

What makes assessment fair?

By

Dr Nyoli Valentine

MBBS, MPH, FRACGP, GradCertHlth, DCH

Thesis

*Submitted to Flinders University
for the degree of*

Doctor of Philosophy

College of Medicine and Public Health

March 2024

TABLE OF CONTENTS

TABLE OF CONTENTS	I
ABSTRACT	V
DECLARATION	VIII
ACKNOWLEDGEMENTS	IX
LIST OF FIGURES	XII
LIST OF TABLES	XIII
CHAPTER ONE: INTRODUCTION	1
Fairness as an ideal	1
How has fairness previously been determined in health professions assessment?	2
Limitations of using objectivity to determine fairness	4
Quantification of quality	5
Reductionism at the expense of learning	8
Assessors are not neutral windows on competence	10
Looking beyond objectivity	15
Aim of this PhD and outline of the thesis	19
CHAPTER TWO: THE PURSUIT OF FAIRNESS IN ASSESSMENT: LOOKING BEYOND THE OBJECTIVE	22
Abstract.....	23
Introduction	24
Objective	27
Limitations of Objectivity.....	27
Moving beyond ‘bias free objectivity’ in the pursuit of fairness.....	35
Conclusion	38
Practice Points.....	40
CHAPTER THREE: THEORETICAL PERSPECTIVES	41
Introduction	41
Health professions assessment	42
Constructivist Paradigm	44
Ontology.....	45
Epistemology.....	46
Methodology.....	47
Axiology.....	47
Theory.....	48
Complexity theory	50
Some key components of complex adaptive systems	50
Complexity theory and health professions education	55
Methodology	57

Reflexivity.....	60
Conclusion	61
CHAPTER FOUR: FAIRNESS IN HUMAN JUDGEMENT IN ASSESSMENT: A HERMENEUTIC LITERATURE REVIEW AND CONCEPTUAL FRAMEWORK.....	63
Abstract.....	64
Introduction	65
Methods	69
Design	69
Focus of the review.....	70
Stages of the review:	71
Stage 2: Data extraction, analysis and interpretation	75
Stage 3: Development of a conceptual model	75
Results:.....	76
Overview: Fairness in human judgement in assessment	83
Values of fair human judgement in assessment	85
Values of fair human judgement: Credibility.....	85
Values of fair human judgement: Defensibility	86
Values of fair human judgement: Fitness for Purpose	87
Values of fair human judgement: Transparency	88
What is needed to create fairness in human judgement in assessment at an individual level?	90
Narratives.....	90
Evidence	93
Boundaries.....	94
Expertise	95
Agility.....	96
What is needed to create fairness in human judgement in assessment at a systems level?	97
Procedural fairness.....	98
Documentation.....	99
Multiple opportunities.....	100
Judgements assessed by multiple assessors.....	102
Validity evidence for judgments.....	104
Discussion	104
Summary of Findings.....	104
Tensions.....	106
Comparison with existing literature	107
Unanswered questions and limitations of the review	108
Conclusion.....	109
CHAPTER FIVE: MAKING IT FAIR: LEARNERS' AND ASSESSORS' PERSPECTIVES OF THE ATTRIBUTES OF FAIR JUDGEMENT.....	111

Abstract.....	113
Introduction	114
Methods	117
Results	121
Individual Characteristics.....	122
System Factors	127
The Environment and Culture.....	134
Discussion	138
Conclusion	144
CHAPTER SIX: FAIRNESS IN ASSESSMENT: IDENTIFYING A COMPLEX ADAPTIVE SYSTEM.....	145
Abstract.....	147
Introduction	148
Methods	152
Reflexivity.....	152
Participants	153
Analysis.....	154
Results	155
Environment and Culture	158
System Factors	161
Individual factors	168
Discussion	172
Implications of viewing fairness through a complexity lens	178
Conclusion	182
CHAPTER SEVEN: WHAT STOPS FAIRNESS FROM EMERGING IN ASSESSMENT? THE FORCES ON A COMPLEX ADAPTIVE SYSTEM.....	183
Abstract.....	184
Introduction	185
Methods	189
Theoretical underpinnings	189
Setting and participants	191
Data analysis.....	192
Results	193
Forces impairing interactivity	196
Forces impairing adaptability	199
Forces impairing embeddedness	203
Discussion	206
CHAPTER EIGHT: DISCUSSION AND CONCLUSIONS.....	214
Introduction to discussion.....	214
Changing from a linear perspective to seeing fairness as a complex adaptive system	216

Practical implications of fairness as a complex adaptive system.....	222
Assessment needs to be adaptable and agile with a focus on connections	222
Fuzzy boundaries contain unpredictability	224
The focus moves from solving problems to identifying patterns.....	225
Simple rules can assist in creating an environment in which fairness can emerge	229
Rich behaviour comes from collaborating and competing agents, and therefore environments and culture needs to support interactions between stakeholders.....	229
Individuals learn through deliberate practice and adapting to prior experience.	232
Strategies to support fairness emerging in practice	235
Forces which can limit fairness emerging	239
Limitations of this research.....	239
Directions for future research	240
Conclusion.....	242
STUDENT PUBLICATIONS DURING HIGHER DEGREE RESEARCH CANDIDATURE ...	244
ADDITIONAL PUBLICATION: USING FAIRNESS TO RECONCILE TENSIONS BETWEEN COACHING AND ASSESSMENT	246
BIBLIOGRAPHY	252
APPENDIX 1: CO AUTHORSHIP APPROVALS FOR HIGHER DEGREE RESEARCH THESIS EXAMINATION	293

ABSTRACT

Introduction

Traditionally, 'objectivity' has been seen as the only approach to fairness. Equating fairness to objectivity may be intuitive, however this can force a quantification of quality which can lead to a neglect of unquantifiable qualities, can hinder learning through the reduction of rich information to a numerical score, and in reality is very challenging to achieve. It also reduces a complex, multi-dimensional and contextual construct to a single linear, non-representative rule with limited fitness for purpose. An ontological shift, looking beyond objectivity, is needed to better understand what makes assessment fair.

Methods

I took a social constructivist stance, assuming fairness as a reality is socially constructed by multiple stakeholders, and that individuals and social groups share interpretations and understandings of fairness. Collecting data from multiple perspectives provided a richer and more nuanced understanding.

