Advances in Health Sciences Education
https://doi.org/10.1007/s10459-020-10002-1

REVIEW

Fairness in human judgement in assessment: a hermeneutic literature review and conceptual framework

Nyoli Valentine[1] · **Steven Durning**[2] · **Ernst Michael Shanahan**[1] · **Lambert Schuwirth**[1]

Received: 13 May 2020 / Accepted: 19 October 2020

Abstract
Human judgement is widely used in workplace-based assessment despite criticism that it does not meet standards of objectivity. There is an ongoing push within the literature to better embrace subjective human judgement in assessment not as a 'problem' to be corrected psychometrically but as legitimate perceptions of performance. Taking a step back and changing perspectives to focus on the fundamental underlying value of fairness in assessment may help re-set the traditional objective approach and provide a more relevant way to determine the appropriateness of subjective human judgements. Changing focus to look at what is 'fair' human judgement in assessment, rather than what is 'objective' human judgement in assessment allows for the embracing of many different perspectives, and the legitimising of human judgement in assessment. However, this requires addressing the question: what makes human judgements fair in health professions assessment? This is not a straightforward question with a single unambiguously 'correct' answer. In this hermeneutic literature review we aimed to produce a scholarly knowledge synthesis and understanding of the factors, definitions and key questions associated with fairness in human judgement in assessment and a resulting conceptual framework, with a view to informing ongoing further research. The complex construct of fair human judgement 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, agility and evidence) and at a systems level by procedures (procedural fairness, documentation, multiple opportunities, multiple assessors, validity evidence) which help translate fairness in human judgement from concepts into practical components.

Keywords Assessment · Fairness · Health professions education · Judgement · Subjective

✉ Nyoli Valentine
vale0046@flinders.edu.au

[1] Prideaux Health Professions Education, Flinders University, Bedford Park 5042, SA, Australia

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

Published online: 29 October 2020

Introduction

Fairness is a fundamental quality of health professions assessment and is commonly accepted as a student's right (Robinson 2002). Traditionally, objectivity has been seen as the predominant way to ensure fairness in assessment and for much of the twentieth century health professions education research and development focussed on construct validity and reliability in assessment (Valentine and Schuwirth 2019; van der Vleuten et al. 1991; ten Cate and Regehr 2019). Over the last few decades, evolving ideas about learning, shifting social ideals and understandings of the limitations of high stakes tests led to many changes within our field. Competency-based education became the dominant approach to medical education in many countries (ten Cate 2017). With this, the role of the clinician has been redefined to include features previously not been emphasised, and learners certified on outcome rather than input (ten Cate and Billett 2014). Competencies have been defined into professional tasks which a learner is entrusted to complete independently (ten Cate and Scheele 2007). Assessment of clinical competence moved from written assessments back into the authentic context of the workplace, and individual assessments made way for programmes of assessment (Dauphinee 1995; van der Vleuten and Schuwirth 2005; Valentine and Schuwirth 2019). Despite these changes, objective approaches have remained a dominant discourse in assessment, with many seeing objectivity as the 'gold standard' to which assessments should be judged (Valentine and Schuwirth 2019; van der Vleuten et al. 1991; Govaerts and van der Vleuten 2013; ten Cate and Regehr 2019). Psychometric models have sought to define fairness from a measurement and quantitative perspective. Workplace based assessments, which utilise human judgement and are designed to assess authentic performance, have been judged using a quantitative framework and therefore criticised for not meeting validity and reliability criteria (Govaerts and van der Vleuten 2013). Using this objective perspective, human judgement is seen by many as too fallible and subjective to be used in high stakes assessment (Valentine and Schuwirth 2019). However an exclusive focus on traditional psychometric approaches can disregard key issues of competence, performance and assessment in complex workplace settings (Govaerts and van der Vleuten 2013; Govaerts et al. 2007), has been thought not be sufficient to capture competence in an academic setting (Boud 1990).

Throughout the literature, many authors have questioned this continued sole focus on objectivity, expressing a desire to better embrace subjective human judgement in assessment not as a 'problem' to be corrected psychometrically but as legitimate perceptions of performance (Jones 1999; Rotthoff 2018; Hodges 2013; ten Cate and Regehr 2019; Bacon et al. 2015; Govaerts and van der Vleuten 2013; Schuwirth and van der Vleuten 2006; Gingerich et al. 2014; Gipps and Stobart 2009). Most recently, in 2020, the Ottawa consensus statement report for performance in assessment specifically called for assessment programs to 're-instate expert judgement' (Boursicot 2020).

Taking a step back and changing perspectives to focus on the fundamental underlying value of fairness in assessment may help re-set the traditional objective approach and provide a more appropriate way to determine the appropriateness of subjective human judgements made in assessment. Changing focus to look at what is 'fair' human judgement in assessment, rather than what is 'objective' human judgement in assessment allows for the embracing of many different perspectives, and allows for the legitimising of human judgement in assessment. However, to do this requires addressing the question: what makes human judgements fair in health professions assessment? This is not a straightforward question with a single unambiguously 'correct' answer. Health professions assessment is

embedded in complex, unpredictable, contextual health care and education environments; it involves patients, institutions, supervisors and learners; and there are multiple, and at times conflicting, facets to both human judgement and fairness.

When faced with a multi-dimensional, complex construct without a simple definition, a shared language and understanding can be helpful. Heifetz noted "When people begin to use the same words with the same meaning, they communicate more effectively, minimize misunderstandings, and gain the sense of being on the same page, even while grappling with significant differences on the issues" (Heifetz et al. 2009). The aim of this literature review was to produce a scholarly knowledge synthesis and understanding of the factors, definitions and key questions associated with fairness in human judgement in health professions assessment, attempting to make ideas about fair human judgement explicit.

To further help manage this complex construct, categories and a resulting conceptual framework was developed, with a view to informing further research, enhancing communication and discussions about fair human judgement and provide assistance in the re-instatement of expert judgement in assessment programs.

Methods

Design

To achieve the aim of this review, we undertook a hermeneutic literature review. Understanding fairness in human judgement requires reviewing and compiling evidence from different disciplines and perspectives, considering unique contexts and complexity, and reviewing implications for many different stakeholders. Not surprisingly, this literature is vast, heterogeneous and without consensus answers from randomised controlled trials. A hermeneutic approach uses as cyclical rather than linear framework, and is concerned with the process of creating interpretive understanding. Papers are interpreted in the context of other papers from the literature and understanding is influenced by each new paper read (Boell and Cecez-Kecmanovic 2010). The popularity of a hermeneutic review is increasing as it has value in generating insights from heterogenous literatures which cannot be synthesised through systematic review methodology, and would otherwise produce inconclusive findings (Greenhalgh and Shaw 2017).

There were two main continuous cyclical processes in the review: the search and acquisition of articles and the analysis and interpretation of the articles obtained to develop an argument as demonstrated in Fig. 1 (Boell and Cecez-Kecmanovic 2014). Throughout the review an interpretive approach was used to meaningfully synthesize and critique the existing literature (Boell and Cecez-Kecmanovic 2014). Consistent with this approach, our literature search was rigorous but flexible and iterative, and as ideas were mapped, classified and critically assessed and the nature of the evidence became more apparent, there was further refinement of the research question (Boell and Cecez-Kecmanovic 2010).

Focus of the review

Following the steps outlined in Fig. 1 as best practice for a hermeneutic review, our literature review started with initial ideas. These formed our initial questions:

The initial questions addressed by our literature review were:

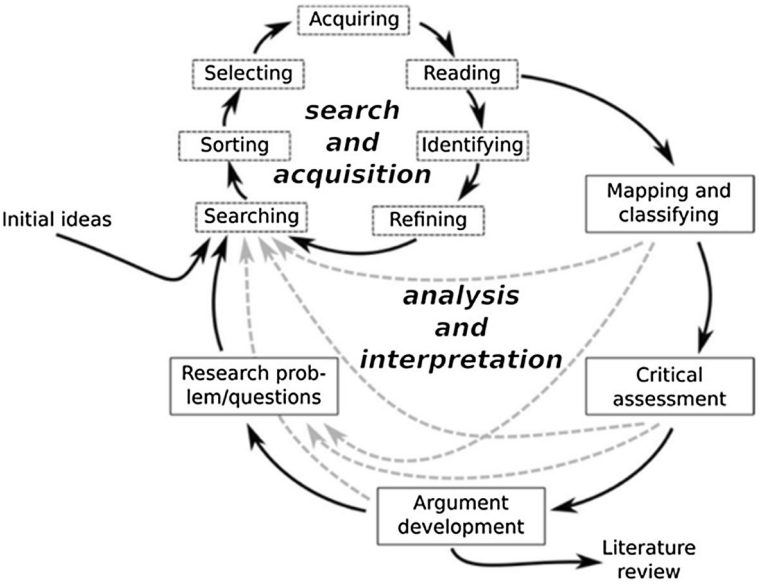

Fig. 1 The hermeneutic circle as a framework for the literature review (Boell and Cecez-Kecmanovic 2014)

1 What are the limitations of "objectivity" in medical assessment?
2 What is fair?
3 Can subjective human judgement in assessment be fair?
4 What is it about human judgement that makes it acceptable and defensible in clinical medicine?
5 What makes an assessor's judgement in assessments legitimate?
6 What are the subdimensions or components of fairness?
7 What is the relationship between these subdimensions?

