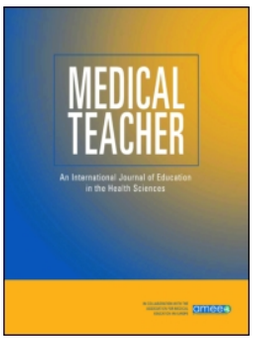

Medical Teacher

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/imte20

Understanding and developing procedures for video-based assessment in medical education

Peter Yeates , Alice Moult , Janet Lefroy , Jacqualyn Walsh-House , Lorraine Clews , Robert McKinley & Richard Fuller

To cite this article: Peter Yeates , Alice Moult , Janet Lefroy , Jacqualyn Walsh-House , Lorraine Clews , Robert McKinley & Richard Fuller (2020): Understanding and developing procedures for video-based assessment in medical education, Medical Teacher, DOI: 10.1080/0142159X.2020.1801997

To link to this article: https://doi.org/10.1080/0142159X.2020.1801997

Published online: 04 Aug 2020.

MEDICAL TEACHER
https://doi.org/10.1080/0142159X.2020.1801997

Understanding and developing procedures for video-based assessment in medical education

Peter Yeates[a,b], Alice Moult[a], Janet Lefroy[a], Jacqualyn Walsh-House[a], Lorraine Clews[a], Robert McKinley[a] and Richard Fuller[c]

[a]School of Medicine, Keele University, Keele, UK; [b]Department of Acute Medicine, Fairfield General Hospital, Pennine Acute Hospital NHS Trust, Bury, UK; [c]School of Medicine, University of Liverpool, Liverpool, UK

ABSTRACT

Introduction: Novel uses of video aim to enhance assessment in health-professionals education. Whilst these uses presume equivalence between video and live scoring, some research suggests that poorly understood variations could challenge validity. We aimed to understand examiners' and students' interaction with video whilst developing procedures to promote its optimal use.

Methods: Using design-based research we developed theory and procedures for video use in assessment, iteratively adapting conditions across simulated OSCE stations. We explored examiners' and students' perceptions using think-aloud, interviews and focus group. Data were analysed using constructivist grounded-theory methods.

Results: Video-based assessment produced detachment and reduced volitional control for examiners. Examiners ability to make valid video-based judgements was mediated by the interaction of station content and specifically selected filming parameters. Examiners displayed several judgemental tendencies which helped them manage videos' limitations but could also bias judgements in some circumstances. Students rarely found carefully-placed cameras intrusive and considered filming acceptable if adequately justified.

Discussion: Successful use of video-based assessment relies on balancing the need to ensure station-specific information adequacy; avoiding disruptive intrusion; and the degree of justification provided by video's educational purpose. Video has the potential to enhance assessment validity and students' learning when an appropriate balance is achieved.

KEYWORDS

Objective Structured Clinical Examinations (OSCEs); performance assessment; video-based assessment; assessor cognition; validity

Introduction

In recent years, several innovations have attempted to address perennial limitations in the consistency (Eva 2018) and educational impact (Harrison et al. 2017) of assessment in medical education by using video technology to: enhance quality assurance of examiners' scoring (Yeates et al. 2019); support assessor training through benchmarking (McManus and Omer 2017), enable remote examining (Vivekananda-Schmidt et al. 2007; Chen et al. 2019), or provide a detailed review of students' performance within feedback conversations (Eeckhout et al. 2016). Whilst these approaches offer significant promise, the implications of using video-based judgements within assessment has received comparatively little scrutiny.

Common to all of these approaches is the assumption that assessors' video-based judgements are equivalent to their judgements of live performances. Several studies have shown equivalent scores between live and video-based modalities (Ryan et al. 1995; Vivekananda-Schmidt et al. 2007; Chen et al. 2019; Yeates et al. 2019). This observation is not, however, universal. Scaffidi et al. (2018) found that video scores in an assessment of colonoscopy skills were systematically higher than live scores, although the two were highly correlated. Conversely, Hance et al. (2005) showed the opposite relationship in an assessment of cardiothoracic surgical skills: video performances received lower scores than live performances, although performance levels were equally well differentiated. The authors speculated that blinding of individuals' identity in the

Practice points

- Using video in assessment relies on equivalence between live and video-based judgements.
- Examiners experience video-based judgements differently to live judgements.
- Ensuring examiners have sufficient video-based information to make equivalent judgements depends on carefully selecting task-specific camera set-up.
- Negative influences of filming on students' performances can be mitigated by careful camera positioning.
- Video use can enhance validity of assessment when video-information adequacy, intrusiveness to students and educational purpose are adequately balanced.

CONTACT Peter Yeates p.yeates@keele.ac.uk School of Medicine, Keele University, David Weatherall Building, Keele University, Keele, ST5 5BG, UK

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

video condition may have mediated this difference. Scott et al. (2000) compared edited videos of the salient portions of laparoscopic surgical skills with live observation scores of the same performances. These measures correlated poorly, and couldn't discriminate trained from untrained surgeons. Consequently, it appears that whilst video and live scores are often the same this similarity is contextually dependent, for reasons which are not clearly established.

Research from the field of social neuroscience suggests that video (compared with live) could influence the *processes* of judgements. When people watch a scene on video, they pay attention to different communication-related features (i.e. the face more than gestures) than someone watching the same scene live (Gullberg and Holmqvist 2006). This may result from a reduced sense of volitional control or reduced expectation of interaction (Foulsham et al. 2011). These effects may not simply emanate from the modality (i.e. video vs live) so much as what information is included in the video. For example, being able to see a person's head and shoulders rather than just their face increases the empathy of watchers (Nguyen and Canny 2009). Collectively these findings suggest that the manner of presentation of video information could importantly influence the processes of assessment judgements.

Extrapolating from these studies', we might posit that a range of factors could influence the way that assessors make video-based judgements compared to live judgements: a narrower focus of vision, reduced interpersonal interaction, or reduced volitional control might all influence assessors' attention, recall, empathy or alter the salience of particular features of performances. As assessment judgements are, at least in part, intuitive (Yeates et al. 2013b), and sensitive to both context (Yeates et al. 2012) and attentional salience (Gingerich et al. 2018), these and possibly other processes have the potential to explain why video scores (despite general similarity) have differed from live judgements on some occasions and to explain whether there are conditions which make it more or less likely for differences to arise.

