Main title:

A realist evaluation of how, why and when Objective Structured Clinical Exams (OSCEs) are experienced as an authentic assessment of clinical preparedness.

Running Head:

Engaging fiction with practice?

Authors

Peter Yeates

Adriano Maluf

Ruth Kinston

Natalie Cope

Kathy Cullen

Aidan Cole

Vikki O’Neill

Ching-wa Chung

Rhian Goodfellow

Rebecca Vallender

Sue Ensaff

Rikki Goddard-Fuller

RK McKinley

Geoff Wong

**Abstract:**

Introduction: Whilst rarely researched, the authenticity with which Objective Structured Clinical Exams (OSCEs) simulate practice is arguably critical to making valid judgements about candidates’ preparedness to progress in their training. We studied how and why an OSCE gave rise to different experiences of authenticity for different participants under different circumstances.

Methods: We used Realist evaluation, collecting data through interviews/ focus groups from participants across four UK medical schools who participated in an OSCE which aimed to enhance authenticity.

Results: Several features of OSCE stations (realistic, complex, complete cases, sufficient time, autonomy, props, guidelines, limited examiner interaction etc) combined to enable students to project into their future roles, judge and integrate information, consider their actions and act naturally. When this occurred, their performances felt like an authentic representation of their clinical practice. This didn’t work all the time: focusing on unavoidable differences with practice, incongruous features, anxiety and preoccupation with examiners’ expectations sometimes disrupted immersion, producing inauthenticity.

Conclusions: The perception of authenticity in OSCEs appears to originate from an interaction of station design with individual preferences and contextual expectations. Whilst tentatively suggesting ways to promote authenticity, more understanding is needed of candidates’ interaction with simulation and scenario immersion in summative assessment.

Key Words 3-5

* Assessment
* OSCE
* Simulation
* Authenticity
* Realist Evaluation

Practice Points (up to 5)

OSCEs comprise a simulated assessment of candidates’ clinical skills. The authenticity with which OSCEs simulate practice arguably contributes to the validity of the assessment.

Authenticity within an OSCE cannot be assumed but appears to emerge from the interaction of station design, individual preferences and contextual expectations of assessment.

Several actions can be tentatively recommended to promote OSCE authenticity, but they require considered application in practice as their influence depends on context.

Enabling candidates to demonstrate an authentic representation of their usual clinical performance within OSCEs may enhance fairness whilst promoting greater engagement in clinical learning.

Introduction

Objective structured clinical exams (OSCEs) are a key component of many assessment programmes within health professionals’ education internationally because of their ability to provide comparable, blueprinted assessments for high-stakes purposes (Boursicot et al., 2021). Whilst beneficial, one observation is inescapable: OSCEs assess candidates’ performance within simulated rather than real practice (Adamo, 2003). Simulated assessment requires that students, simulated patients, and examiners engage in a “fiction contract” (Dieckmann et al., 2007): they need to suspend disbelief in order to inhabit the simulated scenario, particularly if it departs from the reality of their prior clinical experience. When OSCEs are used to determine whether learners are ready to progress into the next phase of training or practice, their validity depends (at least in part) on how authentically they represent the typical work, judgement, skills, and behaviours from the practice into which candidates will progress (Hodges, 2003b). As per Lavoie et al (2020) we assert that whilst authenticity encompasses elements of fidelity (how well the scenario replicates practice), it also includes how learners interact with the scenarios and their subjective responses to them. . Relating this to Kane’s model of validity (Cook et al., 2015), if the OSCE poorly evokes how a candidate will think or behave in practice then the extrapolation domain of the validity argument will be challenged.

Despite the importance of authenticity to OSCEs’ validity, there has been little research on this topic. The majority of OSCE literature is focused on their psychometric properties (Swanson & van der Vleuten, 2013). Commonly, content evidence is sought as part of the evidentiary chain which supports their interpretation for a particular purpose (Downing, 2003; Kane, 2013), but this evidence typically involves establishing how comprehensively the stations within the OSCE sample a blueprint (Kreptul & Thomas, 2016). Whilst content alignment and sampling are critical, it fails to ‘assess authenticity’ at the level of station design.

The small body of literature on authenticity within OSCEs has highlighted a number of themes. Fragmentation of clinical tasks (Nasir et al., 2014; Gupta et al., 2010; van der Vleuten et al., 2010), or scenarios which poorly represent practice (either through unrealistic patient presentations or because the actions expected of candidates are incongruent with usual responses in clinical practice (Hyde et al., 2022)) contribute to inauthenticity. Poor ‘task time agreement’ (i.e. time pressure) causes candidates to hurry, altering their interaction with (simulated) patients (Gormley et al., 2016) and/or simply preventing the testing of many realistic scenarios (Marwaha, 2011). Simulated patients (SP) portrayals of cases may be inauthentic, either because SPs comply with examination instructions more readily than is typical of real patients or because scenario scripting reflects professional stereotypes of patients (Gormley et al., 2016; Hyde et al., 2022). “The OSCE” itself has been conceptualised as theatre (Gormley et al., 2016; Hodges, 2003a), in which the triadic interaction between candidate, examiner and simulated patient produces a ritualised and rehearsed show which interferes with more authentic interaction between the candidate and simulated patient. Moreover, candidates’ expectations around how stations are ‘scored’ may produce perverse incentives to inauthentic behaviour (Gormley et al., 2016). Harrison (2018) describes lack of agency by candidates within observed scenarios which inhibits them from displaying their mastery of a skill. More concerningly, learners’ avoidance of clinical learning prior to OSCEs (Rudland et al., 2008) may emanate from misalignment between authentic practice and their sometimes-formulaic expectations of OSCEs.

