Received: 5 March 2021 | Revised: 19 May 2021 | Accepted: 26 May 2021

DOI: 10.1111/medu.14574

ASSESSMENT

Making it fair: Learners' and assessors' perspectives of the attributes of fair judgement

Nyoli Valentine[1] | Ernst Michael Shanahan[1] | Steven J. Durning[2] | Lambert Schuwirth[1]

[1]Prideaux Discipline of Clinical Education, Flinders University, SA, Australia

[2]Center for Health Professions Education, Uniformed Services University of the Health Sciences, Bethesda, MD, USA

Correspondence
Nyoli Valentine, Prideaux Discipline of Clinical Education, Flinders University, Bedford Park, SA, Australia.
Email: vale0046@flinders.edu.au

Abstract

Introduction: Optimising the use of subjective human judgement in assessment requires understanding what makes judgement fair. Whilst fairness cannot be simplistically defined, the underpinnings of fair judgement within the literature have been previously combined to create a theoretically-constructed conceptual model. However understanding assessors' and learners' perceptions of what is fair human judgement is also necessary. The aim of this study is to explore assessors' and learners' perceptions of fair human judgement, and to compare these to the conceptual model.

Methods: A thematic analysis approach was used. A purposive sample of twelve assessors and eight post-graduate trainees undertook semi-structured interviews using vignettes. Themes were identified using the process of constant comparison. Collection, analysis and coding of the data occurred simultaneously in an iterative manner until saturation was reached.

Results: This study supported the literature-derived conceptual model suggesting fairness is a multi-dimensional construct with components at individual, system and environmental levels. At an individual level, contextual, longitudinally-collected evidence, which is supported by narrative, and falls within ill-defined boundaries is essential for fair judgement. Assessor agility and expertise are needed to interpret and interrogate evidence, identify boundaries and provide narrative feedback to allow for improvement. At a system level, factors such as multiple opportunities to demonstrate competence and improvement, multiple assessors to allow for different perspectives to be triangulated, and documentation are needed for fair judgement. These system features can be optimized through procedural fairness. Finally, appropriate learning and working environments which considers patient needs and learners personal circumstances are needed for fair judgments.

Discussion: This study builds on the theory-derived conceptual model demonstrating the components of fair judgement can be explicitly articulated whilst embracing the complexity and contextual nature of health-professions assessment. Thus it provides a narrative to support dialogue between learner, assessor and institutions about ensuring fair judgements in assessment.

1 | INTRODUCTION

There is broad agreement that assessment in education should be fair.[1] Traditionally, evidence of construct validity and reliability has been central to defend fairness of assessment.[2-4] However, both the notion of validity[5] and medical education itself have undergone a paradigm shift. Competency-based medical education is increasingly seen as being at odds with traditional objective, measurement-based assessments.[3,6-14] This perceived misalignment has led to an increasingly resounding push within the literature to embrace human judgement in assessment and accept its subjective nature.[3,4,8-19] However, in embracing human judgement in assessment, an important question has arisen: 'What makes human judgement "fair"?'. Without insight into this, human judgement will continue to be viewed as too 'subjective' and unfair.

Despite being an essential element of assessment, there is no unanimous agreed understanding of fairness, with 'fair' meaning different things to different stakeholders.[20] The elusiveness of this construct makes it difficult to simply define.[6] One could argue this is perhaps a good thing, as having a simple definition may suggest a complex, diverse, multi-dimensional, context-dependent construct can be reduced to a straightforward rule which is likely to not represent the complexity of the situation. Given that a simple definition will not likely be agreed upon[20] and is potentially not useful, then perhaps changing tack and focussing on the building blocks of fairness may be more fruitful. Better understanding the foundations of fairness can help create a shared narrative to allows for negotiation and agreement between stakeholders of what fair judgement is in complex situations. The underpinnings of fairness are inferred in the medical education and broader education literature. A recent literature review has brought these inferences and underpinnings together to create a theoretically constructed conceptual model.[7] This model identified that fairness could be conceptualised through values (credibility, fitness for purpose, transparency and defensibility) which are upheld at an individual level by characteristics of fair human judgement (narrative, boundaries, expertise, mental agility and evidence) and at a systems level by procedures (procedural fairness, documentation, multiple opportunities, multiple assessors and validity evidence) which help translate fairness in human judgement from concepts into practical components.

Whilst this is helpful, it is merely a literature-derived model. It adds theoretical validity to the conceptualisation of 'fairness'. However, without empirical data, it cannot lend practical validity and thus credibility to its conceptualisation. Understanding the 'on the ground' assessors' and learners' perceptions of what is fair human judgement is therefore necessary.

The purpose of this study is to explore the understanding of fair human judgement from the perspectives of learners and assessors across a continuum of experiences. It seeks to evaluate practical plausibility: To what extent does the literature-derived conceptual model align with the perspectives and experiences of learners and assessors?

This study aimed to address the following research questions:

1. What do assessors and learners perceive to be the characteristics of fair judgement?
2. How do these understandings of fair human judgement of assessors and learners compare with the theoretically constructed conceptual model?