Stages of the review

Stage 1: search and acquisition of evidence

In July 2019 NV began with the search strategy outlined in Fig. 2. Initial inclusion criteria were: peer review papers published prior to March 2020 (including reviews, perspectives, original research and case studies), with abstracts included and written in English, relating to either fairness or judgement within clinical practice, or health professions education, including medical education, or high school/tertiary education. Unlike a formal systematic review, we did not use an explicit strategy of excluding papers from the initial search results but rather a strategy of reading and evaluating papers and including them to build and saturate a development of arguments to address our identified questions. To add rigor, in addition to database searching, snowballing, and seminal searching was utilised. Consistent with the hermeneutic approach, in reviewing each title and abstract, the question was asked: "Is this paper likely to add meaning to our emerging overview of the field?" (Greenhalgh and Shaw 2017). The

Database Search Methods Used:

A comprehensive search was conducted over the databases PubMed & Google Scholar, which included all years until 2019, which was then extended to March 2020, to identify all possibly relevant studies / evidence / perspectives in English language on

- Fair*
- Object*
- Subject*

AND

- Medical education OR
- Education (including high school / university education) OR
- Assess* OR
- Post graduate OR
- Health Professions Education OR
- Portfolio OR
- Learn* OR
- Trainee*

A further search was used across the same databases to identify further relevant evidence / studies / perspectives in English language with regards to:

- Legal*
- Defensib*
- Defensible professional judgement

Subsequent targeted searches were undertaken. These included database search of PsycINFO, and also included other searches to further develop understanding of concepts which had arisen during the initial stages of the literature review. These searches included:

- Value*
- Narrative
- Expertise
- Holistic judgement

Fig. 2 Search strategies used in the literature review

- Opportunity
- Transparency
- Validity

AND

- Medical education OR
- Education (including high school / university education) OR
- Assess* OR
- Post graduate OR
- Health Professions Education OR
- Portfolio OR
- Learn* OR
- Trainee*

Snowballing: The reference lists of included articles were scanned for further relevant articles. The reference lists of these new publications were then reviewed to find yet more relevant titles.

Suggested articles and texts from expert group were also reviewed.

Seminal searching: Using citation tracking in Google Scholar to identify subsequent articles that had cited seminal sources.

Fig. 2 (continued)

literature searching took place over 9 months to allow for subsequent searches as new ideas emerged (Boell and Cecez-Kecmanovic 2010). Consistent with this approach, papers were re-reviewed in light of the new ideas over the search period. In addition, further targeted searches were then made to clarify concepts which had arisen during the review as identified in Fig. 2. There were no existing themes developed prior to starting the search. Having a less structured approached enhanced dialogical interaction between the literature and the researchers, encouraged critical assessment and supported argument development (Kusnanto et al. 2018). The focus of the search was fairness in human judgement in assessment in the context of health professions education rather than fairness in assessment more broadly. References were managed in an EndNote database. The expert authors also selected additional sources which were reviewed.

Stage 2: Data extraction, analysis and interpretation

Throughout the review NV created a narrative synthesis of the key questions, findings and scholarly arguments relevant to the research questions. This narrative synthesis was peer reviewed regularly by all authors throughout the literature review process. It was progressively refined by group discussions as described in Fig. 1. As is required of hermeneutic reviews, there was constant returning to stage 1 for further acquisition of evidence. The hermeneutic cycle was broken and left when a point of saturation was reached.

Stage 3: Development of a conceptual model

During the literature review process, a conceptual model of the definition of fairness in human judgement in health professions assessment was developed based on the literature review (Fig. 3). Initially, concepts and themes were sourced from the literature review which provided input the questions listed above. A conceptual model was developed based on logical inferences derived from the synthesis of the literature, informed by the educational expert authors, our understanding of the assessment literature (individual assessments within programmes of assessment) and our immersion within the identified themes. The initial draft of the conceptual model was very detailed, to help provide a shared narrative for the authors. After the initial draft was developed, the literature was reviewed again, to consider if there were further concepts and themes which were initially overlooked which could improve our understanding of the literature review questions. This re-examination of the literature helped assist in the

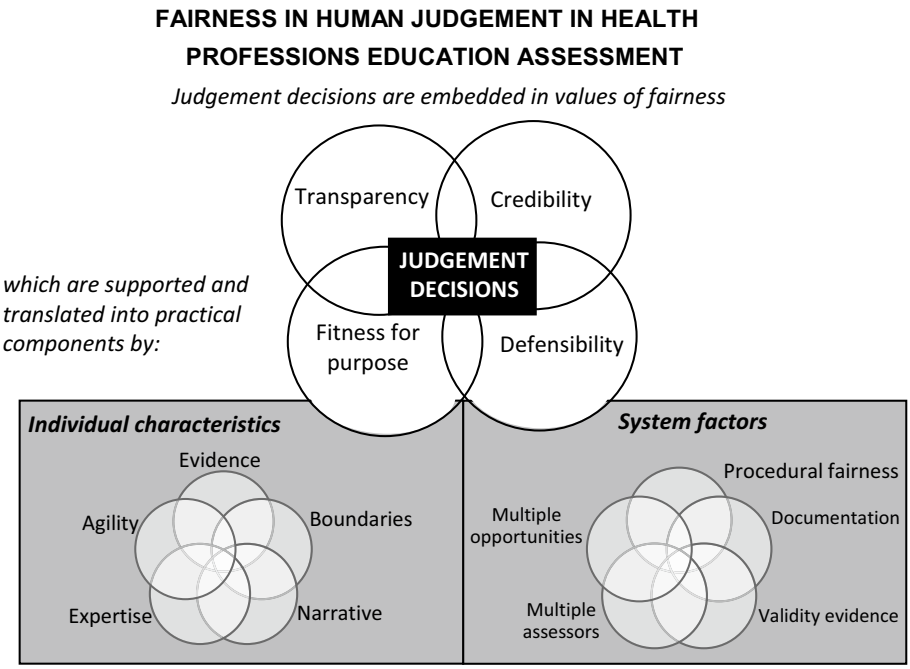

Fig. 3 Conceptual framework of fairness in human judgement in assessment. The values of fairness are supported by individual characteristics and system factors

refinement of the model. Iterations of the model were developed via face to face and Zoom meetings of the authors, with multiple reviews, until complete consensus was reached.

Results

The process 'saturation' on all our questions was reached after the inclusion of 90 papers. These are summarised in Table 1. As a hermeneutic design is cyclical, it precludes a conventional study flowchart. Findings fell into the headings of values of human judgement in assessment, characteristics of fair human judgement as an individual level and procedures and environments required to ensure fair human judgement at a systems level. These headings are expanded in the results section below and displayed in the conceptual model.

Overview: fairness in human judgement in assessment

Fairness is a complex construct with multiple definitions (Tierney 2012). Within the assessment literature, there have been attempts to simplify fairness to "the quality of making judgements that are free from bias and discrimination and requires conformity rules and standards for all students" (Harden et al. 2015), or "absence of bias within the test or assessment processes that give all candidates an equal opportunity to demonstrate their standing on the construct the test is intended to measure" (American Research Association et al. 1999) or as "not a technical psychometric term" (Tierney 2012). However, fairness has also been associated with a wide range of assessment related qualities such as equitable, consistent, balanced, useful and ethically feasible. This breath demonstrates that fairness in assessment is multifaceted and not something which can be reduced to a number, determined dichotomously or a simple definition (Tierney 2012).

To assist in understanding the characteristics of fairness in human judgement, a conceptual framework was derived (Fig. 3) from the results of the literature review. The complex construct of fair human judgement could be conceptualised through values (credibility, fit for purpose, transparency and defensibility) which are supported and translated into practical components at an individual level by characteristics of fair human judgement (narrative, boundaries, expertise, agility and evidence) and at a systems level by procedures and environments (procedural fairness, documentation, multiple opportunities, multiple assessors, validity evidence) which help translate fairness in human judgement from concepts into practical components.

Values of fair human judgement in assessment

The literature review identified four values of fair human judgement in assessment: credibility, fitness for purpose, defensibility and transparency. These values all overlap and relate to each other. At times the values appear to be conflicting, raising tensions which need to be managed. These are described in more detail below.

