Assessors' Engagement with Video-Recorded Performance Assessments in Nursing: A Qualitative Study

Authors:

Conor Scully, PhD.
Dublin City University

Prof. Michael O'Leary, PhD.
Dublin City University

Zita Lysaght, EdD.
Dublin City University

Mary Kelly, PhD.
Dublin City University

ABSTRACT

The use of video in performance assessments has accelerated as a result of the COVID-19 pandemic. Research has tended to focus on the administrative and cost implications of setting up remote assessments: however, few studies have explored how the process of assessing a performance is altered when the assessment is conducted through video. This qualitative paper reports on findings from a simulated video assessment in nursing. As part of the study, 12 nursing assessors watched the same four videos of undergraduate nurses performing either a blood pressure measurement or a naso-gastric tube insertion. They were asked to "think aloud" while doing so, and were also subject to an interview about their assessment practices. Findings revealed that the use of video allows assessors to reduce guesswork when assessing, yet it also limits their field of vision, and in some cases harms the perceived validity of the assessment.

Correspondence E-mail: conor.scully@dcu.ie

Keywords: Video-based performance assessment; Objective Structured Clinical Examination (OSCE); Assessor cognition; Inter-rater reliability in nursing education; Remote clinical skills evaluation

This paper explores how assessors engage with the task of evaluating student performance in a simulated video Objective Structured Clinical Examination (OSCE), in which student nurses were recorded—and graded—ompleting two tasks: blood pressure measurement and naso-gastric tube insertion. The OSCE is a type of performance assessment, an assessment modality in which test-takers have to "construct an answer, produce a product, or perform an activity" (Darling-Hammond & Adamson, 2010, p. 7). In contrast with other assessments such as multiple-choice examinations, performance assessments are distinguished by their closeness to real-world situations that test-takers may encounter (Darling-Hammond & Adamson, 2010).

In nursing, as well as the health sciences more broadly, performance assessments such as the OSCE have been used to evaluate student performances at a range of practical skills that they will go on to use in the "real world" of clinical practice (Rushforth, 2007; Khan et al. 2013). The OSCE was designed due to deficiencies in more traditional assessment formats—notably the short and long cases—that were popular in medicine, particularly the perceived autonomy of assessors to grade students according to personalised or subjective criteria (Harden et al., 1975; Khan et al. 2013). The use of standardised marking tools in the OSCE, as well as the fact that all students who complete the assessment have to perform the same skills, were implemented to ensure that students are graded in a manner that is consistent (Khan et al. 2013).

In common with all assessments, those administering OSCEs need to build a validity argument about the inferences that are to be made on the basis of OSCE scores and provide evidence that these inferences are justified (Fraenkel & Wallen, 2006; Khan et al., 2013; AERA et al. 2014). Two sources of evidence are usually considered important for the validity argument associated with OSCE scores:

1. <u>Fidelity to the real world</u>: OSCEs entail a simulation of a situation that test-takers can be expected to face when they enter the world of clinical practice. As such, they are perceived to entail a high level of real-world fidelity. In general, the closer the approximation between the environment of the test and the real-world situation it is simulating, the stronger the associated validity argument regarding the test scores will be (Darling-Hammond & Adamson, 2010; Eva & Hodges, 2012; Hodges, 2013).
2. <u>Inter-rater reliability (IRR)</u>: OSCEs usually require the use of a range of assessors to grade student performances. It is important that these assessors interpret the performances in the same way, such that students receive the same score regardless of who is assessing them (Khan et al. 2013; Gwet, 2014). If there was a notable variation within a sample of assessors, it would be less defensible to make an inference about a test-taker on the basis of their score, as they may have received a different score had they been assessed by a different assessor. The question of whether OSCE scores do in fact demonstrate high levels of IRR has received significant attention in research literature in both medicine (e.g., Brannick et al., 2011) and nursing (Rushforth, 2007; Navas-Ferrar et al., 2017; Goh et al., 2019).

In part due to these two factors, OSCEs have become a popular assessment format across the Western world, have expanded beyond their original implementation in medicine, and are commonly used in nursing (Rushforth, 2007; Patrício et al., 2013).