Whilst these critiques might challenge the utility of OSCEs, we, like others (Chan and Rashid, 2023), assert that there are several reasons to persevere. Firstly, workplace-based assessments are not without challenges (Watling et al., 2016, Schumacher et al., 2022, LaDonna et al., 2017, Phinney et al., 2022, Melvin et al., 2019, Yepes-Rios et al., 2016, Sebok-Syer et al., 2018). Secondly, simulated assessment enables testing of (particularly undergraduate) preparedness for scenarios which candidates would rarely or never manage in practice. Thirdly, OSCEs demonstrate a commitment to equivalent and therefore comparable testing of candidates. Nevertheless, OSCEs' contribution to programmes of assessment would be enhanced if their authenticity could be improved. Given that relatively little work has considered how the authenticity of OSCEs might be enhanced, we explored how students’ and examiners’ interaction with station design and OSCE conduct may give rise to authenticity (or its absence) within OSCEs. To do this we focused on a comparatively novel approach to station design in use at one of the participating institutions, whilst acknowledging that similar approaches are used in different ways by many institutions. For reference, these were called “authentic OSCE stations” (see methods section). We then asked the question, “How do candidates and examiners use and interact with authentic OSCE stations and how, why, under what circumstances, and for whom does this influence the perception of the OSCE being an authentic test of candidates’ preparedness for their future roles?

Methods:

Data were collected within the AD-Equiv study (Yeates et al., 2022), which aimed to enhance the authenticity, diagnostic accuracy and equivalence of a distributed graduation-level OSCE. We adopted a Realist ontology and epistemology which holds that whilst mind-independent realities exist, knowledge of these realities is constructed by observing and deducing how different contexts interact with particular mechanisms to produce different outcomes for different people (Pawson & Tilley, 1997). We adopted the Realist evaluation approach (Duddy & Wong, 2022; Trisha Greenhalgh et al., 2017) gathering data through realist interviews (Greenhalgh et al., 2017; Manzano, 2016) and focus groups from students, examiners and simulated patients who had recently experienced authentic OSCE stations. Focus groups were used immediately post OSCEs to explore students’ perspectives whilst the experience remained recent, using the interplay to explore why different students responded differently. Interviews, conversely, allowed more in-depth individual responses and reflections with hours to a few weeks hindsight. A comparison of our method with the RAMESES II checklist (Wong et al., 2016) is provided in Appendix 3.

The intervention: introducing ‘authentic’ OSCE stations:

Authentic OSCE stations were designed using the following principles: stations should feature a scenario from the typical work of UK foundation doctors (PGY1-2); (as far as practical) candidates should complete the whole patient encounter; station information should describe the clinical context, and (rather than instructing to perform specific actions) students should use their clinical judgement to decide and then do what they would in practice. Interaction with the examiner is kept to a minimum to encourage agency by students. Pilot work ensured task-time agreement for these newer authentic stations. Appendix 2 contains further details of the OSCE.

Population, sampling and recruitment:

Our study population was participants (candidates, examiners, and simulated patients) in UK undergraduate OSCEs for students in the latter stages of undergraduate training. We sampled purposively (aiming for diversity of geography and schools) from four UK medical schools (one from each of the UK’s nations England, Northern Ireland, Scotland and Wales) including established and comparatively new medical schools, from 4th (penultimate) and 5th (final) year in this these schools. Participants were recruited via email and verbal invitation.

Initial Programme Theory

A central aspect of any theory-driven realist investigation is to develop an initial program theory (IPT). An IPT can be used to frame and understand how, for whom, why, and under what contexts an intervention is expected to work (Duddy & Wong, 2022).

Our IPT posited that the combination of elements in authentic OSCE stations would collectively increase authenticity for candidates because they would enable demonstration of students’ ability to work through a clinical task in a manner more closely representing routine clinical practice. We expected this to be mediated by candidates’ understanding of the task and the degree of challenge the stations presented.

Data collection:

Interviews were performed via Microsoft Teams within 3 months of the OSCE. Student focus groups were performed in-person immediately after the OSCE. All data was audio recorded and transcribed.

Researchers used a question template which explored comparisons between participants’ perceptions of authentic OSCE stations and their usual OSCE station format, focusing on how and why elements of the intervention (content, timing, stimulus materials, whole tasks etc) evoked different responses and produced different outcomes for different stations, different people or within different schools. Consistent with recommendations for realist interviewing (Greenhalgh et al., 2017), researchers adapted their questions to pursue emergent issues and test elements of the evolving programme theory based on their impressions from recent analysis. Throughout interviews, researchers sought to establish a safe, inclusive and confidential environment.

Analysis:

Data were analysed using a realist logic of analysis (Pawson & Tilley, 1997; The RAMESES II Project, 2017). Data analysis progressed in parallel to data collection and informed subsequent interview conduct. Two researchers (PY and AM) read portions of the data for familiarisation and to discuss sensitising concepts (i.e. whole tasks, time, fragmentation, agency etc) in relation to the IPT. One researcher (AM) then coded all data, inductively and theoretically labelling concepts which appeared relevant to the IPT. Ancillary coding was performed by PY to further aid sensitisation. Codes were gathered into similar categories. Researchers scrutinised these categories looking for relevant contexts, mechanisms and outcomes which were then organised into context-mechanism-outcome configurations (CMOCs), using the analytic processes of juxtaposition (comparing data from different sources to find patterns), reconciliation (resolving discrepancies between data sources), adjudication (weighing the evidence to decide on the most plausible explanations), and consolidation (integrating findings into a coherent whole) (Papoutsi et al., 2018). As CMOCs were produced, researchers considered their relevance to the evolving programme theory. GW provided additional, data-driven critique of the CMOCs as they were refined. Analysis progressed until the researchers judged that plausible CMOCs had been developed, which aligned with demi-regularities (semi-recurrent outcome patterns) within the data and a mature programme theory had been produced.