Credibility

Human judgements which are seen as credible, are seen as fair. For learners, a sense of fairness or justice is key to the credibility of the decision, especially in times of uncertainty (Van den Bos and Miedema 2000; Lind and Van den Bos 2002). There is no clear

Table 1 Included papers in the literature review

Summary of included studies in the narrative review	
General background on fairness	Articles from both medical education (Harden et al. 2015) and wider education literature (Tierney 2012; American Research Association et al. 1999)
Values of fair human judgement in assessment	
Credibility	Articles from the social psychology literature (Van den Bos and Miedema 2000; Lind and Van den Bos 2002; Hilligoss and Young Rich 2008), the education literature (Chory 2007; Rieh and Hilligoss 2008; Rodabaugh 1996) as well as perspectives and studies from the medical education literature (Patterson et al. 2011; Govaerts and van der Vleuten 2013; Telio et al. 2016; Watling et al. 2008; Ginsburg et al. 2017a; Watling 2014)
Defensibility	A review from the medical education (Colbert et al. 2017) and legal literature (Upshur and Colak 2003; Groarke 2019; Reid 1850)
Fitness for purpose	Articles from the education literature (Gipps and Stobart 2009; Stobart 2005; Beckett 2008), medical education literature (Eva 2015; Govaerts and van der Vleuten 2013; Duffield and Spencer 2002; Viney et al. 2017), psychology literature (Wolf 1978), legal literature (Stefan 1993; Upshur and Colak 2003; Kaldjian 2010) and a study from the rehabilitation literature (Ståhl et al. 2019)
Transparency	Studies, reviews and viewpoints from the medical education literature (Patterson et al. 2011; Govaerts and van der Vleuten 2013; Dijksterhuis et al. 2009; Colbert et al. 2017; van der Vleuten et al. 2015; Hays et al. 2015; Schuwirth et al. 2002; Duffield and Spencer 2002; Tavares and Eva 2013; Watling 2014) and education literature (Gipps and Stobart 2009; Tierney 2012; Rodabaugh 1996)
Components needed at an individual level	
Narrative	Articles from the clinical medical literature (Greenhalgh and Hurwitz 1999a, b) and the allied health education literature (Bacon et al. 2017), perspectives and studies from the medical education literature (Govaerts and van der Vleuten 2013; Cohen et al. 1993; Durning et al. 2010; Ginsburg et al. 2013, 2015, 2016, 2017a, b; Tavares and Eva 2013; Duffield and Spencer 2002; Kogan et al. 2014; Crossley and Jolly 2012; Weller et al. 2014; Watling et al. 2008; Cleland et al. 2008), the education literature (Rodabaugh 1996; Colbert et al. 2017), a literature review from the nursing education literature (McCready 2007), a legal perspective (Daniels and Sabin 1997) and a study from the rehabilitation literature (Ståhl et al. 2019)
Evidence	Articles from the clinical medicine literature (Upshur and Colak 2003; Downie and Macnaughton 2009), perspectives and studies from the medical education literature (Govaerts and van der Vleuten 2013; Duffield and Spencer 2002; Southgate et al. 2001; Watling et al. 2012, 2013a, 2008; Watling and Ginsburg 2019; Bullock et al. 2019) and papers from the allied health education literature (Bacon et al. 2017) and nursing literature (Webb et al. 2003)

Table 1 (continued)

Summary of included studies in the narrative review	
Boundaries	Conference reports from the education literature (Houston 2002), studies from the medical education literature (Rees and Shepherd 2005; Watling and Ginsburg 2019), the education literature (Rodabaugh 1996) and the health policy literature (Kirkland 2012)
Expertise	Studies, perspectives and a narrative review from the medical education literature (Watling et al. 2012; Telio et al. 2016; Jones 1999; Berendonk et al. 2013; Hauer et al. 2016; Watling et al. 2013b; Govaerts et al. 2011, 2013) and psychology literature (Marewski et al. 2010)
Agility	Studies and perspectives from the medical education literature (Watling 2014; McCready 2007; Govaerts et al. 2013; MacRae 1998; Berendonk et al. 2013; Flin et al. 2007; Crossley and Jolly 2012), papers and reviews from the clinical medical literature (Greenhalgh et al. 2014; Kaldjian 2010; Katerndahl et al. 2010; Plsek and Greenhalgh 2001; Epstein 2013), the education literature (Sadler 2009), the psychology literature (Lipshitz et al. 2001) and legal literature (Stefan 1993)
Components needed at a systems level	
Procedural fairness	Studies, a review and perspectives from the medical education literature (Van der Vleuten et al. 1991; Burgess et al. 2014; Colbert et al. 2017; Ramani et al. 2017; Hays et al. 2015; Watling et al. 2008) and studies from the psychology literature (Van den Bos et al. 1997, 1998; Lind and Tyler 1988)
Documentation	Papers from the medical education literature (Govaerts and van der Vleuten 2013; Webb et al. 2003; McCready 2007; Rees and Shepherd 2005; Hays et al. 2015)
Multiple opportunities	Papers from the clinical medical literature (Hunter 1996), studies, a review and perspectives from the medical education literature (Boulet and Durning 2019; Govaerts and van der Vleuten 2013; Schuwirth et al. 2002; van der Vleuten and Schuwirth 2005; Colbert et al. 2017; Dijksterhuis et al. 2009; Watling et al. 2013a; Hays et al. 2015; Eva 2015; Wycliffe-Jones et al. 2018; Watling et al. 2008) and papers from the education literature (Stobart 2005; Tierney 2012; Gipps and Stobart 2009; Rodabaugh 1996)
Judgements assessed by multiple assessors	Studies and perspectives from the medical education literature (Govaerts and van der Vleuten 2013; Tochel et al. 2009; Hauer et al. 2015; Hauer et al. 2016) perspectives and a study from the allied health professions education literature (Bacon et al. 2015; Krefting 1991; Webb et al. 2003; McCready 2007) and the clinical medicine literature (Ham 1999)
Validity evidence for judgements	Papers from the medical education literature (Govaerts and van der Vleuten 2013; Colbert et al. 2015)

definition of credibility however an overarching view across definitions appears to believability (Hilligoss and Young Rich 2008), and confidence or trustability in the 'truthfulness' of the findings (Govaerts and van der Vleuten 2013).

Credibility assessment is a not dichotomous, nor does it occur at just one point in time. Rather, it is a consideration made throughout the longitudinal process of information seeking (Rieh and Hilligoss 2008). Credibility is related not only to the judgement itself but also to the person making the judgement (Chory 2007). It is an interplay between the credibility of the judgement itself and the person from whom it originates (Chory 2007). Past experience impacts credibility judgements. For example, if a learner questions the credibility of the source, all information from that source is "second guessed" from that point forward (Rieh and Hilligoss 2008).

Interpersonal or interactional fairness, is an important component of credibility and fairness (Rodabaugh 1996; Patterson et al. 2011). Most learners respect their teachers and wanted to be treated with respect also (Rodabaugh 1996). A dominant theme of several studies in medical education is the importance of assessor engagement in learner's credibility judgements. Studies have noted learners make credibility judgements regarding the assessors' apparent enthusiasm, dedication and motivation for teaching, and their apparent feelings towards the learner in regards to trust, respect and fondness (Telio et al. 2016; Watling et al. 2008; Ginsburg et al. 2017a). Prolonged observation, a positive learning culture, and multiple opportunities for evidence support development of this credibility judgement (Watling et al. 2008; Watling 2014).

Defensibility

Judgement decisions in assessment need to be (legally) defendable as learners may seek legal redress with the concept of fairness often forming the basis of claims (Colbert et al. 2017). In legal terms, a judgement is an assertion made with some evidence or for good reason (Reid 1850). Judgements in complex, uncertain environments such as medical education are difficult to categorise as true or false and rest more on plausibility, or acceptability rather than certainty (Upshur and Colak 2003; Groarke 2019). Within medical education, no matter the assessment, there will always be uncertainty. No assessment method is ever conclusive proof that a trainee will be able to fulfil the expectations of being a doctor in all circumstances. Individual characteristics and system procedures such as procedural fairness, documentation, expertise and boundaries build the defensibility of judgements.

Fitness for purpose

Many authors have argued that fairness is a social construct (Stobart 2005; Ståhl et al. 2019; Wolf 1978; Eva 2015; Gipps and Stobart 2009). Gipps et al. argue that assessment is a socially embedded activity that can only be fully understood by taking account of the social and cultural contexts within which it operates, alongside the technical characteristics (Gipps and Stobart 2009; Stobart 2005). Medical education occurs in diverse, clinical contexts, with learning produced by engagement in unpredictable tasks of authentic health care practice and shaped by unique physical, social and organisational contexts (Govaerts and van der Vleuten 2013). Therefore, what is fair and credible in a judgement must be determined by the context of the clinical encounter, and the environment and culture, not just by the existence of other evidence (Upshur and Colak 2003). Within the US legal system there is general consensus if the intent is inappropriate, such as punishment, administrative convenience, or budgetary constraints/availability of resources then the professional judgement is disregarded (Stefan 1993).

Fair judgement decisions also need to relate to the work of a health care professional and the needs of the patient. Studies have noted that learners perceived assessment that, among other things, had clinical relevance was fair (Duffield and Spencer 2002; Viney et al. 2017). Context dependent and fit for purpose fair judgements are holistic. Patients are not neatly broken down into measurable units and neither can the work of a health professional. Integrated or holistic competence advocates a selective accessibly of evidence, which is sensitive the to the context of the workplace and patient situation, from which competence is inferred (Beckett 2008).

