Shadow systems in assessment: how supervisors make progress decisions in practice.

Damian J. Castanelli[1, 2], Jennifer M. Weller[3, 4], Elizabeth Molloy[5], Margaret Bearman[6].

[1]School of Clinical Sciences at Monash Health, Monash University, Clayton, Victoria, Australia.
[2]Department of Anaesthesia and Perioperative Medicine, Monash Health, Clayton, Victoria, Australia
[3]Centre for Medical and Health Sciences Education, School of Medicine, University of Auckland, Auckland, New Zealand
[4]Department of Anaesthesia, Auckland City Hospital, Auckland, New Zealand
[5]Department of Medical Education, Melbourne Medical School, University of Melbourne, Melbourne, Victoria, Australia
[6]Centre for Research and Assessment in Digital Learning (CRADLE), Deakin University, Geelong, Victoria, Australia

Address for correspondence: Damian.Castanelli@monash.edu

Abstract

Medical educators need to make decisions on trainee progress through training, and credentialing for independent practice, based on robust evidence from workplace-based assessment. We do not know how the current emphasis on assessment for learning in workplace-based assessment has impacted this decision-making; meeting both these purposes may present unforeseen challenges. In this study we explored how supervisors make decisions on trainee progress in practice.

We conducted semi-structured interviews with 19 supervisors of postgraduate anesthesia training across Australia and New Zealand. We used thematic analysis of verbatim transcripts to produce our findings.

Supervisors looked beyond the formal assessment portfolio when making high-stakes performance decisions. They instead used assessment 'shadow systems' based on their own observation and confidential judgements from trusted colleagues. Decision making involved expert judgement of the salient aspects of performance and the standard to be attained, with supervisors conscious of the opportunities and constraints of the local learning environment. Supervisors found making progress decisions an emotional burden. When faced with difficult decisions, supervisors found ways to share the responsibility and balanced the potential consequences for the trainee with the need to protect their patients.

From a community of practice perspective, the development of assessment 'shadow systems' indicates a lack of alignment between local workplace assessment practices and the advocated programmatic assessment approach to high-stakes progress decisions. Avenues for improvement include cooperative development of formal assessment processes to better meet local needs or acknowledging the value of the information in 'shadow systems' and formally incorporating it into the suite of assessment processes.

Introduction

Medical educators bear a heavy responsibility to ensure those who graduate from our programs provide quality care that meets community expectations (Watling, 2016) but they must also simultaneously support learning, through feedback and coaching. Programmatic assessment has been developed to assist with this responsibility, providing "an integral approach to the design of an assessment program with the intent to optimize its learning function, its decision-making function and its curriculum quality-assurance function" (van der Vleuten et al., 2014). Ideally, programmatic assessment assists educators by separating the 'high stakes' judgement about progression from 'low stakes' judgements about particular tasks, promoting honest and constructive feedback (van der Vleuten et al., 2012). However, in practice, there is a difference between work as prescribed, or what people ideally do, and work as enacted, or what they actually do (Allard & Bleakley, 2016; ~~Braithwaite, 2018~~; Wenger, 1998). Assessment practices inevitably stray from theoretical intentions (Bearman et al., 2016), and judging learner progress may be more difficult than it sounds in theory (Bok et al., 2013). Questions remain about how educators can best make "high stakes" decisions about progress in programmatic assessment systems in the real world.

A fundamental tenet of programmatic assessment is that any individual workplace assessment provides insufficient evidence for decision-making on learner progression (van der Vleuten et al., 2012). Educators use each assessment to observe the trainee, provide feedback information, and help the trainee reflect on and improve their performance. The judgement of trainee progression or graduation is postponed until a decision-maker has gathered a broad range of information from multiple assessments. Hence, while workplace-based assessments (WBAs) contribute to 'high-stakes' decisions in programmatic assessment, each individual assessment contributes only a small amount.

The use of WBAs for assessment *for* learning and assessment *of* learning produces two distinct purposes. Functionally, the *feedback* moment is separated in time from the *decision* moment, and those clinical supervisors providing feedback are not necessarily those making decisions (Schuwirth & Ash, 2013; Schuwirth & van der Vleuten, 2012; van der Vleuten et al., 2012). At some point, based on review of the entire assessment portfolio, a decision about progression must be made. In some contexts, an individual program director may be responsible, however, more recently, competence committees or review panels often perform this function (Colbert et al., 2015; Fitzgerald et al., 2016; Hauer et al., 2015). There is, metaphorically speaking, a 'firewall' (van der Vleuten et al., 2012) between those making the many 'low stakes' judgements and those making a singular 'go/ no-go' judgement. Separating the assessors' functions into 'feedback' and 'review' is intended to establish this 'firewall'. This allow assessors to provide 'low stakes' moments with coaching and feedback information, thereby promoting trainee engagement with learning without the pressure of a high stakes assessment.

Combining these two purposes of assessment and separating the functions and roles has proven difficult. While curriculum designers and administrators may exhort trainees to adopt a learning focus in individual assessments, many trainees have instead viewed individual assessments as high-stakes (Bindal et al., 2013; Bindal et al., 2011; Bok et al., 2013; Castanelli et al., 2016; Massie & Ali, 2016). These trainees reportedly aim to maximize scores by selecting easy cases and lenient assessors (Castanelli et al., 2016; Harrison et al., 2017; Massie & Ali, 2016), choosing to 'play the game' so they 'look good' (Gaunt et al.,

2017). Clinical teachers also struggle to focus on learning in individual assessments, wrestling with an entrenched attachment to summative assessment (Bok et al., 2013; Castanelli et al., 2016; Harrison et al., 2017).

While these 'on the ground' challenges for work-based assessment in promoting learning are well described, there is considerably less research into how the individual assessments are subsequently translated into a decision point, and available evidence suggests that synthesizing multiple low stakes assessments into a high stakes judgement is also more challenging in practice than in theory**.** Learners have questioned the credibility of decision-makers who have not observed their practice (Driessen et al., 2012), and decision-makers have expressed reservations concerning their capability to make the necessary judgements when they may have minimal personal knowledge of the trainee (Bok et al., 2013). Understanding how decision-making in programmatic assessment is working in practice might help us to overcome the challenges experienced thus far in implementation.