Reflexivity: Our research team included clinical (PY, RK, KC, AC, CC, RG, SE, RF, RMK, GW) and non-clinical (NC, VON, RV) OSCE experts, including people who have developed the authentic OSCE format at Keele (RK, NC) and people from medical schools which use different OSCE formats (KC, AC, VON, CC, RG, RV, SE). It also included Realist evaluation experts (GW) and other researchers with little experience of OSCEs (AM). Consequently, whilst the team was interested in the potential benefits of authentic OSCEs, this diversity produced contrasting views of the benefits and challenges of elements of the intervention (for example the balance of giving students instructions vs enabling autonomy) and challenges and refinements to CMOC description.

Ethics:

All participation in the study was voluntary, and participants had the right to withdraw. Informed consent was documented through electronic consent forms. All data were pseudonymised and treated as confidential. Neither of the interviewers (PY or AM) had direct course responsibility for students or their progression. Ethical approval for the study was granted by Keele Research Ethics Committee (ref MH-210209).

**Results**

The relevant portions of 35 interviews and 3 focus groups were included, comprising a total of 15 candidates, 13 examiners and 7 simulated patients. Our analysis produced 13 CMOCs, each providing a data-driven theoretical claim about the mechanistic operation of a facet of the intervention in a given context. Individual CMOCs with supporting evidence can be seen in Appendix 1 and are summarised in Table 1. Here we present a narrative of the programme theory which they support, referencing the relevant CMOCs within the text along with some abbreviated illustrative quotes.

In this study, many of the scenarios we created enabled students to demonstrate what felt, for them, like a very authentic representation of how they would typically behave in practice. This therefore felt like a fair means to assess their readiness to progress in practice. For them, this improved on existing assessments. This didn’t, however, work for everyone or in every station.

The degree to which students and examiners experienced the scenarios as authentic was influenced by a number of factors which interacted with each other differently in each scenario. Several of these were characteristics of the station design and emerged from the principles of authentic OSCE stations, which underpinned the intervention. Some were person-related and others related to peoples’ norms and expectations, or more broadly the culture within the assessment.

**When Authenticity was enhanced**

When scenarios effectively represented typical work of F1 doctors, this gave the scenario credibility (CMOC\_1a\_TW) and situated students thinking in the clinical environment because it aligned with students’ experiences and expectations of their future roles (CMOC\_1b\_TW).

*I also liked the presentations that the patients had were things that you would have to manage as an F1 [okay], it wasn’t like spotting signs of something you know, it was very much like acute presentation, managing the patient you know, things that you do to stabilise them before you would refer them onto the next team so, I thought that was quite useful. (24th AD equiv, Focus Group Site4, Students)*

This in turn caused them to project themselves into that role which enabled them to think and behave as they would in practice (CMOC\_1c\_TW). This helped them to act naturally. When clinical problems were rich and complex enough to be unlikely to be rehearsed, and to require nuanced approaches, students found that that standard answers weren’t appropriate, and students had to draw from their experience to propose solutions (CMOC\_1b\_TW and CMOC\_2\_WT). Conversely, less complex or more predictable scenarios were more likely to evoke rehearsed responses. Asking students to complete the whole clinical task gave them space to integrate and follow through on the information they chose to gather. This tapped into their evolving clinical acumen, whilst avoiding fragmentation, so offered a better representation of what students would actually do in practice (CMOC\_2\_WT).

*This* [referring to the authentic OSCE] *is the sort of task that I’m going to have to be able to do altogether in a few months’ time… it’s, it’s reflected in kind of like my ward placements as well. … I know how to take a, a fairly comprehensive history. I know how to examine a patient. It’s – but it’s the – it’s putting it altogether and it’s the, the decision-making that comes during that and afterwards as well; … it’s the putting it altogether and the, the, the synthesising all of the information.* (23rd AD-Equiv, interview, site3 Stud6).

This was further aided by both the comparatively limited direction about what actions to perform and by avoiding prompting by examiners; students had to work out what to do, and their choices had consequences for the outcome of the station. This forced students to make judgements about what was appropriate (CMOC\_3a\_LP), which advantaged students who had engaged more in the clinical environment as they could draw from their experience to integrate information and direct their actions (CMOC\_3b\_LP).

*it helped me see the station more as something like real-life … I could go in and kind of make my own plan and, and go in my own direction. I suppose it maybe adds to a little bit of uncertainty because you don’t know if you’re going down the right track or not. … knowing that even if I was doing something wrong or, … no one would interrupt to tell me … so it was definitely a lot more like, like real-life* (23rd AD-Equiv, interview, Site3 Stud6)

Providing sufficient time aided this as it enabled students to think and integrate without either panicking or defaulting to rehearsed routines (CMOC\_4\_Ti).

*it made me kind of be a bit more realistic because it’s kind of like, ‘Okay. I don’t have to worry about, you know, seven-minute timer, eight-minute timer’. It’s like I can actually think and not just regurgitate* (16th AD-Equiv, interview, site 3 stud3)

Supplying relevant resources (such as guidelines) avoided artificial reliance on memory and let students show their familiarity with accessing information. This further aligned with what they would do in practice (CMOC\_5\_Tr). Use of simple props (i.e. a telephone) made scenarios look and feel more realistic which further aided students' immersion and therefore helped them to act naturally (CMOC\_6\_Pr). When scenarios required students to demonstrate actions (for example using a probe to measure oxygen saturations rather than simply stating an intention to measure them), this added further to the assessments’ authenticity because it tested relevant workplace skills rather than just knowledge (CMOC\_7\_Dem). This contributed to students’ projection into their future roles and their ability to demonstrate their preparedness.

