; Ilgen et al. 2015). Consequently, it is not unusual for global ratings to also be included in judgements by examiners, enabling a more complete judgment of the performance to be recorded (Schuwirth & Ash 2013; Wood & Pugh 2019). This inclusion seeks to redress the balance between measures of the objective and subjective elements of the observed performance (Hodges 2013). However, in many contexts, global scores are used primarily only for standard setting purposes (e.g. in borderline regression) rather than for directly judging performance. This presents a quandary on the part of the assessment designers as to how best to guarantee an overriding perception of the adequacy (or otherwise) of the performance in the final summative judgement made based on accumulated scores, not on holistic grades. And so once again, as with the apparent weakness of the blueprinting process discussed above, attention of the designers returns to the use of a 'safety net' of a conjunctive MNSP standard to articulate that a 'critical threshold' has been surpassed.

We suggest, then, that the introduction of an additional hurdle such as a MNSP by assessment designers partially addresses their desire to document an additional and important element of the performance, one that is not automatically reflected in the checklist or domain score alone. Using a conjunctive standard in this way may be perceived as addressing the assessment designers' fears that candidates might have been able to 'cheat' the test - by appearing 'on paper' to have met the itemised-criteria of the station-tasks by 'putting on' a performance that might have felt inauthentic to their examiners (Gormley et al. 2016). Requiring candidates to pass a MNSP (in addition to achieving the cut-score) may therefore serve to further reassure stakeholders that successful candidates have certainly achieved the required standard, and have not been able to simply 'act' their way through the test. The extent to which this reassurance reflects the reality is, in general, an open question, and depends on the quality of the OSCE in question, and on how 'high' the MNSP standard is set. Requiring merely a few stations to be passed in addition to the cut-sore might not be sufficient. We turn to the issue of setting this standard in the next section.

Setting Minimum Number of Stations Passed standards

Having discussed the motivations behind the application of a conjunctive MNSP standard in OSCEs, in this section we outline four possible approaches to actually setting a defensible MNSP, and briefly consider relevant psychometric issues.

Four potential MNSP methods

We describe four *post hoc* methods for setting the MNSP standard appropriate to the borderline OSCE candidate (Cizek & Bunch 2007, p. 48). We assume that each OSCE station is scored using a checklist/domain score and a global grade, and that there is robust method already in place for setting the overall exam-level cut-score – for example, borderline regression (McKinley & Norcini 2014).

We present these methods in a tentative order of increasing defensibility – based on our theoretical judgement of their relative merits. One difficulty here is that in almost all OSCE standard setting, the conception of the borderline candidate is made at the station, not the test, level (McKinley & Norcini 2014; Homer et al. 2017), so this makes the setting of an MNSP standard at the test-level a challenge.

1. *Set percentage of stations required equal to aggregate cut-score percentage*

This is the easiest method to describe and, perhaps because of this, the most intuitively appealing. If the cut-score is, for example, 55% then 55% of stations must be passed. An exam with a higher pass mark therefore requires a higher proportion of stations to be passed, and there is an elegance in the simplicity of this approach. However, it is not immediately obvious why a borderline cut-score performance should translate directly to percentage of stations that must be passed for the borderline candidate.

2. *Identify a 'borderline group' at exam level and calculate their typical number of stations passed*

Use the aggregate cut-score to define a borderline group of candidates based on performance 'close' to this (e.g. within a standard error of measurement). Then take the average (e.g. median) number of stations passed for this group – and this would be the requisite MSNP standard. This has an advantage over the first option in that it is clearly based on a 'borderline' group's performance, defined at the exam level. There is, however, at least one immediate problem with this approach – the somewhat arbitrary definition of the borderline group - how many SEMs should the borderline group include? It is also possible that there are no candidates in this group – although in that case, all overall decisions will presumably be passes or fails.

3. Use logistic regression in each station and aggregate across exam

Use logistic regression to predict pass/fail decisions at the station level based only on global grade performance in the station. Use the model to calculate the probability that a candidate with a borderline grade would pass the station. Aggregate this up to the station level to give the MNSP standard. This method mirrors, to an extent, the borderline regression method (Pell et al. 2010; McKinley & Norcini 2014) by using regression modelling, but rather than predicting scores based on grades in each station, we are predicting the probability of passing the station based on grades, with an ultimate focus on the 'borderline' grade. This produces a theoretically defensible standard, but requires additional calculations at both station and exam level. This method clearly demands some quantitative expertise, and so its feasibility is questionable when such support, and indeed time, is often limited.