2 | METHODS

As this study focussed on practical plausibility, we used a thematic analysis approach. Thematic analysis focuses on meanings across a data set and allows researchers to make sense of collective or shared meanings and experiences.[21] Thematic analysis is flexible and able to conduct in many different ways.[21] In this study, we used an inductive, emergent and constant comparative approach to assist in understanding the complex and non-uniform perceptions and experiences of fair judgement. As developers of the previous conceptual model, we were aware that we were not without prior knowledge of the topic. Therefore, we balanced our approach between a thematic approach and a more inductive approach to ensure the perceptions of the participants were not interpreted in a desired direction. We undertook open coding prior to mapping to the existing model. Mapping involved a deliberate intent to uncover dissent between the participants' perception and the existing model. As such, we sought to explore four types of outcomes:

- perceptions voiced that were not in the model
- aspects of the model that were not reflected in the data
- perceptions voiced that existed in the model but with different or additional connotations
- perceptions voiced which aligned with the model

A purposive sample of assessors and trainees was recruited from universities and post-graduate colleges in Adelaide, Australia. Potential participants were emailed, introduced to the study and invited to participate. Specialty, years of experience, supervisor position within a hospital or community and gender were considered in the purposeful sampling, aiming for variation in these characteristics which might be anticipated to influence responses. No incentive was provided to participate. Semi-structured interviews occurred via Zoom (due to the pandemic) lasting up to 60 minutes. Interviews were recorded and transcribed verbatim without any identifying data. NVIVO software system was used to assist with data management.

Vignettes were chosen as the starting points for the interviews as these are multivalent representations embedded in concrete realistic context.[22] This reduces the abstract nature of the concept, in our case of fairness, but still allows for simultaneous investigation of factors and their relationships.[22] Three vignettes were presented during the interview (see Appendix S1). To ensure the vignettes reflected realistic assessment scenarios, we drew

on the experience of the authors to initially develop 6 vignettes. These were mapped against the theoretically derived conceptual model, and therefore, they stimulated discussion on a broad range of issues related to fair judgement, including at an individual and system level. Through discussion with the authors, the vignettes were reduced to three, deliberately representing different stages of training, under-graduate, post-graduate and post fellowship. The vignettes were also chosen to represent high-stakes judgements, as this was anticipated to promote more discussion and also have more practical applicability. At the end of the three vignettes, participants were asked to share their own stories to identify further concepts relating to the research question which may not have been identified in the literature review. As the aim of the study was to understand the participants' perceptions of the characteristics of fair judgement, no information or introduction was given about what the researchers meant by fairness, to ensure interviewees were not unduly influenced.

The study was undertaken from July 2020 until December 2020. Collection, analysis and coding of the data occurred simultaneously in an iterative manner, each informing the other. Initially, the data were read to ensure familiarisation with the data, and reflective memoing was used to improve immersion and engagement with of data and to document decision-making throughout the research process. Initial codes were generated, and earlier transcripts were repeatedly re-examined following the completion of each further interview to allow for ongoing comparisons across the dataset. A code book was created to allow for discussion between authors about the codes.

The initial coding scheme was constantly refined during the data collection and analysis phase. Once the coding was refined, all codes were analysed and categorised into potential themes. Finally, the data were analysed to elaborate the relationships between the codes and categories, with the raise to the analytical level from categorical to conceptual. These themes were then reviewed and refined. It was at this point that the data were then considered in light of the existing model. We refined the conceptual model based on our study findings, examining how these study data elaborated or contradicted these theoretical findings. Throughout the collection and analysis process, the authors met regularly to discuss the codes, themes and interpretative models. A complete consensus was achieved. Ethics approval was obtained from Flinders University (ID:2379).

3 | RESULTS

Twenty interviews were undertaken, 12 assessors and 8 post-graduate trainees. There were 11 females and 9 males from a variety of specialties (General Practice, n = 10, internal medicine, n = 5, surgery, n = 4, obstetrics and gynaecology, n = 1). The post-graduate trainees ranged from first to final year of training, and assessors ranged from 5 to 28 years of experience. All of the assessor participants were involved in on-the-ground supervision. Nineteen of the interviewees shared at least one personal story of perceived unfairness in addition to the vignettes. The data from the vignettes and stories were coded together.

Saturation was reached after 19 interviews. After initially being coded into 115 codes, the participants' perceptions of fair judgement are characterised by 3 main themes, with 9 sub-themes. These themes were organised into individual (evidence, narrative, boundaries, agility and expertise), system (multiple assessors, multiple opportunities, documentation and procedural fairness) and environmental factors and compared with the theoretically derived conceptual model from our literature review.[7] The perspectives of the assessors and learners supported the literature model and added further detail. The relationship between different components was also established and the conceptual model modified accordingly (see Figure 1).

3.1 | Individual characteristics

3.1.1 | Fair judgement decisions need to contain meaningful and constructive narratives

A narrative was seen to be essential for a judgement to be fair; as narratives allow for learner reflection and improvement through feedback. A judgement was only considered fair if there was a clear, meaningful feedback narrative about how a learner could improve their performance. And as such it automatically signals that the learner's best interest is at the centre.

> It's unfair because everybody needs communication to continue to enhance your performance and help you grow and you develop… So the unfairness is that you're not going to learn here.

Furthermore, a narrative is needed to align the learner and assessor's perspectives on how the learner is performing. It is the responsibility of the assessor to ensure they have attempted to inform the learner of how they are performing against expectations. A surprise judgement is considered unfair.

> I did have some issues … but it wasn't brought to my attention when it happened. Because everything just went on, so I didn't think it was a big deal.

Furthermore, fair judgements need to be equitable in that all learners have the opportunity to be genuinely judged and provided with feedback, not just those who are struggling.