A hermeneutic literature review was undertaken for a scholarly knowledge synthesis of the definitions, factors and key questions associated with fairness in assessment. Two studies then explored how supervisors, learners and assessment leads conceptualise fair judgement. The first study used semi-structured interviews with vignettes, and the second engaged online focus groups with assessment leads from Australian and New

Zealand medical schools. Initial analysis of study two revealed fair judgement is best studied as a complex adaptive system and so data analysis proceeded using a complexity lens. In a third study, online focus groups with academic leaders from the Netherlands explored how external systems' forces on the complex adaptive system can impair fairness from emerging.

Results

In line with complexity theoretical notions, the same four elements of fairness (transparency, fitness for purpose, accountability and credibility) occurred throughout the data. These elements interacted with each other at all levels in the assessment program and behaved like a fractal. Within a complex adaptive system, a system's behaviour relies less on the mere presence of the individual components but more on the dynamic strength and nature of the interactions between them. In line with this, people seek to create fairness through managing the interplay between fitness for purpose, credibility, transparency and accountability when interacting with others rather than using them as a tick box list.

Assessors used different strategies to influence the interactions between fairness components, including utilising narratives, aggregating evidence from multiple sources, procedural strategies, enabling a learning culture allowing for learner agency, articulating reasonable expectations of learners and ensuring a sound theoretic basis of assessment design.

Discussion

Considering fairness as a complex adaptive system changes our views about how we both approach and seek to improve assessment as well as a perspective to navigate the tensions of unpredictable real-world clinical and learning situations. In line with a complexity perspective, fairness can only emerge through the interaction of its components. This requires agile assessment, with assessors and learners who are able to adapt to different contexts using a variety of different strategies. Refraining from using strict regulations to supporting interactions and allowing learning by action is more likely to support the emergence of fairness. Viewing fairness through a lens of complexity rather than as a linear, causal model will enable better understanding of what is fair assessment and lead to more purposeful, meaningful changes which are more aligned with 21st century assessment.

DECLARATION

I certify that this thesis:

1. does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university
2. and the research within will not be submitted for any other future degree or diploma without the permission of Flinders University; and
3. to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.

Signed..

A handwritten signature in cursive script that reads "Valentine". The signature is written in black ink and is positioned to the right of the "Signed.." label.

Date.....7/12/23

ACKNOWLEDGEMENTS

I started this PhD approximately 5 years ago, at the same time I fell pregnant with my third child. Many would question this sense of timing, but then, I have never been one to shy away from a challenge. Watching my youngest daughter grow up has been a great motivator to remind me of the need to complete this PhD!

I would like to express my sincere gratitude to my supervisors who embraced this journey with me. You have shown me endless patience as I've struggled to connect the academic dots, much like a child desperate to build Lego but who hasn't yet worked out how the pieces go together, let alone orientate the instruction book. I've had so many false starts - pulling pieces off, turning them upside down, finding a new piece and trying again. Thank you for all your amazing wisdom, for challenging me to think in new ways, to use parts of my brain I did not know existed, reading my endless rambling notes and guiding my ideas. Your energy and enthusiasm never wavered, even in the eleventh hour! Thanks for hanging in there with me and always making me laugh; this was so appreciated during moments of stress and the odd tantrum. I am so grateful for this incredible opportunity to learn from you all.

Thank you also to all the delightful Prideaux staff and PhD students for your support and providing a sense of camaraderie throughout this journey. The research meetings and journal clubs are what motivated me return to research and started me on this

PhD road many years ago, and the writing sessions exactly what I needed to get me over the finish line.

One of the advantages of being married to a surgeon whose day job involves dealing with difficult situations is that it takes a fair amount to surprise him. So when I came home and said that as a busy practicing clinician, trying to get pregnant, with two kids, I was contemplating commencing a PhD, he took it in his stride and said, 'sure, why not?!'. And I'm indebted to his support and encouragement over the last five years, especially over the last few months. Thanks for listening to my complaining, taking the kids so I could work, coming to my presentations, pretending you understanding what I'm talking about, filling the fridge with diet coke and tolerating my ups and downs. It means the world to me.

Zahlia, Henley & Anderley thanks for being so patient with me while I have been doing this PhD. You were only 5, 3 and not yet born when I started, so all you remember is mum going to Flinders Uni to work on this crazy project of hers. Thanks for indulging my dream of completing this PhD and enduring the chaos and lack of routine which has come with this. Your willingness to go with the flow, be picked up from school by anyone (and everyone), and eat endless pasta at all hours of the evening has contributed greatly to my productivity. Thank you so much my darlings; I'm so grateful for your sacrifices too.

A massive debt of gratitude also goes to my amazing parents for raising me with the determination which has seen me through the other side of this PhD. I also would not have made it without the hours of babysitting they have provided and unwavering support to keep me afloat. Similarly, to my extended family, to my sisters, Kirrilie and Jenelle, and their families, your support of Rowan and I and the kids while I've been busy working has been amazing. Thanks for tolerating this crazy idea of mine to complete a PhD. I could not have done it without you!

LIST OF FIGURES

Figure 1	The McNamara fallacy	Chapter 2
Figure 2	The hermeneutic circle as a framework for the literature review (Boell & Cecez-Kecmanovic, 2014)	Chapter 4
Figure 3	Search strategies used in the literature review	Chapter 4
Figure 4	A conceptual framework of fairness in human judgement in assessment	Chapter 4
Figure 5	Vignettes used in semi-structured interviews	Chapter 5
Figure 6	A conceptual model of the components of fair judgement in assessment	Chapter 5
Figure 7	The components of fair judgement	Chapter 6
Figure 8	Implications of considering fairness with a complexity lens	Chapter 8
Figure 9	Strategies to facilitate the emergence of fairness through supporting the interaction of its components	Chapter 8

LIST OF TABLES

Table 1	Included papers in the literature review	Chapter 4
Table 2	Academic titles of focus group participants	Chapter 6
Table 3	Fair judgement demonstrated as a complex adaptive system	Chapter 6
Table 4	Key features of a complex adaptive system	Chapter 7
Table 5	The forces preventing fairness emerging from the complex adaptive system	Chapter 7