The significance of context has been underlined by Govaerts, who notes: "if we want to understand assessors' judgements in the real world, the unit of analysis should be the *assessor–in–situation*" (Govaerts, 2016) [italics ours]. In this instance, we take context to mean the social world in which their actions occur (Lave, 1993), which includes the workplace, those who work there, the work done, social relationships, and the local culture and its values (Bates & Ellaway, 2016). Postgraduate training is inherently social, mediated through a vast array of interactions in the diverse environments of the workplace. We believe examining the experiences of educators in making progress decisions within a social world, may enhance our understanding of how assessment is practiced rather than designed. Our research question in this study is therefore: "How are low-stakes judgements synthesized into high stakes decisions in a programmatic assessment system in postgraduate medical education?" We focus on the particular setting of anesthesia training in Australia and New Zealand.

Methods

Context

Postgraduate anesthesia training in Australia and New Zealand is overseen by the Australian and New Zealand College of Anaesthetists (ANZCA). As of June 2018, there were 1537 trainees across 160 accredited training hospitals. ANZCA has a competency-based curriculum which incorporates WBAs with examinations into a programmatic approach to assessment (Australian and New Zealand College of Anaesthetists, 2012). ANZCA uses four different types of WBA: Mini-Clinical Evaluation Exercise (mini-CEX), Direct Observation of Procedural Skills (DOPS), case-based discussion (CbD), and multisource feedback (MSF). A mini-CEX assesses management of a patient for all or part of an observed case (Weller et al., 2014), while a DOPS assesses an observed technical procedure (Wragg et al., 2003). The CbD form, which requires a trainee to discuss a previous case with a clinical teacher, assesses clinical reasoning. MSF is completed by clinical teachers and other work colleagues including surgeons, nurses and anesthesia assistants, and raters are selected by the trainees. All trainees are required to complete the same minimum number of each WBA.

ANZCA training is overseen by one or more Supervisors of Training (Supervisors) in each department. While all clinical teachers in a department supervise trainees in their work, and perform individual WBAs, Supervisors are specialist anesthesiologists with responsibility for

training in their department, officially appointed by ANZCA. They oversee each trainee's clinical performance, perform regular clinical placement reviews, and adjudicate on satisfactory progression. ANZCA does not use competence committees; Supervisors are individually responsible for summative assessment decisions in the workplace, described as 'high-stakes' decisions in programmatic assessment. According to the ANZCA handbook, the primary source of information for the judgements made by Supervisors on progression through training is the workplace-based assessment portfolio (Australian and New Zealand College of Anaesthetists, 2017).

Ethics

Ethics approval was granted by Monash University Human Research Ethics Committee (Reference: 2016000919) and the University of Auckland Human Participants Ethics Committee (Reference: 017408).

Methodological Positioning

We have adopted a qualitative methodology, using an interpretivist paradigm to guide our selection of research methods. We position this work mostly within the tradition of 'qualitative description', which seeks to interpret the experiences of the participants while remaining close to their own frame of reference (Sandelowski, 2000). As we, and our participants, bring our own beliefs and experience to the research process, we provide some background for readers. The first author (DC) is a clinician-educator who works as a specialist anesthesiologist and has more than ten years' experience as a Supervisor. JW is a specialist anesthesiologist and medical education researcher, while EM and MB are medical education researchers with expertise in workplace learning and qualitative methods. The research team thus includes varying degrees of 'insider' and 'outsider' perspectives.

Sampling

We sampled purposefully, ensuring diversity in characteristics we anticipated might influence our participants' experiences and views (Patton, 2015). We considered gender, geographic spread, rurality, experience, and hospital size. ANZCA staff approached potential participants on our behalf to preserve their privacy. Eleven of nineteen participants were women. The median experience in the role was four years, with a range from one to eleven years. There was at least one participant from each Australian state and territory, and four from New Zealand. Seven participants were from large metropolitan hospitals, and four each were from small metropolitan, large regional, and small regional hospitals.

Interviews

We developed initial broad interview questions based on the rationale for using workplace-based assessment for making progress decisions articulated by van der Vleuten et al. (van der Vleuten et al., 2012) and the ANZCA handbook (Australian and New Zealand College of Anaesthetists, 2017). We refined the interview questions after trialing them with Supervisors in the first author's institution. We iteratively revised the schedule as analysis proceeded and we e discovered which questions best prompted our participants' descriptions of how they made their decisions. A single interviewer, either the research assistant or the first author, conducted each interview. We used the interview questions as a starting point; exploring participant responses determined the subsequent course of the interviews (Kvale &

Brinkmann, 2009). Interviews were transcribed verbatim (46 to 70 mins in length). The final interview schedule is in Appendix 1.

Analysis

We used the methods of thematic analysis described by Braun and Clark (Braun & Clarke, 2006; Braun et al., 2015). The process involved moving back and forth between analytic phases as the research progressed, with our initial analysis recursively informing subsequent data collection. Initially, familiarization with the data involved the first author listening to the interviews with the transcripts and recording initial insights and reflections on how participants' responses added to, reinforced or contrasted with those of previous participants (Braun et al., 2015). This phase of familiarization overlapped with and informed the interviewing and coding.

The next step in the analysis was descriptive coding and generation of a coding framework, which occurred in parallel with the interviews. DC developed the initial coding framework with input from the research team after four interviews, and after a number of iterations, with further discussion amongst the research team, it stabilized after 12 interviews. DC used this framework to code all the interviews, with other team members also coding selected interviews for comparison and discussion. Coding was facilitated using NVivo software (version 11, QSR International Pty Ltd, Melbourne, Vic., Australia.)