3.1.2 | Fair judgements fall within boundaries

Fair judgement decisions are based on evidence which is 'within scope' and what is 'out of scope'; or in other words what is in or out of bounds.. It is considered unfair to be assessed as 'competent

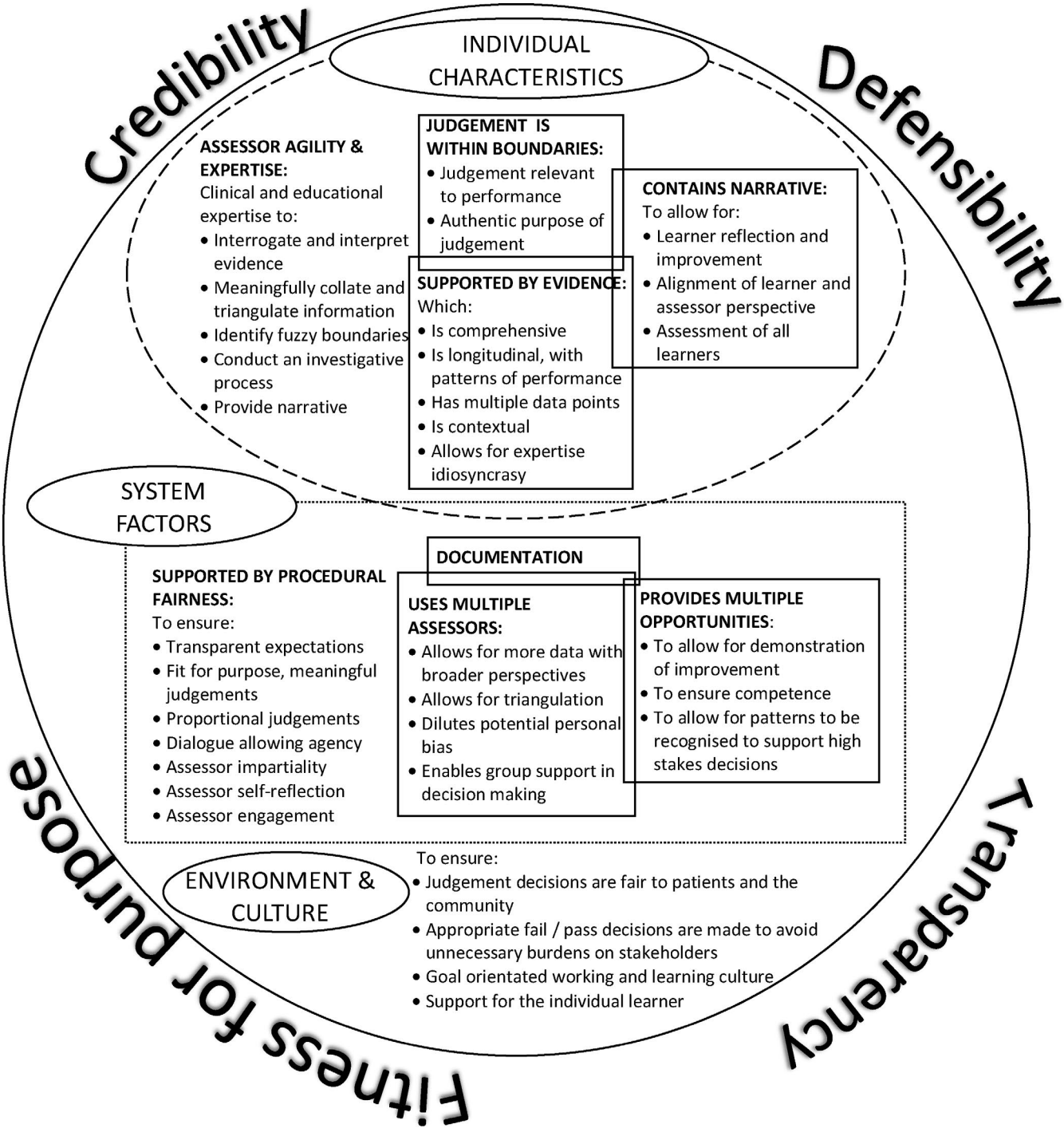

FIGURE 1 A conceptual model of the components of fair judgement in assessment

or incompetent by proxy'; when factors other than clinical performance are used in making assessment judgements. The boundaries of fair judgement also help determine the credibility of the assessors because the credibility of the judgement 'message' is seen as a function of both the message itself and the 'sender'. This study highlighted several sub-themes related to boundaries.

Firstly, judgement decisions need to be relevant to remain within boundaries. As supported by the literature review, factors such as gender, race, family, likability and social connections are not considered relevant to competence and are considered unfair.

> …keeping that boundary which can be a little bit trickier… I have to be very conscious then about separating this is a particularly lovely person and I've seen photos of their kids… from their clinical performance.

Secondly, judgement decisions which had a misplaced purpose, where the decision was not made in the best interests of the learner or patients, were considered outside of the boundaries of what is fair. It was considered reasonable to have high expectations of a learner and to fail if needed, but judgements need to be made in the light of having

an authentic, genuine aim of wanting learners to improve and succeed, to ensure they are able to provide excellent health care. Any other aim, such as assessor self-interest including an unwillingness to share their private judgement decisions, gossiping about learners, pushing their own agenda or abusing their role as an assessor is considered out of the boundaries of a fair judgement.

> If you've got somebody who is interested in helping that junior doctor become a better doctor and who actually wants to intervene not because they're interested in tearing someone apart, but because they go okay… if you can help them then we get a better doctor at the end of it

> I absolutely know for a fact that some registrars will be given borderline passes rather than fails because it's easier.

3.1.3 | Fair judgement decisions are supported by supporting evidence

The literature review noted evidence was a means of supporting judgements and suggested that having multiple sources of evidence improved the perception of fairness. In this study, participants agreed with these premises and provided detail about what this means in practice. Evidence in this context was considered to include such things as rationale, artefacts or observation.

For judgement decisions to be fair, there needs to be comprehensiveness of evidence. Multiple competencies are needed to be a competent clinician and fair judgement decisions consider all of these competences, not just knowledge.

> In order for me to feel that I'm being treated fairly I need to feel that they've assessed my different skills that I have, not I'm being judged on one skill and that's it

Evidence was expected to be longitudinal and consider patterns of performance to be considered fair. Having multiple pieces of evidence allows for triangulation.