In the wake of the COVID-19 pandemic, universities across the world were forced to suspend in-person teaching and implement new or altered assessment modalities. In a traditional OSCE, test-takers, assessors, and—in some cases—Standardised Patients are required to be in the same room (Khan et al. 2013). As such, the onset of the pandemic meant that educators were forced to come up with alternative ways of conducting assessments of students' clinical skills (Lara et al., 2020). Some researchers have described the administration of fully remote OSCEs, in which the OSCE is conducted "live", via an online platform such as Zoom or Microsoft Teams (e.g., Hopwood et al., 2020; Lara et al., 2020). Others have discussed the possibility of allowing students to record their own videos and upload them to an online platform for grading by assessors, a scenario that had been described elsewhere in the literature prior to the pandemic (Purpora & Prion, 2018).

In all cases, the role of the assessor shifts from one where they are able to observe a student's performance in-person, and in real time, to one where their grading of a performance takes place through video. The issue of how this affects the process of judging and grading performances is central to this paper, which addresses the following research question: *How does the use of video in performance assessments in nursing affect assessors' engagement with the task of assessment?*

This issue has implications for the validity argument associated with test scores, as evidence that assessors use meaningfully different strategies when assessing through video compared with in-person assessments may affect the IRR of test scores and may reduce the comparability of test scores across different administrations of the assessment.

Literature Review

Assessor cognition

The issue of what specifically happens when examiners are given the task of assessing student performance at clinical tasks is one that has received significant attention in the medical assessment literature, particularly in the last decade and a half. This emergent field, known as "assessor cognition" or "rater cognition", broadly seeks to understand how assessors "interpret and construct their own personal reality of the assessment context" (Govaerts & van der Vleuten, 2013, p. 1169). This area of inquiry is underpinned by the recognition that because all assessors are individuals, it is likely that they bring with them their own ideas and perspectives that may influence how they engage with the task of assessment (Eva & Hodges, 2012).

Researchers in the field of assessor cognition have used methods such as interviews and "think alouds" in order to develop models that map assessors' thoughts as they watch and grade student performances (e.g., Kogan et al., 2011; Yeates et al., 2013; Roberts et al., 2020). These methods have revealed numerous ways in which assessors may diverge in their approach to the assessment task. For instance, some assessors may be prone to making inferences about a student beyond what is directly visible to them when the student is completing an assigned task, or might compare a student's performance with performances graded immediately before (Kogan et al., 2011; Yeates et al., 2015; St-Onge et al., 2016). Gauthier et al. (2016) conducted a review of assessor cognition literature and developed a model for understanding what happens when assessors watch and interpret student performances: their model conceptualises judgement formation as a three-stage process of observation, processing, and integration.

Perhaps unsurprisingly given the focus inherent in assessor cognition research on divergences in judgement formation, studies on assessor cognition have often taken an inter-rater reliability (IRR) perspective. Seen this way, differences in how assessors approach the task of assessment may result in divergences in the scores they award, even when they are watching the same performance (Gingerich et al., 2014a). This poses a threat to the IRR of assessment scores and may ultimately undermine the validity of decisions made on the basis of these scores (AERA et al., 2014). Multiple studies have attempted to measure the amount of score variance that can be attributed to divergences in assessors' patterns of judgement (e.g., Gingerich et al., 2014b; Roberts et al. 2020). For example, a study by Gingerich et al. (2017) found that assessors' varied cognitive processes accounted for between 21% and 53% of score variance in scores awarded to students in a recorded performance assessment.

It is notable that the vast majority of published studies on the subject of assessor cognition have taken place in the field of medicine, rather than other healthcare disciplines. This is in spite of the documented popularity of performance assessments (especially the OSCE) in other fields, particularly nursing (Rushforth, 2007; Navas-Ferrar et al., 2017; Goh et al., 2019). An exception to this is a 2014 study by East et al., in which nursing assessors were interviewed about their assessment practices, and a recent study (Scully et al., 2024) which mapped nursing assessors' cognitive processes. The issue of how nursing assessors specifically engage with the task of assessment is one that is under-researched.

Use of Video in Performance Assessments

Video has been incorporated into performance assessments for many years in various different ways. For example, numerous studies have described the creation of video exemplars, the purpose of which is to demonstrate optimum performance at an assessed task (or series of tasks) that test-takers will be instructed to complete (e.g., Barratt, 2010; Massey et al., 2017). Researchers have used the recording of performances as a means of providing formative feedback to test-takers, who are able to understand comments from examiners by looking at a video of their own performance (e.g., Paul, 2010). Researchers have also described studies in which students had recorded themselves completing specific tasks and uploaded these videos to an online platform for grading by assessors, an approach deemed to be beneficial for those participating in distance learning (Purpora & Prion, 2018).