Transparency

Throughout the literature, there is an emphasis on fair assessments demonstrating openness to build a shared understanding with learners (Dijksterhuis et al. 2009; Colbert et al. 2017; van der Vleuten et al. 2015; Hays et al. 2015; Schuwirth et al. 2002), with some authors arguing transparency is the best defence against unfair assessment (Gipps and Stobart 2009). This includes explicit communication about what judgements will be made, who will make them, the purpose, criteria, and results of the judgement decisions (Tierney 2012). Research has demonstrated communication interventions to improve transparency can improve candidate perceptions of overall fairness (Patterson et al. 2011). Transparency brings out into the open the values and biases of the judgement process and provides an opportunity for debate about the influences on this (Gipps and Stobart 2009).

Transparency also includes a narrative which focuses on performance improvement and feedback (Rodabaugh 1996; Colbert et al. 2017). One study noted 'more feedback' as a common response in a survey of medical students about fairness. Several respondents noted that without adequate feedback, they could continue to make the same mistakes in the future, and this was considered unfair (Duffield and Spencer 2002). High quality, appropriate judgements about a performance which provide feedback build the credibility, transparency and thus fairness of a judgement decisions (Tavares and Eva 2013; Govaerts and van der Vleuten 2013).

However, transparency as a value can conflict with other values of fairness (Tierney 2012). For example, transparency provides learners with a framework and an understanding of expectations, but this can restrict opportunities for individualised, contextual assessment which is more credible, fit for purpose and defensible. Transparency can lead to checklists, rubrics and judgement aids which aim to be context independent. Watling (2014) noted predetermined assessment forms, where assessors are forced to make judgements on a wide range of competencies not observed or in context of the clinical situation can diminishes the learners' trust in the assessor and process, and hides potentially credible decisions in a mountain of meaningless platitudes. Furthermore, there are many individualised, tacit values and personal characteristics which come into play when making judgements which cannot be explicitly expressed. To ensure transparency can occur in symbiosis with credibility, defensibility and fit for purpose in fairness in human judgement, many characteristics such as expert abilities, boundaries, narrative and agility of assessors are needed as demonstrated in Fig. 3.

What is needed to create fairness in human judgement in assessment at an individual level?

If judgement decisions are embedded in the values of fairness in human judgement in assessment, then these will need to be supported by components at an individual level, including narrative, evidence, boundaries, expertise and agility.

Narratives

Narratives provide transparency, credibility, defensibility, context, boundaries and perspective to human judgement. It intentionally captures context-specific aspects of performance (Govaerts and van der Vleuten 2013; Bacon et al. 2017; Ginsburg et al. 2015), allows for capturing of non-linear assessment by defining how, why and in what way a learner has been judged, allows for the construction of meaning and encourages reflection (Greenhalgh and Hurwitz 1999a, b) which can improve defensibility and ensure the judgements remain fit for purpose.

Some authors propose that expert subjective narrative comments are 'indispensable for trustworthy decision making in summative assessments', and thus credibility of judgements (Ginsburg et al. 2015; Marjan Govaerts and van der Vleuten 2013). Allowing assessors to articulate their thinking, may be more credible and defensible than reductionism which occurs when assessments rely on numerical scores which mask assessors' thinking (Govaerts and van der Vleuten 2013; McCready 2007). The use of descriptive narratives in assessment has been shown to identify at-risk learners earlier (Cohen et al. 1993; Durning et al. 2010; Ginsburg et al. 2017b; Ginsburg et al. 2013) and contributes to predicting future performance or need for remediation (Cohen et al. 1993). Narratives also lead assessors to more holistic judgements (Bacon et al. 2017) and allow for feedback which learners see as essential for a fair judgement (Rodabaugh 1996; Colbert et al. 2017; Duffield and Spencer 2002; Govaerts and van der Vleuten 2013; Tavares and Eva 2013; Watling et al. 2008). Furthermore, within the return-to-work literature, perceptions of the fairness of the judgements was at least partly dependent on the communication skills of the professionals involved (Ståhl et al. 2019).

Narratives also add to defensibility at a systems level by facilitating group decision making, allowing assessors to articulate assumptions, discuss disconfirming views and learn from the observations of others (Bacon et al. 2017). When a person is required to use narratives to articulate the reasons for their decisions they become more focused in their decision making ensuring they remain fit for purpose (Daniels and Sabin 1997).

Whilst assessors' language may be vague and indirect, requiring faculty and learners to guess what assessors intended by their comments (finding a 'hidden code') there is surprising consistency amongst faculty and learners in interpreting this code (Ginsburg et al. 2015, 2016, 2017a). However, due to multiple factors, including 'hedging' to save face, narrative often focuses on how hard a learner works which can be unhelpful in judging performance (Ginsburg et al. 2016, 2017a), although learners often see this recognition of effort as fair (Rodabaugh 1996). Furthermore, some assessors feel they lack the training and narrative to give negative messages effectively (Cleland et al. 2008). To overcome these limitations, many have called for narratives which fit clinical practice to be used when asking assessors to make judgement (Kogan et al. 2014; Crossley and Jolly 2012). Aligning rating scales to the construct of clinical independence or entrustment has been shown to improve score reliability and assessor discrimination (Crossley and Jolly 2012; Weller et al. 2014). This also allows for clinical evidence to be form the basis of the narrative of the judgement which improves credibility (Watling et al. 2012). Furthermore, it also is fairer to patients, as the judgements are focused on high quality clinical care rather than rating scales (Kogan et al. 2014).

Evidence

Evidence is offered as a means of supporting judgements (Upshur and Colak 2003), and is essential for creating a validity argument (Govaerts and van der Vleuten 2013). Without evidence, it is not a judgement but a guess (Downie and Macnaughton 2009). Evidence itself is often subjective. There is no universal standard to adjudicate evidence that can be applied in each context, and the type of evidence needed will therefore vary accordingly (Upshur and Colak 2003). It has also been demonstrated that in high stakes assessment, the data gathering phase and evidence collected is more often challenged than actual judgement itself (Southgate et al. 2001).

Watling et al. (2012) noted evidence for judgements that were embedded into the actual work of a doctor, such as patient clinical outcomes and feedback from patients was seen by learners as being intrinsically credible. Having the opportunity to be directly observed by the assessor making judgement decisions is fundamental to the trustworthiness and perception of fairness of the assessment (Watling and Ginsburg 2019; Watling et al. 2013a; Watling et al. 2008), and this perception of the fairness is enhanced by prolonged observation (Duffield and Spencer 2002; Bullock et al. 2019). System procedures such as having multiple sources of evidence in a variety of clinical settings (triangulation), continuous collection of evidence and tripartite meetings (peer debriefing and member checks) is also seen to improve the perception of fairness of evidence (Webb et al. 2003; Bacon et al. 2017; Watling et al. 2013a).

Boundaries

Fair judgement decisions can be seen as having boundaries. These are boundaries between what is acceptable/not acceptable, what is relevant/not relevant or what is fit for purpose/not fit for purpose in the process of arriving at and communicating a judgement. Such boundaries are social constructs, connected with values and thus assessors construct boundaries in different places (Houston 2002). By their very nature, boundaries are fuzzy. Learners are concerned about where boundaries lie, and what is "assessable" (Rees and Shepherd 2005). Continuous observation may mean every observation is an opportunity for learners to lose face and impact their assessment outcome (Watling and Ginsburg 2019). One study noted students felt a faculty member's partiality to some students on the basis of race, gender or age was unfair, (Rodabaugh 1996) and in many countries this is also illegal. Implicit shared values, standard documents assist in creating boundaries of what is able to be evidence for judgement decisions. Holding extreme views, at the edge of boundaries also tends to lower the credibility of the person and the judgements they make (Kirkland 2012).

Expertise

Within medical education, there are two types of expertise, clinical and educational (Jones 1999). Assessors perceive that credibility as an expert clinician is required if one is to have credibility as an assessor (Watling et al. 2012, 2013b; Telio et al. 2016; Berendonk et al. 2013). Decision making committees also value expertise, relying on

faculty members' qualifications via their perceived status as expert to help ensure fairness and credibility (Hauer et al. 2016).

Learners value clinical expertise over educational expertise (Watling et al. 2013b). However, experts in medical education in general make more inferences on information, cluster sets of information into meaningful patterns and abstractions (Govaerts et al. 2011). They have a well-developed set of personal schemas, and are able to choose a schema used based on the specific problem or context they are assessing, which is effective for facilitating judgement in unpredictable contexts (Watling et al. 2012; Govaerts et al. 2013; Marewski et al. 2010). They also are more likely to make evaluative judgements, combining various context specific information into meaningful patterns, providing richer and more interpretive descriptions of trainee performance as compared to novices who mostly provide literal, superficial descriptions of what they had seen (Govaerts et al. 2011).