> …you'd have a look at the morbidity/mortality meetings. Is he over represented in that? What's his approach to when something goes wrong and what are his communication skills like with the families? Have any of the families complained?

Importantly, evidence needs to be contextual to be considered fair. An important role of an assessor is to interpret evidence in light of the context. This is explored further when considering expertise and agility.

> …was it an emergency after hours where if you didn't give it a go, the person was going to die, versus there was someone in the next room who could've helped you and you didn't ask

Finally, evidence for judgement decisions should allow for expertise idiosyncrasy. Different clinicians will have different individual ways of practising and this variation is not necessarily incompetence, so to judge someone as such is considered unfair.

> I can say you know … I think you managed that differently to how I would've but you did really well.

3.1.4 | Assessors making judgement decisions need agility, and content and assessment expertise

All participants highlighted the need for assessor expertise and agility. Lombardo and Eichinger coined the phrase mental agility to describe the degree to which individuals think through problems from fresh points of views are comfortable with complexity, ambiguity and explaining their thinking to others[23] Interviewees noted that to make fair judgements, assessors have multiple tasks for which they need agility and expertise to complete. These include embracing the complexity of the situation and meaningfully collating and triangulating pieces of evidence that cannot be added numerically through interpreting and weighing up evidence presented and considering the quality and context of the evidence, within identified fuzzy boundaries. This was considered a key role of an assessor, and if this was not done, the judgement decision was considered unfair. This also often occurs with time pressures as assessment usually occurs in real life, and judgement is needed to be made in real time to ensure patient safety.

> Sometimes the trainees are not very good in terms of professionalism but then the patients love them. So it is a matter of interpreting that comprehensive assessment

> He wrote something on it like this has never been my impression of you [name removed] in any of my interactions… at least it made me feel… maybe he realised it wasn't a reflection of me after all.

To be able to adequately interpret, interrogate and combine the evidence presented in a fair way, an investigative process is needed. This may involve collecting more evidence, or identifying more information about the evidence presented.

> I grill the consultants a bit more and find out what's the underlying issue and I get them to try and describe the scenario, what was the situation, what happened and who was there… I just go and chat to the people in that situation… and find out what people's version of events were

Furthermore, assessors need educational expertise to ensure they are able to provide narrative feedback which can allow for improvement.

3.2 | System factors

3.2.1 | Fair judgement decisions have allowed for multiple opportunities

Fair judgments about progression in training programmes need to have provided multiple opportunities for learners to demonstrate competence over a period of time to allow for multiple data points to be collected, patterns of performance to be recognised and to reduce the chance of an external factor (ie unwell on the day of an assessment) influencing their ability to demonstrate competence. Specifically, this study emphasised that learners need to also have a time and work opportunity to respond to narrative feedback and demonstrate improvement before the next assessment or the end of term.

> …it's almost like two strikes and you're out, but they've only had one shot to improve themselves so I think that it's unfair in that aspect.

Having multiple opportunities also was seen as possibly making the task of failing a candidate easier, because there were multiple data points and check points to support the decision.

> Failing someone is much harder than passing them in terms of actually the workload… the cognitive load, the emotional load, but actually the documentation and the conversations and those sorts of things are much bigger and I guess if there were more perhaps slightly smaller check points and processes built in all the way through for everybody then perhaps it's not as big of a monumental job to fail someone.

3.2.2 | Multiple assessors are used in fair judgement decisions

This study confirmed the findings of the literature review that using multiple assessors is perceived to contribute to fairness, because it enables more data to be collected which allows for triangulation and for a broader range of competencies to be assessed.

…you really do have to triangulate and get different points of view.

> In fact, even more important than medical staff is non-medical staff. So, it's often nursing staff, allied health staff, patients, that will give a much more true [sic] picture of an individual's performance rather than medial staff.

Multiple assessors also allow for diverging perspectives and dilutes any one individual assessor's single perspective. This is not to necessarily ignore the judgement of any individual assessor but rather to consider this in the light of other judgement decisions. As such, it relates to the issue of allowing for expertise idiosyncrasy as described above.

> it's not just one person's opinion. I think that's really important, failing a term, that it's not just a personality clash or something… So, in essence that is fair.

Having multiple assessors also allows for group support in making judgement decisions, particularly difficult decisions.

> I think it was very much a team decision… we all felt that we'd reached the limit of what we could offer him

3.2.3 | Documentation

To ensure transparency, all facets of the judgement need to be documented. There was minimal discussion by participants on documentation, so details of what and how documentation should occur are uncertain.

3.2.4 | Procedural fairness supports fair judgement decisions

The literature review identified the importance of procedural fairness in fair judgements, but the concept was not further defined conceptually. This study helped provide detail about what procedural fairness may look like from the perspective of the learner and assessor.

An important component of procedural fairness is transparency of expectations of the learner. Transparency relies on the information to be explicit and comprehensive; a lack of information can mean learners are required to guess what is expected of them and may use their previous experience as a guide. Judging a learner on unwritten or uncommunicated expectations is therefore seen as unfair, even when only part of the expectations were not explicitly communicated.

> I wasn't oriented to the unit and what was expected… coming from India… when the registrar is talking about the patient you just stay quiet…[I was told] you do not contribute to ward rounds and I said… I don't know what I need to say, I can just give you the results and give you what information you require but I'm not going to butt in and that was a cultural shock to me… Now they have made it very transparent, now they have made it necessary we have job assessment.

Procedural fairness includes ensuring judgements are fit-for-purpose. Arbitrary rules or judgements lacking a meaningful rationale are seen as procedurally unfair. Examples are rigid, predetermined assessment forms which do not allow for assessor agility and expertise or judgements about elements that do not intuitively contribute to becoming a better practitioner. Typically, such unfairness can lead to gaming of the assessment and learners feeling forced to focus on passing the assessment rather than becoming the best possible healthcare professional, which is not seen as fair.