Since the COVID-19 pandemic, numerous studies have described remote performance assessments which take place "live" over Zoom or other online platforms. (e.g., Hopwood et al., 2020; Lara et al., 2020; Major et al., 2020). These assessments were usually developed and administered as a result of the prohibition on in-person teaching and learning that was implemented during the pandemic (Felthun et al., 2021). As of 2024, enough studies describing the development of these remote assessments (usually, though not always, classified as OSCEs) have been published so as to allow for several review pieces. These papers have the aim of drawing conclusions about the feasibility of conducting live performance assessments, and the potential barriers to their effective implementation.

Felthun et al. (2021) reviewed 11 studies of video-based performance assessments (which they broadly label "teleOSCEs") administered during the pandemic. Of these 11, seven entailed a “live” examination, while four involved the submission of student-produced videos, which were uploaded to a portal for grading by assessors. More recently, Giri and Stewart (2023) reviewed 28 studies describing the use of video in performance assessments in medicine, nursing, and dentistry, again finding a mix of studies which described “live” assessments and those involving student-produced videos. Both reviews found that performance assessments conducted through video were perceived by stakeholders as being viable, particularly when there is a lack of in-person assessment possibilities.

However, in both reviews, the authors described a lack of insight into how assessors engage with the task of assessment when an examination is conducted through video as opposed to in-person. As noted by Felthun et al. (2021, p. 4), their review “revealed little about whether examiners extract different information about student performance from teleOSCEs and in-person assessments”. They mention that further research into the use of video in performance assessments should “focus on how the online platform impacts… examiners’ judgments” (p. 4). This lack of information might have implications for the reliability of assessment scores, especially if test administrators need to compare scores from a remote administration of the assessment with previously determined scores from an in-person administration of the same assessment. As discussed by Giri and Stewart (2023, p. 14): “remote assessment of practical skills should be interpreted with caution because of a lack of correlation between the assessment scores of the face-to-face examiner and remote examiner”. As such, while researchers have published many studies describing the administration of remote performance assessments, these studies have generally failed to examine in detail the potential changes in assessors’ cognition that may take place when an assessment is conducted through video, and the resultant reliability implications.

The present study, therefore, sits at the intersection of two areas of inquiry: research into assessors’ cognitive processes and research into the use of video in performance assessments. This paper is among the first to investigate the specifics of how assessors approach the task of assessment when the assessment takes place through video, and whether there are observable differences in their approach vis-a-vis in-person assessment.

Methods

This qualitative paper reports on findings from a larger study that had the aim of exploring assessors’ cognitive processes as they watched and discussed the same OSCE performances (Scully et al., 2024) As part of the study, the first and fourth authors filmed six videos of three students each completing two OSCEs: blood pressure measurement (BP) and naso-gastric tube insertion (NG). The students were at different stages in completing their General Nursing programme (one student from years 1, 2 and 3, respectively) which offered the opportunity for some divergence in performance levels to be assessed. Four of these videos recorded were used in the study: the first-year student completing the NG OSCE (P01NG), the second-year student completing the BP OSCE (P02BP) and both third-year student videos (P03BP and P03NG). The BP OSCE was performed on a real person, while the NG OSCE was performed on a mannequin. The videos were recorded using UniCam, with a camera and microphone embedded into the ceiling that the researchers could operate with their phones.

Having completed the video recording, 12 assessors were recruited —using convenience sampling methods —to participate in the study. All 12 were employed in the same university nursing department, either as lecturers or clinical skills nurses (who have the responsibility of teaching clinical skills to nursing students). All participants had experience of assessing undergraduate nursing OSCEs, with nine of the 12 having over five years of experience. When asked to rate their proficiency as assessors, one selected *Advanced Beginner,* three selected *Competent,* six selected *Proficient*, and two selected *Expert* (Benner, 1982).

The 12 participants engaged in a one-to-one, semi-structured interview with the first author, to discuss how they perceived their roles as assessors, and the processes they employed when making judgements about students' OSCE performances (East et al., 2014). Additionally, each assessor participated in a cognitive interview (Ericsson & Simon, 1980; Willis, 2015), which provided insight into their interpretations of the two OSCE marking guides used in the study.