Whilst variations in examiners' judgements between video and live modalities could threaten the validity of resulting video-based scores, any undue influence of filming on students' performances would also constitute a source of construct irrelevant variance (Amin et al. 2011). In sports science, the combination of an audience and video recording has been shown to reduce performance for self-conscious individuals, whilst improving it for others (Wang et al. 2004). Whilst test anxiety in OSCEs due to an awareness of examiners is well described (Harrison et al. 2015), it is unclear whether videoing students' assessment performances might add to this sense, thereby unduly altering performance for some students.

As a result, whilst video use has the potential to enhance assessment in several ways, there may be a variety of largely unexplored influences on, or implications for, video-based performance judgements in health professions education. As these poorly characterised processes could have important unintended consequences for the validity, fairness or acceptability of video-based assessment judgements, we sought to understand the process of video-based judgements whilst seeking to establish whether there are conditions which will help to ensure their optimal use, by addressing the following related research questions:

1. How do examiners' judgemental processes compare when judging video-based performances and live performances?
2. What filming procedures are needed for different types of assessment tasks to maximise the likelihood of the processes of video-based judgements being equivalent to live judgements?
3. How do students and examiners experience and interact with video in assessment, and what conditions are needed to minimise any resulting threats to assessment validity?

Methods

We used design-based research (Baumgartner et al. 2003) to explore and develop a theory of video-based assessment whilst iteratively developing our filming approach. Design-based research enables development of a learning environment through continuous cycles of design, enactment, analysis, and redesign, whilst simultaneously developing educational theory. Data are typically collected through a mixture of methods, which may include surveys, measurements, observations, field notes, brief conversations with participants, think-aloud, interviews or focus groups (Cobb et al. 2003). In order to manipulate both assessment scenarios and filming conditions without prejudicing actual examinations, we used simulated Objective Structured Simulated Examinations (OSCE) stations (Newble 2004) which were both videoed and examined live. We principally collected data through participant interviews with examiners and students and documentation of researchers' observations. We additionally performed a number of examiner and student focus groups to determine whether additional perceptions were co-constructed within resulting dialogues.

Population, sampling and recruitment

Our study populations were undergraduate OSCE examiners and clinical years medical students from Keele School of Medicine. We purposively sampled participants from a variety of ethnic backgrounds, with English as either a first or second language, and from different regions of the UK. We sampled novice and experienced examiners from a variety of specialities.

Recruitment was performed via email and announcements at meetings. Participation was voluntary and all participants signed a consent form. Ethical approval for the study was granted by Keele University Ethical Review Panel (ref ERP2379).

Simulated OSCE stations and data gathering

Simulated OSCE stations mimicked typical OSCE stations in terms of rooms, furniture, equipment and the presence of a timer. An examiner and simulated patient were present within each station. Station content was developed by experienced clinical educators (PY & JL) and varied across iterations (see Table 2). Examiners were provided with

detailed station information (marking criteria, simulated patient scripts, student instructions).

After reading station instructions for the OSCE scenario, students were asked to enter the OSCE station and perform as they would in a real OSCE. In most iterations, two examiners were present: one in the room (the 'live' examiner), and one outside (the 'video' examiner). The video examiner was provided with the same station information as the live examiner, and asked to judge the video performance in the same manner as they would a live performance. The video examiner watched the same performance the live examiner had judged via video immediately (within 20 mins) after the live performance whenever possible or following a delay (up a few days) if the examiner's availability made this pragmatically necessary. Whenever participants' availability allowed, examiners crossed over between the live and video roles and judged a second student performing at the same station. Both live and video examiners scored the performance and considered the feedback they would give.

Filming approaches

Researchers filmed the simulated OSCE stations using a variety of filming methods. This included fixed, wide angle ceiling cameras (identical to filming in Yeates et al. (2019); tripod based camcorders, positioned in various places within the room, and using varied degrees of zoom; head cameras worn by the examiner; and wall mounted pan/tilt/zoom CCTV cameras positioned in various places within the room (see Figure 1). Sound was collected variously using ceiling hanging microphones; focused microphones placed on the camcorders; and table-top condenser microphones. Camera positions and settings were documented and iteratively developed (see Figure 1 for an example).

Examiner interviews

Examiners were asked to persist in the frame of mind evoked by examining whilst they completed score sheets. Although both 'live' and 'video' examiners were asked to note the feedback they would give, the 'live' examiner verbally communicated their feedback to the student so that the student's learning was supported by study participation. Both 'live' and 'video' examiners were asked to perform retrospective think-aloud (Van Den Haak et al. 2003) describing all aspects of performance which were salient to their judgements. Next, researchers used semi-structured interviews (Galletta 2013) to explore examiners' perceptions of: judgemental influences, simulation authenticity, encountered difficulties, information management strategies and their judgemental certainty. Topic guides were derived from our initial literature review and evolved to test emergent theory. Researchers probed 'video' examiners' perceptions of the availability of salient visual and audio information and 'live' examiners' perceptions of differences between modalities. Further questions explored all examiners' comfort making video-based judgements, perceptions of the acceptability and intrusiveness of filming and potential implications of uses of the videos.

Student interviews and focus groups

Semi-structured interviews with students explored their awareness of cameras, or any perceived influence of cameras on their performances, along with their perceptions of the acceptability, challenges or potential educational benefits of video within assessment. Focus groups (Gill et al. 2008) explored issues at the intersections of students' and examiners' perspectives by discussing a similar range of issues. Focus groups were conducted on the same day as a filming iteration, and involved examiners and students who had been present for that iteration.

Data analysis

Analysis in design-based research can draw from an array of methods (Anderson and Shattuck 2012), but analysis methods derived from grounded theory (Guba and Lincoln 1982; Charmaz 2006) have been recommended for interview and focus group data to ensure rich theory development (Bakker and van Eerde 2015). Using these analysis methods, interview and focus group data were analysed iteratively, interspersed with new data collection. Two researchers (PY, a clinician educator, and AM a post-doctoral health psychologist) independently performed both inductive and theoretical open coding (Bryant and Charmaz 2019) of early iterations' data. Through frequent discussion, analysts agreed a coding frame which evolved as analysis progressed. AM coded all data whilst PY additionally coded all think-aloud data and six interviews. Discrepancies were resolved through discussion.