**When Authenticity was challenged**

This didn’t work for everyone. For example, a minority of students and examiners found it hard to move beyond the fact the actor was not a patient and that the students’ actions would have no impact on them. This (or any) focus on unavoidable differences from practice reduced their engagement in the scenarios’ fiction and therefore made the assessment feel like an artificial test rather than a useful representation of the students’ practice (CMOC\_8\_DFC).

*You know there’s someone watching you, so you’re kind of doing things to look like you’re doing things, rather than to really be looking for what the problem is, whereas in a real patient, you don’t bother with any of the OSCE faff. You just get straight to what’s important* (11th AD-Equiv, interview Site3 st1)

The way in which scenarios were enacted also influenced students’ immersion. When elements were incongruous with real practice (i.e. a simulated patient looked comfortable when they were supposed to be seriously unwell), this could lead students to make assumptions or feel confused, which could impact their performance (CMOC\_9\_Inc), but perhaps more importantly presented a dilemma about where the limits lay of doing what they should in practice versus demonstrating a skill in an artificial way. Where scenarios could legitimately be approached in different ways (for example a focused versus more systematic physical examination; different reasonable management approaches) students and examiners were unsure what to expect or how to behave, because they struggled to reconcile differences between approaches which may be acceptable in practice and typical expectations of formal exams (CMOC\_10\_LDP). This tended to further reduce students’ inclination to show their authentic clinical performance.

*I love the idea of talking to the patient naturally and naturally leading onto an examination without being artificially put into that scenario … [but] there is a set marking scheme, there is a set thing we need to do … I was trying to think, okay, what do the examiners want from me?**[okay] if I don’t do the clubbing, would I get marked down?* (24th AD Equiv focus group, site4 students)*.*

Some students struggled with the lack of precise direction about what actions to perform; they were used to demonstrating specified skills and the need to use judgement made them anxious and made it hard for them to behave authentically (CMOC\_11a&b\_Anx).

Students described OSCEs which impose ritualistic inauthentic expectations as reducing their engagement with clinical learning environments because their clinical learning isn’t rewarded in OSCEs (CMOC\_12\_Rit). Conversely, students perceived that the more that OSCEs reward demonstration of authentic practice, the more they would prepare by spending time in clinical environments resulting in greater preparedness for practice (CMOC\_13a&b\_IOL).

*So maybe like by making the OSCE a bit more realistic and sort of something we can learn by being on the ward. You know, it might encourage us to prepare for practice a bit more. (8th AD-Equiv, focus group, site2 students)*

**Discussion:**

Summary of programme theory

Most students, when asked to respond to “authentic” scenarios recognised the applicability of the scenarios to their future roles and were able to act naturally and use their evolving clinical acumen to demonstrate an authentic representation of their preparedness for practice. This recognition was promoted when scenarios were typical of the work of new doctors, presented as complete tasks, and which required students to use their judgement to gather and integrate information to reach judgements and then communicate with and manage the patient. Students needed sufficient time to think and avoid rushing. Props and resources which were typical of the situation aided immersion. Students who had been more engaged in clinical practice recognised the applicability of scenarios to their future roles and were able to act more naturally.

Some circumstances challenged authenticity: when students or examiners focused on unavoidable differences with practice, or when incongruous elements of station implementation produced assumptions or confusion for students, or where students were preoccupied with the examiners’ expectations, this tended to inhibit immersion. These contexts caused students to question how naturally to behave, which produced anxiety and reduced the authenticity of the assessment for both students and examiners.

Theoretical Interpretation

Our findings suggest that authenticity in OSCEs critically depends on how effectively candidates (and to a lesser, but still important degree) examiners and simulated patients suspend their disbelief and immerse themselves in the “fiction contract” (Dieckmann et al., 2007) required by the scenario. Several of our findings resonate with literature on fidelity in simulation: the use of cases drawn from real practice (Maclean et al., 2019), the need to actually perform (rather than state) tasks (Engström et al., 2016), the use of environmental cues (i.e. props) (Nanji et al., 2013) and the chance for the scenario to evolve based on the student’s actions (Marei et al., 2018). What our work adds is the observation that these can combine within an OSCE scenario in a way which enables students to project into their future role and act (and therefore demonstrate) their readiness to progress to the next stage of clinical practice. Moreover, students’ desire to do well in an assessment context sometimes appeared to conflict with their ability to be immersed, suggesting that immersion is more complex in assessment scenarios than when simulation is focused on learning. This may partly be because people find it hard to behave naturally whilst being observed (i.e. the Hawthorne Effect) (Adair, 1984), a phenomenon which also occurs in workplace-based assessment (Watling et al., 2016). The assessment context may also make it harder for students to judge how far to suspend disbelief (can I trust my visual gestalt about a “patient” or do I need to demonstrate the steps I would take with a sicker patient in order to obtain marks?). Whilst this relates to Dieckmann’s (2007) idea of shared rules it shows that in an assessment context a conflict may arise between allowing enough autonomy to behave authentically and fairly communicating expectations. Consequently, whilst many principles of simulation fidelity are well established, we have shown how their application in an assessment must be more complex.