> 10% is actually really not meaningful when it's just a rule for the sake of a rule

Importantly, fair judgements have to be proportional, with alignment of the stakes of the decisions and the richness of the information on which they are based.

> …why would one exam constitute a failure in the whole year?… this is the whole year of somebody's life… This is high stakes, is it fair that somebody has to do a whole year because they failed one exam?… There has to be some rationale behind why does this particular segment of the exam carry with it such an important predictor of future professional competence or capability.

Procedural fairness importantly included allowing learners to speak and provide their perspective to the situation. This dialogue and perspective need to be considered by assessors to make fair judgements. Or in other words, the learner feels that they can assume agency over their own learning and a dialogue is a way to enable this.

> …then, as part of any kind of fair trial the accused should have an opportunity to defend themselves… present the complaints… and hear the junior consultant's side of the story

> … during that time I had been sexually harassed, I had been told was I sure I wanted to be a doctor, I hear you like baking are you sure you just don't want to spend your time in the kitchen… I was devastated that the Head of the Rural School hadn't said to me [name removed] what's your opinion on this? I was never given the opportunity to say.

Procedural fairness needs to ensure hierarchical power differentials do not hinder the provision of information, judgement or feedback to the learner, or if the learner is unable to respond as this is seen as unfair. Such power differential could flow from the assessor to the learner or from learner to assessor. Furthermore, an important dilemma in procedural fairness is deciding between assessors having prior knowledge about a candidate which may provide useful information for a more balanced judgement on the one hand and the notion of remaining objective on the other. From a perspective of fairness, judgements can be fair in both circumstances. Whilst assessors may have a genuine need to discuss learners from a continuity of care perspective, this clearly needs to be balanced with the risk of creating a 'reputation' for the learner that may bias future judgements. It was seen as unfair if a learner was prejudged and their assessment considered on hand-over factors rather than their clinical performance as this was outside the boundaries of fair assessment.

> I think in some ways it can be helpful if they know you well, they can give you constructive feedback and constructive views of your strengths. But I think also as the person being supervised, you need to feel like you can talk to your supervisor about things that you're struggling with and so if you then feel like the supervisor is going to flip it back on you and assess you poorly because you've sought their help and support, I think that's unfair.

> There's a colleague… who has made a very bad impression to one or two of the consultants and word of mouth has spread and I think a lot of the other teams are then very very carefully watching this person and putting them under scrutiny… it's a bit unprofessional and unfair because… the whole division is biased against this particular trainee.

Procedural fairness also includes assessor self-reflectivity. This might include being aware of their own susceptibility to biases and how personality characteristics can impact judgement decisions. This is seen as an unfair influence that can be mitigated if the assessor makes the effort of reflection.

> when I'm doing an assessment I have to think to myself… am I being too hard on them because I have a tendency to be hard on myself and therefore I expect it from others too. I think you have to have an understanding of your own interpretation of the world to be a fair assessor of others

Finally, judgement decisions from assessors only marginally engaged in assessment are considered unfair. Engagement includes spending sufficient time on the assessment, making the effort to observe learners in the assessment process and taking responsibility for a learner's assessment, having their best interest at heart. Furthermore, all staff within the assessment system, not just those directly responsible for assessment, have a responsibility to communicate with the learner if they have any concerns with their performance.

> I've had a lot of generic assessments.. from assessors who haven't really taken the effort to actually go speak to the [junior doctor] supervising me

> I personally think that the Head of Unit is just as much fault if not more than the junior consultant… because if you don't have a Head of Unit willing to take responsibility [for assessment]… then that is going to cause a big systemic problem

3.3 | The environment and culture

This study highlighted another component to fair judgement that is the environment in which the judgement decisions are made. Learners are future health professionals, and there is community expectation they are well trained. Judgement decisions are, therefore, seen as fair if they consider the impact on patient care and the community, including their working community. To be fair to patients, learners need to meet expectations or earn the right to further opportunities. If there was a tension between fairness to the patient and fairness to the learner, fairness to the patient was seen as more important.

> …but ultimately the person at the centre of this is the patients… So that's how I would actually view this whole thing.

> you start to wonder how many opportunities the trainee will have despite feedback and is it unfair let's say on the program, the taxpayer, or patients to expect the institution to constantly support someone who may never have shown the aptitude.

Furthermore, not making difficult judgements was seen as unfair as it may deny learners opportunities to improve earlier in training with less high-stakes consequences. It also may lead to unnecessary burdens for colleagues who are required to work with an unidentified struggling learner, and future assessors who have to make even higher stakes decisions with graver ramifications.

> There is a [doctor specialty removed] who very famously got through her training by involving lawyers. So she gave feedback that her assessments were unfair and she got lawyers involved and she ended up passing… when I was a very junior registrar… there was a day where it was horrendously unsafe… I was not supported by a consultant [the one mentioned earlier] who had adequate skills. And so she [the consultant] got into a job that there were very clear red flags she was not going to be able to do, it put me in a situation where I was having to act above my skillset, I ended up going into the toilet calling a [speciality removed] consultant and saying you need to come… she ended up getting fired

> I know that that person had difficulty with getting jobs in advanced training. I think it's a bit unfortunate to be told oh yeah you're fine, you're fine, you're -fine, and then oh yeah you haven't got a job [because we failed to fail you]

Judgement of learners is only considered fair if the learning environment allows for learning and has a culture of wanting the learner to improve for the sake of patient care and the learner themselves. This includes ensuring relevant skills and knowledge are taught, an appropriate workload, an opportunity to express learning needs and a culture of feedback.

> …that junior consultant might be very competent and very good at their job and just not in an environment that makes that possible for them to achieve.