4. Regress stations passed on total score

Regress the total number of stations passed for each candidate on their total exam score. Use this regression model to find the predicted value of the number of stations corresponding to the extant exam cut-score - this is the MNSP standard. Figure 1 gives an illustration of this approach, indicating in this example that the MNSP standard would be (with rounding) 10 stations.

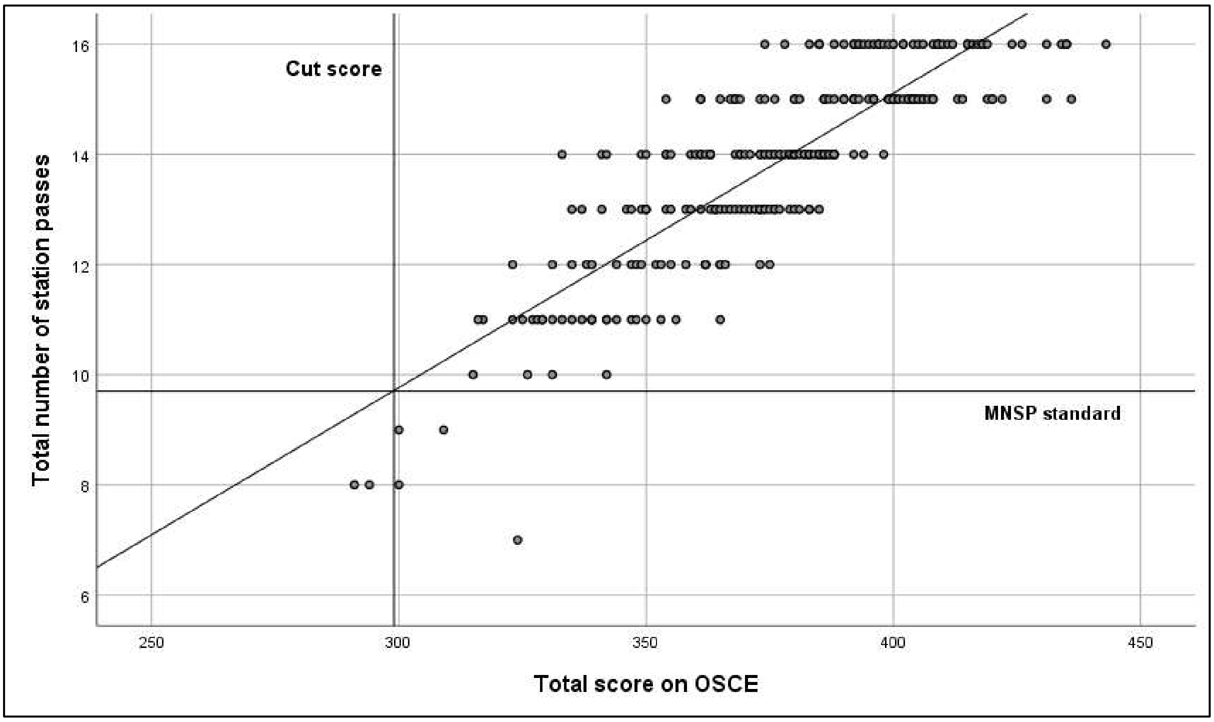

Figure 1: Setting the MNSP by regressing total stations passed on total score

In essence, this method is a subtle development of method 2 in the mathematical sense of being equivalent to shrinking the width of borderline group to zero. This method has some advantages over the others – it is relatively simple to calculate, and it uses the clearly defined and defensible aggregate cut-score. It is also a method working mainly at the exam level – so draws on data at that level (compare with 3). Further, the method itself provides important insight into the relationship between the two exam-level scores – stations passed and total score (e.g. via a scatter plot of these) – this could be useful as part of the evidence when considering assessment quality (Pell et al. 2010).For example, if the relationship between these is relatively weak that might be a cause for concern.