Immersion in the scenarios appeared to work less well for some people than others. Our data aren’t able to explain this observation fully and further investigation is warranted. Further explanations could arise from individual differences in abstract thinking ability or imagination. In other literature, differences in peoples’ ability to be immersed in virtual reality has been attributed to variation in spatial awareness (Coxon et al., 2016), whilst differences in engagement with a story’s narrative has been attributed to differences in emotional sensitivity (Samur et al., 2021). Consequently, there may be some individuals for whom engagement with the fiction contract is simply more difficult. This suggests the need for further exploration around the validity implications of OSCEs for such people.

Practical Implications:

Based on our findings, whilst acknowledging the need for replication in summative contexts, we tentatively suggest a number of principles which may be used to enhance the authenticity of assessment in OSCEs (see box 2). As with all complex interventions, implementing these principles will require careful consideration within the context they are used. The suggestion to allow sufficient time may result in fewer, longer stations. This may reduce sampling of the OSCE blueprint, which has itself been proposed as a major source of validity evidence (Wass et al., 2001). The comparative reliability of these stations is not yet well established and is hard to predict. On one hand, fewer independent observations may reduce reliability (Eva, 2018) but judgements which align more with clinicians’ experience of practice could support reliability (Crossley, 2010) whilst focusing the assessment more on complex problem solving skills that are typical features of contemporary taxonomies of learning (Dubas & Toledo 2016). This debate is not new; Cook et al (2015) note a tension between what Kane (2013) describes as extrapolation (using the score[s] as a reflection of real-world performance) and generalisation (using the score[s] as a reflection of expected performance in all possible test settings), or what has sometimes historically been referred to simply as validity versus reliability (van der Vleuten 1996). Authenticity of assessment tasks is key to extrapolating from the current observation to likely behaviour in practice. Consequently, we assert that by thoughtfully applying the principles we have proposed, educators may increase the ability to extrapolate from OSCEs to likely performance in real life. Further work will be required to balance this against impact on generalisation.

Given the practice-focused nature of these questions, we recommend that both station writing and examination should ideally be performed by people who are highly familiar with current work and expectations of new doctors’ roles.

Strengths and Limitations:

Our findings are supported by data which was sampled from four different institutions and included a large number of participants using Realist evaluation to study the operation of this station format across different contexts. Despite this, the study has some limitations. Firstly, it is limited by the voluntary nature of students’ and examiners’ participation whose perspectives may not have fully represented wider participants. The assessment was conducted in a formative context which could have produced different behaviour or reactions by participants compared with a summative setting. Consequently, it would be useful to determine if our findings hold true in routine summative settings. Our findings are limited by the sample of stations we used. Whilst these encompassed a representative range of tasks, further samples of stations designed along the same principles could potentially produce different results. Students could potentially have encountered the simulated patients in this study at previous times during their studies when the simulated patients were playing other roles. As we did not collect data on this, we cannot comment on the degree to which this may have influenced authenticity. This would benefit from exploration in future work.

Future Research:

Future research should examine how these principles operate across different contexts, including summative and postgraduate assessments and when applied to a wide range of other OSCE station formats. Further research should seek further contexts where the principles may achieve different outcomes or provide further insight into some of the remaining challenges, such as how to manage legitimately different approaches to the same task, different formats of station instruction, or how to avoid incongruous elements in station design. Theoretical work should explore how candidates, examiners and simulated patients engage with the fiction contract within summative assessment and its implications for assessment validity and outcomes for individuals.

Conclusions:

The validity with which OSCEs can inform decisions on students’ preparedness to enter practice is (at least partly) dependent on how authentically the OSCEs simulate the practice into which students will progress. Responding to challenges to OSCEs’ authenticity, we have shown how the combination of design features used within “authentic OSCE stations” operated to increase assessment authenticity for the majority of students across a range of scenarios. The principles which we have developed represent a further advancement in our shared understanding of good assessment design and we encourage their use to enhance assessment authenticity within OSCEs.

Acknowledgements:

We would like to acknowledge and thank all students, simulated patients, examiners, assessment and administrative staff in each of the four participating medical schools who contributed to the study.

Conflicts of Interest:

All authors confirm that they have no conflicts to declare

Funding

Peter Yeates was funded through a National Institute for Health Research (NIHR) Clinician Scientist award. The study constitutes independent research and does not represent the views of the NIHR, the NHS or the department of health and social care.

Author contributions

Peter Yeates: substantial contribution to study conception, design, data collection analysis and drafting

Adriano Maluf: substantial contribution to study conception, design, data collection and analysis

Ruth Kinston: substantial contribution to study conception, design, data collection and analysis

Natalie Cope: substantial contribution to study conception, design, data collection and analysis

Kathy Cullen: substantial contribution to study conception and data collection

Aiden Cole: substantial contribution to study conception and data collection

Vikki O’Neill: substantial contribution to study conception and data collection

Ching-wa Chung: substantial contribution to study conception and data collection

Rhian Goodfellow: substantial contribution to study conception and data collection

Rebecca Vallender: substantial contribution to study conception and data collection

Sue Ensaff: substantial contribution to study conception and data collection

Richard Fuller: substantial contribution to study conception and data analysis

RK McKinley: substantial contribution to study conception and data analysis

Geoff Wong: substantial contribution to study conception and data analysis

Data sharing:

Data are not available in a repository

References

Adair, J. G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. Journal of Applied Psychology, 69(2), 334–345.

Adamo, G. (2003). Simulated and standardized patients in OSCEs: Achievements and challenges 1992-2003. Medical Teacher, 25(3), 262–270.