In the final stage of the process, each assessor watched the four recorded videos and, using a think-aloud protocol, shared their opinions as to how well or badly the students were performing. "Thinking aloud" is a technique that has been used in numerous studies of assessors' cognitive processes (e.g., Kogan et al., 2011; Yeates et al., 2013; St-Onge et al., 2016), and provides a rich insight into how assessors engage with the task of assessment. During this section of the study, assessors were given autonomy to pause and rewind the video as they wished. When this happened, the first author probed them by asking why they had done so.

Qualitative data from the study were analysed according to the principles of thematic analysis, a six-step process of *familiarisation, coding, theme search, theme review, theme refinement* and *write-up* (Braun & Clarke, 2006; 2021). Coding largely focused on the semantic meaning of the words, however in some instances latent codes were identified that indicated participants' views about video assessment. The data allowed for a range of insights into the cognitive processes of the assessors who participated in the study (reported more widely in Scully et al., 2024). This paper reports on aspects of the data that relate to how assessors engage with the task of assessment in ways that are specific to the video format, and how these processes of engagement differ compared to when the OSCE is administered in-person.

Results

Analysis of the data revealed three ways in which the use of video affects assessors' engagement with the assessment task: *reducing guesswork, obstructed vision,* and *reduction in perceived validity*. These factors are discussed in turn below, augmented with illustrative quotes from participants.

Reducing guesswork

When assessments happen in real life, events unfurl quite quickly and assessors have to make snap decisions as to what took place in front of them. When video is used — as in the present study —assessors have the opportunity to pause or rewind the video as needed, in order to be sure that they are correctly observing a student's performance and, therefore, making the correct decision about how well that student executed the required tasks. Ten of the 12 assessors in the sample spoke about this phenomenon as being a key benefit of video OSCEs. These assessors noted that the use of video allowed them to make decisions with more confidence, and therefore rely less on instinct or guessing, which they admitted had happened occasionally when OSCEs were administered and assessed live:

> This just reminded me of when I'm doing it, that when you're doing so many of them in real life, sometimes you're going "Did she do that?" and that's where the video is helpful because you can actually stop it and look back, and you can't do that in real life.
> - Assessor 10

> I would probably watch the whole thing through and then I'd go back. I might highlight something that might pop up, that I need to look a little bit more carefully. And then I would go back and I might stop it a few times if I'm not sure, especially with the first few until I become familiar with exactly what the student is doing.
> - Assessor 1

For these assessors, the use of video technology within the OSCE allows them to view the same performance multiple times to ensure they have not missed anything and can confidently award a grade to the student. As such, the use of technology affected their process of coming to a judgement about a student. Seen this way, using video technology should improve score reliability, as assessors would be less reliant on guesswork or short-term memory when judging a student.

Shielded vision

When assessment takes place through video, assessors are unable to move around in order to see something with more clarity, as they would in real life. As a result, their vision is bounded by what the students have filmed, and they may be unable to see specific parts of the procedure that the student is performing. Although students are usually given specific instructions to make all aspects of their performance visible on the video, the reality expressed by participants in this study is that this was not always the case. Assessors mentioned that, when assessing OSCEs in the past they have had to guess whether a student had completed a specific step on the guide:

> Sometimes it can be due to a camera angle as well, and the quality of the video sometimes isn't that good. So you are kind of guessing, “did they or didn't they?”
> - Assessor 11

This phenomenon was reported to be more pronounced for minute or intricate tasks such as —in the case of this study —locating the brachial artery. Assessors discussed how it was impossible to tell through video if the student had correctly located the artery on the patient. As a result, they had to choose whether or not to either give the benefit or the doubt to students regarding their completion of this step on the marking guide. In this sample, different assessors reported different strategies for what to do when something was not clearly visible to them: some would fail to award a mark for an item that they could not see clearly, while others would award the mark. The implications in terms of IRR are perhaps obvious, with the clear possibility that the same performance would be graded differently depending on the assessor. Indeed, for one of the recorded performances in this study, there was widespread disagreement within the sample of 12 assessors as to whether the student had correctly located the brachial artery during a blood pressure measurement.