Psychometric considerations

A count of stations passed will necessarily have lower reliability than a total test score – in essence, dichotomising a continuous station score into a pass/fail decision provides a less reliable measure of performance (MacCallum et al. 2002; Kuss 2013). This is one of the reasons that the MNSP is often conceptualised as a secondary hurdle, with the pass/fail decision resting, for most candidates, on their total score relative to the overall cut-score – a more reliable, and therefore more defensible, measure. Ideally, an MNSP standard should therefore only be employed if it is 'of sufficient reliability to add information to the resulting decision' (Clauser et al. 1996). One way to partially manage the lack of reliability is to increase the MNSP standard using the standard error of measurement for the total number of stations passed (Hays et al. 2008), but this does of course increase false negatives, so is not cost free.

In the end, the need for an MNSP-type standard will be stronger in contexts where the traits being assessed are clearly multi-dimensional. In a less complex OSCE, measuring a narrowly defined set of closely related skills, the relationship between total scores and total stations passed will be strong, and any issue of excessive compensation across stations much less important (Clauser et al. 1996; Ben-David 2000). Where domains are sampled more widely, such as in sequential testing (Pell et al. 2013), it maybe that the need for a conjunctive standard is less important. Under such models, borderline candidates are less likely to pass based on strong performance in a small proportion of stations, and the wider sampling is likely to provide additional reassurance to all stakeholders as to the robustness of the decision-making (Homer et al. 2018).

The need for more empirical work

We have deliberately not reported on any empirical work here, partly for reasons of space, but also because any such research is very likely to be highly context specific – so building evidence across a range of settings would be most welcome. There are a number of key questions that need careful consideration before implementing a varying MNSP standard:

- How do the new standards differ from those currently employed in particular contexts?
- To what extent do each of these methods give appropriate standards – do they seem to survive a 'reality check' (Zieky & Perie 2006)?
- Does the standard set by each method (1 to 4) vary systematically in and across contexts?
- Does the fact that the MNSP is largely considered a secondary hurdle fit with the empirical evidence – in other words, do these standards fail only small proportion of candidates relative to the 'main' standard?
- What is the reliability of the total number of stations passed, and how does this compare to the reliability of the total score?
- Is it clear that the MNSP adds meaningful information to the pass/fail decision-making in a defensible way?

Conclusion

In this paper, we have argued that the widespread practice of setting conjunctive MNSP standards in performance assessments like OSCEs is not sufficiently evidence-based, and is under-theorised. Whilst there are psychometric arguments for and against such MNSP standards, it is clear that many stakeholders approve and value their application (Ben-David 2000; General Medical Council 2019). Since defensibility of decision-making in high stakes assessments is an absolute necessity (Cizek & Bunch 2007, chap. 1), we have attempted to add to the limited literature on this important topic in two main ways. Firstly, we have tried to develop better understanding, and provide a more nuanced view, of why MNSP-type standards are appealing to assessment designers and others. We have also tentatively introduced specific examinee-centred methods for setting such standards – to our knowledge no-one has attempted to do this in the past.

The common perception is that the inclusion of MNSP standards serves to prevent excessive compensation by candidates that would otherwise allow false-positive test outcomes. According to this narrative, the additional hurdle is perceived as a necessity in the context of an ongoing emphasis on patient care and the requirement for no compromise on issues of patient safety (General Medical Council 2018). However, we suggest an alternative rationale for the continued support for conjunctive standards - that the MNSP also compensate for a lack of an ideal blueprinting process. This is as a direct consequence of suboptimal assessment practice in a world where busy clinicians 'double-up' as assessment designers and writers, and time and resources are in short supply. In addition, concerns about the adequacy of OSCE scoring instruments also suggest the need for the additional reassurance that an MNSP standard can provide about those passing the exam. One should, however, remember that a single OSCE usually forms a part of a wider programme of assessment (van der Vleuten & Schuwirth 2005) , and that it is generally unrealistic for decision-making to rest on a single instrument.

Our work is largely theoretical at this stage, and additional empirical, psychometric work is needed to answer some of the questions posed earlier regarding how the methods for setting the MNSP play out in practice. Ideally, such analysis would be across a range of different settings to contribute to more universal understanding of this important, but somewhat neglected area of medical education practice. Further theoretical engagement and development is also encouraged.

Practice points

- In many summative OSCEs, achieving a minimum number of stations passed is an additional requirement for progression beyond the cut-score alone.
- There is little extant literature on the theoretical or practical considerations behind the application of such a conjunctive standard.
- Assessment designers employ these standards because of their awareness of the considerable challenges they face in assessing clinical performance as robustly as possible.
- Potential methods for deriving defensive examinee-centred minimum stations passed standards are presented.
- Context-specific additional work is needed to confirm that these methods provide robust and appropriate standards.