CHAPTER ONE: INTRODUCTION

Fairness as an ideal

It is generally agreed that fairness is a desirable quality of education and assessment. (Green, Johnson, Kim & Pope, 2007; Tierney, 2013) Both students and academics alike have commonly upheld fairness as a basic right. (Robinson, 2002) The notion of fairness has been associated with a wide range of qualities pertinent to assessment such as equitable, consistent, balanced, useful and ethically feasible. (Tierney, 2013) However, despite the relatively widespread agreement on fairness as an ideal, what fairness looks like in practice is far more contentious, and concerns about fairness in assessment are a constantly reoccurring theme in the health professions education literature. (Tierney, 2013)

This does not just pertain to education. What is considered *fair to patients*, for example, has also changed in recent decades. In much of the world there has been a change from a paternalism-based view of the doctor-patient relationship to one which centres around the principle of respect for autonomy. (Lazcano-Ponce, Angeles-Llerenas, Rodríguez-Valentín, Salvador-Carulla, Domínguez-Esponda, Astudillo-García, Madrigal-de Leon & Katz, 2020) Society, therefore, expects health professionals to have a range of skills which include not only knowledge of established and evolving biomedical and clinical sciences but also the skill to apply of that knowledge to patient care, to provide care which is compassionate, appropriate and effective, and to demonstrate interpersonal and communication skills that result in the

effective exchange of information and collaboration with patients, all of which allows for greater patient autonomy. (Norcini, Anderson, Bollela, Burch, Costa, Duvivier, Galbraith, Hays, Kent, Perrott & Roberts, 2011) This task is further complicated by patients presenting with medical problems in many different ways due to the multiple dimensions of human experience (biological, psychological, social, spiritual) and because patients all respond differently to a vast array of therapeutic options. (Kaldjian, 2010)

There is also an ethical commitment for medical training to ensure graduates can meet communities' changing needs and expectations. (Frenk et al., 2010; Hauer, Chesluk, Iobst, Holmboe, Baron, Boscardin, ten Cate & O'Sullivan, 2015) As a result and to recognise this broadening scope, competency-based medical education has become the dominant approach to medical education in many countries. (Ten Cate, 2017) Universities and specialty training organisations too are required to have processes in place to ensure that future health professionals meet these expectations and remain socially accountable and fair to society. (Hauer, Chesluk, Iobst, Holmboe, Baron, Boscardin, ten Cate & O'Sullivan, 2015)

How has fairness previously been determined in health professions assessment?

Fairness in assessment is often implicitly implied rather than explicitly articulated. There are multiple definitions of fairness, and as an adjective or adverb its meaning

changes according to the noun or verb it describes. (Tierney, 2013) But the common use of the word generally conveys a sense of openness, neutrality or balance.

(Tierney, 2013) Within the assessment literature, some have attempted to simplify the concept as “the quality of making judgements that are free from bias and discrimination and require conformity to rules and standards for all students.”

(Harden, Lilley & Patricio, 2015) However such a prescriptive description of a complex phenomenon such as ‘fairness’ is likely to be non-representative and too much of a reductionist approach.

Most commonly, objectivity has been seen as the predominant way to ensure fairness in assessment. (Ten Cate & Regehr, 2019; Hodges, 2013; Valentine, Durning, Shanahan & Schuwirth, 2021) In this view, objectivity can be defined as the absence of bias or without influence of personal opinions, preferences, views, interests or sentiments. (Ten Cate & Regehr, 2019; Park, Konge, & Artino, 2020) Subjectivity, framed in opposition to objectivity was then seen to mean biased and unfair. (Hodges, 2013) A common example of this can be seen in the assessment of competence. Previously, competence has been seen as a series of stable individual traits such as ‘skills’, ‘knowledge’, ‘problem solving ability’ and ‘attitudes’. (Hodges, 2013; Morcke, Dornan & Eika, 2013; Schuwirth & van der Vleuten, 2006) It was then logically assumed that these traits could and should be objectively measured and expressed as numerical values. Objectivity, from such a positivist perspective, suggests that for each item being measured, a ‘true’ score exists. Any deviation from this true score is a measurement error. (Ten Cate & Regehr, 2019) Generally, a positivist paradigm emphasises objective observation of data and states that only observable and measurable phenomena can be considered valid sources of knowledge. It is based on

the assumption that a single tangible reality which can be identified and measured exists. (Park, Konge, & Artino, 2020) But when competence is defined as such a combination of traits, it resides in the minds of the candidates and cannot be observed directly. Validity theory provides an avenue to reconcile what can be directly observed with what is to be assumed to exist in the minds of the candidates. (Kane, 2001) Therefore, construct validity, consistency, reliability, consensus and reproducibility of scores across items, cases and examiners were seen as a defining feature of the quality of an assessment, as well as the test's ability to discriminate between 'high' competence learners and 'low' competence learners. (Schuwirth & van der Vleuten, 2011a; Schuwirth & van der Vleuten, 2020) This desire for one correct judgement or a 'single truth,' free of any personal biases or judgements has been used to justify the fairness of assessment and objectivity could thus be construed as the hallmark of high-quality assessment. (Govaerts & van der Vleuten, 2013; McGuire, 1993; Ten Cate & Regehr, 2019; Valentine & Schuwirth, 2019)

Limitations of using objectivity to determine fairness

However, over time, research and other insights have led to cracks appearing in the argument that objectivity always leads to fairness and as a result there have been consistent and repeated calls in the literature to move away from an objectivity paradigm. (Bacon, Williams, Grealish & Jamieson, 2015; Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014; Govaerts, van de Wiel, Schuwirth, van der Vleuten & Muijtjens, 2013; Hodges, 2013; Jones, 1999; Rotthoff, 2018; Schuwirth & van der Vleuten, 2006; Ten Cate & Regehr, 2019)