Researchers used constant comparison involving challenge and search for discrepancy in both new and existing data (Lincoln and Guba 1985), micro-analysis (Engward 2013), and memo-writing (Montgomery et al. 2007). Consistent with our design-based research approach(Cobb et al. 2003), data from interviews, think-aloud and focus groups were integrated with researchers field notes of practical adaptions to filming conditions and observations of effects of particular modifications as we developed axial (Strauss and Corbin 1998) and then selective codes (Holton 2010) which were used to organise the final theory. Data sufficiency (Varpio et al. 2017) was deemed to have occurred when the developed theory adequately described all observations within the 9th and 10th iterations.

In line with the approach adopted in prior design-based research(Koivisto et al. 2018; Papavlasopoulou et al. 2019) some of the reported results are illustrated by verbatim quotes from participants whilst other findings are drawn from researchers observations across multiple iterations and are therefore not illustrated with quotes.

Results

Sixteen students and fourteen examiners participated across 10 iterations of data collection. Participants represented diversity of nationality and UK regions, ethnicity and people for whom English was not their first language (see Table 1). Iteration details are presented in Table 2. Interview, think aloud, and focus group data comprised approximately 28 h corresponding to 313 pages of data, which sat alongside notes summarising observations and modifications at each stage.

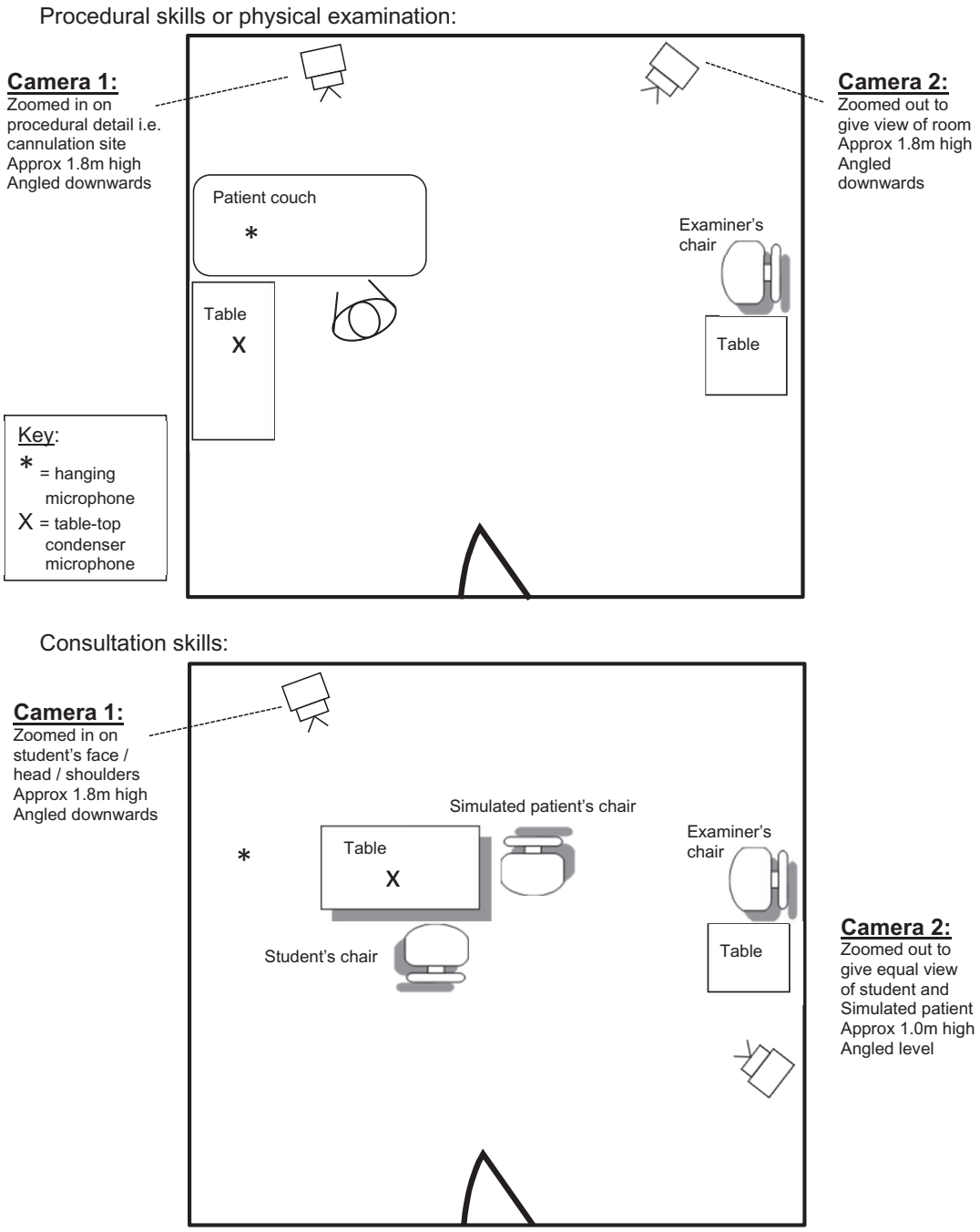

Figure 1. Examples of typical room and camera orientation for procedural skills and physical examination stations.

Making judgements from videos

Examiners described a sense of detachment when judging performances by video which made them less immersive than live judgements.

> … it's very hard to pinpoint what I may have lost from watching it on the video, I'm not sure if there was anything specific that was lost. But it just felt very different …
>
> Examiner 1, interview

Examiners described reduced volitional control over what they saw which contributed to this sense of distance. Despite this, they were usually able to comfortably make judgements on performances from videos.

> I feel overall, again I feel confident of my judgment compared to if I was in the room… I actually feel quite happy I've got adequate amount of information to make a judgement.
>
> Examiner 11, interview

Video was perceived to be capable of enhancing assessment, but this capability depended on achieving a sufficient compromise between three inter-related themes: ensuring information adequacy in videos; interaction with examiners' judgemental tendencies; and balancing acceptability and purpose.

Ensuring information adequacy in videos

Broadly speaking, examiners commented on similar aspects of performances whether examining live or video performances. Whilst in some instances, pairs of examiners varied in their focus or interpretation, these appeared to emanate from individual differences between examiners rather than the modality. There did not appear to be any overall systematic influence of the modality (video vs live) on

examiners' focus, judgemental processes, or interpretation of behaviour.