Some assessor participants noted that in spite of the potential limitations of the video format in terms of allowing them to observe all aspects of a performance, the students could make up for this by narrating what they were doing. In this way, students could still communicate to assessors what they were doing, even if the video could not show it in detail:

> With the patients [in the NG OSCE], sometimes you can get curling of the tubes at the back of the throat and so forth. You might not see that all the time with the mannequin, it might just have one way down and it just goes down. So that can be a little bit difficult to see. But if they're vocalising that to you, they may say, "well, when I'm sliding it in, I want it to go upwards and backwards, inwards and backwards", and then they know the basis of it.
>
> - Assessor 9

However, the potential for narration to make up for the lack of detail in the recorded performance is not one that was discussed by all assessors in the sample, which again indicates differences in how assessors engage with the task of assessment and has potential implications for IRR.

Reduction in perceived validity

As noted at the beginning of this paper, performance assessments are perceived to be effective in part because of fidelity to the real world. Three of the assessors in the current study discussed how the use of self-recorded videos affected their perception of the assessment. These assessors emphasised that when OSCEs are administered through video, students can record their attempt at the skill multiple times; as such, the disconnect between the "real world" of clinical practice, where a nurse may only have one attempt at performing a procedure, and the testing environment, is increased. For these assessor participants, the use of video may decrease the objectivity of the OSCE:

> But of course, if it's a physical assessment, you will probably have gotten more information... And I would have been able to assess more objectively because it's real time. And whatever mistakes they're making, they're making it in real time... Whatever thing they're doing, it's real time. And it gives you more objectivity for sure.
>
> - Assessor 2

A reduction in the perceived validity of the assessment on the part of the assessors has implications for how they engage with the task of assessment. The assessors who discussed how they preferred in-person assessment were more likely to note that they intentionally deviated from the marking scheme in order to reward students that they perceived to be competent.

Discussion

The present study sought to determine whether there are differences in assessors' cognitive processes when an assessment takes place through video, rather than in-person. In order to address this question, 12 assessors of undergraduate nursing OSCEs were interviewed about their assessment practices and participated in a "think aloud" during which they vocalised their thought processes while watching and grading four videos of students completing OSCE stations. Analysis of the data resulted in three themes being identified: *reducing guesswork, obstructed vision,* and *reduction in perceived validity*. These themes speak to the specific ways that the use of video in performance assessments affects how assessors engage with the task of assessment.

As noted in several review articles about the use of video in performance assessments (Felthun et al., 2021; Giri & Stewart, 2023), such studies have lacked a focus on assessors and how the use of video affects their cognitive processes. In spite of this, several findings in the present study have been noted elsewhere. Specifically, Chen et al. (2018) discussed how assessors may be limited by the use of video, as they are unable to walk around the room to improve their ability to see certain aspects of a student's performance. Relatedly, Chan et al. (2014) found that while videos in general could be used for assessment of clinical performance, caution should be exercised when using video for intricate physical tasks, as these were much less likely to be visible on a video. As determined in the present study, this has IRR implications, as some assessors may resort to guessing whether a student completed a task correctly or allowing them to compensate by narrating what they're doing.

A vocal minority of assessors in the present study expressed that the use of video reduced the objectivity of the assessment, as it decreased the fidelity of the OSCE in relation to the real world of clinical practice. The issue of whether OSCEs, and other performance assessments, are perceived as valid by assessors is one that has recurred in the literature on such assessments (e.g., Roberts et al., 2020; Hyde et al., 2022). Indeed, Hyde et al. (2022) found that, driven by a lack of belief in the utility of the OSCE, experienced assessors were more likely to intentionally deviate from marking guide designed for an OSCE station, and judge students according to what they personally believe to be important. As such, any adjustment to the OSCE, such as the incorporation of video, is likely to increase the risk of assessors intentionally deviating from the marking guide, which would affect the IRR of assessment scores.

There are two notable implications of the present study for those using video-based performance assessments. The first is to ensure that there is a robust procedure in place for conducting reliability checks within a pool of examiners. Calculating the IRR of awarded scores should be a step in any assessment, particularly one that is high-stakes (Khan et al., 2013; AERA et al., 2014). The present study indicates that the use of video can bring about specific threats to IRR, and as such the need for reliability checks may be amplified. Secondly, this study has indicated that certain tasks, especially those which are physically intricate, such as the location of a brachial artery, may be difficult or impossible to assess consistently through video. As a result, test administrators should consider the type of skills they wish to assess, and whether it is feasible to do so through video (Chan et al., 2014; Giri & Stewart, 2023). In spite of the specificity of the present study, these general principles will apply in any domain where practical skills are assessed through video: someone carrying out a video assessment in dentistry should also have a strict procedure in place to ensure that IRR levels are sufficient and would also need to think about what skills are feasible to be assessed through video.