Quantification of quality

Whilst there are benefits to objectivity in assessments, it can lead to an overreliance on quantitative and psychometric data, and ignore important aspects of assessment which cannot be measured. The dangers of this have been well illustrated through non-clinical examples of what is known as 'The McNamara fallacy'. Robert McNamara was considered a Harvard 'Whiz Kid' who worked at the Ford Motor company in the 1950s. He and his colleagues achieved a dramatic turnaround at the Ford Motor company taking the company from disarray to profit in a very short period of time by applying rational statistical analysis and planning. After his success at Ford, McNamara became the US Secretary for Defence in 1961 and held this position during the Vietnam war. (O'Mahony, 2017) He used his same trademark reproducible, quantitative and psychometric data analysis, such as body counts and territory gained to make strategic wartime decisions. (O'Mahony, 2017) Unfortunately this process of reducing complex human processes to numbers failed as it ignored the highly unconventional, highly motivated people's movements and the chaos and destruction of war. (McNamara, 2017) McNamara was blindsided by the data, computer models and statistics and was convinced his side had 'as good as won' despite his commanders and people in the field telling him the exact opposite. (Carmody, 2019) Following the war, the economist reported "He was haunted by the thought that amid all the objective-setting and evaluating, the careful counting and the cost-benefit analysis stood ordinary human beings. They behaved unpredictably." (O'Mahony, 2017)

Daniel Yankelovich coined the McNamara fallacy as: “The first step is to measure whatever can be easily measured. This is OK as far as it goes. The second step is to disregard that which can't be easily measured or to give it an arbitrary quantitative value. This is artificial and misleading. The third step is to presume that what can't be measured easily really isn't important. This is blindness. The fourth step is to say that what can't be easily measured really doesn't exist. This is suicide.” (Yankelovich, 1972)

There are also clinical examples of this fallacy. Cancer pharmaceutical trials frequently report progression-free survival while ignoring more meaningful measurements such as overall survival or quality of life. As progression-free survival does not always correlate with overall survival, this measure has been criticised as being not being clinically significant for either doctors or patients. Some authors have concluded: “Let us not assign meaning to something that is merely measurable, while failing to measure, or failing to make decisions based on, those things that are truly important.” (Booth & Eisenhauer, 2012)

Health professions education has struggled to reconcile assessment based on what we can easily measure, and what we should measure; acknowledging that indeed there is so much more to competence than just knowledge and skills. (Boulet & Durning, 2019; Rotthoff, 2018) There has always been a concern that ‘we start with the intent of making the important measurable and end up making the measurable important.’ (William, 2001) This is Goodheart’s law. Anthropologist Marilyn Strathern generalised this law as ‘When a measure becomes a target, it ceases to be a good

measure.’ (Strathern, 1997) Any effort to prioritise and measure one aspect of trainees’ qualities, such as knowledge, will inevitably reduce emphasis on other aspects which might be deemed important. (Eva, 2015) In its attempt to ensure fairness, it could be argued that objectivity actually reduces fairness because it only measures what can be measured by a quantitative value. It rewards learners who can remember and recall facts or follow protocols and is unfair to those with strengths in other skills such as communicating with patients, making decisions in difficult situations or reacting to changing circumstances. This may be unfair to patients and society, because society highly values unquantifiable competencies such as compassion, kindness and courage in their health care professionals. (O'Mahony, 2017)

Another consequence of reducing complex processes to numbers, is that it reduces the whole to its individual components making the assumption we can manage complex interactions by separating the whole into parts, analysing these parts, then putting them back together without significant loss. (Plsek & Greenhalgh, 2001) The faulty parts can even be replaced, or in the setting of education, trained. (Periyakoil, 2008) A further assumption often is that if we can comprehend the workings of each piece, the whole can be understood. (Plsek & Greenhalgh, 2001) Unfortunately, it can be argued that these assumptions do not hold true for health professions education. We know that the sum of what a competent clinician does is far greater than the sum of what can be measured in competence terms. (Grant, 1999; ten Cate, 2006) The competencies required of a health professional, such as communication skills, professionalism, collaborating with other team members, problem solving and so on are so intertwined that assessing these individually has no practical value. (Ten Cate, 2006) The interactions between the components add to the system, so that the whole

is more than the sum of the parts. Reed illustrates it as, 'life is more than molecules and atoms – it is the complex patterns of organisation that emerge between them'. (Reed, Howe, Doyle & Bell, 2018)

Thus, one of the limitations of objectivity is that it forces a quantification of quality. In the complex field of health professions education, this is not only impractical, but it may lead to skewed priorities and a dangerous neglect of important unquantifiable qualities. Furthermore, a deconstructive reductionist approach ignores the rich interactions which occur in a system. These interactions contribute to the outcomes which emerge. Thus, equating fairness solely with objectivity may actually inadvertently undermine fairness.

Reductionism at the expense of learning

Another limitation of objectivity is the inevitable reductionism and likely irrelevance which must occur to obtain a numerical score. Whilst this has allowed assessment to be able to discriminate between 'high' competence learners and 'low competence' learners (Schuwirth & van der Vleuten, 2020) there are several limitations to this approach. In addition to promoting competition between colleagues, this approach can be argued to be at the detriment of learning, which in itself is unfair to both learners and society as it denies learners the opportunity to improve. (Cilliers, Schuwirth, Adendorff, Herman & van der Vleuten, 2010; Cilliers, Schuwirth, Herman, Adendorff & van der Vleuten, 2012)

Reducing the rich assessment information to a score, and confining expert assessors to a pre-determined marking grid can limit the ability to provide credible and meaningful feedback to the learner. Credible and meaningful information to help the trainee learn may be missing because it has been reduced to a number, or it does not fit within the predetermined marking criteria. (Schuwirth & van der Vleuten 2011) A number in itself, without at least an explanatory narrative is meaningless. Schuwirth and van der Vleuten note some literature on scoring rubrics and standard setting methods is basically literature on how best to throw away assessment information. (Schuwirth & van der Vleuten, 2011a) Rich assessment information is not captured and cannot be provided to the learner to help them improve and address their individual strengths and weaknesses. Alternatively, pre-determined assessment forms require assessors to make judgements which may not be appropriate for the context of the clinical situation and which may reduce the credibility of the information provided. (Watling, 2014b)

Reducing information to a numerical score also necessitates making arbitrary decisions, such as a cut off score set at 50%. (Schuwirth & van der Vleuten, 2020) Conversations to arrive at these arbitrary decisions are rarely straightforward and often require complex negotiation among experts, and perhaps could be better consider a (negotiated) shared subjectivity rather than a 'true' objective score. (Ten Cate & Regehr, 2019) Once a pass score has been reached, it can be argued that there is no encouragement for learners to continue to improve, especially if there is no feedback provided.