After 19 interviews, we conducted a provisional analysis, generating candidate themes and a thematic structure. These themes and thematic structure were then reviewed in detail for their coherence, application to the research question, and the density of the descriptive data set and its boundaries (Braun & Clarke, 2012). At this point, we decided the thematic structure and themes were sufficient to give a rich description of the phenomenon, and ceased data collection. Our analysis then continued with further refinement of the thematic categories achieved through consensus.

Results

Our findings describe the process of how Supervisors synthesized low stakes judgements into high stakes decisions in a programmatic assessment system in postgraduate medical education. In our analysis, we identified two major themes ~~themes in two broad areas~~: 1) the conflicted nature of the Supervisors' dual role; and 2) assessment 'shadow systems'. In the former, participants' roles as both coach and judge amplified the discomforts of anticipating consequences of high-stakes decisions. In the latter, participants used informal or 'shadow systems' of assessment tailored within their own context in parallel with the official assessment system.

The conflicted nature of the Supervisors' dual role

As described above, our participants were cast in the decision-making role in programmatic assessment while also coaching trainees and continuing to work as their clinical teachers. This direct personal involvement coupled with a dual judge-coach role led to negative emotional burdens and challenged the Supervisor-trainee relationship. However, as explored in the next theme, the burdens of this dual role were balanced by the opportunity it afforded for continued involvement with other clinical teachers and direct observation of trainees.

Making progress decisions was an emotional burden

Supervisors found communicating poor performance and managing the anticipated trainee responses stressful. These conversations risked eliciting an emotional response that might be difficult to handle, particularly if they concerned professional and inter-personal issues in contrast to knowledge or technical deficits:

"It's difficult to predict how someone will react… when you deliver that message … my hands were physically shaking" (Supervisor 16)

The trainee response might include threats of legal action or accusations of bullying:

"He'd had a go at us earlier saying we were bullying him… we were a little bit in fear of what might happen." (Supervisor 15)

Supervisors were also genuinely concerned with the ramifications for trainees of performance decisions, even though these were anticipated more than realised. Many participants felt that recording adverse judgements of trainee performance in the official training record would stigmatize the trainee. While some participants emphasized their responsibility to record issues, others sought to avoid this and find other ways to manage the issue. Participants identified that poor performance threatens the trainee's career progress, their self-image, and their future livelihood: "it ruins their lives basically" (Supervisor 7). They reported that "you have to check the trainee's psychological state" (Supervisor 8), and that "in the back of my mind is the suicide rate amongst registrars [residents] and what can tip people over the edge." (Supervisor 2)

Making progress decisions challenged the Supervisor–trainee relationship

Supervisors acknowledged that their dual role created a tension between supporting trainee development and providing judgement. They described different mechanisms for managing this tension. Some acknowledged they held two potentially incompatible roles, which meant they would be seen as "gatekeepers" (Supervisor 19) when delivering an adverse judgement and hence limit their future supporting role. Other participants reconciled the tension by framing the provision of honest judgement as an extension of their developmental role and ultimately in the best interests of the trainee:

"I have always believed that people are better off to know, to acknowledge that there are problems and then to look for solutions. We are not the policeman, but we are the support." (Supervisor 3)

Some Supervisors emphasized the importance of building rapport and emphasizing their investment in a trainee's learning as a mechanism for maintaining the coaching role when adverse decisions or remediation were required, for example:

"I think we do get the trainee to see that we care about [their] training; 'if we didn't care, we wouldn't highlight this. We're highlighting this because we want you to improve, to become a better anesthetist'… we're doing this for their benefit, and they come along for that journey" (Supervisor 16)

Regardless of how the tension between supporting and judging was handled when an adverse decision was made, participants recognized there would be consequences when subsequently directly supervising a trainee as a clinical teacher:

"To be managing a trainee's performance and then working with them in a stressful situation, with a difficult clinical case, that interaction is very challenging." (Supervisor 10)

Assessment 'shadow systems'

The assessment 'shadow systems', largely invisible to the governing body, were formed from the particular workplace contexts of the Supervisors. While noting that each was unique, we outline their general features here using the following short subthemes.

Supervisors used local systems instead of WBAs to inform their decisions

Supervisors reported that the collected and synthesized formal workplace assessments made little, if any, contribution to the evidence of performance used to inform their decisions:

"To form a global assessment, we are getting very little useful information from the WBAs." (Supervisor 8)

Supervisors preferred to collect their own informal records of trainee performance, including written accounts of trainee progress, from their colleagues. In contrast to the episodic nature of WBAs, these reports were generally based on longitudinal observation of trainee development over time and in varied situations. Unlike the WBAs, these reports were usually not visible to the trainee or ANZCA, and participants thought this meant assessors were more willing to provide honest information. This desire for anonymous reports contrasted with their beliefs about records in the official system, where:

"Everyone is naturally disinclined to say or to write down hard truths" (Supervisor 4).

These unofficial written reports generally focused on aspects of performance valued by the Supervisor, and predominantly concerned clinical expertise and conscientiousness. Another consistent feature of these systems was that reports were completed by clinical teachers who were not selected by the trainee, which avoided trainees' strategic choice of case and assessor in WBAs.

Supervisors pooled global judgements rather than specific performance information

Supervisors wanted more from their colleagues than to document feedback for learning. They described how they depended on colleagues' honest views in order to make global judgements of trainee performance, in particular whether the trainees were working at the expected standard. Rather than processing performance information about a series of singular performances and coming to a judgement, our Supervisors preferred to aggregate and evaluate global judgments from others to inform their own.

"I think the most helpful feedback is the *combined opinions of my colleagues*, in terms of 'is this person doing as they should for their age and stage?'" (Supervisor 13) [italics ours]

Supervisors already had access to MSF, which might have been expected to fill this role of collecting judgements. Some Supervisors did value MSF, however others discounted the value of the MSF as there were few, and because they reportedly focused on the 'intrinsic' CanMEDS roles (Sherbino et al., 2011) rather than the medical expert role.