Boursicot, K., Kemp, S., Wilkinson, T., Findyartini, A., Canning, C., Cilliers, F., & Fuller, R. (2021). Performance assessment: Consensus statement and recommendations from the 2020 Ottawa Conference. Medical Teacher, 43(1), 58–67.

Chan, S. C. C., & Rashid, M. A. (2023). The art of reinvention: The remarkable longevity of the OSCE. *Medical Education*. On-line ahead of print. https://doi.org/10.1111/medu.15266

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane’s framework. Medical Education, 49(6), 560–575.

Coxon, M., Kelly, N., & Page, S. (2016). Individual differences in virtual reality: Are spatial presence and spatial ability linked? Virtual Reality, 20(4), 203–212.

Crossley, J. (2010). Good Question, Good Answer: Construct alignment improves the performance of workplace based assessment scales. Medical Education.

Dieckmann, P., Gaba, D., & Rall, M. (2007). Deepening the theoretical foundations of patient simulation as social practice. In Simulation in Healthcare (Vol. 2, Issue 3, pp. 183–193).

Downing, S. M. (2003). Validity: on meaningful interpretation of assessment data. Medical Education, 37(9), 830–837.

Dubas, J. M., & Toledo, S. A. (2016). Taking higher order thinking seriously: Using Marzano’s taxonomy in the economics classroom. *International Review of Economics Education*, *21*, 12–20.

Duddy, C., & Wong, G. (2022). Grand rounds in methodology: when are realist reviews useful, and what does a ‘good’ realist review look like? BMJ Quality & Safety, bmjqs-2022-015236. https://doi.org/10.1136/bmjqs-2022-015236

Engström, H., Andersson Hagiwara, M., Backlund, P., Lebram, M., Lundberg, L., Johannesson, M., Sterner, A., & Maurin Söderholm, H. (2016). The impact of contextualization on immersion in healthcare simulation. Advances in Simulation, 1(1). https://doi.org/10.1186/s41077-016-0009-y

Eva, K. W. (2018). Cognitive Influences on Complex Performance Assessment: Lessons from the Interplay between Medicine and Psychology. Journal of Applied Research in Memory and Cognition, 7(2), 177–188.

Gormley, G. J., Hodges, B. D., McNaughton, N., & Johnston, J. L. (2016). The show must go on? Patients, props and pedagogy in the theatre of the OSCE. Medical Education, 50(12), 1237–1240.

Greenhalgh, T., Pawson, G., Westhorp, G., Greenhalgh, J., Manzano, A., & Jagosh, J. (2017). The Realist Interview: The RAMESES II Project. https://www.ramesesproject.org/media/RAMESES\_II\_Realist\_interviewing.pdf

Harrison, C. (2018). Aiming for agency and authenticity in assessment. In Perspectives on Medical Education; 7(6),348–349.

Hodges, B. (2003a). OSCE! Variations on a theme by Harden. Medical Education, 37, 1134–1140.

Hodges, B. (2003b). Validity and the OSCE. In Medical Teacher (Vol. 25, Issue 3, pp. 250–254). Carfax Publishing Company.

Hyde, S., Fessey, C., Boursicot, K., MacKenzie, R., & McGrath, D. (2022). OSCE rater cognition – an international multi-centre qualitative study. BMC Medical Education, 22(1).

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement, 50(1), 1–73. https://doi.org/10.1111/jedm.12001

Kreptul, D., & Thomas, R. E. (2016). Family medicine resident OSCEs: A systematic review. Education for Primary Care, 27(6), 471–477.

LaDonna, K. A., Hatala, R., Lingard, L., Voyer, S., & Watling, C. (2017). Staging a performance: learners’ perceptions about direct observation during residency. Medical Education, 51(5), 498–510.

Lavoie, P., Deschênes, M. F., Nolin, R., Bélisle, M., Blanchet Garneau, A., Boyer, L., Lapierre, A., & Fernandez, N. (2020). Beyond Technology: A Scoping Review of Features that Promote Fidelity and Authenticity in Simulation-Based Health Professional Education. In Clinical Simulation in Nursing Vol. 42, pp. 22–4).

Maclean, S., Geddes, F., Kelly, M., & Della, P. (2019). Realism and presence in simulation: Nursing student perceptions and learning outcomes. Journal of Nursing Education, 58(6), 330–338.

Manzano, A. (2016). The craft of interviewing in realist evaluation. Evaluation, 22(3), 342–360.

Marei, H. F., Al-Eraky, M. M., Almasoud, N. N., Donkers, J., & Van Merrienboer, J. J. G. (2018). The use of virtual patient scenarios as a vehicle for teaching professionalism. European Journal of Dental Education, 22(2), e253–e260.

Marwaha, S. (2011). Objective Structured Clinical Examinations (OSCEs), psychiatry and the Clinical assessment of Skills and Competencies (CASC)Same Evidence, Different Judgement. BMC Psychiatry, 11.

Melvin, L., Rassos, J., Panisko, D., Driessen, E., Kulasegaram, K. M., & Kuper, A. (2019). Overshadowed by Assessment: Understanding Trainee and Supervisor Perspectives on the Oral Case Presentation in Internal Medicine Workplace-Based Assessment. Academic Medicine, 94(2), 244–250.

Nanji, K. C., Baca, K., & Raemer, D. B. (2013). The effect of an olfactory and visual cue on realism and engagement in a health care simulation experience. Simulation in Healthcare, 8(3), 143–147.

Nasir, A. A., Yusuf, A. S., Abdur-Rahman, L. O., Babalola, O. M., Adeyeye, A. A., Popoola, A. A., & Adeniran, J. O. (2014). Medical students’ perception of objective structured clinical examination: A feedback for process improvement. Journal of Surgical Education, 71(5), 701–706.