Fair judgements can only occur in an environment which considers learners' personal unique circumstances, particularly when learners are not meeting expectations.

> What I think we should do with the struggling registrar is decide whether it's fair to compare their progress… with the registrar who is flying, I think that's probably unfair. Then what we've got to decide is whether they need more training, and we need to give them more opportunities to improve.

4 | DISCUSSION

The findings from this study support the conceptual model previously derived from the literature.[7] This study noted that fair judgement in assessment is multi-dimensional, complex and contextual. It highlighted the`re are individual characteristics to fair judgement, specifically narrative, evidence, boundaries, agility and expertise. But, where the literature review suggested these characteristics are interlinked, parallel characteristics, in this study we found a different relationship. This study highlighted that agility and expertise were encompassing of the other characteristics, as agility and expertise were essential to provide narratives, to consider available and possible missing evidence and interpret this within boundaries.

Judgement decisions are always made within assessment and educational systems, and systems can both enable and restrict fair judgement decisions such as through infrastructure, time, resources, rules, cultures and regulations. In considering the impact of system factors on fair judgement in this study, the relationship between the different components was also refined compared to the outcome of the literature review. We identified that multiple assessors, multiple opportunities and documentation are needed for fair judgement decisions and procedural fairness provides the framework to allow these system components to occur. But procedural fairness can be difficult to define, and this study provided a clearer idea of what this means in practice when related to fairness of assessment

judgements. Notably, 'documentation' was only scarcely and superficially mentioned by the study participants, whereas it was more prominent in the literature review. However, program designers may have a different perspective on this, and this is an area for future research.

This study also highlighted more clearly the role of the environment in judgement decisions. Training of health professionals does not occur in a vacuum and fair judgement decisions must consider the impact on patients, colleagues and the wider community. Whilst there were some inferences of this within the literature, the concept of environmental culture was much more prominent in this study. The breadth and frequency of codes related to this theme far greater than in the literature review and the passion with which the learners and assessors spoke about the environmental culture were unexpected. We interpret this as being a representation of their lived experience of judgement in busy workplace-based environments, and their ability to see the impact of these environments first hand. All of the study findings helped to further refine and build the conceptual model.

Our findings have relevance in the perspective of modern ideas about assessment. Workplace-based assessment has been recognised by many authors as a complex system.[11,26] Where the system is complex, the solution likely needs to be as complex as the problem itself[27] and the dynamic and unpredictable nature of complex systems logically precludes the effective use of reductionist values and methods.[28] But despite the non-linear dynamics of complex systems, there are still boundaries, internalised rules and a requirement for constant adaption to the changes within the system.[29] With prolonged observation, patterns and networks can still be revealed.[24,29] Our model aims to allow stakeholders to navigate complexity by identifying rules or definitions of approaches, networks and patterns, and highlight relationships between different components without reducing the complexity.

This links to another predominant idea in medical education; programmatic assessment. Programmatic assessment principles include the use of multiple pieces of data, longitudinal assessment, proportionality and meaningful triangulation of data allowing for rich information-based decision-making and meaningful feedback to the learner.[30] This study's data supports all of these premises. Having multiple assessments and assessors allows for more data and perspectives to be collected, patterns to be identified, member checking and triangulation to take place, and to allow for a broader range of competencies to be assessed.[31-33] In programmatic assessment, it is acknowledged that data cannot be simply numerically collated or even that it will be contextually similar, and that easy addition of assessment components is not valid for the assessment of complex competence. On the contrary, data which are heterogenous need to be meaningfully triangulated, considering the context of the judgement. Within the literature, it has been recognised that specific expertise is needed to consider context in the combination of data.[24,34-36] Additional tools such as narrative, boundaries and assessor agility are needed to do this, as noted in the model.

This study particularly emphasised that fair judgement is not a one-size-fits-all; the specific situational characteristics and the context must be included for it to be considered fit-for-purpose. Expert and agile assessors are required to collate, interrogate, interact with and interpret the evidence within fuzzy boundaries and context of the situation. This was one of the most prominent codes present in this study and voiced in all 20 interviews. Surprisingly, this is so fundamentally—one would say epistemologically—at odds though with the idea of a standardised, measurement-based assessment. Van der Vleuten noted that rather than striving for perfect reliability among raters, a more appropriate goal would be to develop rigorous methods of collecting and synthesising assessment data in a program of assessment.[30] Perhaps, this study's finding suggests stakeholders recognise this and the need to move forward from the idea that performance rating in the workplace is not as much about measurement as it is about expert 'judgement' in a dynamic system environment.[11,34] The corollary of this is that inter-judge disagreement is not necessarily unfair as long as each judge has sufficient expertise to add a fair and valuable perspective.[15]

The need for meaningful and actionable feedback and agreement between the assessor and learner is an important aspect in an assessment for learning philosophy.[37-40] Lee argues that the use of specific narratives and contextual comments may be more informative for trainees than the judgement itself.[41] Our study supported these ideas. Both learners and assessors perceived judgements to be only fair if they allowed for learning, through the provision of feedback about how the learner could improve. Assessment for learning can only occur in a learning and working culture, where learners can practice purposefully, and errors typically become learning opportunities.[42,43] This study also noted such an environment was essential for judgement decisions to be accepted as fair.