Ethics

No data was generated or analysed in this study. Hence, ethical approval was not required.

Acknowledgments

The authors would like to thank David B. Swanson from the American Board of Medical Specialties for discussions on early aspects of some of this work, particularly his initial outline of method 3. They would also like to thank Godfrey Pell for various discussions about the underlying philosophical aspects of this work. Anything erroneous remains, of course, the authors' responsibility.

Declaration of interest

The authors have no declarations of interest to report.

Notes on contributors

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.

Jen Russell, MBChB, MEd, FHEA is a Consultant in Emergency Medicine at Leeds Teaching Hospitals NHS Trust and Honorary Senior Lecturer in the School of Medicine at the University of Leeds. Her research interests centre on developing frameworks for personalising feedback in OSCEs.

Glossary – conjunctive standard

In a summative assessment such as an OSCE, candidates are usually required to achieve an aggregate score in order to pass. If this is the sole passing requirement, then students can (fully) compensate between elements of the assessment – for example, they can fail individual stations and still pass. A conjunctive standard is an additional requirement, such as passing a minimum number of stations or passing separate domains (e.g. management, communication skills, clinical skills) (Ben-David 2000). Careful consideration of impacts on pass rates, and the degree of potential error in decision-making, is needed when implementing such conjunctive standards.

Ben-David, M.F. 2000. AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher*. **22**(2), pp.120–130.

Highlights of paper

We develop understanding of minimum stations passed hurdles in OSCEs, and argue these are motivated by concerns over the limits of what an OSCE can measure robustly. We develop methods for setting such standards, but more research is needed to assess their psychometric properties.

References

Ben-David MF. 2000. AMEE Guide No. 18: Standard setting in student assessment. Med Teach. 22(2):120–130.

Biggs J. 1996. Enhancing teaching through constructive alignment. High Educ. 32(3):347–364.

Boursicot K, Etheridge L, Setna Z, Sturrock A, Ker J, Smee S, Sambandam E. 2011. Performance in assessment: consensus statement and recommendations from the Ottawa conference. Med Teach. 33(5):370–383.

Boursicot KAM, Roberts TE, Burdick WP. 2010. Structured Assessments of Clinical Competence. In: Underst Med Educ [Internet]. [place unknown]: John Wiley & Sons, Ltd; [accessed 2020 Aug 12]; p. 246–258. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444320282.ch17

Carraccio CL, Benson BJ, Nixon LJ, Derstine PL. 2008. From the educational bench to the clinical bedside: translating the Dreyfus developmental model to the learning of clinical skills. Acad Med J Assoc Am Med Coll. 83(8):761–767.

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.

Clauser BE, Clyman SG, Margolis MJ, Ross LP. 1996. Are fully compensatory models appropriate for setting standards on performance assessments of clinical skills? Acad Med J Assoc Am Med Coll. 71(1 Suppl):S90-92.

Coderre S, Woloschuk W, McLaughlin DK. 2009. Twelve tips for blueprinting. Med Teach. 31(4):322–324.

Daniels VJ, Pugh D. 2018. Twelve tips for developing an OSCE that measures what you want. Med Teach. 40(12):1208–1213.

Downing SM. 2003. Validity: on the meaningful interpretation of assessment data. Med Educ. 37(9):830–837.

Dreyfus HL, Dreyfus SE, Athanasiou T. 1988. Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer. Reprinted Ed edition. New York: Free Press.

Eilström P-E, Kock H. 2008. Competence development in the workplace: concepts, strategies and effects. Asia Pac Educ Rev. 9(1):5–20.

Ericsson KA. 2004. Deliberate Practice and the Acquisition and Maintenance of Expert Performance in Medicine and Related Domains. Acad Med. 79(10):S70.