In contrast, research has demonstrated that narrative feedback provided in workplace - based assessment provides additional interpretive information, suggesting there is perhaps a richness of narrative comments which scores do not capture and which cannot be achieved with numbers or psychometrics. (Ginsburg, Eva & Regehr, 2013) Focusing on objectivity ignores this information. Even in assessment situations where feedback is provided, if the dominant culture is quantitative summative assessment, then the feedback is not used effectively for learning. (Harrison, Könings, Schuwirth, Wass & van der Vleuten, 2015) It has also been argued that documentation of subjective experiences is more likely to be defensible, such as “I’m comfortable with you doing x now” as opposed to “meets expectations”. (Ten Cate & Regehr, 2019)

In summary, a second limitation of objectivity is the reduction of rich information and valuable insights as a consequence of the desire to provide numerical scores. This is likely to hinder credible and meaningful feedback and to overlook contextual appropriateness which, in turn, can lead to competition and impair learners’ ability to improve.

Assessors are not neutral windows on competence

It has been assumed that expert assessors and/or clinicians have a shared understanding on what competence-based assessment is and the criteria for a competent performance. (Apramian, Cristancho, Sener & Lingard, 2018) However in

reality, every assessor, trainee, patient and family member will have different cultural and social positions and lived experiences, that will therefore influence their belief about what is appropriate clinician performance. (Kuper, Reeves, Albert & Hodges, 2007) As a result, individual views of competences will be inevitably different depending on one's background and culture. (Kuper, Reeves, Albert & Hodges, 2007) Assessors have been shown to have different interpretations of individual performances, (Apramian, Cristancho, Watling, Ott & Lingard, 2016a; Govaerts, Schuwirth, van der Vleuten, & Muijtjens, 2011; Govaerts, van de Wiel, Schuwirth, van der Vleuten & Muijtjens, 2013; ten Cate & Regehr, 2019) to have different perceptions of whether a performance upholds competence principles, (Apramian, Cristancho, Sener & Lingard, 2018; Bacon, Williams, Grealish, & Jamieson, 2015) and make different inferences during assessment about knowledge, skills and attitudes which can't be directly observed. (Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014; Kogan, Conforti, Bernabeo, Iobst, & Holmboe, 2011) Furthermore, it has been noted that it is rare for assessors to explicitly apply criteria of best practice when assessing clinical performances (Kogan, Conforti, Bernabeo, Iobst & Holmboe, 2011), but rather attach varying importance to different aspects of assessments or clinical procedures (Apramian, Cristancho, Watling, Ott, & Lingard, 2015; Apramian, Cristancho, Watling, Ott & Lingard, 2016a; Apramian, Cristancho, Watling, Ott & Lingard, 2016b) as well as attaching importance to factors outside of competency frameworks, which is another source of variability. (Oudkerk Pool, Govaerts, Jaarsma & Driessen, 2018) Assessors draw from multiple frames of reference (i.e. comparing the trainee's performance to oneself, or to other learners, or the assessment prior), (Apramian, Cristancho, Sener & Lingard, 2018; Bacon, Williams, Grealish & Jamieson, 2015; Kogan, Conforti, Bernabeo, Iobst & Holmboe, 2011; Kogan, Hess, Conforti & Holmboe, 2010; Yeates, O'Neill, Mann & Eva, 2013) use variable methods to synthesise judgements into

numerical ratings (Kogan, Conforti, Bernabeo, Iobst & Holmboe, 2011) and may modify assessment judgements to avoid unpleasant repercussions. (Berendonk, Stalmeijer & Schuwirth, 2013; Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014) Supervisors' personal tendency to trust also influences their ability to identify a trainees' level of competence. (Apramian, Cristancho, Sener & Lingard, 2018; Kogan, Conforti, Bernabeo, Iobst & Holmboe, 2011; Lipshitz, 2001)

Historically, multiple different strategies have been made to overcome these variabilities. (Bacon, Williams, Grealish & Jamieson, 2015; Eva & Hodges, 2012; Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014; Govaerts & van der Vleuten, 2013; Hodges, 2013; Jones, 1999; Rotthoff, 2018; Schuwirth & van der Vleuten, 2006; ten Cate & Regehr, 2019) These strategies have included ensuring using sufficiently large samples to ensure sufficient reliability of the assessment. (Schuwirth, Southgate, Page, Paget, Lescop, Lew, Wade & Baron-Maldonado, 2002) Another approach was to minimise the influence of human judgement through ensuring an objective design of the assessment, such as the objective structured clinical examinations (OSCE) and the mini-clinical evaluation exercise (mini-CEX). (Harden & Gleeson, 1979; Norcini, Blank, Arnold & Kimball, 1995; ten Cate & Regehr, 2019; Valentine & Schuwirth, 2019) Furthermore, attention was turned to examiner training, with rater discrepancy considered an error to be corrected, and so there was a focus on ensuring examiners were more consistent in their judgement. (Boursicot, Kemp, Wilkinson, Finyartini, Canning, Cilliers & Fuller, 2021) It has been hypothesised that assessor variability is the result of not knowing or not correctly applying the assessment criteria and thus could be improved by training, using relevant guidelines, performance criteria and an agreed upon frame of reference to assess performance. However, judgement and

decision making are highly complex, subject to multiple influences and idiosyncratic (Bacon, Williams, Grealish & Jamieson, 2015; Durning, Artino, Schuwirth, & van der Vleuten, 2013) and extensive attempts at different types of examiner training across many different institutions has not led to widespread rater-agreement, improved test psychometrics and reduced measurement “error.” (Downing, 2004; Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014; Gingerich, Regehr & Eva, 2011; Newble, Hoare & Sheldrake, 1980) One reason for this is that humans have limited capacity within their short-term memories requiring them to observe with individual pre-existing memories and experiences which vary significantly. Cognitive load theory states that assessors cannot simply passively observe and capture performance as human working memory and processes are limited in both capacity and duration allowing for maintaining and processing only a few pieces of information at any time. (Van Merriënboer & Sweller, 2005; Van Merriënboer & Sweller, 2010) To overcome this limited cognitive capacity, one method of adaption is through the development of nonanalytic resources. (Moulton, Regehr, Mylopoulos, & MacRae, 2007) Information is linked to a person’s pre-existing knowledge structures to allow it to be retained and used. For example, people activate schemas or networks of information which are used to judge the “new” information being observed and influence what judgements they reach and their recall of what occurs. (Boshuizen & Schmidt, 1992; Schmidt & Boshuizen, 1993) These processes are often not under conscious control and people are often unaware of their unconscious thoughts that influence either their cognition or behaviour which makes them hard to predict. (Boreham, 1994; Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014)

Even within health professions education, many authors have stressed the need to see assessors as active information processors and highlighted a complex interaction of impression formation, interpretation, memory recall and judgement in assigning ratings. (Gingerich, Regehr & Eva, 2011) In making judgements, assessors recognise and select relevant information, interpret and organize information in memory, search for additional information, and retrieve and integrate relevant information. (Govaerts, Schuwirth, van der Vleuten & Muijtjens, 2011) This process is not linear but is based on the assessor's past experiences, pattern recognition as well as effortful processing of information and understandings of their social, cultural and contextual surroundings. Different assessors spot different things within a complex performance and construct different interpretations of them.