Supervisors valued their own direct observation of trainees

Supervisors' use of the feedback and judgements of others was supplemented by their own exposure to the trainee. Supervisors saw their own direct observation as a high-quality source

of information, which did not require the same evaluation as information from other sources. They found it difficult to make performance judgements without direct exposure to the trainee. Supervisors in small training centers were able to rely on their own exposure to a greater extent. However, in recognition of its value, Supervisors in larger centers manipulated rosters to increase their exposure to a trainee if they had concerns about their performance.

Supervisors made credibility judgements about clinical colleagues

In contrast to their confidence in their own observations of trainees, our Supervisors did not take performance information or judgements at face value; rather, they judged the credibility of their clinical colleagues. Supervisors reported that they routinely evaluated the worth of feedback or an assessment based on their knowledge of the person responsible for generating the information. Supervisors judged clinical teachers' willingness to provide honest feedback and also considered their relative leniency or stringency. Supervisors also recognized that some informants were more capable judges of trainee performance, and one aspect of this capability was acknowledging the broad range of acceptable practice:

"some of my colleagues are better at assessing trainees… they come to realize that actually, you can be very different but still do a good job as an anesthetist." (Supervisor 13)

Supervisors attended to idiosyncratic and contextual ~~context-relevant~~ features of performance

Individual Supervisors emphasized different constituents of trainee performance relevant to their particular contexts. Supervisors referenced diverse external authorities to support their personal views of the important facets of trainee performance. More generally, their own conceptualization of performance, which had emerged through their own experience, dominated:

"I think you could break it down in many different ways; technical abilities, non-technical skills, and interpersonal skills. And I guess just how organized they are" (Supervisor 5)

The relevance and relative weighting of constituents of trainee performance appeared to vary greatly among participants. Supervisors adapted these to their circumstances, allowing for the particular nature of the trainee's current clinical work and learning opportunities. Technical skill proficiency was required but was seen as more readily remediated and hence of less significance. In contrast, perceived deficits in professional attributes were more concerning, particularly a lack of conscientiousness or clinical judgement:

"She cannot make a competent judgment, rather than she can't put a tube in." (Supervisor 17)

Supervisors conceptions of the required standard varied

The required standard of performance was conceptualized differently by participants. Performance benchmarks were noted to be 'grey' (Supervisor 3) and not always clear, with one participant wishing for a "Yes with an asterisk" (Supervisor 4) option to convey a more nuanced judgement. Supervisors recognized that competence fluctuates and that variability in performance is to be expected, which increased the complexity of the decision-making:

"The performance was so inconsistent: there'd be periods of time when she would be performing well, and then periods of time when she may as well not have been at work." (Supervisor 14)

Some participants were also conscious of the trainee's trajectory toward the ultimate aim of independent practice when determining the standard of performance, asking themselves:

'Are they heading in the right direction?' 'Do we think this person is going to make a good anaesthetist?' (Supervisor 8)

Some Supervisors described using a norm-referenced standard, for example:

"I compare them to a similar trainee at a similar level of training and if they are performing alike in terms of their ability to manage cases of a similar complexity and similar clinical situations." (Supervisor 1)

Other Supervisors had a conception of minimal competence; that you could be practising at a level below the standard of your peers but that this could still be acceptable. This idea of minimal competence overlapped with the concept of maintaining safe practice, which served as the ultimate determinant of the acceptable minimum level of performance for many participants:

"I will have trainees who I can see are not very slick… But they are okay. They are not going to be the world's best anesthetist but… they are safe, and their patients are fine." (Supervisor 4)

Time played a central role in making a progress decision

Participants described looking for "patterns" of trainee performance emerging after repeated interactions, so that:

"A dot becomes two dots, becomes a line; another dot becomes a shape" (Supervisor 16).

The importance of time was also apparent from the multiple comments about the difficulty in making judgements on short placements. There was a broad recognition that trainees ought to have the opportunity to address performance concerns before a high-stakes judgement is made. However, in a number of cases, participants reported making an adverse decision where it was definitely warranted in spite of a compressed time period.

Supervisors shared the decision-making

All participants found it difficult to describe how they made the final decision. They spoke of the complexity involved in dealing with "so many subjective things" (Supervisor 18) and the tailoring of judgement to the individual circumstance:

"It is very, very difficult to quantify and codify… It is subjective; it is based on lots of interactions, phone calls, conversations… You get a bit of a vibe about it… there's far more than just individual cases and individual procedures. It is a global assessment that is far more telling" (Supervisor 8)

Supervisors often managed the uncertainty around this complex judgement by sharing the responsibility.

"I am not going to be the one person who decides that this person is not performing up to standard... it's not something I will do independently." (Supervisor 7)

One way they did this was by incorporating the collective judgement of their colleagues. They used this to a varying extent, from one extreme where they acted as a messenger

delivering their colleagues' collective judgement, to another where they privileged their own judgement but used concordance with their colleagues to reinforce their resolve:

"The consensus from the consultants in the department was that they were not yet performing at a level where we were confident in them managing cases independently" (Supervisor 12)

Alternatively, some Supervisors depended less on the collective judgement of colleagues in the broad sense, but relied on one or more other Supervisors in their department, or trusted confidantes to review the same information and concur in their judgement.

"If there's going to be a fail, it's a joint decision between both of us, as Supervisors." (Supervisor 15)

Discussion

We have examined how high-stakes decisions are made in practice in a programmatic assessment system. Although our participants recorded their decisions in the official assessment system, we discovered that they made little use of the documented low-stakes assessments and generally did not follow the prescribed process in coming to their decision. Instead, our participants used assessment 'shadow systems'(Shaw, 1997), privileging social contexts above formal systems. At the same time, these assessment 'shadow systems' relied on Supervisors being both clinical teachers and college assessors, which caused tensions with trainees and emotional discomfort.