Agility

Govaerts et al. (2013) noted that assessors consider multiple performance dimensions when assessing performance. For example, when assessing performance during history taking, physical examination or patient management, raters assessed not only students' ability to adequately handle the 'medico-technical' aspects of the problem, but also communication, interpersonal and time management skills. In contrast, many assessment forms aim to be context independent and list performance dimensions as separate distinct entities which all need to completed regardless of the clinical situation. Although this is transparent, it is not credible or fit for purpose (Watling 2014; McCready 2007) and does not recognise assessors' agility to make contextually appropriate, holistic and individualised judgement decisions (Govaerts et al. 2013). Equating "quality" with someone who strictly adheres to guidelines or protocols, is to overlook the evidence on the more sophisticated process of expertise (Greenhalgh et al. 2014). From a fairness perspective, these fit-for-purpose, individualised holistic judgements demonstrate at least as much, if not more, assessor agreement and performance discrimination than checklists of actual items (Crossley and Jolly 2012; MacRae 1998; Sadler 2009) and are fairer to society because patients need a health professional who can approach them as a whole person, in their psychosocial environment, not one who can do 'parts' of an consultation. From a legal perspective, in medicine there is increasing recognition that the context strongly influences the adjudication of argument adequacy and if a clinical judgement is not made on an individualised basis, it constitutes a departure from professional judgement (Stefan 1993).

Furthermore, because assessment often occurs in real life, uncertain situations where issues only become apparent as the consult evolves in real time, assessors need to make judgements in real time to ensure patient fairness and safety (Katerndahl et al. 2010; Plsek and Greenhalgh 2001; Kaldjian 2010; Berendonk et al. 2013; Lipshitz et al. 2001; Flin et al. 2007; Epstein 2013). A continuous cycle of monitoring to assess the situation, taking appropriate actions and re-evaluating the results is required (Flin et al. 2007). This requires agility. This agility, combined with expertise allows for trainees to engage in workplace based learning, gaining clinical experiences on real life patients to maximise learning whilst still ensuring patient safety (Flin et al. 2007).

What is needed to create fairness in human judgement in assessment at a systems level?

Individual assessment judgements are not independent, rather they are part of an assessment system. Utilising a systems thinking lens enables a richer examination of individual characteristics and values of fair human judgement than would be possible from simply examining fairness at an individual level alone (Colbert et al. 2015). At a systems level, systems and environments which are able to support the values and individual characteristics of fairness include procedural fairness, documentation, multiple opportunities, multiple assessors and validity evidence.

Procedural fairness

Procedural fairness is an amorphous concept. There is no clear definition of procedural fairness within education. However, the importance of this amorphous concept is clear. People are more willing to voluntarily accept outcomes given to them by an authority if they perceive there is fair procedures in deciding the outcomes (Van den Bos et al. 1998; van der Vleuten et al. 1991). This is one of the most frequently replicated findings in social psychology, found in in laboratory experiments, survey studies and real world environments (Van den Bos et al. 1997). Procedural fairness plays an important role in the credibility of high stakes decisions such as selection and assessment, for both candidates and institutions (Burgess et al. 2014; Colbert et al. 2017).

There are several things which have been shown to positively influence the perception of procedural fairness which such as explicitly describing the process by which judgements are made (Lind and Tyler 1988), by formal, regular inclusive reviews of the judgement process, and provision of an appeals process (Hays et al. 2015). Also important for procedural fairness is to ensure the learner is explicitly told of their expectations and what else is required if they did not meet these expectations (Colbert et al. 2017). Providing learners with information as early as possible has been shown to positively impact perceptions of fairness, as has allowing learners to voice their opinion (Van den Bos et al. 1997). The timing of assessment is another relevant aspect; judgements provided at the end of a rotation are less well received, as there is no opportunity for learners to modify their behaviour which is seen as unfair (Ramani et al. 2017; Watling et al. 2008).

Documentation

Documentation of rich, meaningful information about judgements made, and documentation of values and standards expected allows for external audit, reconstruction, evaluation and quality assurance and thus transparency, credibility and defensibility (Govaerts and van der Vleuten 2013; Webb et al. 2003; McCready 2007). Furthermore, procedural fairness as described above needs clear and comprehensive documentation outlining assessment policies and procedures (Hays et al. 2015).

The detail of the documentation required depends on the context. One study noted a learner questioned the credibility of a judgement because the assessor only provided a global competency grade. Although this could potentially be seen as more credible because

the assessor did not meaninglessly tick boxes, the lack of complete documentation led to the opposite effect (Rees and Shepherd 2005).

Multiple opportunities

Diseases are most useful when they are thought of not as objects but instead seen as plots that unravel over time requiring physicians to interpret signs, symptoms and progression (Hunter 1996). Similarly, it has been suggested a single point in time assessment judgement is not adequate to predict future performance, and longitudinal assessment is needed to allow for a more continuous evaluation of knowledge, skills and attitudes (Boulet and Durning 2019). Because competencies are not generic and stable traits that apply in any given situation, a broad range of tasks, contexts, and assessors are needed to gain an in-depth understanding of a person's performance and capability to adapt to various task requirements (Govaerts and van der Vleuten 2013; Schuwirth et al. 2002; van der Vleuten and Schuwirth 2005). Several authors suggest that a fair and defensible assessment program utilising human judgement should be comprehensive, multimodal, incorporate factual knowledge, sufficiently large samples of direct observation, multisource feedback, and a portfolio to monitor progress and to develop learning plans and self-reflection (Dijksterhuis et al. 2009). However, obtaining multiple pieces of evidence can be problematic as in some training programs a low return rate for trainee assessment is not uncommon (Colbert et al. 2017).

Fair human judgement in assessment is inseparable from fairness in access to opportunities (Stobart 2005). Supervisors are able to influence the quality of the learner's opportunities to learn, both through physical opportunities, or when uniformly low expectations are held for student learning (Tierney 2012). Students' sense of fairness has been found to be more closely related to opportunities afforded to them by teaching practices such as review sessions and study guides, than scoring modifications or manipulations that have the effect of raising grades (Rodabaugh 1996). The medical literature suggests all learners should have opportunities to experience all assessment types prior to major assessments (Hays et al. 2015), and to allow learners alternative opportunities to demonstrate evidence of expertise, which is especially important for those who are disadvantaged on one type of assessment (Gipps and Stobart 2009). Furthermore, learners value opportunities to demonstrate they have understood and incorporated feedback they have received (Watling et al. 2013a, 2008).

Fairness has often been viewed as 'equal' treatment or practice (Colbert et al. 2017). However, countless philosophers and mathematicians have argued that equal treatment does not always ensure fairness (Eva 2015; Stobart 2005). For example, Eva asks: 'is it fair to give two medical students equal remediation for missing a mandatory education session when one was absent because he had a migraine headache, whereas the other had a hangover (Eva 2015)?' Neutrality, consistency and avoidance of favoritism is one on hand fair, however, treating all learners the same be it in terms of the methods used, or the feedback given, is on another hand unfair because it is reducing the opportunity of some students to learn (Tierney 2012). Neutrality is often context independent, and in this sense is unfair. For example, a quiet learner who does not speak up during ward rounds could be incorrectly inferred as having deficits in medical knowledge (Colbert et al. 2017). This is further conflicted by the fact that learners themselves see fairness as related to effort. For example they consider it unfair if most students receive high grades because input does not match output and no distinction is made between those

who worked hard and those who did not (Rodabaugh 1996) or if judgements are not aligned with the inputs that the students brings (Wycliffe-Jones et al. 2018).

Judgements assessed by multiple assessors

Group decision making is now a standard mechanism for assessment decisions in many countries around the world (Hauer et al. 2016; Bacon et al. 2015; Govaerts and van der Vleuten 2013). Creating groups to critically review evidence through open deliberative and critical dialogue is seen as defensible, credible and fair by both learners and assessors because there is a concept of shared subjectivity about learners (Tochel et al. 2009; Hauer et al. 2015; Bacon et al. 2015; Govaerts and van der Vleuten 2013; Krefting 1991; Webb et al. 2003; Ham 1999). Dialogue allows for member checking, verification with secondary assessors, prolonged engagement in the assessment process through review and discussion, articulation of different interpretations or assumptions, triangulation of evidence and analysis and reconciliation of disconfirming evidence and judgements. All of these things allow for diversity prior to agreement, which can be used to improve the defensibility of the professional judgements (Bacon et al. 2015; Govaerts and van der Vleuten 2013; Krefting 1991; Webb et al. 2003; Ham 1999). These qualitative methods of assessing evidence also allow for less tangible learning outcomes such as professional values to be captured (McCready 2007).

Diversity of group members can positively influence group functioning by increasing the number of perspectives considered by group members (Hauer et al. 2016). This needs to be coupled with strategies to facilitate information sharing, to overcome tendances of the group to prioritise information known to more group members or information shared first (Hauer et al. 2016).

However, it has been noted that judgement decisions from assessment panels may focus on only a few sources of evidence despite the widespread availability of multiple data points from multiple different assessment tools (Hauer et al. 2015). Furthermore, an absence of concern was taken to imply readiness for advancement in a review of some panel decisions, and often the data regarding a majority of residents wasn't discussed (Hauer et al. 2015).

Validity evidence for judgments

Evidence is needed to create validity argument. Using a wide range of evidence from multiple sources and contexts is need to ensure the validity of performance appraisals (Colbert et al. 2015). Judgement decisions involve a series of inferences and assumptions leading from the observed performances to conclusions and decisions. In essence, validity refers to the degree to which the interpretations are adequate and appropriate, as justified by evidence or theoretical rationales (Govaerts and van der Vleuten 2013). Evaluation of the plausibility of the inferences and assumptions made by assessors using appropriate evidence is needed to create a validity argument (Govaerts and van der Vleuten 2013). Validity inferences are therefore not procedural per se, but must play a role in the whole system of judgement and decision-making.