Furthermore, it has become clear that context specificity is an additional important source of inter-rater variability in assessment. (Eva, 2003; Norman, Tugwell, Feightner, Muzzin & Jacoby, 1985; Swanson & Norcini, 1989) Assuming for one moment that a health professional could be trained to have these competencies such as professionalism, communication, collaboration, medical knowledge and so on, then research has highlighted that the ability to demonstrate these skills in one situation does not mean a learner will be able to demonstrate these skills in another situation. (Norman, Tugwell, Feightner, Muzzin & Jacoby, 1985; ten Cate, 2006)

So, in summary, although it was assumed that expert assessors could be trained to share a common understanding of competence - based assessment and performance criteria, in reality, differing personal situations, backgrounds, cultural and social

perspectives result in individual views of competence making rater agreement very challenging to consistently achieve.

Looking beyond objectivity

In response to many of the challenges described above, several changes have been made in the way clinical educators approach assessment. For example, although the reduction of data to numerical scores for the sake of discriminating between 'high' competence learners and 'low competence' learners (Schuwirth & van der Vleuten, 2020) is still common practice, changing views on learning have suggested that being able to assess an individual's progress over time, (Schuwirth & van der Vleuten, 2011b) and differentiate abilities within individuals (Hodges, 2013) is seen as more useful in an educational context than discriminating between individuals. There has been a shift towards longitudinal assessment which includes triangulation and shared subjectivity, allowing for meaningful feedback and targeted learning activities. (Schuwirth & van der Vleuten, 2020) Treating assessment data with an integrative rather than reductionist perspective ensures that valuable data allowing for learning is more likely to be prioritised. This is fairer to learners as it allows them to improve and likely to be fairer to society because facilitating continuous and meaningful learning for trainees in the long term leads to better outcomes. (Soderstrom & Bjork, 2015)

Furthermore, instead of striving for consistency, reliability, consensus and reproducibility in assessors; research has demonstrated that assessor diversity can

contribute to validity evidence. (Boursicot, Kemp, Wilkinson, Finyartini, Canning, Cilliers & Fuller, 2021; Gingerich, 2015) Gingerich et al. argue that clinical tasks are complex, and expert assessors may differ in their perspectives because they observe different aspects of a multifaceted phenomenon. (Gingerich, 2015) Assessors' past experiences and understanding of their social, cultural and contextual surroundings all contribute to these perspectives. (Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014) In this sense, these dissenting perspectives are not something to be 'calibrated' but are considered to have validity evidence if they add meaningful perspective to the assessment and add to the bigger picture being created about the learner. (Boursicot, Kemp, Wilkinson, Finyartini, Canning, Cilliers & Fuller, 2021)

Ten Cate and Regehr also argue that a skilful practitioner is required to be alert to the various ways in which their behaviour can be interpreted and adjust their behaviour based on feedback. Ensuring a diverse interpretation of performance and feedback will help prepare trainees for the real world clinical environment unlike assessments which suggest a single 'best' approach. (Ten Cate & Regehr, 2019)

As a consequence, assessor expertise is increasingly considered important, especially in terms of content expertise, assessment expertise and awareness of the goal of the assessment. Expert assessors are better able to take a broader, more holistic review in interpreting learner behaviour and integrating different aspects of performance. (Govaerts, Schuwirth, van der Vleuten & Muijtjens, 2011)

Despite these changes and many other transformations to how assessment has been viewed over time, (Schuwirth & van der Vleuten, 2020) tensions and debates regarding subjective judgement in assessment persist. (Bacon, Williams, Grealish & Jamieson 2015; Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014; Govaerts & van der Vleuten, 2013; Hodges, 2013; Jones, 1999; Rotthoff, 2018; Schuwirth & van der Vleuten, 2006; ten Cate & Regehr, 2019) Within the field of health professions education, questions arise about where do expert supervisor judgements fit alongside standardised testing? How do debates about psychometrics fit with debates about the credibility of human judgement? What is good, fit for purpose assessment?

Furthermore, there has been little research done on fairness. Within the literature, fairness is often implied rather than explicitly considered. One of the reasons for the lack of literature and research on this topic may be because fairness is not a simple construct to define. Equating fairness to objectivity may be intuitive, but this reduces a complex, multi-dimensional and contextual construct to a single linear, non-representative rule and is therefore likely to reduce fairness rather than promoting it.

Given the arguments against objectivity as the sole way to determine fairness, the consideration that workplace-based assessments typically occur in authentic, unpredictable clinical environments which suggests the implementation of standardised, reproducible, measurement-based assessments is often neither fair nor feasible, and society's changing expectations of health professionals (Lazcano-Ponce, Angeles-Llerenas, Rodríguez-Valentín, Salvador-Carulla, Domínguez-Esponda, Astudillo-García, Madrigal-de Leon & Katz, 2020) perhaps the time has come to

consider reconsider what is fairness in assessment. An ontological shift, looking beyond objectivity, is needed to better understand fairness in assessment.

And in light of the changing views on assessment, taking a step back and considering *what is fair assessment* may help provide valuable insight what the next era of assessment should look like. Changing focus to ask “what is fair human judgement in assessment”, rather than “what is ‘objective’ human judgement in assessment” allows us to embracing different perspectives.

Understanding of what makes assessment fair going beyond objectivity is needed. If we are able to look beyond the paradigm of objectivity, other factors can be considered; for example, creating a shared narrative and negotiating a shared subjectivity. But we need to understand how fairness is conceptualised and defined to successfully navigate this shift away from relying on objectivity to ensure fairness. Research looking explicitly into fairness in assessment is needed to help inform further development in assessment.