We aimed to focus on the ~~'assessor in situation'(Govaerts, 2016)~~ Supervisor in their context and we found that Supervisors relied on their knowledge of both their particular workplace and the credibility of their colleagues to provide the judgements and information about trainee performance they wanted. The perceived limited value of WBAs for decision-making is surprising as they have previously been reported to have sufficient reliability to support robust decisions in studies in the ANZCA environment (Castanelli et al., 2019; Weller et al., 2017; Weller et al., 2014). The Supervisors' actions contrast with the intentions of programmatic assessment to separate the feedback moment from the decision moment (van der Vleuten et al., 2014) through clinical teachers providing feedback to trainees in individual WBAs and Supervisors making judgements of performance at a later time. This separation is designed to manage the intersection between coaching and assessing trainees (Dijkstra et al., 2010; Schuwirth & van der Vleuten, 2012). Because our participants predominantly used assessment 'shadow systems' rather than WBAs to inform their decisions, WBAs were effectively 'feedback only' assessments. However, a previous study in the ANZCA environment has shown that trainee and Supervisor perceptions of the summative purpose of WBAs have potentially undermined their role in encouraging feedback and learning (Castanelli et al., 2016). Combining these findings suggests the possibility that the WBAs were not optimally used either for feedback or for assessment. At the same time, the use of assessment 'shadow systems' effectively circumvented the so-called 'firewall' between the people providing feedback and those making decisions (van der Vleuten et al., 2012).

Given the importance of 'shadow systems' in our findings, we wondered why 'shadow systems' had not been more frequently reported. Once we began to look, we saw traces elsewhere. For example in Ginsberg et al's description of program directors' 'reading between the lines' in narrative comments on official assessments (Ginsburg et al., 2015) or how informal assessment messages can be the trigger for recognizing the learner in difficulty

(Hauer et al., 2015; Schumacher et al., 2018). These traces give some indication of an absence in the health professions education literature (Paradis & Whitehead, 2015). It may be that these ~~horizontal~~ undocumented accounts are a form of what Bourdieu would call *doxa*, the "naturalised arbitrariness" (Bourdieu, 1977) of the way things are done that " leaves unsaid all that goes without saying" (Bourdieu, 1977). That is, assessment 'shadow systems' are so familiar they become taken for granted to the point of being unspoken and unseen.

We now turn to practice theory (Hager et al., 2012) to explain why our participants developed 'shadow systems' and did not conform to the method for making progress decisions prescribed in the programmatic assessment design. Practice theories are concerned with the "everyday activities of assessment as conducted" (Boud et al., 2016), the *modus operandi* rather than the *opus operatum* (Bourdieu, 1977), and hence align well with the focus of our analysis. Although there are multiple practice theories, we think community of practice theory (Wenger, 1998), with its emphasis on the negotiated nature of practice situated within a community, can provide useful insights.

Community of practice theory describes 'vertical' systems of accountability, where a community of practice is accountable to a higher authority (Wenger, 2010). In our study this is the official assessment system whereby the local department is accountable to the wider ~~network of the~~ profession through ANZCA. A community of practice's response to vertical accountability depends upon the extent of enforcement and the degree of alignment with local practice but is nevertheless characterized by the appearance of compliance (Wenger, 1998). All our participants ensured the requirements of the official assessment system for the recording of WBAs and high-stakes decisions were met. Hence at a superficial level, the available documentation would indicate that the implementation of programmatic assessment is working as intended. However, our participants' assessment 'shadow systems' illustrate their relative freedom to maintain their own preferred practice while presenting this appearance of compliance.

In addition to vertical accountability, a local system of 'horizontal accountability' also exists within a community of practice (Wenger, 1998). Features of these horizontal systems include a reliance on the commitment of colleagues, locally negotiated standards of practice, peer recognition of competence, and collective responsibility for learning. The assessment 'shadow systems' we discovered share many of these characteristics; the idiosyncratic nature of the standard required of trainees, the reliance on the collective judgement of colleagues, and evaluation of individual contributions based on their perceived competence as assessors.

According to Wenger, these vertical and horizontal mechanisms often do not interact and may be invisible to one another (Wenger, 2010). In our study, the information in the 'shadow system' only became visible to trainees or included in the official assessment system at the discretion of the individual Supervisor. The Supervisor fulfilled the role of a 'broker', bringing the information from the horizontal accounting system to the attention of the vertical accounting system (Wenger, 2010).

Wenger contends that while horizontal structures embedded in the workplace, such as the 'shadow systems' we have identified, better reflect the reality of practice, both modes of accountability are required and we should not 'romanticize' local mechanisms (Wenger, 2010). Risks arise from a lack of vertical accountability, such as the reproduction of undesirable practice or acceptance of inadequate standards, and these may endanger the reputation of the profession as a whole. Additionally in our own findings, relying on 'shadow

systems' increased the emotional burden for Supervisors and led to difficult interactions with trainees. Ideally, the two levels of accountability would work together rather than in isolation, and vertical accountability measures would provide structure to and simplify the work involved in local processes.

Reflecting on the implementation of new assessment practice, a key message from community of practice theory that is consistent with our findings is that "no matter how much external effort is made to shape, dictate, or mandate practice, in the end it reflects the meanings arrived at by those engaged in it."(Wenger, 1998). Accordingly, rather than rely on further direction or punitive mechanisms to increase the fidelity of implementation, community of practice theory would suggest more nuanced negotiated approaches to improving high-stakes decision making in programmatic assessment. We think the impetus for Supervisors to improvise 'shadow systems' reflects the extent to which the official system fails to account for local assessment practice. The options in this view are two-fold. In the first instance, there appears to be scope to review and adapt the WBA system in coordination with Supervisors to better reflect the realities of the workplace and emphasize aspects of performance Supervisors find most salient. Based on the preferences of our participants, such a review might lead to the development of new tools that collect judgements of the areas of practice of most interest based on observations of performance over time from respected colleagues.