Our data suggest that embracing fair, subjective judgements can present challenges. For many institutions, this may be a cultural change[44,45] and there may be faculty skill gaps and difficulty in making adaption to new and epistemological unfamiliar methods of assessment.[41,44] This being said, however, many of the components of fair human judgement identified by this literature review are not necessarily new. The use of multiple assessors, longitudinal assessments and collection of multiple pieces of evidence is common in many institutions.[46] Transparent expectations, orientations, procedures and documentation are also common in most training programmes. The importance of feedback is increasingly recognised in assessment and the role of narrative has become more prominent as many acknowledge that numbers alone are not sufficient for learning.[47-51] And finally, the learning environment has been gaining increasing attention in the medical education literature.[42] From a practical point of view, specifically ensuring assessment programmes require contextual evidence as justification for decisions, have provision for feedback narrative throughout the programme, identify what is considered to be 'within scope' for judgement decisions and engage expert assessors to meaningfully collate and triangulate information will help to ensure judgement decisions are considered 'fair'. Furthermore, institutions can ensure multiple

assessors are used in assessment programmes, decisions are well documented, expectations of candidates are transparent and the environment in which the decisions is made considers patient needs and learner circumstances.

There are limitations to this study. Our study focussed on stakeholder conceptualisation of learners and assessors. It, therefore, did not include medical students or program designers who are also important stakeholders in the conversation of fair judgement decisions. It is likely that program designers and academics particularly would have an additional perspective, and follow-up studies with such groups may highlight further important aspects or shed new perspectives on those already identified. Any further, important caveat is the fact that this study was done from within a Western-oriented cultural context. It is plausible to assume that certain cultural dimensions have been so implicit in the literature and interview data that they may put a limit on the generalisability of our model. We would not only argue for further studies with different stakeholders in our own cultural context but also for replication in different cultural contexts.

5 | CONCLUSION

Woodruff noted that the challenge for medical education researchers is to not be distracted by 'solutions' but to look at problems more deeply.[28] Whilst a simple, universally agreed upon definition of fairness may at first glance appear to be desirable, delving deeper to better understand what the foundations of fair judgement are may allow for a more useable narrative for training institutions to negotiate what fair judgement actually is. This study builds on the theoretically derived conceptual model and demonstrates that components of fair human judgement can be explicitly articulated whilst still embracing the complexity and contextual nature of health-professions assessment. Thus, it provides a narrative to support dialogue between learner, assessor and institutions about ensuring fair judgements in assessments. This model is not to be considered yet another checklist, but rather creating a shared understanding about what fairness of human judgement in assessment is.

ACKNOWLEDGEMENTS

The views expressed herein are those of the authors and not necessarily those of the Department of Defense or other federal agencies.

CONFLICT OF INTEREST

No competing interests.

AUTHOR CONTRIBUTIONS

All authors contributed to the design of the research and drafting of the manuscript. All authors gave approval and agree with the final publication.

ETHICAL APPROVAL

Ethical review was obtained from Flinders University (ID:2379).

ORCID

Nyoli Valentine https://orcid.org/0000-0002-3526-5012
Lambert Schuwirth https://orcid.org/0000-0002-6279-5158

REFERENCES

1. Green SK, Johnson RL, Kim DH, Pope NS. Ethics in classroom assessment practices: issues and attitudes. *Teach Teach Educ*. 2007;23(7):999-1011.
2. Valentine N, Schuwirth L. Identifying the narrative used by educators in articulating judgement of performance. *Perspect Med Educ*. 2019;8(2):83-89.
3. Ten Cate O, Regehr G. The power of subjectivity in the assessment of medical trainees. *Acad Med*. 2019;94(3):333-337.
4. Van der Vleuten CP, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ*. 1991;25(2):110-118.
5. Kane M. Validation. In: Brennan RL, ed. *Educational Measurement*. Westport, CT: ACE/Praeger; 2006:17-64.
6. Desy J, Coderre S, Davis M, Cusano R, McLaughlin K. How can we reduce bias during an academic assessment reappraisal? *Med Teach*. 2019;41(11):1315-1318.
7. Valentine N, Durning S, Shanahan EM, Schuwirth L. Fairness in human judgement in assessment: a hermeneutic literature review and conceptual framework. *Adv Health Sci Educ Theory Pract*. 2021;26(2):713-738.
8. Hauer KE, Lucey CR. Core clerkship grading: the illusion of objectivity. *Acad Med*. 2019;94(4):469-472.
9. Rotthoff T. Standing up for subjectivity in the assessment of competencies. *GMS J Med Educ*. 2018;35(3).Doc29
10. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach*. 2013;35(7):564-568.
11. Schuwirth LW, van der Vleuten CP. A history of assessment in medical education. *Adv Health Sci Educ Theory Pract*. 2020;25(5):1045-1056.
12. Schuwirth LW, van der Vleuten CP. A plea for new psychometric models in educational assessment. *Med Educ*. 2006;40(4):296-300.
13. Eva KW, Hodges BD. Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? *Med Educ*. 2012;46(9):914-919.
14. Govaerts M, van der Vleuten CP. Validity in work-based assessment: expanding our horizons. *Med Educ*. 2013;47(12):1164-1174.
15. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the 'black box' differently: assessor cognition from three research perspectives. *Med Educ*. 2014;48(11):1055-1068.
16. Jones A. The place of judgement in competency-based assessment. *J Vocat Educ Train*. 1999;51(1):145-160.
17. Bacon R, Williams L, Grealish L, Jamieson M. Credible and defensible assessment of entry-level clinical competence: Insights from a modified Delphi study. *FoHPE*. 2015;16(3):57.
18. Boursicot K. Consensus Statement Reports: Performance Assessment. Ottawa 2020. 2020. Kuala Lumpar, Malayasia.
19. Muller J. The Tyranny of Metrics: On the Use and Misuse of Metrics in Medicine and Education. Paper presented at: AMEE Conference2020; Online.
20. Tierney RD. Fairness in classroom assessment. In: McMillan JH , ed. *SAGE Handbook of Research on Classroom Assessment*. Thousand Oaks, CA: SAGE Publications; 2013;125.
21. Braun V, Clarke V. Thematic analysis. In: Cooper H, PM Camic, DL Long, AT Panterc, D Rindskopf, & KJ Sher, eds. *APA handbooks in psychology®*. APA handbook of research methods in psychology, Vol. 2. Research designs: Quantitative, qualitative, neuropsychological, and biological. Worcester, MA: American Psychological Association; 2012:57-71.