Whilst we may never achieve perfectly fair assessment, we can make it fairer. (Gipps & Stobart, 2009) This may not be a straightforward task, but specifically investigating what makes assessment fair for both learners and patients may lead to more purposeful, meaningful changes to our assessment systems and make them more aligned with the next era in assessment.

Aim of this PhD and outline of the thesis

Originally, the focus of inquiry of this PhD was to determine what constitutes fair judgement in health professions assessment. The hope was that by understanding what made subjective human judgements fair, it would enable these judgements to be legitimately incorporated into assessment programs. This expansion of our assessment programs to include expert supervisor judgements alongside traditional knowledge tests would ensure a broader range of capabilities are assessed to ensure learners are better equipped to be the health professionals required of the 21st century.

However, it was soon appreciated that for subjective human judgements in assessment to be fair, the entire assessment system in which they were made needed to be fair. Isolating judgement from the systemic context is not possible if fairness is to be achieved. Consequently, the research's trajectory evolved, and the scope was expanded to the overarching question: what is fairness in assessment? The original question: What makes human judgement fair is now integrated into this wider question.

This changing perspective is perhaps not surprising when another change in the way assessment has been perceived is considered. More recent ideas see assessment treated as a whole system problem. (Van Der Vleuten & Schuwirth, 2005) This

involves the meaningful integration and triangulation of assessment information from a systems perspective. (Schuwirth & van der Vleuten, 2020)

This PhD thesis contains 8 chapters. Chapter 1, is a general introduction and provides the background to this PhD. Chapter 2 is a published position paper which sets the scene of the PhD research question. This paper was written to assist with the conceptualisation of ideas. The aims of this position paper were to focus on fairness as a fundamental quality of assessment, by synthesising and linking literature, identifying established knowledge and perspectives, highlighting gaps in understanding, and providing direction on what remains to be understood. (Eva, 2008) I sought to challenge the often-held assumption that objectivity always leads to (and is the only way to) achieving fairness in assessment and I posit that subjective human judgement has a legitimate place alongside objectivity in fair assessment. Chapter 3 is a theoretical chapter which outlines the research paradigm and theories used in this PhD thesis. Chapter 4 is a published literature review. I chose a hermeneutic literature review for a scholarly knowledge synthesis and understanding of the factors, definitions and key questions associated with fairness in human judgement in assessment. (Boell & Cecez-Kecmanovic, 2010) The initial questions of our literature review were:

- What are the limitations of “objectivity” in medical assessment?
- What is fair?
- Can subjective human judgement in assessment be fair?
- What is it about human judgement that makes it acceptable and defensible in clinical medicine?

- What makes an assessor's judgement in assessments legitimate?
- What are the subdimensions or components of fairness?
- What is the relationship between these subdimensions?

Chapters 5 and 6 are published qualitative studies which sought to explore different stakeholders' conceptualisations on the characteristics of fair judgement. The first study (Chapter 5) used semi-structured interviews, using vignettes, of Australian supervisors and post-graduate trainees, and the second study (Chapter 6) consisted of online focus groups with assessment leaders from Australian and New Zealand medical schools. The specific aims of the first study (Chapter 5) were: how do assessors and learners conceptualise the characteristics of fair judgement and how do these understandings of fair human judgement of assessors and learners compare with our theoretically-constructed conceptual model. The aims for study two (Chapter 6) were: what are the characteristics of fair judgement from an assessment leaders' perspective, and to compare and contrast these understandings with our previously reported theoretically-constructed conceptual model. The study was also designed to better understand how these theoretical aspects translate to practice; and suggest design principles to assist in the practical application of the theory-derived conceptual model. Chapter 7 is a published qualitative study of assessment leaders from medical schools in the Netherlands. The aim of this study was to understand how external forces on a complex adaptive system can disrupt fair judgement emerging. Chapter 8 is a general discussion and the conclusions of the PhD.

CHAPTER TWO: THE PURSUIT OF FAIRNESS IN ASSESSMENT: LOOKING BEYOND THE OBJECTIVE

This is an accepted manuscript of an article published by Taylor & Francis in Medical Teacher in April 2022, available at:

<https://www.tandfonline.com/doi/full/10.1080/0142159X.2022.2031943>

Valentine N, Durning S, Shanahan EM, Van der Vleuten CM, Schuwirth L. The pursuit of fairness in assessment: Looking beyond the objective. *Med Teach.* 2022;44(4):353-9.

This article was co-authored with Professor Steven Durning, Professor Michael Shanahan, Professor Cees van der Vleuten and Professor Lambert Schuwirth. My contribution to this article was approximately 80% of the completed work. For this article, I formulated and refined the ideas and perspectives presented by engaging in collaborative discussions with my fellow authors. Additionally, I was responsible for writing the initial draft manuscript, and incorporating the suggested edits and revisions provided by the other authors. I was also responsible for the submission process, ensure the article adhered to the publication guidelines. Professor Steven Durning, Professor Michael Shanahan and Professor Lambert Schuwirth each contributed approximately 4.5%, and Cees van der Vleuten was responsible for the remaining 2% of the publication.

This manuscript sets the scene for this PhD and articulates the rationale for the need to investigate what is fair human judgement in assessment. It encourages readers to take a step back and change perspectives to focus on the fundamental underlying value of fairness in assessment. By shifting the focus to the core value of fairness in assessment, this paper lays the groundwork for the upcoming series of studies. It also played a key role in ensuring clarity of research ideas prior to commencing the literature review.

Abstract

Health professions education has undergone significant changes over the last few decades, including the rise of competency based medical education, a shift to authentic workplace-based assessments and increased emphasis on programmes of assessment. Despite these changes, there is still a commonly held assumption that objectivity always leads to and is the only way to achieve fairness in assessment. However, there are well documented limitations to using objectivity as the “gold standard” to which assessments are judged. Fairness, on the other hand, is a fundamental quality of assessment and a principle which almost no one contests. Taking a step back and changing perspectives to focus on fairness in assessment may help re-set a traditional objective approach and identify an equal role for subjective human judgement in assessment alongside objective methods.