Despite this general overall impression, there were clearly instances where video-based examiners weren't entirely satisfied with the information which videos presented: sometimes this related to information excluded from the shot ('*so we don't have his head this time … so absolutely no facial expressions*', examiner 5, viewing tripod camera); sometimes about the clarity of detail within the shot ('*So I can't actually see what she's prescribing at all, so I don't feel able to actually give that a mark*.' Examiner 11, viewing video); sometimes something in the shot was obscured (i.e. the student had faced away from or blocked the camera). In some instances the sound was indistinct ('*I could see him talking to the student but I couldn't actually hear exactly what he was asking*', Examiner 4 watching headcam video) or the lighting was poor. Rarely examiners described an overall sense that communication simply had not transmitted as well as in the live scenario. These challenges were more common in early iterations as researchers developed the filming approach.

Table 1. Frequencies of examiners' and students' self-reported demographic characteristics, accent and sensory impairments.

Characteristic	Frequency
Examiners	
Ethnicity	
Indian	2
British	10
Spanish	1
Accent	
Neutral	4
Mild	9
Sensory impairments	
Mild hearing loss	1
Mild speech impairment	1
No sensory impairment	12
Students	
Ethnicity	
British	8
African	1
Russian	1
Pakistani	3
Indian	2
Accent	
Neutral	11
Mild	4
Moderate	1
Sensory impairments	
Mild hearing loss	2
No sensory impairment	14

Factors which mediated information adequacy

Despite specifically comparing across different participants and multiple iterations, we did not find evidence within our sample that students' or examiners' ethnicity, accent or the presence or absence of sensory impairments mediated information adequacy. Several equipment-related factors critically mediated video adequacy: the horizontal and vertical position of cameras within the room; the angulation of the cameras; the degree of image zoom; lighting (neither too bright which caused glare, nor too dim); the type and positioning of microphones within the room; and the number of views provided to examiners

> I actually felt comfortable with that because the two views that were selected were complimentary. They gave me enough of a view that I could see everything that was going on within the room.
>
> Examiner 11, wide-angle and zoomed tripod camera views, Iteration 6

We found that the optimal combination of these factors varied for different types of OSCE task. Examiners generally preferred viewing consultations from a seated eye-level (0.8–1.0m), whilst procedural skills were seen clearly using a zoomed-in camera 1.8 m from the ground. Examiners were satisfied with a single perpendicular view of a consultation (student facing the SP). Conversely procedural or examination skills required two views: one oblique wide-angle view to give a sense of interaction with the SP, and a reverse angle view to give close-up procedural detail (see Figure 1). Particular requirements emerged for specific tasks: the need to see the simulated patient's back during a respiratory examination; the need to see a close up of the student's writing in a prescribing task. Consequently, we found that it was necessary to be able to move the position of cameras within the room for different station set ups.

The type of video cameras influenced information adequacy. Tripod-mounted camcorders were flexible, but it was difficult to get sufficient height to see procedures without being obtrusive. Conversely, ceiling camera views

Table 2. Details of station content and camera selection in each iteration of the study.

Iteration	OSCE station(s)	Cameras used
1	Asthma history	Tripod camera Ceiling camera
2	Venepuncture	Ceiling camera Head camera
3	ECG skill	Two tripod cameras Head camera Ceiling camera
4	Arterial blood gas sampling	Tripod cameras Head camera Ceiling camera
5	Diabetes history	Head camera Ceiling camera
6	IV Fluids for sepsis	Two tripod camera Ceiling camera
7	History for acute Cholecystitis	Two tripod cameras
8	Insertion of nasogastric tube	Tripod camera Wall-mounted camera
9	History of a migraine Arterial blood gas sampling Respiratory examination	Two wall-mounted cameras
10	History of migraine	Two wall-mounted cameras

seemed 'flattened' and made it difficult to see facial expressions:

> the format we viewed which was the sort of birds eye view. That didn't feel anything like examining in a normal OSCE … you feel much further removed and I think it affected the experience.
>
> Examiner 1, describing ceiling camera

Wearable head cameras obtained similar views to live examiner's vision, but head movements by the live examiner produced numerous issues for the video examiner: motion-sickness, lost information (i.e. looking away) or cueing (i.e. by nodding). Wall-mounted CCTV cameras were less intrusive and enabled real-time video access but were inflexible if fixed in a single location.

Working from these observations, we found the optimal balance was achieved by using two wall-mounted CCTV cameras. These gave excellent image quality, were unobtrusive to students and enabled rapid video processing. To enable flexible positioning we developed movable frames which let the cameras be set at various heights and positions within a room. We chose camera positions, angulation and zoom for each station based on analysis of its layout and tasks (see Figure 1).

Consequently, we found that videos were capable of providing examiners with sufficient information to make dependable judgements but required both technical audiovisual expertise and analysis of station content by someone with clinical/educational expertise to choose station-specific camera positions and settings.

Interaction with examiners' judgemental tendencies

Despite not having the immersive, three dimensional immediacy of live examining and there being occasional details which examiners could not see or hear, examiners were, for the most part, comfortable to judge performances via video. Examiners described (or displayed) a range of judgemental tendencies which enabled them to manage the limitations of videos. Examiners described a tendency (in both live and video scenarios) to make global judgements of candidates, or to be guided by the candidate's fluency.

> there was something about his overall approach, it was … he could have been slightly slicker and more fluent but he was, he clearly knew what he was doing.
>
> Examiner 5, interview (video)

This enabled examiners to make judgements even if some specific details were missing. Examiner 10 commented that not all information in the performance was salient; that small aspects of performance were 'not a deal breaker' (Examiner 10, Iteration 6). Examiners sometimes made inferences about specific aspects of a candidate's performance which they had been unable to see. Occasionally examiners would do this for aspects of performance which were extremely important:

> So I felt I've made a big assumption that she has primed the line correctly which I think for this particular skill is quite a big assumption to make. Because if you had flushed a line full of air into someone, that's a "never event". But I think given her overall demeanour and the confidence, I could tell there was fluid in the chamber, I feel confident I've probably made the correct decision there.
>
> Examiner 11, interview (video)

Notably, despite offering a judgement at the time, this examiner expressed further doubts about the clarity of their observation later in the interview. A few examiners described having an instinct about when it was ok to make inferences. Examiners sometimes referred to proxy information when trying to interpret situations where there was something which they couldn't see or hear:

> It looks as though he's going to be doing the ankles but you can't actually see the ankles but you sort of move down that end
>
> Examiner 5, Iteration 3 (video)

Only occasionally did an examiner state that the video did not offer sufficient information to enable them to make a reasonable judgement. Notably, these judgemental tendencies by examiners' were not limited to video-based judgements. Examiners described making inferences during real OSCE examining, sometimes in response to fatigue or brief lapses in concentration, or their judgements being influenced by a global sense of performance.