Secondly, it may be useful to accept that the variety of practices in local workplaces may circumscribe a limit to the ability for any assessment tool to be adapted to meet local needs in all instances. Hence, another approach would be to negotiate with Supervisors for the information in any 'shadow systems' to ultimately be reflected in some form in the official system. In this way, the information in the 'shadow systems' would become transparent to trainees in line with accepted principles of natural justice and transparency, and the risks posed by 'shadow systems' identified above would be minimized. We anticipate that further work will be required to see if these two approaches are complementary or whether one might preclude the need for the other.

Although our Supervisors commonly worked as individual decision-makers, committees are increasingly preferred for this role in programmatic assessment (Harris et al., 2017). Our Supervisors found ways to share decision-making when it was difficult, and we think our findings should encourage the adoption of group decision-making for progress decisions. A group provides a mechanism to pool judgements and share the emotional burden and responsibility inherent in high-stakes decisions. However, there is no reason to assume that group decision-making would prevent the misalignment of official assessment design and local assessment practice; Committees will need to account for 'shadow systems' operating in parallel to any official system. The challenges of using personal knowledge of a trainee's performance while avoiding bias (Ekpenyong et al., 2017; Hauer et al., 2015), and interpreting information sourced from others (Hauer et al., 2015) have already been acknowledged. Our findings suggest direct exposure to trainees and contextual knowledge of the workplace are important assets that should be incorporated into a competence committee's deliberations, however we recognize the need to maintain transparency in assessment and procedural fairness for the trainee.

Limitations and further research opportunities

Our study participants and investigators are all from Australia and New Zealand, and our participants work in postgraduate anesthesia, so our findings may not transfer to other specialty training programs or cultures. However, we think the deep exploration in our context of an issue that arises in many assessment systems may be of value to others when they contemplate their own context.

In undertaking an interview-based study, we are reliant on self-report, which can sometimes reflect what people think they do rather than what they actually do; future research in this area could include observation of practice. As discussed above, further research could also evaluate how to collect the information favored by decision-makers in a way that was fair and transparent to trainees, either by modifying the official assessment system or incorporating the information from 'shadow systems'. Also, our study was designed to capture the Supervisors' perspective, and exploring how trainees view these decisions would be a valuable avenue of further inquiry.

Although community of practice theory has been widely used in health professions education research (Buckley et al., 2019), it is not generally applied in the context of assessment. Our use of it here is consistent with the call to use practice theories in assessment research in higher education (Boud et al., 2016). Focusing on the practices of assessment may allow health professions education researchers to examine what "goes without saying and therefore goes unquestioned" (Bourdieu, 1977). Moving beyond measurement to more fully account for why things work the way they do may open new avenues for improving assessment and learning.

Conclusion

In this study of how Supervisors synthesize low stakes judgements into high stakes decisions in a programmatic assessment system in postgraduate medical education, we found that Supervisors used assessment 'shadow systems' while maintaining a façade of compliance with the prescribed assessment system. We think the development of these assessment 'shadow systems' stems from a lack of alignment between the programmatic assessment design implemented from above and the local practice of learning and assessment in the workplace. The extent to which Supervisors' judgements were influenced by their relationships with trainees and colleagues and their knowledge of their working environments has significant implications for assessment systems. Community of practice theory provides useful insights into the nature of the assessment 'shadow systems' reported by our participants and allows us to suggest ways to improve the implementation of programmatic assessment in practice.

Appendix 1: Interview Schedule

1. What do you see as your role as an SoT (Supervisor of Training)?
2. What do you think of the CPR* (Clinical Placement Review) and CUR* (Core Unit Review) assessments?
3. Can you think back to a recent CPR and tell me about it?
4. What decision did you make?
5. Thinking about that recent (CPR) interview as an example, how do you make this decision?
6. Have you ever made a decision about a trainee where it was difficult, or you felt uneasy about it?
7. How did you feel making this decision?
8. What was your experience in communicating the decision to the trainee?
9. What happened afterward?
10. Is there anything more you'd like to tell me?

*N.B. Clinical Placement Reviews occur at three to six monthly intervals and require the supervisor to make a decision whether the trainee's progress is satisfactory, borderline, or unsatisfactory. Core Unit Reviews take place between stages of training and the end of training and require the supervisor to make a decision on whether the trainee's performance is sufficient to justify promotion to the next stage or graduation from training. Both reviews include an interview between the supervisor and trainee.

Acknowledgements

The authors wish to thank the ANZCA Clinical Trials Network for their assistance in recruiting participants for this study.

Funding/Support

This study was supported by an untied grant from the ANZCA Research Foundation (Grant No: S16/043).

Declaration of Interest

DC and JW hold voluntary positions on ANZCA Educational Committees. Otherwise, all authors report no declarations of interest in this work.

Ethical Approval

Ethics approval was granted by Monash University Human Research Ethics Committee (Reference: 2016000919) and the University of Auckland Human Participants Ethics Committee (Reference: 017408). Informed consent was obtained from all individual participants included in the study. All procedures performed were in accordance with the ethical standards of the institutional ethics committees and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

Allard, J., & Bleakley, A. (2016). What would you ideally do if there were no targets? An ethnographic study of the unintended consequences of top-down governance in two clinical settings. *Adv Health Sci Educ Theory Pract, 21*(4), 803-817. doi: 10.1007/s10459-016-9667-8

Australian and New Zealand College of Anaesthetists. (2012). Anaesthesia training program curriculum. Retrieved 15 Nov 2017, from http://www.anzca.edu.au/documents/anaesthesia-training-program-curriculum.pdf

Australian and New Zealand College of Anaesthetists. (2017). Handbook for Training and Accreditation. Retrieved 8 Nov 2018, from http://www.anzca.edu.au/documents/training-accreditation-handbook.pdf

Bates, J., & Ellaway, R. H. (2016). Mapping the dark matter of context: a conceptual scoping review. *Med Educ, 50*(8), 807-816. doi: 10.1111/medu.13034

Bearman, M., Dawson, P., Boud, D., Bennett, S., Hall, M., & Molloy, E. (2016). Support for assessment practice: developing the Assessment Design Decisions Framework. *Teaching in Higher Education, 21*(5), 545-556. doi: 10.1080/13562517.2016.1160217