Discussion

Summary of findings

To continue to utilise human judgement in assessment, the fairness of these expert judgements needs to be considered. This literature review has demonstrated that fairness is a complex construct which cannot be simplistically defined. Furthermore, context is essential in determining fairness and no one definition will fit across different environments. Learning from the professionalism literature, it is important to frame the problem as the complex problem it is, rather than as a technical or simple problem which can be addressed through checklists (Lucey and Souba 2010). The Ottawa recommendations for the assessment of professionalism embraced complexity and considered professionalism to be multi-dimensional with intrapersonal, interpersonal and macro-societal (public) themes, and interactions between these themes (Hodges et al. 2011). Greenhalgh and Papoutsi (2018) supported this holistic, systems approach, noting that health professions education needed research designs and methods which foreground dynamic interactions and narratives which paid attention to how systems come together as a whole from different perspectives. Whilst there is no simple definition of fair human judgement in assessment, the underpinning foundations of fairness are inferred in the medical education and broader education literature. In this review we have attempted to bring these inferences, studies and perspectives together to create a conceptual model which can be used as a guide to help further discussions of fairness in human judgement and guide research and exploration in this area. This conceptual model aims to embrace complexity, and present fair human judgement in assessment as multi-dimensional with values, individual characteristics and system procedures. The model aims to facilitate internal and external conversations by institutions and academics about fair human judgement in assessment by providing a shared narrative and understanding. Moore noted that creating shared understanding between stakeholders about the problem was key. This is not necessarily complete agreement, but that "the stakeholders understand each other's positions well enough to have intelligent dialogue about the different interpretations of the problem, and to exercise collective intelligence about how to solve it" (Moore 2011).

Tensions

We have revealed several tensions in the development of this conceptual model which add to the complexity of fairness. For example, transparency as a value of fairness can conflict with other values such as credibility, defensibility and fitness for purpose (Tierney 2012). Transparency requires assessment to be known to learners and documented in advance, but clinical work is never predictable and so complete transparency is challenging. If assessment is fit for purpose, it needs to be agile and flexible to respond to the changing clinical situation, however this can limit transparency.

Another example of a tension is providing 'equal' treatment to all learners. Neutrality, consistency and the providing the same opportunities to all learners is on one hand fair, however neutrality is context independent, and this sense is unfair (Eva 2015; Stobart 2005; Tierney 2012). Every learner is entitled to the same quality of judgement and decision making in their assessment, but this should not mean the same process.

A further tension is balancing the need for multiple pieces of evidence with expert, holistic judgements. Expert assessors typically make contextually appropriate, holistic and individualised judgement decisions (Govaerts et al. 2013) which from a fairness perspective are fit for purpose. However, these holistic judgements may provide fewer pieces evidence to a committee who are making decisions on a learner's progression, which on the other hand is unfair.

At times, there is also a tension between what is fair to patients and what is fair to learners. Almost all individual and system components of fairness in human judgement require time and training for assessors, especially for novice assessors. As most assessors are busy clinicians, this can take time away from treating patients. Professional development in education for assessors can also come at a cost to clinical professional development which has the potential to impact patients.

These tensions and seemingly conflicting values or components need to be managed. Govaerts and colleagues note that assessment systems are rife with tensions, and fairness in human judgement in assessment is no different. They suggest that these tensions need to managed not in a traditional 'fix the problem, either-or solutions' but suggest understanding and engaging with the tensions and seeing them as polarities to be leveraged to maximum advantage (Govaerts et al. 2019).

Comparison with existing literature

We found no in-depth examination of fairness in human judgement in our literature search. Throughout this paper we have cited multiple studies and perspectives which have considered human judgement in assessment, its role, benefits and limitations. We believe we have added to this work by using formal, hermeneutic methodology to create a review which incorporates a wide range of literature.

Unanswered questions and limitations of the review

This is not an exhaustive literature review, but rather an attempt to produce a parsimonious synthesis of a complex construct. It is also important to note that our topic was confined to fairness in human judgment in assessment not fairness in assessment in general. No literature review is free from bias (Eva 2008) and we do not claim this review is either. Indeed, this review only included English language papers which may limit the reviews applicability. This literature review also does not aim to reduce the complexity of the literature but rather help provide a way forward in our common aim of continuing to improve the way we undertake and utilise human judgement in assessment. Whittly noted "it is rare that all the evidence needed for a moderately complex policy problem comes from a single discipline, and rarer still that it comes from a single study" and suggested one of the most useful offerings academics can make to policy makers and institutions is to produce a succinct and integrative synthesis of existing information, incorporating quantitative and qualitative, and make sense of the topic area (Whitty 2015; Greenhalgh and Shaw 2017). This is what we have attempted to do here with our conceptual model.

As is to be expected, despite this extensive review, there are still many unanswered questions. Firstly, do the stakeholders in this area hold a different perspective to that of the literature? Expert assessors, university academics and others are currently navigating the use of human judgement in many assessment programs round the world. Is there unspoken tacit knowledge about human judgement in assessment which is not documented

or published? What are the practical implications of fair human judgement within their assessment program? Does it match the literature and if not, why not?

Secondly, how can this conceptual framework be used in a practical manner given the complexity of workplace-based assessment? If assessment programs further utilise human judgement in assessment, then can this conceptual framework be used as a guide? What are the implications for learners, institutions and supervisors?

Thirdly, how can we reconcile the tensions between different values? What is needed to achieve symbiosis of these values, to ensure maximal benefit? How can we also ensure fairness to patients, whilst trying to achieve fairness for learners?

Conclusion

In 2009 Gipps and Stobart said: "The challenge for twenty-first-century assessment is to broaden our views of fairness to take fuller account of social and cultural contexts. The temptation, however, is to back away from the larger social issues because they are difficult, and to concentrate on the assessment itself, for example, in relation to bias" (Gipps and Stobart 2009). Broadening our view of fairness to consider fairness as it relates to both the learner and to the patient, to look beyond just objectivity and consider all facets and complexity of fairness in human judgement in assessment is likely to be beneficial in our ongoing use of human judgement in assessment programs. In this literature review we have highlighted fair human judgement as a multi-dimensional complex concept with values, individual characteristics and system procedures. This model can be used to help the implementation of human judgement in assessment and further research in this area.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest. The views expressed herein are those of the authors and not necessarily those of the Department of Defense or other federal agencies.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education & Joint Committee on Standards for Educational Psychological Testing. (1999). *Standards for educational and psychological testing*. Washington: American Educational Research Association.

Bacon, R., Holmes, K., & Palermo, C. (2017). Exploring subjectivity in competency-based assessment judgements of assessors. *Nutrition & Dietetics, 74*(4), 357–364.

Bacon, R., Williams, L., Grealish, L., & Jamieson, M. (2015). Credible and defensible assessment of entry-level clinical competence: Insights from a modified Delphi study. *Focus on Health Professional Education: A Multi-Disciplinary Journal, 16*(3), 57.

Beckett, D. (2008). Holistic competence: Putting judgements first. *Asia Pacific Education Review, 9*(1), 21–30.

Berendonk, C., Stalmeijer, R. E., & Schuwirth, L. W. (2013). Expertise in performance assessment: Assessors' perspectives. *Advances Health Sciences Education: Theory and Practice, 18*(4), 559–571.

Boell, S., & Cecez-Kecmanovic, D. (2010). Literature reviews and the hermeneutic circle. *Australian Academic & Research Libraries, 41*(2), 129–144.

Boell, S. K., & Cecez-Kecmanovic, D. (2014). A hermeneutic approach for conducting literature reviews and literature searches. *Communications for the Association of Information Systems, 34,* 12.