Consequently, a number of well described judgemental tendencies (global judgements, inferences, differential salience) along with some previously undescribed processes (using proxy information) appeared to enable examiners to manage most challenges which emanated from videos. Whilst often reflective of authentic live examining or an examiner's expertise, in some instances examiners might have tended towards making video-based judgements which were not adequately informed to ensure safe assessment decisions. Consequently, examiner's judgemental tendencies appeared to interact with the adequacy of information in the videos to either enhance or detract from the overall quality of video-based judgements. This was particularly the case for stations where it was harder to ensure information quality, for example in procedural skills stations where fine detail or particular movements had the potential to importantly influence examiners' judgements.

Balancing acceptability and purpose

In the majority of iterations, students perceived that cameras were only minimally intrusive. Several commented that they rapidly forgot about cameras as they engaged in the task.

> I think with the camera positioning, both of us agreed that when we were talking to the patient, the camera wasn't in our face. It wasn't even in our vision.
>
> Student 1, focus group

In one instance, Student 5 who was simultaneously being filmed by four cameras (two tripod cameras, a ceiling camera and an examiner head-camera) commented that they were only passingly aware of one tripod camera. Another student (student 10) commented that despite passing awareness of the cameras, they didn't feel the cameras increased their sense of scrutiny beyond that caused by the examiner.

Despite these general assurances, there were instances where both students and examiners perceived that the cameras had disrupted a student's performance

> when you asked me the questions and I was answering and I looked at the camera, I forgot the question.
>
> Student 1, focus group

> I didn't mind the camera at all but I didn't like what the camera did to the student 'cos she was fine until she turned around to answer my questions, saw the camera and panicked.
>
> Examiner 2, interview

Whilst this only occurred in a small minority of cases it indicated the potential for students to freeze or lose their train of thought in response to seeing cameras. Some students perceived that other students within their year (outside of our sample) who were more prone to assessment anxiety would be at greater risk of disruption.

We found that negative influences on students' performance could be prevented by careful positioning of equipment. For example, we found that positioning cameras 1/ where students would look whilst performing the task, or 2/where they would look whilst talking to the examiner could potentially be disruptive. Despite this, positioning a camera in the arc between points 1 and 2 it did not appear to be disruptive, despite students' moving their gaze across this arc whilst turning to face the examiner. As a result a tension existed between, on one hand, optimising camera and microphone placement to maximise information adequacy for examiners and, on the other hand, minimising the potential for the presence of cameras or microphones to unduly influence students' performance.

For both students and examiners, the acceptability of any potential intrusion by cameras was balanced against the potential benefits of videoing ('It depends on the goal' Student 2). Most students were clear that they cared greatly about standardisation in OSCEs and believed that there was room within current practice for it to be enhanced. As long as cameras were not unduly obtrusive, students perceived that the potential for video to enhance standardisation offset any sense of intrusion which cameras caused. The intended use and distribution of videos was important to students. Student 2 commented that restricting access to their videos to just a few members of staff (rather than wider availability) was an acceptable degree of exposure given the potential for the videos to enhance standardisation.

Examiners and students described numerous ways videos might enhance OSCE standardisation: to facilitate examiner score comparisons, benchmarking or training, or mediation of appeals. Participants described potential enhancement of students' learning through video-based feedback**:**

> to actually see your performance and think 'Oh okay yes, I can see I really didn't do well there' … that would be very helpful … you'd get so much more out of the OSCE experience rather than just 'It's an exam'.
>
> Student 8, interview

Some participants suggested assessments would be more authentic with just the student and simulated patient in the room.

Students and examiners perceived that cameras might help to prevent examiners deviating from assessment instructions.

> [referring to being videoed] You have to actually listen carefully and use the mark scheme that everybody else is going to be using, otherwise you will stand out like a sore thumb. So I think it's good for examiners.
>
> Examiner 15, interview

Whilst potentially beneficial to standardisation, examiners and students perceived that videoing could detrimentally influence examiners' interactions with candidates. Both examiners and students suggested examiners may legitimately encourage students, especially when flustered or nervous, but that perceived scrutiny of their behaviour by video might make examiners stricter or colder. Student 1 commented that they look for indications of examiners' approval, which they felt they would be lacking if video were being used.

Consequently, both students and examiners perceived that video could enhance assessments without unduly influencing students' performance or being unacceptable to participants.

Discussion

Summary of results

Examiners experienced video-based performances differently to live performances. Whilst judgemental processes were similar between video and live modalities, specific combinations of station content and filming conditions limited the adequacy of information availability or interacted with examiners' judgemental processes to produce judgements which may not have been fully representative of live judgements. Students rarely perceived cameras as intrusive and performance disruption could be avoided through thoughtful camera placement. Video was perceived to enhance assessment when a sufficient balance was achieved between: supplying examiners with enough information; minimising intrusion; and ensuring the purpose of videoing provided adequate justification (see Figure 2 for illustration).

Relationship to existing literature and theory

We observed a number of well-described features of examiners' judgements including global judgements (Yeates et al. 2013a), first impressions (Wood 2014) and the use of inferences (Govaerts et al. 2011; Kogan et al. 2011). Whilst these processes are collectively presumed to emanate from automatic (or system 1) judgements or schema-based processing (Bargh and Chartrand 1999), the degree to which examiners' judgements rely on automatic as opposed to conscious deliberate processing remains debated (Gauthier et al. 2016). The observation that examiners in our study were often comfortable making judgements despite limitations in visual information, and, moreover, that examiners described missing details within their observations of real OSCEs due to fatigue or lapses in concentration, tends to suggest that the role of automatic (system 1) reasoning in examiners' judgements could be substantial. Whilst automatic judgements may provide an efficient judgemental means of managing the mental workload of assessments (Byrne et al. 2014; Tavares and Eva 2014), they also enable the Halo effect (the tendency for an impression created in one area to influence opinion in another) (Gingerich et al. 2011). Our study suggested that this could be a particular concern in video-based assessment when critical details are not captured (i.e. fine detail of equipment handling, detail of written information) if they contradict the more general impression created by the performance.