Papoutsi, C., Mattick, K., Pearson, M., Brennan, N., Briscoe, S., & Wong, G. (2018). Interventions to improve antimicrobial prescribing of doctors in training (IMPACT): a realist review. Health Services and Delivery Research, 6(10), 1–136.

Pawson, R., & Tilley, N. (1997). Realistic Evaluation (1st ed.). Sage Publications Ltd.

Peter Yeates, Adriano Maluf, Ruth Kinston, Natalie Cope, Gareth McCray, Kathy Cullen, Vikki O’Neill, Aidan Cole, Rhian Goodfellow, Becky Vallander, Ching-Wa Chung, Lauren Cotes, Bob McKinley, Richard Fuller, & Geoff Wong. (2022). Enhancing Authenticity, Diagnosticity and Equivalence (AD-Equiv) in multi-centre OSCE exams in Health Professionals Education. Protocol for a Complex Intervention Study. BMJ Open.

Phinney, L. B., Fluet, A., O’Brien, B. C., Seligman, L., & Hauer, K. E. (2022). Beyond Checking Boxes: Exploring Tensions With Use of a Workplace-Based Assessment Tool for Formative Assessment in Clerkships. Academic Medicine, 97(10), 1511–1520.

Piyush Gupta, Pooja Dewan, & Tejinder Singh. (2010). Objective Structured Clinical Examination (OSCE) Revisited. Indian Pediatrics, 47, 911–920.

Rudland, J., Wilkinson, T., Smith-Han, K., & Thompson-Fawcett, M. (2008). “You can do it late at night or in the morning. You can do it at home, I did it with my flatmate” The educational impact of an OSCE. Medical Teacher, 30(2), 206–211.

Samur, D., Tops, M., Slapšinskaitė, R., & Koole, S. L. (2021). Getting lost in a story: how narrative engagement emerges from narrative perspective and individual differences in alexithymia. Cognition and Emotion, 35(3), 576–588.

Schumacher, D. J., Michelson, C., Winn, A. S., Turner, D. A., Elshoff, E., & Kinnear, B. (2022). Making prospective entrustment decisions: Knowing limits, seeking help and defaulting. Medical Education, 56(9), 892–900.

Sebok-Syer, S. S., Chahine, S., Watling, C. J., Goldszmidt, M., Cristancho, S., & Lingard, L. (2018). Considering the interdependence of clinical performance: implications for assessment and entrustment. Medical Education, 1–11.

Swanson, D. B., & van der Vleuten, C. P. M. (2013). Assessment of Clinical Skills With Standardized Patients: State of the Art Revisited. Teaching and Learning in Medicine, 25(S1), S17–S25.

The RAMESES II Project. (2017). Retroduction in realist evaluation. Nihr, p 207, 1–3. http://www.ramesesproject.org/media/RAMESES\_II\_Retroduction.pdf

Trisha Greenhalgh, Ray Pawson, Geoff Wong, Gill Westhorp, Joanne Greenhalgh, Ana Manzano, & Justin Jagosh. (2017). Quality Standards for Realist Evaluation For evaluators and peer-reviewers The RAMESES II Project. www.ramesesproject.org

van der Vleuten, C. P. M. (1996). The assessment of professional competence: Developments, research and practical implications. Advances in Health Sciences Education, 1(1), 41–67.

van der Vleuten, C. P. M., Schuwirth, L. W. T., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: Building blocks for theory development. In Best Practice and Research: Clinical Obstetrics and Gynaecology (Vol. 24, Issue 6, pp. 703–719).

Wass, V., Vleuten, C. Van Der, Shatzer, J., & Jones, R. (2001). Medical education quartet Assessment of clinical competence. The Lancet, 357, 945–949.

Watling, C., LaDonna, K. A., Lingard, L., Voyer, S., & Hatala, R. (2016). ‘Sometimes the work just needs to be done’: socio-cultural influences on direct observation in medical training. Medical Education, 50(10), 1054–1064.

Wong G, Westhorp G, Manzano A, Greenhalgh J, Jagosh J, Greenhalgh T. (2016) RAMESES II reporting standards for realist evaluations. BMC Medicine, 14(96), 1-18. DOI 10.1186/s12916-016-0643-1

Yepes-Rios, M., Dudek, N., Duboyce, R., Curtis, J., Allard, R. J., & Varpio, L. (2016). The failure to fail underperforming trainees in health professions education: A BEME systematic review: BEME Guide No. 42. Medical Teacher, 38(11), 1092–1099.