In terms of theoretical implications, the present paper adds to the assessment literature by contributing to a growing body of research on assessors' cognitive processes (e.g., Kogan et al., 2011; Yeates et al., 2013; Roberts et al., 2020), and is one of the few studies to focus on nursing assessors specifically (e.g., East et al., 2014; Scully et al., 2024), as well as one of the first to focus on how the use of video affects assessors' judgements. As noted elsewhere, incipient research on video assessment has tended to focus on the cost and feasibility of setting up such assessments (e.g., Felthun et al., 2021; Giri & Stewart, 2023). This paper refocuses the literature by examining the assessors themselves, and whether they extract different information from a performance when it takes place through video.

A clear next step for researchers is to determine the comparability of assessors' grades when an assessment is conducted through video, as opposed to in-person. As noted by Giri and Stewart (2023, p. 14), there may be "a lack of correlation between the assessment scores of the face-to-face examiner and remote examiner". A study by Dagnaes-Hansen et al. (2018) is one of the few pieces of research to measure this specifically (using assessment scores of a cystoscopy exam), and their research should be extended into other contexts where possible.

The content of this study should be noted in light of its limitations. Firstly, a sample of 12 assessors working at the same institution necessarily limits the generalisability of the findings beyond this context. Ideally, researchers wishing to apply these findings in other contexts (especially outside the domain of nursing and the health sciences more broadly) should exercise caution, and conduct a comparable study where possible. Secondly, the assessments used in the study were simple, first- and second-year nursing OSCEs. It is possible that the effects of the use of video may be different if assessors were tasked with the interpretation of more complex tasks. As noted above, research has indicated that specific skills (such as communication) may be better suited to assessment through video. This study does not allow for generalisability beyond the specific, technical competencies assessed during a blood pressure measurement or a gastric tube insertion. Finally, the lack of score data in the current study prevents measurable links being made between assessors' cognitive processes and the scores they award. In other words, it was not possible to determine whether the changes in assessors' engagement as a result of the incorporation of video led to measurable effects regarding their awarded scores (e.g., Gingerich et al., 2014b; Chahine et al., 2016). Ideally, score data would be obtained when an assessment is conducted in-person, and compared when the same assessment is conducted through video (e.g., Dagnaes-Hansen et al., 2018). Such an approach would allow for a quantifiable measure of the effect of an assessment taking place through video. In spite of these limitations, this study will be of interest to researchers in nursing assessment, as well as those pursuing remote assessments more generally.

References

American Educational Research Association., American Psychological Association., National Council on Measurement in Education., & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for Educational and Psychological Testing*.

Barratt, J. (2010). A focus group study of the use of video-recorded simulated objective structured clinical examinations in nurse practitioner education. *Nurse Education in Practice, 10*(3), 170–175. https://doi.org/10.1016/j.nepr.2009.06.004

Benner, P. (1982). From Novice to Expert. *American Journal of Nursing, 82*, 402–07.

Brannick, M. T., Erol-Korkmaz, H. T., & Prewett, M. (2011). A systematic review of the reliability of objective structured clinical examination scores. *Medical Education, 45*(12), 1181–1189. https://doi.org/10.1111/j.1365-2923.2011.04075.x

Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–01. https://doi.org/10.1191/1478088706qp063oa

Braun, V. & Clarke, V. (2021). *Thematic Analysis: A Practical Guide*. Thousand Oaks, CA: Sage Publications.

Chahine, S., Holmes, B., & Kowalewski, Z. (2016). In the minds of OSCE examiners: uncovering hidden assumptions. *Advances in Health Sciences Education, 21*(3), 609–625. https://doi.org/10.1007/s10459-015-9655-4

Chan, J., Humphrey-Murto, S., Pugh, D. M., Su, C., & Wood, T. (2014). The objective structured clinical examination: can physician-examiners participate from a distance?. *Medical Education, 48*(4), 441–450. https://doi.org/10.1111/medu.12326

Chen, T. C., Lin, M. C., Chiang, Y. C., Monrouxe, L., & Chien, S. J. (2018). Remote and onsite scoring of OSCEs using generalisability theory: A three-year cohort study. *Medical Teacher, 41*(5), 578–583. https://doi.org/10.1080/0142159X.2018.1508828