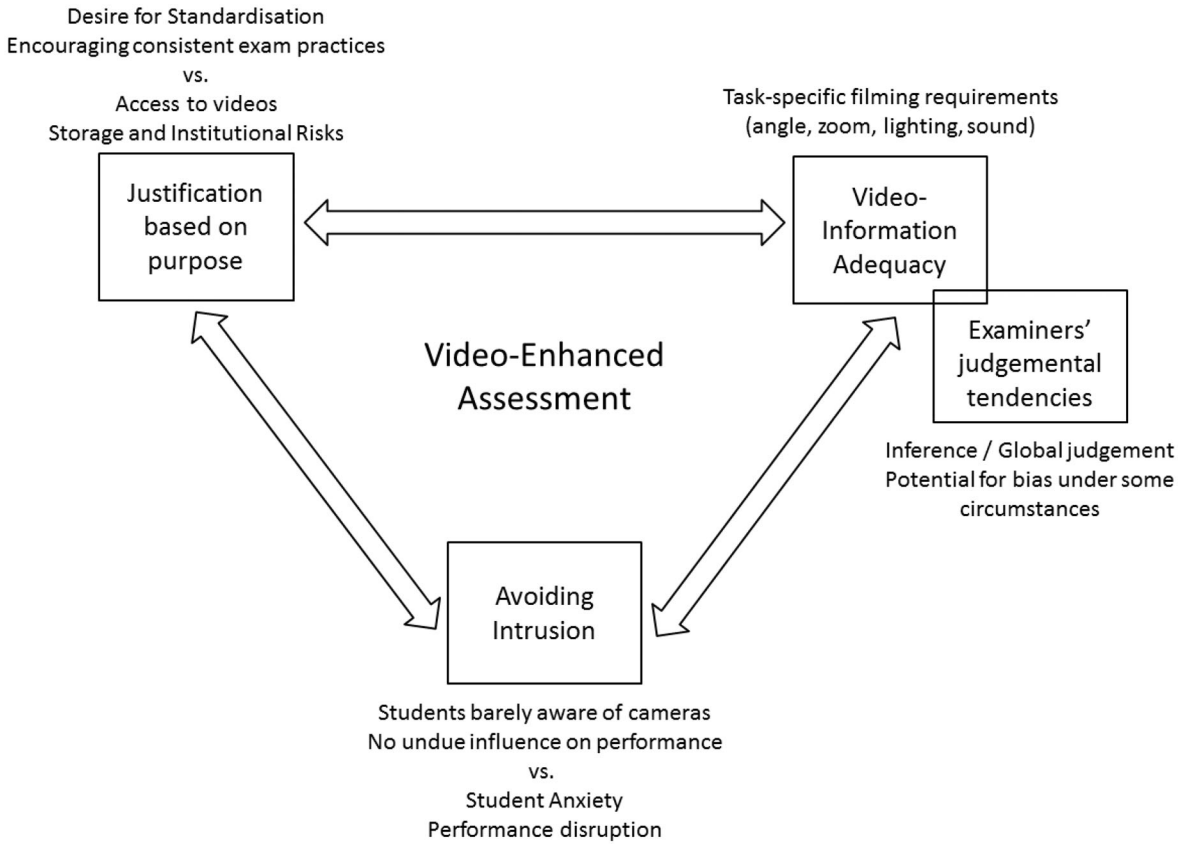

Figure 2. Illustration of tensions in using video to enhance assessment.

Our concern that students' performance might be impeded by the presence of video cameras was only occasionally realised. Students attributed their limited awareness of video cameras to focusing their attention on the assessment task. This may be an example of 'inattentional blindness' (Simons and Chabris 1999) in which people fail to perceive a clearly visible object (for example a gorilla walking amongst people playing basketball (Simons 2010)) due to actively focusing on something they consider important. As a result, video cameras may be less intrusive than anticipated, as long as students are actively focused on a task. Nonetheless, the fact that cameras occasionally caused students to freeze underscores the importance of mitigating this risk by careful camera placement.

Implications for practice

Based on our findings, we recommend that whilst video has the potential to enhance assessment in several ways, careful set up of video equipment (position, zoom, focus) must involve someone who understands the clinical and educational content of each station. Consideration should be given on a station by station basis to the likely actions of students, their predictable gaze patterns, the movement of examiners, and the features of performances for which close-up detail may be required. Video-capture should ensure that examiners have adequate views of critical performance elements to avoid the use of potentially detrimental inferences. This risk appears to be content-dependent and may pose greater risks in procedural skills stations.

Many of the compromises we have described (information vs intrusion; acceptability vs purpose) will vary for different assessment purposes. Students may be less concerned about intrusiveness in an assessment which has few consequences. Situations where examiners need to view videos immediately after candidates' performances may preclude video-processing to provide zoomed and wide-angle views. Students may agree to their videos being used to help standardise the exam in which they have participated (Yeates et al. 2019), but not agree to their broader use in faculty development. As a result different choices are likely to be appropriate in different contexts.

Whilst video may seem to offer an attractive means of settling appeals, reviewing them could impose substantial institutional time demands or produce vulnerability to legal challenges. Consequently institutions should think carefully about the duration of video storage and how they communicate with students about the purpose, use and access of videos.

As a result, whilst we anticipate that the filming solution we described in the results (based on dual CCTV cameras which can be moved to different positions and heights within the room) will be suitable for many assessment situations, use of this equipment in each specific assessment context should be tailored to balance the competing tensions we have described.

Reflexivity

Four researchers (PY, AM, RF, RMK) are involved in researching video-based methods to enhance OSCE standardisation (Yeates et al. 2019). These researchers acknowledge a motivation to attempt to ensure equivalence between live and video-based scores. Including two other researchers (JL, LC) who are heavily involved in teaching and assessing clinical skills, and 1 undergraduate medical student, brought balance to the research team.

Limitations

Whilst our study had significant strengths in terms of diverse sampling, careful data collection and rigorous analysis, it nonetheless has some limitations. The design of our study prevented us from determining the influence of modality (video versus live) on the scores of individual stations as these comparisons were confounded by inter-examiner variability. Whilst these questions have already been addressed by large quantitative comparisons (Chen et al. 2019), our purpose was to understand how the modality might influence judgemental *processes*. By exploring participants' experiences and perceptions and comparing across several iterations our method was able to offer insight into this phenomenon.