Tables and Figures

**Table 1: Titles of CMOCs**

|  |  |
| --- | --- |
| CMOC 1a-c | **Typical Work of new doctors (TW)**a/. When scenarios were drawn from the typical work of new doctors (C), students experienced them as more credible assessment of their clinical skills (O) because they aligned with their experiences of working in practice (M). b/. When OSCE stations depict scenarios which are typical of new doctors’ work (C) this tends to situate students’ thinking in their experience of practice (O) because students have to draw from observations and experiences they have learnt from in clinical workplaces to demonstrate how they might act in that scenario as a new doctor (M).c/. When OSCE stations depict clinical scenarios drawn from the typical work of new doctors (C), students project themselves into their future roles (O), because the scenarios evoke a sense of working as a junior doctor. |
| CMOC 2 | **Use of whole tasks (WT)**When stations present whole-tasks rather than fragments of clinical tasks (C), these provide a more challenging assessment of students’ preparedness for practice (O) because they require students to gather and then integrate information in a manner which aligns more closely with the requirements of their future roles (M).  |
| CMOC 3a&b | **Limited Prompting (LP)**a/. When whole-task OSCE stations based on typical work of new doctors limit both the guidance to students (C) and the amount of examiner intervention (C) students experience a challenging and authentic test of their clinical skills (O) because they are forced to judge which approach and problem solving skills are needed for the station (M).b/. When experienced students are given OSCE scenarios without prompting by examiners about which actions to perform (C) they are more able to navigate and manage the scenario (O) because clinical experience has taught them how to integrate information and direct their own actions (M).  |
| CMOC 4 | **Sufficient Time (Ti)**When OSCE stations present students with sufficient time for the tasks which are required (C), they are more likely to reflect and plan their responses rather than relying on rehearsed routines (O) because they are not panicked by time-pressure (M). |
| CMOC 5 | **Use of typical resources (TR)**When OSCE stations enable students to use typical resources (e.g. relevant guidelines/prescribing resources) in a way that reflects routine clinical practice (C), this aids the authenticity of the station (O) because it resonates with what students do and see others do in practice (M) and reduces artificial reliance on memory (M). |
| CMOC 6 | **Use of Props (Pr)**When OSCE stations utilise relevant props within the simulation of a clinical scenario (C), these help students to become immersed in the scenario (O) because it makes the scenario look and feel more realistic (M).  |
| CMOC 7 | **Demonstrate rather than describe (Dem)**When OSCE stations require students to demonstrate rather than describe their actions when managing patients (C) it produces a more authentic test of clinical skills (O) because it tests ability to perform the skill rather than knowledge that the action is required (M). |
| CMOC 8 | **Disruption to Fiction Contract (DFC)** When students and examiners focus on unavoidable differences from practice (C), they experience the assessment as an artificial test of the students’ clinical skills (O), because this focus reduces their engagement in a fiction contract (M).Examples of things about the simulated context which make this challenging: time, observation, doing things to look like you’re doing them, the patient doesn’t have physical signs, the patient doesn’t really need treatment |
| CMOC 9 | **Impact of incongruity (Inc)** When students are asked to perform in simulated scenarios where specific elements of the scenario are incongruous with authentic practice (C) their performance may be detrimentally influenced (O) because of disruption to the students’ sense of immersion in the scenario (M) or confusion or misleading assumptions (M). |
| CMOC 10 | **Legitimate differences in clinical practice (LDP)**When OSCE stations depict scenarios where legitimate variations in clinical practice exist (C) students and examiners are unsure about which behaviours should be performed (O) because they don’t know how to resolve differences between approaches which may be acceptable in clinical practice and the typical expectations of formal exams (M).  |
| CMOC 11a&b | **Student anxiety (Ax)**a/. When OSCE stations rely on students deciding how to structure their time and activity, students who are less proficient or less familiar with the format (C), may experience uncertainty as to what they need to do (O) because they are over-reliant on getting external reassurance that they are demonstrating the expected skills (M).b/. When students in an OSCE station are uncertain as to what they need to do (C), they may become anxious (O), because they are afraid of choosing actions which will reduce their score (M). |
| CMOC 12 | **Ritualised behaviours in OSCEs (Rit)**When assessments impose ritualistic, inauthentic expectations of performance (C), this reduces students’ engagement with practice-based learning (O) because they realise that the skills they learn in practice are not rewarded in exams (M).  |
| CMOC 13a&b | **Impact of authentic stations on students’ learning (IOL)**a/. When medical schools examine students using OSCE stations that closely replicate real-life clinical learning opportunities (C), students are likely to seek out and learn from their clinical encounters (O) because they know doing so will help them with their assessments (M).b. When the format of OSCE assessments encourage students to engage more deeply in practice-based learning (C), students will be more prepared to transition into working as doctors (O) because their focus has been on demonstrating the skills and judgement they will need in that role (M). |

Box 1:

Glossary of Realist Evaluation Terminology:

Context: Any condition that triggers and/or modifies the behaviour of a mechanism. Context refers to the important feature(s) of the circumstances in which an intervention ‘works’ (or a phenomenon happens) which ‘trigger’ the mechanisms that generate outcomes.

Mechanism: The underlying process by which outcomes are caused. Mechanisms are usually descriptions of the tendencies, reasoning and behaviour of agents involved in a process or participants in an intervention and their response to the important context(s) in which they exist.

CMOC: ‘Context–mechanism–outcome configuration’; a diagrammatic or narrative description offering an explanation of the relationship between some particular context(s), mechanism(s) and outcome(s). Multiple CMOCs may exist within a single programme theory.

Programme theory: A theory that describes what an intervention comprises and how it is expected to work, or the process by which the outcomes of interest are thought to come about (expressed as a narrative description or in a diagram). A realist programme theory is expressed in terms of the relationships between relevant context(s), mechanism(s) and outcome(s) (CMOCs)—and the relations between CMOCs.

See (Duddy & Wong, 2022) for reference.

Box 2: Summary recommendations to enhance authenticity in OSCEs

1. Select tasks typical of new doctors’ work.
2. Ask students to complete the whole clinical encounter as far as possible.
3. Encourage students’ judgement: limit prompting and intervention by examiners.
4. Provide sufficient time to avoid rushing. This enables clinical reasoning.
5. Provide access to realistic knowledge resources, such as guidelines or prescribing resources.
6. Use props to aid immersion in scenarios.
7. Require students to demonstrate skills rather than just verbalizing them.
8. Avoid incongruent elements which can disrupt scenario immersion.
9. Prepare students and examiners for uncertainty, acknowledging multiple legitimate approaches to safe, effective outcomes.