Dagnaes-Hansen J., Mahmood O., Bube S., Bjerrum, F., Subhi, y., Rohrsted, M. & Konge, L. (2018). Direct observation vs. video-based assessment in flexible cystoscopy. *Journal of Surgical Education, 75*(3), 671–677. https://doi.org/10.1016/j.jsurg.2017.10.005

Darling-Hammond, L. & Adamson, F. (2010). Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning. Stanford Center for Opportunity Policy in Education. Available at: https://globaled.gse.harvard.edu/files/geii/files/beyond-basic-skills-role-performance-assessment-achieving-21st-century-standards-learning-report_0.pdf

East, L., Peters, K., Halcomb, E., Raymond, D., & Salamonson, Y. (2014). Evaluating Objective Structured Clinical Assessment (OSCA) in undergraduate nursing. *Nurse Education in Practice, 14*(5), 461–467. https://doi.org/10.1016/j.nepr.2014.03.005

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215–251. https://psycnet.apa.org/doi/10.1037/0033-295X.87.3.215

Eva, K., & D Hodges, B. (2012). Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? *Medical Education, 46*(9), 914–919. https://doi.org/10.1111/j.1365-2923.2012.04310.x

Felthun, J. Z., Taylor, S., Shulruf, B., & Allen, D. W. (2021). Assessment methods and the validity and reliability of measurement tools in online objective structured clinical examinations: a systematic scoping review. *Journal of Educational Evaluation for Health Professions, 18*, 11. https://doi.org/10.3352/jeehp.2021.18.11

Fraenkel, J.R. & Wallen, N.E. (2006). *How to Design and Evaluate Research in Education*. New York, USA: McGraw Hill.

Gauthier, G., St-Onge, C., & Tavares, W. (2016). Rater cognition: Review and integration of research findings. *Medical Education, 50*(5), 511–522. https://doi.org/10.1111/medu.12973

Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014a). Seeing the "black box" differently: Assessor cognition from three research perspectives. *Medical Education, 48*(11), 1055–1068. https://doi.org/10.1111/medu.12546

Gingerich, A., Van Der Vleuten, C. P. M., Eva, K. W., & Regehr, G. (2014b). More consensus than idiosyncrasy: Categorizing social judgments to examine variability in Mini-CEX ratings. *Academic Medicine, 89*(11), 1510–1519. https://doi.org/10.1097/acm.0000000000000486

Gingerich, A., Ramlo, S. E., van der Vleuten, C. P. M., Eva, K. W., & Regehr, G. (2017). Inter-rater variability as mutual disagreement: identifying raters' divergent points of view. *Advances in Health Sciences Education, 22*(4), 819–838. https://doi.org/10.1007/s10459-016-9711-8

Giri, J., & Stewart, C. (2023). Innovations in assessment in health professions education during the COVID-19 pandemic: A scoping review. *The Clinical Teacher, 20*(5), e13634. https://doi.org/10.1111/tct.13634

Goh, H. S., Zhang, H., Lee, C. N., Wu, X. V., & Wang, W. (2019). Value of nursing objective structured clinical examinations: A scoping review. *Nurse Educator, 44*(5), E1–E6. https://doi.org/10.1097/nne.0000000000000620

Govaerts, M., & van der Vleuten, C. P. (2013). Validity in work-based assessment: Expanding our horizons. *Medical Education, 47*(12), 1164–1174. https://doi.org/10.1111/medu.12289

Gwet, K.L. (2014). *Handbook on Inter-rater Reliability*. Gaithersburg, MD, USA: Advanced Analytics.

Harden, R. M., Stevenson, M., Downie, W. W., & Wilson, G. M. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal, 1*(5955), 447–451. https://doi.org/10.1136%2Fbmj.1.5955.447

Hodges, B. (2013). Assessment in the post-psychometric era: learning to love the subjective and collective. *Medical Teacher, 35*(7), 564–568. https://doi.org/10.3109/0142159X.2013.789134

Hopwood, J., Myers, G., & Sturrock, A. (2020). Twelve tips for conducting a virtual OSCE. *Medical Teacher, 43*(6), 633–636. https://doi.org/10.1080/0142159X.2020.1830961

Hyde, S., Fessey, C., Boursicot, K., MacKenzie, R. & McGrath, D. (2022). OSCE rater cognition – an international multi-centre qualitative study. *BMC Medical Education, 22*(6). https://doi.org/10.1186/s12909-021-03077-w