Participants (both students and examiners) were self-selected volunteers, the scenarios were simulated (and therefore low-stakes) and no students reported significant assessment anxiety. We can't exclude the possibility that videoing would be more intrusive in higher stakes settings. Further research is needed to explore this potential. We sampled across a diverse range of students and examiners. Whilst we didn't find any suggestion of differential effects of videoing on any group of students or examiners, we can't exclude the possibility that video-based assessment could operate differently for groups of students or examiners outside of our sample.

Suggestions for future research

Further research should seek to replicate these findings in other contexts, with other groups of participants. Survey research could determine whether our participants' perceptions are shared more widely amongst students and examiners. Given that some prior research has suggested that video-based performances may be remembered differently to live performances (Ihlebaek et al. 2003; Landström et al. 2005), future research could determine whether video-based performances obtain greater prominence in assessors recollections. This could be important in longitudinal forms of assessment. Research should determine how emerging uses of video in assessment (remote examining, benchmarking, video-based feedback, video-based score comparisons) enhance assessments or contribute evidence towards their validity.

Conclusions

Whilst video offers the potential to enhance assessment through several novel means, its implementation requires care. Educators should thoughtfully balance the intended purpose of videoing; the content-specific need to ensure information adequacy for examiners; and the potential for filming to disrupt assessment performance, to produce a compromise which enhances assessment validity and supports students' learning.

Acknowledgements

We would like to acknowledge the support of the Faculty of Health Information Technology team, the Clinical Skills team, and the assessments team at Keele University School of Medicine, as well as the examiners and students who participated in the study.

Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

Funding

Peter Yeates is funded by a National Institute for Health Research (NIHR) Clinician Scientist Award. This paper presents independent research funded by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Glossary

Video-based assessment: Observing a student or trainees' performance on a screen after it has been filmed and video recorded, rather than observing it in-person whilst it occurs.

Equivalence: The process of reaching the same assessment judgement under one set of circumstances as would have been reached under a different set of circumstances. In some circumstances this may account for contextual influences on performance.

Notes on contributors

Peter Yeates, MRCP, PhD, is a senior lecturer in medical education research and a consultant in acute and respiratory medicine. His interests focus on assessor cognition and technology-enhanced assessment.

Alice Moult, MSc, PhD, is a post-doctoral research assistant. Her interests focus on assessment in medical education and patient inclusion within education.

Janet Lefroy, MRCGP, PhD, is a senior lecturer and lead for consultation skills at Keele School of Medicine, and a general practitioner. She is interested in developing students' consultation skills, transitions between phases of healthcare professionals' careers, feedback and assessment.

Jacqualyn Walsh-House, BSc, is a year 4 undergraduate medical student. Her interests include assessment in medical education, and surgical education.

Lorraine Clews, MRCM, MA, PGCME, is a consultations and clinical skills tutor at Keele School of Medicine, who was voted clinical teacher of the year in 2019. Her interests focus on approaches to clinical skills teaching and student motivation.

Robert McKinley, MD, FRCGP, FRCP, is an emeritus professor of education in primary care and retired general practitioner. His interests include education in primary care, feedback, consultation skills, transitions between phases in healthcare professionals' education, assessment and assessor cognition.

Richard Fuller, MA, FRCP, is a professor of medical education and Deputy Dean of the School of Medicine at University of Liverpool and an Honorary Consultant Stroke Physician. His interests include innovation in assessment and technology-enhanced assessment methods.

ORCID

Peter Yeates http://orcid.org/0000-0001-6316-4051
Alice Moult http://orcid.org/0000-0002-9424-5660
Janet Lefroy http://orcid.org/0000-0002-2662-1919
Robert McKinley http://orcid.org/0000-0002-3684-3435
Richard Fuller http://orcid.org/0000-0001-7965-4864

References

Amin Z, Boulet JR, Cook DA, Ellaway R, Fahal A, Kneebone R, Maley M, Ostergaard D, Ponnamperuma G, Wearn A, et al. 2011. Technology-enabled assessment of health professions education: Consensus statement and recommendations from the Ottawa 2010 conference. Med Teach. 33(5):364–369.

Anderson T, Shattuck J. 2012. Design-based research: a decade of progress in education research? Educ Res. 41(1):16–25.

Bakker A, van Eerde D. 2015. An introduction to design-based research with an example from statistics education. In Bikner-Ahsbahs A, Knipping C, Presmeg N, editors. Approaches to qualitative research in mathematics education. Advances in mathematics education. Dordrecht: Springer; p. 429–466.

Bargh JA, Chartrand TL. 1999. The unbearable automaticity of being. Am Psychol. 54(7):462–479.

Baumgartner E, Bell P, Brophy SP, Hoadley C. 2003. Design-based research: an emerging paradigm for educational inquiry. Educ Res. 32(1):5–8.

Bryant K, Charmaz A. 2019. The SAGE handbook of current developments in grounded theory. London: SAGE Publications.

Byrne A, Tweed N, Halligan C. 2014. A pilot study of the mental workload of objective structured clinical examination examiners. Med Educ. 48(3):262–267.

Charmaz K. 2006. Constructing grounded theory: a practical guide through qualitative analysis. London: SAGE Publications.

Chen T-C, Lin M-C, Chiang Y-C, Monrouxe L, Chien S-J. 2019. Remote and onsite scoring of OSCEs using generalisability theory: a three-year cohort study. Med Teach. 41(5):578–583.

Cobb P, Confrey J, diSessa A, Lehrer R, Schauble L. 2003. Design experiments in educational research. Educ Res. 32(1):9–13.

Eeckhout T, Gerits M, Bouquillon D, Schoenmakers B. 2016. Video training with peer feedback in real-time consultation: acceptability and feasibility in a general-practice setting. Postgrad Med J. 92(1090): 431–435.

Engward H. 2013. Understanding grounded theory. Nurs Stand. 28(7): 37–41.

Eva KW. 2018. Cognitive influences on complex performance assessment: lessons from the interplay between medicine and psychology. J Appl Res Memory Cogn. 7(2):177–188.

Foulsham T, Walker E, Kingstone A. 2011. The where, what and when of gaze allocation in the lab and the natural environment. Vision Res. 51(17):1920–1931.

Galletta A. 2013. Mastering the semi-structured interview and beyond: From research design to analysis and publication. New York (NY): NYU press.

Gauthier G, St-Onge C, Tavares W. 2016. Rater cognition: review and integration of research findings. Med Educ. 50(5):511–522.

Gill P, Stewart K, Treasure E, Chadwick B. 2008. Methods of data collection in qualitative research: interviews and focus groups. Br Dent J. 204(6):291–295.