Berendonk, C., Stalmeijer, R. E., & Schuwirth, L. W. (2013). Expertise in performance assessment: assessors' perspectives. *Adv Health Sci Educ Theory Pract, 18*(4), 559-571. doi: 10.1007/s10459-012-9392-x

Bindal, N., Goodyear, H., Bindal, T., & Wall, D. (2013). DOPS assessment: a study to evaluate the experience and opinions of trainees and assessors. *Med Teach, 35*(6), e1230-1234. doi: 10.3109/0142159X.2012.746447

Bindal, T., Wall, D., & Goodyear, H. M. (2011). Trainee doctors' views on workplace-based assessments: Are they just a tick box exercise? *Med Teach, 33*(11), 919-927. doi: 10.3109/0142159X.2011.558140

Bok, H. G., Teunissen, P. W., Favier, R. P., Rietbroek, N. J., Theyse, L. F., Brommer, H., . . . Jaarsma, D. A. (2013). Programmatic assessment of competency-based workplace learning: when theory meets practice. *BMC Med Educ, 13*, 123. doi: 10.1186/1472-6920-13-123

Boud, D., Dawson, P., Bearman, M., Bennett, S., Joughin, G., & Molloy, E. (2016). Reframing assessment research: through a practice perspective. *Studies in Higher Education*, 1-12. doi: 10.1080/03075079.2016.1202913

Bourdieu, P. (1977). *Outline of a theory of practice*. Cambridge: Cambridge University Press.

Braithwaite, J. (2018). Changing how we think about healthcare improvement. *Bmj, 361*, k2014. doi: 10.1136/bmj.k2014

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. doi: 10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods*

in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological (Vol. 2, pp. 57-71). Washington, DC, US: American Psychological Association.

Braun, V., Clarke, V., & Terry, G. (2015). Thematic Analysis. In P. Rohleder & A. C. Lyons (Eds.), *Qualitative Research in Clinical and Health Psychology* (pp. 95-114). New York: Palgrave Macmillan.

Buckley, H., Steinert, Y., Regehr, G., & Nimmon, L. (2019). When I say ... community of practice. *Med Educ*. doi: 10.1111/medu.13823

Castanelli, D. J., Jowsey, T., Chen, Y., & Weller, J. M. (2016). Perceptions of purpose, value, and process of the mini-Clinical Evaluation Exercise in anesthesia training. *Canadian Journal of Anesthesia, 63*(12), 1345-1356. doi: 10.1007/s12630-016-0740-9

Castanelli, D. J., Moonen-van Loon, J. M. W., Jolly, B., & Weller, J. M. (2019). The reliability of a portfolio of workplace-based assessments in anesthesia training. *Canadian Journal of Anesthesia/Journal canadien d'anesthésie, 66*, 193-200. doi: 10.1007/s12630-018-1251-7

Colbert, C. Y., Dannefer, E. F., & French, J. C. (2015). Clinical Competency Committees and Assessment: Changing the Conversation in Graduate Medical Education. *J Grad Med Educ, 7*(2), 162-165. doi: 10.4300/JGME-D-14-00448.1

Cuncic, C., Regehr, G., Frost, H., & Bates, J. (2018). It's all about relationships : A qualitative study of family physicians' teaching experiences in rural longitudinal clerkships. *Perspect Med Educ, 7*(2), 100-109. doi: 10.1007/s40037-018-0416-y

Dijkstra, J., Van der Vleuten, C. P., & Schuwirth, L. W. (2010). A new framework for designing programmes of assessment. *Adv Health Sci Educ Theory Pract, 15*(3), 379-393. doi: 10.1007/s10459-009-9205-z

Driessen, E. W., van Tartwijk, J., Govaerts, M., Teunissen, P., & van der Vleuten, C. P. (2012). The use of programmatic assessment in the clinical workplace: a Maastricht case report. *Med Teach, 34*(3), 226-231. doi: 10.3109/0142159X.2012.652242

Ekpenyong, A., Baker, E., Harris, I., Tekian, A., Abrams, R., Reddy, S., & Park, Y. S. (2017). How do clinical competency committees use different sources of data to assess residents' performance on the internal medicine milestones?A mixed methods pilot study. *Med Teach, 39*(10), 1074-1083. doi: 10.1080/0142159X.2017.1353070

Fitzgerald, J. T., Burkhardt, J. C., Kasten, S. J., Mullan, P. B., Santen, S. A., Sheets, K. J., . . . Gruppen, L. D. (2016). Assessment challenges in competency-based education: A case study in health professions education. *Med Teach, 38*(5), 482-490. doi: 10.3109/0142159X.2015.1047754

Gaunt, A., Patel, A., Rusius, V., Royle, T. J., Markham, D. H., & Pawlikowska, T. (2017). 'Playing the game': How do surgical trainees seek feedback using workplace-based assessment? *Medical Education, 51*(9), 953-962. doi: 10.1111/medu.13380

Ginsburg, S., Regehr, G., Lingard, L., & Eva, K. W. (2015). Reading between the lines: faculty interpretations of narrative evaluation comments. *Med Educ, 49*(3), 296-306. doi: 10.1111/medu.12637

Ginsburg, S., van der Vleuten, C. P., Eva, K. W., & Lingard, L. (2017). Cracking the code: residents' interpretations of written assessment comments. *Med Educ*. doi: 10.1111/medu.13158

Govaerts, M. J. (2016). Competence in assessment: beyond cognition. *Med Educ, 50*(5), 502-504. doi: 10.1111/medu.13000

Gruppen, L., Frank, J. R., Lockyer, J., Ross, S., Bould, M. D., Harris, P., . . . Collaborators, I. (2017). Toward a research agenda for competency-based medical education. *Med Teach, 39*(6), 623-630. doi: 10.1080/0142159X.2017.1315065

Hager, P., Lee, A., & Reich, A. (2012). Problematising Practice, Reconceptualising Learning and Imagining Change. In P. Hager, A. Lee, & A. Reich (Eds.), *Practice, Learning and Change: Practice-Theory Perspectives on Professional Learning* (Vol. 8). Dordrecht: Springer Science+Business Media.