Khan, K. Z., Gaunt, K., Ramachandran, S., & Pushkar, P. (2013). The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An historical and theoretical perspective. *Medical Teacher, 35*(9), 1437-46. https://doi.org/10.3109/0142159x.2013.818634

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: A conceptual model. *Medical Education, 45*(10), 1048–1060. https://doi.org/10.1111/j.1365-2923.2011.04025.x

Lara, S., Foster, C., Hawks, M. and Montgomery, M. (2020). Remote assessment of clinical skills during COVID-19: A virtual, high-stakes, summative paediatric objective structured clinical examination. *Academic Paediatrics, 20*(6), 760–761. https://doi.org/10.1016%2Fj.acap.2020.05.029

Major, S., Sawan, L., Vognsen, J., & Jabre, M. (2020). COVID-19 pandemic prompts the development of a Web-OSCE using Zoom teleconferencing to resume medical students' clinical skills training at Weill Cornell Medicine-Qatar. *BMJ Simulation & Technology Enhanced Learning*, *6*(6), 376–377. https://doi.org/10.1136/bmjstel-2020-000629

Massey, D., Byrne, J., Higgins, N., Weeks, B., Shuker, M. A., Coyne, E., Mitchell, M., & Johnston, A. N. B. (2017). Enhancing OSCE preparedness with video exemplars in undergraduate nursing students. A mixed method study. *Nurse Education Today, 54*, 56–61. https://doi.org/10.1016/j.nedt.2017.02.024

Navas-Ferrer, C., Urcola-Pardo, F., Subiron-Valera, A.B., & German-Bes, C. (2017). Validity and reliability of Objective Structured Clinical Evaluation in nursing. *Clinical Simulation in Nursing, 13*(11), 531–543. https://doi.org/10.1016/j.ecns.2017.07.003

Patrício, M. F., Julião, M., Fareleira, F., & Carneiro, A. V. (2013). Is the OSCE a feasible tool to assess competencies in undergraduate medical education? *Medical Teacher, 35*, 503–514. https://doi.org/10.3109/0142159X.2013.774330

Paul, F. (2010). An exploration of student nurses' thoughts and experiences of using a video-recording to assess their performance of cardiopulmonary resuscitation (CPR) during a mock objective structured clinical examination (OSCE). *Nurse Education in Practice, 10*(5), 285–290. https://doi.org/10.1016/j.nepr.2010.01.004

Purpora, C., & Prion, S. (2018). Using student-produced video to validate head-to-toe assessment performance. *Journal of Nursing Education, 57*(3), 154–158. https://doi.org/10.3928/01484834-20180221-05

Roberts, R., Cook, M., & Chao, I. (2020) Exploring assessor cognition as a source of score variability in a performance assessment of practice-based competencies. *BMC Medical Education, 20*(1), 168. https://doi.org/10.1186/s12909-020-02077-6

Rushforth, H. E. (2007). Objective structured clinical examination (OSCE): Review of literature and implications for nursing education. *Nurse Education Today, 27*(5), 481–490. https://doi.org/10.1016/j.nedt.2006.08.009

Scully, C., Kelly, M., Lysaght, Z., & O'Leary, M. (2024). The cognitive processes employed by undergraduate nursing OSCE assessors: A qualitative research study. *Nurse Education Today, 134*, 106083. https://doi.org/10.1016/j.nedt.2023.106083

St-Onge, C., Chamberland, M., Lévesque, A., & Varpio, L. (2016). Expectations, observations, and the cognitive processes that bind them: expert assessment of examinee performance. *Advances in Health Sciences Education, 21*(3), 627–642. https://doi.org/10.1007/s10459-015-9656-3

Willis, G.B. (2015). *Analysis of the Cognitive Interview in Questionnaire Design*. Oxford: Oxford University Press.

Yeates, P., O'Neill, P., Mann, K., & Eva, K. (2013). Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. *Advances in Health Sciences Education: Theory and Practice, 18*(3), 325–341. https://doi.org/10.1007/s10459-012-9372-1

Yeates, P., Moreau, M., & Eva, K. (2015). Are Examiners' Judgments in OSCE-Style Assessments Influenced by Contrast Effects? *Academic Medicine: Journal of the Association of American Medical Colleges, 90*(7), 975–980. https://doi.org/10.1097/acm.0000000000000650