Boud, D. (1990). Assessment and the promotion of academic values. *Studies in Higher Education, 15*(1), 101–111.
Boulet, J. R., & Durning, S. J. (2019). What we measure… and what we should measure in medical education. *Medical Education, 53*(1), 86–94.
Boursicot, K. (2020). *Consensus statement reports: Performance assessment*. Paper presented at Ottawa 2020, Kuala Lumpar, Malayasia.
Bullock, J. L., Lai, C. J., Lockspeiser, T., O'Sullivan, P. S., Aronowitz, P., et al. (2019). In pursuit of Honors: A multi-institutional study of students' perceptions of clerkship evaluation and grading. *Academic Medicine, 94*(11S), S48–S56.
Burgess, A., Roberts, C., Clark, T., & Mossman, K. (2014). The social validity of a national assessment centre for selection into general practice training. *BMC Medical Education, 14*(1), 261.
Chory, R. M. (2007). Enhancing student perceptions of fairness: The relationship between instructor credibility and classroom justice. *Communication Education, 56*(1), 89–105.
Cleland, J. A., Knight, L. V., Rees, C. E., Tracey, S., & Bond, C. M. (2008). Is it me or is it them? Factors that influence the passing of underperforming students. *Medical Dducation, 42*(8), 800–809.
Cohen, G. S., Blumberg, P., Ryan, N. C., & Sullivan, P. L. (1993). Do final grades reflect written qualitative evaluations of student performance? *Teaching and Learning in Medicine: An International Journal, 5*(1), 10–15.
Colbert, C. Y., Dannefer, E. F., & French, J. C. (2015). Clinical competency committees and assessment: Changing the conversation in graduate medical education. *Journal of Graduate Medical Education, 7*(2), 162–165.
Colbert, C. Y., French, J. C., Herring, M. E., & Dannefer, E. F. (2017). Fairness: The hidden challenge for competency-based postgraduate medical education programs. *Perspectives on Medical Education, 6*(5), 347–355.
Crossley, J., & Jolly, B. (2012). Making sense of work-based assessment: Ask the right questions, in the right way, about the right things, of the right people. *Medical Education, 46*(1), 28–37.
Daniels, N., & Sabin, J. (1997). Limits to health care: Fair procedures, democratic deliberation, and the legitimacy problem for insurers. *Philosophy & Public Affairs, 26*(4), 303–350.
Dauphinee, W. D. (1995). Assessing clinical performance: Where do we stand and what might we expect? *Journal of the American Medical Association, 274*(9), 741–743.
Dijksterhuis, M. G. K., Voorhuis, M., Teunissen, P. W., Schuwirth, L. W. T., ten Cate, O. T. J., et al. (2009). Assessment of competence and progressive independence in postgraduate clinical training. *Medical Education, 43*(12), 1156–1165.
Downie, R., & Macnaughton, J. (2009). In defence of professional judgement. *Advances in Psychiatric Treatment, 15*(5), 322–327.
Duffield, K., & Spencer, J. (2002). A survey of medical students' views about the purposes and fairness of assessment. *Medical Education, 36*(9), 879–886.
Durning, S. J., Hanson, J., Gilliland, W., McManigle, J. M., Waechter, D., et al. (2010). Using qualitative data from a program director's evaluation form as an outcome measurement for medical school. *Military Medicine, 175*(6), 448–452.
Epstein, R. M. (2013). Whole mind and shared mind in clinical decision-making. *Patient Education and Counselling, 90*(2), 200–206.
Eva, K. W. (2008). On the limits of systematicity. *Medical Education, 42*(9), 852–853.
Eva, K. W. (2015). Moving beyond childish notions of fair and equitable. *Medical Education, 49*(1), 1–3.
Flin, R., Youngson, G., & Yule, S. (2007). How do surgeons make intraoperative decisions? *Quality & Safety in Health Care, 16*(3), 235–239.
Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014). Seeing the 'black box' differently: Assessor cognition from three research perspectives. *Medical Education, 48*(11), 1055–1068.
Ginsburg, S., Eva, K., & Regehr, G. (2013). Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Academic Medicine, 88*(10), 1539–1544.
Ginsburg, S., Regehr, G., Lingard, L., & Eva, K. W. (2015). Reading between the lines: Faculty interpretations of narrative evaluation comments. *Medical Education, 49*(3), 296–306.
Ginsburg, S., van der Vleuten, C. P. M., & Eva, K. W. (2017a). The hidden value of narrative comments for assessment: A quantitative reliability analysis of qualitative data. *Academic Medicine, 92*(11), 1617–1621.
Ginsburg, S., van der Vleuten, C., Eva, K. W., & Lingard, L. (2016). Hedging to save face: A linguistic analysis of written comments on in-training evaluation reports. *Advances in Health Science Education : Theory and Practice, 21*(1), 175–188.

Ginsburg, S., van der Vleuten, C. P., Eva, K. W., & Lingard, L. (2017b). Cracking the code: Residents' interpretations of written assessment comments. *Medical Education, 51*(4), 401–410.
Gipps, C., & Stobart, G. (2009). Fairness in assessment. In *Educational assessment in the 21st century* (pp. 105–118). Springer.
Govaerts, M. J., Schuwirth, L. W., Van der Vleuten, C. P., & Muijtjens, A. M. (2011). Workplace-based assessment: Effects of rater expertise. *Advances in Health Science Education: Theory and Practice, 16*(2), 151–165.
Govaerts, M. J., Van de Wiel, M. W., Schuwirth, L. W., Van der Vleuten, C. P., & Muijtjens, A. M. (2013). Workplace-based assessment: Raters' performance theories and constructs. *Advances in Health Science Education: Theory and Practice, 18*(3), 375–396.
Govaerts, M., & van der Vleuten, C. P. (2013). Validity in work-based assessment: Expanding our horizons. *Medical Education, 47*(12), 1164–1174.
Govaerts, M. J. B., van der Vleuten, C. P. M., & Holmboe, E. S. (2019). Managing tensions in assessment: Moving beyond either-or thinking. *Medical Education, 53*(1), 64–75.
Govaerts, M. J., van der Vleuten, C. P., Schuwirth, L. W., & Muijtjens, A. M. (2007). Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. *Advances in Health Science Education: Theory and Practice, 12*(2), 239–260.
Greenhaigh, T., & Hurwitz, B. (1999). Why study narrative? *Western Journal of Medicine, 170*(6), 367–369.
Greenhalgh, T., Howick, J., & Maskrey, N. (2014). Evidence based medicine: A movement in crisis? *British Medical Journal, 348,* g3725.
Greenhalgh, T., & Hurwitz, B. (1999). Narrative based medicine: Why study narrative? *British Medical Journal, 318*(7175), 48–50.
Greenhalgh, T., & Papoutsi, C. (2018). Studying complexity in health services research: Desperately seeking an overdue paradigm shift. *BMC Medicine, 16*(1), 95.
Greenhalgh, T., & Shaw, S. (2017). Understanding heart failure; explaining telehealth—A hermeneutic systematic review. *BMC Cardiovascular Disorders, 17*(1), 156.
Groarke, L. (2019). 'Informal logic' Summer 2019. In E. N. Zalta (Ed.), *The Standford encyclopedia of philosophy*.
Ham, C. (1999). Tragic choices in health care: Lessons from the child B case. *British Medical Journal, 319*(7219), 1258–1261.
Harden, R. M., Lilley, P., & Patricio, M. (2015). *The Definitive Guide to the OSCE: The Objective Structured Clinical Examination as a performance assessment*. Philadelphia: Elsevier Health Sciences.
Hauer, K. E., Cate, O. T., Boscardin, C. K., Iobst, W., Holmboe, E. S., et al. (2016). Ensuring resident competence: A narrative review of the literature on group decision making to inform the work of clinical competency committees. *Journal of Graduate Medical Education, 8*(2), 156–164.
Hauer, K. E., Chesluk, B., Iobst, W., Holmboe, E., Baron, R. B., et al. (2015). Reviewing residents' competence: A qualitative study of the role of clinical competency committees in performance assessment. *Academic Medicine, 90*(8), 1084–1092.
Hays, R. B., Hamlin, G., & Crane, L. (2015). Twelve tips for increasing the defensibility of assessment decisions. *Medical Teacher, 37*(5), 433–436.
Heifetz, R. A., Heifetz, R., Grashow, A., & Linsky, M. (2009). *The practice of adaptive leadership: Tools and tactics for changing your organization and the world*. Boston: Harvard Business Press.
Hilligoss, B., & Young Rich, S. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics and interaction in context. *Information Processing and Management, 44,* 1467–1484.
Hodges, B. (2013). Assessment in the post-psychometric era: Learning to love the subjective and collective. *Medical Teacher, 35*(7), 564–568.
Hodges, B. D., Ginsburg, S., Cruess, R., Cruess, S., Delport, R., et al. (2011). Assessment of professionalism: Recommendations from the Ottawa 2010 Conference. *Medical Teacher, 33*(5), 354–363.
Houston, D. (2002). Quality and the University: Stakeholders, boundary judgements and systems. In *Change management: Proceedings of the 7th International Conference on ISO9000 and TQM* Melbourne, RMIT University.
Hunter, K. (1996). "Don't think zebras": Uncertainty, interpretation, and the place of paradox in clinical education (journal article). *Theoretical Medicine, 17*(3), 225–241.
Jones, A. (1999). The place of judgement in competency-based assessment. *Journal of Vocational Education and Training, 51*(1), 145–160.
Kaldjian, L. C. (2010). Teaching practical wisdom in medicine through clinical judgement, goals of care, and ethical reasoning. *Journal of Medical Ethics, 36*(9), 558–562.
Katerndahl, D., Parchman, M., & Wood, R. (2010). Trends in the perceived complexity of primary health care: A secondary analysis. *Journal of Evaluation in Clinical Practice, 16*(5), 1002–1008.

Kirkland, A. (2012). The legitimacy of vaccine critics: What is left after the autism hypothesis? *Journal of Health Politics, Policy and Law, 37*(1), 69–97.

Kogan, J. R., Conforti, L. N., Iobst, W. F., & Holmboe, E. S. (2014). Reconceptualizing variable rater assessments as both an educational and clinical care problem. *Academic Medicine, 89*(5), 721–727.

Krefting, L. (1991). Rigor in qualitative research: The assessment of trustworthiness. *American Journal of Occupational Therapy, 45*(3), 214–222.