Federation of State Medical Boards, National Board of Medical Examiners. 2020. United States Medical Licensing Examination | Step 2 CS (Clinical Skills). USMLE [Internet]. [accessed 2020 Nov 17]. https://www.usmle.org/step-2-cs/

General Medical Council. 2017. Generic professional capabilities framework [Internet]. London: GMC; [accessed 2020 Aug 12]. https://www.gmc-uk.org/education/standards-guidance-and-curricula/standards-and-outcomes/generic-professional-capabilities-framework

General Medical Council. 2018. Outcomes for graduates [Internet]. London: GMC; [accessed 2020 Aug 12]. https://www.gmc-uk.org/education/standards-guidance-and-curricula/standards-and-outcomes/generic-professional-capabilities-framework

General Medical Council. 2019. Requirements for the Medical Licensing Assessment Clinical and Professional Skills Assessment [Internet]. London: GMC; [accessed 2020 Aug 12]. https://www.gmc-uk.org/education/standards-guidance-and-curricula/standards-and-outcomes/generic-professional-capabilities-framework

Gormley GJ, Hodges BD, McNaughton N, Johnston JL. 2016. The show must go on? Patients, props and pedagogy in the theatre of the OSCE. Med Educ. 50(12):1237–1240.

Gormley GJ, Johnston JL, Cullen KM, Corrigan M. 2020. Scenes, symbols and social roles: raising the curtain on OSCE performances. Perspect Med Educ [Internet]. [accessed 2020 Jun 29]. https://doi.org/10.1007/s40037-020-00593-1

Govaerts MJB, van der Vleuten CPM, Schuwirth LWT, Muijtjens AMM. 2007. Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. Adv Health Sci Educ Theory Pract. 12(2):239–260.

Greenaway D. 2013. Shape of training: Securing the future of excellent patient care [Internet]. London: GMC; [accessed 2020 Aug 12]. https://www.gmc-uk.org/education/standards-guidance-and-curricula/standards-and-outcomes/generic-professional-capabilities-framework

Haladyna T, Hess R. 1999. An Evaluation of Conjunctive and Compensatory Standard-Setting Strategies for Test Decisions. Educ Assess. 6(2):129–153.

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.

Hays R, Gupta TS, Veitch J. 2008. The practical value of the standard error of measurement in borderline pass/fail decisions. Med Educ. 42(8):810–815.

Hodges B. 2013. Assessment in the post-psychometric era: learning to love the subjective and collective. Med Teach. 35(7):564–568.

Hodges B, McIlroy JH. 2003. Analytic global OSCE ratings are sensitive to level of training. Med Educ. 37(11):1012–1016.

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. 2018. The benefits of sequential testing: Improved diagnostic accuracy and better outcomes for failing students. Med Teach. 40(3):275–284.

Homer M, Pell G, Fuller R. 2017. Problematizing the concept of the "borderline" group in performance assessments. Med Teach. 39(5):469–475.

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.

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):e1447-1463.

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.

Kuss O. 2013. The danger of dichotomizing continuous variables: A visualization. Teach Stat. 35(2):78–79.

MacCallum RC, Zhang S, Preacher KJ, Rucker DD. 2002. On the practice of dichotomization of quantitative variables. Psychol Methods. 7(1):19–40.

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

Messick S. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol. 50(9):741–749.

Miller GE. 1990. The assessment of clinical skills/competence/performance. Acad Med J Assoc Am Med Coll. 65(9 Suppl):S63-67.

Newble D. 2004. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ. 38(2):199–203.

Norcini J, Burch V. 2007. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Med Teach. 29(9):855–871.

Norman G. 2002. The long case versus objective structured clinical examinations. BMJ. 324(7340):748–749.

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.

Raymond MR, Grande JP. 2019. A practical guide to test blueprinting. Med Teach. 0(0):1–8.

Schauber SK, Hecht M, Nouns ZM. 2018. Why assessment in medical education needs a solid foundation in modern test theory. Adv Health Sci Educ Theory Pract. 23(1):217–232.

Schuwirth L, Ash J. 2013. Assessing tomorrow's learners: In competency-based education only a radically different holistic method of assessment will work. Six things we could forget. Med Teach. 35(7):555–559.

Schuwirth LWT, van der Vleuten CP. 2006. A plea for new psychometric models in educational assessment. Med Educ. 40(4):296–300.

van der Vleuten CPM, Schuwirth LWT. 2005. Assessing professional competence: from methods to programmes. Med Educ. 39(3):309–317.

White CB, Ross PT, Haftel HM. 2008. Assessing the Assessment: Are Senior Summative OSCEs Measuring Advanced Knowledge, Skills, and Attitudes? Acad Med. 83(12):1191–1195.

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.

Zieky M, Perie M. 2006. A Primer on Setting Cut Scores on Tests of Educational Achievement. Princeton: Educational Testing Service.