Gingerich A, Regehr G, Eva KW. 2011. Rater-based assessments as social judgments: rethinking the etiology of rater errors. Acad Med. 86(10 Suppl):S1–S7.

Gingerich A, Schokking E, Yeates P. 2018. Comparatively salient: examining the influence of preceding performances on assessors' focus and interpretations in written assessment comments. Adv Health Sci Educ Theory Pract. 23(5):937–959.

Govaerts MJB, Schuwirth LWT, Van der Vleuten CPM, Muijtjens AMM. 2011. Workplace-based assessment: effects of rater expertise. Adv Health Sci Educ Theory Pract. 16(2):151–165.

Guba EG, Lincoln YS. 1982. Epistemological and methodological bases of naturalistic inquiry. Educ Commun Technol. 30(4):233–252.

Gullberg M, Holmqvist K. 2006. What speakers do and what addressees look at: visual attention to gestures in human interaction live and on video. P&C. 14(1):53–82.

Van Den Haak M, De Jong M, Schellens PJ. 2003. Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. Behav Inf Technol. 22(5):339–351.

Hance J, Aggarwal R, Stanbridge R, Blauth C, Munz Y, Darzi A, Pepper J. 2005. Objective assessment of technical skills in cardiac surgery. Eur J Cardio-Thoracic Surg. 28(1):157–162.

Harrison CJ, Könings KD, Schuwirth L, Wass V, van der Vleuten C. 2015. Barriers to the uptake and use of feedback in the context of summative assessment. Adv Health Sci Educ Theory Pract. 20(1): 229–245.

Harrison CJ, Könings KD, Schuwirth LWT, Wass V, van der Vleuten CPM. 2017. Changing the culture of assessment: the dominance of the summative assessment paradigm. BMC Med Educ. 17(1):1–14.

Holton JA. 2010. The coding process and its challenges. Ground Theory Rev. 9(1):21–40.

Ihlebaek C, Løve T, Eilertsen DE, Magnussen S. 2003. Memory for a staged criminal event witnessed live and on video. Memory. 11(3): 319–327.

Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. 2011. Opening the black box of clinical skills assessment via observation: a conceptual model. Med Educ. 45(10):1048–1060.

Koivisto J-M, Haavisto E, Niemi H, Haho P, Nylund S, Multisilta J. 2018. Design principles for simulation games for learning clinical reasoning: a design-based research approach. Nurse Educ Today. 60: 114–120.

Landström S, Granhag PA, Hartwig M. 2005. Witnesses appearing live versus on video: effects on observers' perception, veracity assessments and memory. Appl Cognit Psychol. 19(7):913–933.

Lincoln YS, Guba EG. 1985. Naturalistic inquiry. Thousand Oaks (CA): Sage Publications.

McManus B, Omer S. 2017. A tiered approach in mandatory assessment training. Med Educ. 51(5):548–549.

Montgomery P, Bailey PH, Bailey PH. 2007. Field notes and theoretical memos in grounded theory. West J Nurs Res. 29(1):65–79.

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

Nguyen DT, Canny J. 2009. More than face-to-face: empathy effects of video framing. Proceedings of the 27th International Conference on Human Factors in Computing Systems - CHI '09; Boston (MA), USA. p. 423.

Papavlasopoulou S, Giannakos MN, Jaccheri L. 2019. Exploring children's learning experience in constructionism-based coding activities through design-based research. Comput Hum Behav. 99(7491): 415–427.

Ryan AM, Daum D, Bauman T, Grisez M, Mattimore K, Nalodka T, McCormick S. 1995. Direct, Indirect, and Controlled Observation and Rating Accuracy from continual attention to the observation of others. J Appl Psychol. 80(6):664–670.

Scaffidi MA, Grover SC, Carnahan H, Yu JJ, Yong E, Nguyen GC, Ling SC, Khanna N, Walsh CM. 2018. A prospective comparison of live and video-based assessments of colonoscopy performance. Gastrointest Endosc. 87(3):766–775.

Scott DJ, Rege RV, Bergen PC, Guo WA, Laycock R, Tesfay ST, Valentine RJ, Jones DB. 2000. Measuring operative performance after laparoscopic skills training: edited videotape versus direct observation. J Laparoendosc Adv Surg Tech. 10(4):183–190.

Simons D. 2010. Selective attention test, YouTube. [accessed 2020 Mar 11]. https://www.youtube.com/watch?v=vJG698U2Mvo.

Simons DJ, Chabris CF. 1999. Gorillas in our midst: sustained inattentional blindness for dynamic events. Perception. 28(9):1059–1074.

Strauss A, Corbin J. 1998. Basics of qualitative research techniques and procedures for developing grounded theory. 2nd ed. Thousand Oaks (CA): Sage.

Tavares W, Eva KW. 2014. Impact of rating demands on rater-based assessments of clinical competence. Educ Prim Care. 25(6):308–318.

Varpio L, Ajjawi R, Monrouxe LV, O'Brien BC, Rees CE. 2017. Shedding the cobra effect: problematising thematic emergence, triangulation, saturation and member checking. Med Educ. 51(1):40–50.

Vivekananda-Schmidt P, Lewis M, Coady D, Morley C, Kay L, Walker D, Hassell AB. 2007. Exploring the use of videotaped objective structured clinical examination in the assessment of joint examination skills of medical students. Arthritis Rheum. 57(5):869–876.

Wang J, Marchant D, Morris T, Gibbs P. 2004. Self-consciousness and trait anxiety as predictors of choking in sport. J Sci Med Sport. 7(2): 174–185.

Wood TJ. 2014. Exploring the role of first impressions in rater-based assessments. Adv Health Sci Educ Theory Pract. 19(3):409–427.

Yeates P, O'Neill P, Mann K, Eva KW. 2012. Effect of exposure to good vs poor medical trainee performance on attending physician ratings of subsequent performances. JAMA. 308(21):2226–2232.

Yeates P, O'Neill P, Mann K, Eva K. 2013a. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. Adv Health Sci Educ Theory Pract. 18(3):325–341.

Yeates P, O'Neill P, Mann K, Eva KW. 2013b. "You're certainly relatively competent": assessor bias due to recent experiences. Med Educ. 47(9):910–922.

Yeates P, Cope N, Hawarden A, Bradshaw H, McCray G, Homer M. 2019. Developing a video-based method to compare and adjust examiner effects in fully nested OSCEs. Med Educ. 53(3): 250–263.