Kusnanto, H., Agustian, D., & Hilmanto, D. (2018). Biopsychosocial model of illnesses in primary care: A hermeneutic literature review (Review Article). *Journal of Family Medicine and Primary Care, 7*(3), 497–500.

Lind, E., & Tyler, T. (1988) Critical issues in social justice. In *The social psychology of procedural justice*. New York: Plenum Press.

Lind, E. A., & Van den Bos, K. (2002). When fairness works: Toward a general theory of uncertainty management. *Research in Organizational Behavior, 24,* 181–223.

Lipshitz, H. D., Klein, G., Orasanu, J., & Salas, E. (2001). Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making, 14*(5), 331–352.

Lucey, C., & Souba, W. (2010). Perspective: The problem with the problem of professionalism. *Academic Medicine, 85*(6), 1018–1024.

MacRae, R. G., MacRae, H., Reznick, R. K., et al. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine, 73,* 993.

Marewski, J. N., Gaissmaier, W., & Gigerenzer, G. (2010). Good judgments do not require complex cognition. *Cognitive Processing, 11*(2), 103–121.

McCready, T. (2007). Portfolios and the assessment of competence in nursing: A literature review. *International Journal of Nursing Studies, 44*(1), 143–151.

Moore, T. (2011). Wicked problems, rotten outcomes and clumsy solutions. Children and families in a changing world. In *NIFTeY/CCCH Conference 2011. Children's place on the agenda… past, present and future*, pp. 28–29.

Patterson, F., Zibarras, L., Carr, V., Irish, B., & Gregory, S. (2011). Evaluating candidate reactions to selection practices using organisational justice theory. *Medical Education, 45*(3), 289–297.

Plsek, P. E., & Greenhalgh, T. (2001). Complexity science: The challenge of complexity in health care. *British Medical Journal, 323*(7313), 625–628.

Ramani, S., Post, S. E., Konings, K., Mann, K., Katz, J. T., et al. (2017). "It's just not the culture": A qualitative study exploring residents' perceptions of the impact of institutional culture on feedback. *Teaching and Learning in Medicine, 29*(2), 153–161.

Rees, C., & Shepherd, M. (2005). The acceptability of 360-degree judgements as a method of assessing undergraduate medical students' personal and professional behaviours. *Medical Education, 39*(1), 49–57.

Reid, T. (1850). *Essays on the intellectual powers of man*. Cambridge: J. Bartlett.

Rieh, S. Y., & Hilligoss, B. (2008). *College students' credibility judgments in the information-seeking process. Digital media, youth, and credibility* (pp. 49–72). Cambridge, MA: The MIT Press.

Robinson, J. M. (2002). In search of fairness: An application of multi-reviewer anonymous peer review in a large class. *Journal of Further and Higher Education, 26*(2), 183–192.

Rodabaugh, R. C. (1996). Institutional commitment to fairness in college teaching. *New Directions for teaching and learning, 1996*(66), 37–45.

Rotthoff, T. (2018). Standing up for subjectivity in the assessment of competencies. *GMS Journal for Medical Education, 35*(3), Doc29.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159–179.

Schuwirth, L., Southgate, L., Page, G., Paget, N., Lescop, J., et al. (2002). When enough is enough: A conceptual basis for fair and defensible practice performance assessment. *Medical Education, 36*(10), 925–930.

Schuwirth, L. W., & van der Vleuten, C. P. (2006). A plea for new psychometric models in educational assessment. *Medical Education, 40*(4), 296–300.

Southgate, L., Cox, J., David, T., Hatch, D., Howes, A., et al. (2001). The General Medical Council's Performance Procedures: Peer review of performance in the workplace. *Med Education, 35*(Suppl 1), 9–19.

Ståhl, C., Seing, I., Gerdle, B., & Sandqvist, J. (2019). Fair or square? Experiences of introducing a new method for assessing general work ability in a sickness insurance context. *Disability and Rehabilitation, 41*(6), 656–665.

Stefan, S. (1993). What constitutes departure from professional judgment? *Mental and Physical Disability Law Reporter, 17*(2), 207–213.
Stobart, G. (2005). Fairness in multicultural assessment systems. *Assessment in Education: Principles, Policy & Practice, 12*(3), 275–287.
Tavares, W., & Eva, K. W. (2013). Exploring the impact of mental workload on rater-based assessments. *Advances in Health Science Education: Theory and Practice, 18*(2), 291–303.
Telio, S., Regehr, G., & Ajjawi, R. (2016). Feedback and the educational alliance: Examining credibility judgements and their consequences. *Medical Education, 50*(9), 933–942.
ten Cate, O. (2017). Competency-based postgraduate medical education: Past, present and future. *GMS Journal for Medical Education, 34*(5), Doc69.
ten Cate, O., & Billett, S. (2014). Competency-based medical education: Origins, perspectives and potentialities. *Medical Education, 48*(3), 325–332.
ten Cate, O., & Regehr, G. (2019). The power of subjectivity in the assessment of medical trainees. *Academic Medicine, 94*(3), 333–337.
ten Cate, O., & Scheele, F. (2007). Competency-based postgraduate training: Can we bridge the gap between theory and clinical practice? *Acadamic Medicine, 82*(6), 542–547.
Tierney, R. D. (2012). *Fairness in classroom assessment*. Thousand Oaks: Sage.
Tochel, C., Haig, A., Hesketh, A., Cadzow, A., Beggs, K., et al. (2009). The effectiveness of portfolios for post-graduate assessment and education: BEME Guide No 12. *Medical Teacher, 31*(4), 299–318.
Upshur, R. E., & Colak, E. (2003). Argumentation and evidence. *Theoretical Medicine and Bioethics, 24*(4), 283–299.
Valentine, N., & Schuwirth, L. (2019). Identifying the narrative used by educators in articulating judgement of performance. *Perspectives Medical Education, 8*(2), 83–89.
Van den Bos, K., Lind, E. A., Vermunt, R., & Wilke, H. A. (1997). How do I judge my outcome when I do not know the outcome of others? The psychology of the fair process effect. *Journal of Personality and Social Psychology, 72*(5), 1034.
Van den Bos, K., & Miedema, J. (2000). Toward understanding why fairness matters: The influence of mortality salience on reactions to procedural fairness. *Journal of Personality and Social Psychology, 79*(3), 355.
Van den Bos, K., Wilke, H. A., & Lind, E. A. (1998). When do we need procedural fairness? The role of trust in authority. *Journal of Personality and Social Psychology, 75*(6), 1449.
van der Vleuten, C. P., Norman, G. R., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: Issues of reliability. *Medical Education, 25*(2), 110–118.
van der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*(3), 309–317.
van Der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Govaerts, M. J. B., & Heeneman, S. (2015). Twelve tips for programmatic assessment. *Medical Teacher, 37*(7), 641–646.
Viney, R., Rich, A., Needleman, S., Griffin, A., & Woolf, K. (2017). The validity of the Annual Review of Competence Progression: A qualitative interview study of the perceptions of junior doctors and their trainers. *Journal of the Royal Society of Medicine, 110*(3), 110–117.
Watling, C. J. (2014). Unfulfilled promise, untapped potential: Feedback at the crossroads. *Medical Teacher, 36*(8), 692–697.
Watling, C., Driessen, E., van der Vleuten, C. P., & Lingard, L. (2012). Learning from clinical work: The roles of learning cues and credibility judgements. *Medical Education, 46*(2), 192–200.
Watling, C., Driessen, E., van der Vleuten, C. P., Vanstone, M., & Lingard, L. (2013a). Beyond individualism: Professional culture and its influence on feedback. *Medical Education, 47*(6), 585–594.
Watling, C., Driessen, E., van der Vleuten, C. P. M., Vanstone, M., & Lingard, L. (2013b). Music lessons: Revealing medicine's learning culture through a comparison with that of music. *Medical Education, 47*(8), 842–850.
Watling, C. J., & Ginsburg, S. (2019). Assessment, feedback and the alchemy of learning. *Medical Education, 53*(1), 76–85. https://doi.org/10.1111/medu.13645.
Watling, C. J., Kenyon, C. F., Zibrowski, E. M., Schulz, V., Goldszmidt, M. A., et al. (2008). Rules of engagement: Residents' perceptions of the in-training evaluation process. *Academic Medicine, 83*(10 Suppl), S97–S100.
Webb, C., Endacott, R., Gray, M. A., Jasper, M. A., McMullan, M., et al. (2003). Evaluating portfolio assessment systems: What are the appropriate criteria? *Nurse Education Today, 23*(8), 600–609.
Weller, J. M., Misur, M., Nicolson, S., Morris, J., Ure, S., et al. (2014). Can I leave the theatre? A key to more reliable workplace-based assessment. *British Journal of Anaesthesia, 112*(6), 1083–1091.
Whitty, C. J. (2015). What makes an academic paper useful for health policy? *BioMed Central, 13,* 301.

Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart 1. *Journal of Applied Behavior Analysis, 11*(2), 203–214.
Wycliffe-Jones, K., Hecker, K. G., Schipper, S., Topps, M., Robinson, J., et al. (2018). Selection for family medicine residency training in Canada: How consistently are the same students ranked by different programs? *Canadian Family Physician, 64*(2), 129–134.