22. Steiner PM, Atzmüller C, Su D. Designing valid and reliable vignette experiments for survey research: a case study on the fair gender income gap. *J Methods Meas Soc Sci*. 2016;7(2):52-94.
23. Lombardo MM, Eichinger RW. High potentials as high learners. *Hum Resour Manag*. 2000;39(4):321-329.
24. Cleland J, Durning SJ, eds. *Researching Medical Education*. Chichester, UK: John Wiley & Sons; 2015.
25. de Feijter JM, de Grave WS, Dornan T, Koopmans RP, Scherpbier AJ. Students' perceptions of patient safety during the transition from undergraduate to postgraduate training: an activity theory analysis. *Adv Health Sci Educ Theory Pract*. 2011;16(3):347-358.
26. Durning SJ, Artino AR Jr, Pangaro LN, Van Der Vleuten C, Schuwirth L. Perspective: redefining context in the clinical encounter: implications for research and training in medical education. *Acad Med*. 2010;85(5):894-901.
27. Glouberman S, Zimmerman B. Complicated and Complex Systems: What Would Successful Reform of Medicare Look Like? *Commission on the Future of Healthcare in Canada: Dsiscussion Paper No 8. 2002*.
28. Woodruff JN. Solutionism: A study of rigour in complex systems. *Med Educ*. 2021;55(1):12-15.
29. Rosas SR. Systems thinking and complexity: considerations for health promoting schools. *Health Promot Int*. 2017;32(2):301-311.
30. van der Vleuten CPM, Schuwirth LWT, Driessen EW, et al. A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34(3):205-214.
31. Dijkstra J, Galbraith R, Hodges BD, et al. Expert validation of fit-for-purpose guidelines for designing programmes of assessment. *BMC Med Educ*. 2012;12:20.
32. Dijkstra J, Van der Vleuten CP, Schuwirth LW. A new framework for designing programmes of assessment. *Adv Health Sci Educ Theory Pract*. 2010;15(3):379-393.
33. Driessen E, van der Vleuten C, Schuwirth L, van Tartwijk J, Vermunt J. The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. *Med Educ*. 2005;39(2):214-220.
34. Govaerts MJB, Schuwirth LWT, Van der Vleuten CPM, Muijtjens AMM. Workplace-based assessment: effects of rater expertise. *Adv Health Sci Educ Theory Pract*. 2011;16(2):151-165.
35. Marewski JN, Gaissmaier W, Gigerenzer G. Good judgments do not require complex cognition. *Cogn Process*. 2010;11(2):103-121.
36. Govaerts M, Van de Wiel M, Schuwirth L, Van der Vleuten C, Muijtjens A. Workplace-based assessment: raters' performance theories and constructs. *Adv Health Sci Educ Theory Pract*. 2013;18(3):375-396.
37. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. *Med Teach*. 2011;33(6):478-485.
38. Watling CJ. Unfulfilled promise, untapped potential: feedback at the crossroads. *Med Teach*. 2014;36(8):692-697.
39. Lockyer J, Carraccio C, Chan M-K, et al. Core principles of assessment in competency-based medical education. *Med Teach*. 2017;39(6):609-616.
40. Cantillon P, Sargeant J. Giving feedback in clinical settings. *BMJ*. 2008;337:a1961.
41. Lee V, Brain K, Martin J. Factors influencing mini-CEX rater judgments and their practical implications: a systematic literature review. *Acad Med*. 2017;92(6):880-887.
42. Young J, Williamson M, Egan T. Students' reflections on the relationships between safe learning environments, learning challenge and positive experiences of learning in a simulated GP clinic. *Adv Health Sci Educ Theory Pract*. 2016;21(1):63-77.
43. Turner S, Harder N. Psychological safe environment: a concept analysis. *Clin Simul Nurs*. 2018;18:47-55.
44. McDonald JA, Lai CJ, Lin MYC, O'Sullivan PS, Hauer KE. "There Is a Lot of Change Afoot": a qualitative study of faculty adaptation to elimination of tiered grades with increased emphasis on feedback in core clerkships. *Acad Med*. 2021;96(2):263-270.
45. Bullock JL, Lai CJ, Lockspeiser T, et al. In pursuit of honors: a multi-institutional study of students' perceptions of clerkship evaluation and grading. *Acad Med*. 2019;94(11S):S48-S56.
46. Hauer KE, Cate OT, Boscardin CK, et al. Ensuring resident competence: a narrative review of the literature on group decision making to inform the work of clinical competency committees. *J Grad Med Educ*. 2016;8(2):156.
47. Watling C. Cognition, culture, and credibility: deconstructing feedback in medical education. *Perspect Med Educ*. 2014;3(2):124-128.
48. Konopasek L, Norcini J, Krupat E. Focusing on the formative: building an assessment system aimed at student growth and development. *Acad Med*. 2016;91(11):1492-1497.
49. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. *Med Educ*. 2019;53(1):76-85.
50. Eva KW, Bordage G, Campbell C, et al. Towards a program of assessment for health professionals: from training into practice. *Adv Health Sci Educ Theory Pract*. 2016;21(4):897-913.
51. Ericsson KA. An expert-performance perspective of research on medical expertise: the study of clinical performance. *Med Educ*. 2007;41(12):1124-1130.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Valentine N, Shanahan EM, Durning SJ, Schuwirth L. Making it fair: Learners' and assessors' perspectives of the attributes of fair judgement. *Med Educ*. 2021;55:1056–1066. https://doi.org/10.1111/medu.14574