Harris, P., Bhanji, F., Topps, M., Ross, S., Lieberman, S., Frank, J. R., . . . Collaborators, I. (2017). Evolving concepts of assessment in a competency-based world. *Med Teach, 39*(6), 603-608. doi: 10.1080/0142159X.2017.1315071

Harrison, C. J., Könings, K. D., Schuwirth, L. W. T., Wass, V., & van der Vleuten, C. P. M. (2017). Changing the culture of assessment: the dominance of the summative assessment paradigm. *BMC Medical Education, 17*(1). doi: 10.1186/s12909-017-0912-5

Hauer, K. E., Chesluk, B., Iobst, W., Holmboe, E., Baron, R. B., Boscardin, C. K., . . . O'Sullivan, P. S. (2015). Reviewing residents' competence: a qualitative study of the role of clinical competency committees in performance assessment. *Acad Med, 90*(8), 1084-1092. doi: 10.1097/ACM.0000000000000736

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

Kvale, S., & Brinkmann, S. (2009). *InterViews: Learning the craft of qualitative research interviewing* (2nd ed.). Thousand Oaks, CA: Sage.

Lave, J. (1993). The practice of learning. In S. Chaiklin & J. Lave (Eds.), *Understanding Practice Perspectives on Activity and Context* (pp. 3-32). Cambridge, England: Cambridge University Press.

Massie, J., & Ali, J. M. (2016). Workplace-based assessment: a review of user perceptions and strategies to address the identified shortcomings. *Advances in Health Sciences Education, 21*(2), 455-473. doi: 10.1007/s10459-015-9614-0

Paradis, E., & Whitehead, C. R. (2015). Louder than words: power and conflict in interprofessional education articles, 1954-2013. *Med Educ, 49*(4), 399-407. doi: 10.1111/medu.12668

Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and practice* (4th ed.). Thousand Oaks, CA: Sage.

Sandelowski, M. (2000). Whatever happened to qualitative description? *Research in Nursing & Health, 23*, 334-340.

Schumacher, D. J., Michelson, C., Poynter, S., Barnes, M. M., Li, S. T., Burman, N., . . . Frohna, J. G. (2018). Thresholds and interpretations: How clinical competency committees identify pediatric residents with performance concerns. *Med Teach, 40*(1), 70-79. doi: 10.1080/0142159X.2017.1394576

Schuwirth, L., & Ash, J. (2013). Assessing tomorrow's learners: In competency-based education only a radically different holistic method of assessment will work-Six things we could forget. *Medical Teacher, 35*(7), 555-559. doi: http://dx.doi.org/10.3109/0142159X.2013.787140

Schuwirth, L. W., & van der Vleuten, C. P. (2012). Programmatic assessment and Kane's validity perspective. *Med Educ, 46*(1), 38-48. doi: 10.1111/j.1365-2923.2011.04098.x

Shaw, P. (1997). Intervening in the shadow systems of organizations. *Journal of Organizational Change Management, 10*(3), 235-250. doi: 10.1108/09534819710171095

Sherbino, J., Frank, J. R., Flynn, L., & Snell, L. (2011). "Intrinsic Roles" rather than "armour": renaming the "non-medical expert roles" of the CanMEDS framework to match their intent. *Adv Health Sci Educ Theory Pract, 16*(5), 695-697. doi: 10.1007/s10459-011-9318-z

Telio, S., Ajjawi, R., & Regehr, G. (2015). The "educational alliance" as a framework for reconceptualizing feedback in medical education. *Acad Med, 90*(5), 609-614. doi: 10.1097/ACM.0000000000000560

van der Vleuten, C. P., Schuwirth, L. W., Driessen, E. W., Govaerts, M. J., & Heeneman, S. (2014). 12 Tips for programmatic assessment. *Med Teach*, 1-6. doi: 10.3109/0142159X.2014.973388

van der Vleuten, C. P., Schuwirth, L. W., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: building blocks for theory development. *Best Pract Res Clin Obstet Gynaecol, 24*(6), 703-719. doi: 10.1016/j.bpobgyn.2010.04.001

van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D. E., Baartman, L. K. J., & Tartwijk, J. v. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher, 34*(3), 205-214. doi: 10.3109/0142159X.2012.652239

Watling, C. (2016). The uneasy alliance of assessment and feedback. *Perspect Med Educ, 5*, 262-264. doi: DOI 10.1007/s40037-016-0300-6

Weller, J. M., Castanelli, D. J., Chen, Y., & Jolly, B. (2017). Making robust assessments of specialist trainees' workplace performance. *British Journal of Anaesthesia, 118*(2), 207-214. doi: 10.1093/bja/aew412

Weller, J. M., Misur, M., Nicolson, S., Morris, J., Ure, S., Crossley, J., & Jolly, B. (2014). Can I leave the theatre? A key to more reliable workplace-based assessment. *Br J Anaesth, 112*(6), 1083-1091. doi: 10.1093/bja/aeu052

Wenger, E. (1998). *Communities of Practice: learning, meaning and identity*. Cambridge: Cambridge University Press

Wenger, E. (2010). Communities of Practice and Social Learning Systems: the Career of a Concept. In C. Blackmore (Ed.), *Social Learning Systems and Communities of Practice*, (pp. 179-198). London: Springer.

Wenger, E. (2010). Conceptual Tools for CoPs as Social Learning Systems: Boundaries, Identity, Trajectories and Participation. In C. Blackmore (Ed.), *Social Learning Systems and Communities of Practice* (pp. 125-144). London: Springer.

Wragg, A., Wade, W., Fuller, G., Cowan, G., & Mills, P. (2003). Assessing the performance of specialist registrars. *Clinical Medicine, 3*, 131-134.