This paper explores fairness as a fundamental quality of assessments. This approach legitimises human judgement and shared subjectivity in assessment decisions alongside objective methods. Widening the answer to the question: “What is fair assessment” to include not only objectivity but also expert human judgement and shared subjectivity can add significant value in ensuring learners are better equipped to be the health professionals required of the 21st century.

Introduction

Fourteen years ago, Schuwirth and van der Vleuten made a plea for new psychometric models in education assessment. (Schuwirth & van der Vleuten, 2006) Their paper argued “Assessment should be fair, honest and defensible...but the strict operationalisation of these values is—in our humble opinion—currently of limited value”. (Schuwirth & van der Vleuten, 2006) They made an appeal for a major revision of statistical concepts, approaches to assessment and the development of a new model that fits current assessment developments better. (Schuwirth & van der Vleuten, 2006) Indeed, around the turn of the century many changes were made to medical education assessment. Competency-based medical education became the dominant approach to medical education in many countries. (Ten Cate, 2017) With this, the role of the doctor was redefined to include features which had previously not been emphasised, and learners were certified based on outcome rather than input. (Ten Cate & Billett, 2014) 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; Valentine & Schuwirth, 2019; van der

Vleuten & Schuwirth, 2005) More recently, competencies have been defined into professional tasks which a learner is entrusted to complete independently. (Ten Cate & Scheele, 2007)

Throughout these times of change, objective approaches have still remained a dominant discourse in assessment, with many seeing objectivity as the “gold standard” to which assessments should be judged. (Govaerts & van der Vleuten, 2013; ten Cate & Regehr, 2019; Valentine & Schuwirth, 2019; van der Vleuten, Norman & De Graaff, 1991)

More recently, there has been an increasing push in the literature to better utilise the role of human judgement and subjectivity in assessment (Bacon, Williams, Grealish & Jamieson 2015; Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014; Govaerts & van der Vleuten, 2013; Hodges, 2013; Jones, 1999; Rotthoff, 2018; Schuwirth & van der Vleuten, 2006; ten Cate & Regehr, 2019) and in 2020, the Ottawa consensus statement report for performance in assessment specifically called for assessment programs to ‘re-instate expert judgement’. (Boursicot, 2020)

Perhaps with the benefit of a decade and half of hindsight, we could say this 2006 paper (Schuwirth & van der Vleuten, 2006) didn’t go quite far enough as it was still looking for new psychometric methods with an ‘objective’ mindset. Objectivity which De Groot defined as “judgement without interference or even potential interference of personal opinions, preferences, modes of observation, views, interests or sentiments”

(Ten Cate & Regehr, 2019) has frequently become synonymous with fairness and is often used to determine the quality of the assessment. Workplace-based assessments, designed to assess authentic performance, are often judged using a quantitative psychometric framework and therefore criticised for not meeting validity and reliability criteria. (Govaerts & van der Vleuten, 2013) However an exclusive focus on traditional psychometric approaches can disregard key issues of competence, performance and assessment in complex workplace settings. (Govaerts & van der Vleuten, 2013) 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.

Fairness is a fundamental quality of assessment and a general principle which no one contests. (General Medical Council, 2017; Green, Johnson, Kim & Pope, 2007; Tierney, 2013) However, fairness is not a simple construct which is easy to define or conclusively described in the literature. There is no simple, all-encompassing consensus definition of fairness. Fairness has been associated with a wide range of assessment related qualities such as equitable, consistent, balanced, useful and ethically feasible. This breadth demonstrates that fairness in assessment is a comprehensive term and not something which can be reduced to a number, determined dichotomously or a straight forward process. (Tierney, 2013, 2014) More recently, Shaw and Nisbet considered fair assessment in the light of COVID-19 and similarly identified multiple challenges to consider in defining fairness. (Shaw & Nisbet, 2021) Simply equating fairness to objectivity, can reduce a complex, diverse multi-dimensional, context dependent construct to a single linear, likely non-representative rule. If the lens of improvement remains on optimising objectivity, then the focus is on

better psychometric techniques, but if we return to the principle of fairness, then the focus becomes much wider.

Objective

In this paper we seek to focus more broadly on fairness as a fundamental quality of assessment, not in a traditional systematic review but rather by synthesising and linking literature, identifying established knowledge and perspectives, highlighting gaps in understanding and providing direction on what remains to be understood. (Eva, 2008) We look to challenge the often-held assumption that objectivity always leads to, and is the only way to achieve fairness in assessment, and suggest that subjective human judgement has a legitimate place alongside objectivity in fair assessment. Two arguments will be put forward to challenge this assumption. Firstly, we consider the contention that objectivity can comprehensively assess the complexity of clinical practice to be a fallacy, and secondly, that true objectivity in assessment is unobtainable. Finally, from the collation of the perspectives in the literature, we will suggest that focusing on fairness rather than objectivity ensures that expert judgement, and shared subjectivity can be seen as at least equal to and used in combination with objectivity. We will also discuss opportunities for future research from a fairness lens.

Limitations of Objectivity

Whilst there are many benefits to objectivity in assessment, it can lead to a naïve trust in linear causality, a reliance on reproducible, quantitative and psychometric data, and to reification. During the Vietnam War, McNamara, the US Secretary of Defence, quantified the war effort into metrics such as body counts and territory gained. As a past chairman of the Ford Motor Company, McNamara applied objective, quantified metrics to improve production lines with great success. However, war is a complex non-linear and largely unpredictable process and the approach of reducing complex human processes to body counts and territory gained failed as it ignored the actions of highly motivated people and the chaos and destruction of war. (McNamara, 2017) This lead to the McNamara fallacy, which was coined by Yankelovich (figure 1). (Yankelovich, 1972)

McNamara Fallacy

- The first step is to measure whatever can be easily measured.
This is ok as far as it goes
- The second step is to disregard that which can't be easily measured or to give it an arbitrary quantitative value.
This is artificial and misleading
- The third step is to presume that what can't be measured easily really isn't important.
This is blindness
- The fourth step is to say that what can't be easily measured really doesn't exist.
This is suicide

Daniel Yankelovich 1972

Figure 1: The McNamara fallacy

Medicine, like war, is also complex, non-linear and to a certain extent unpredictable. Complex systems are characterised as having multiple, dynamic components interacting in non-linear and unpredictable ways, where the whole system is more than the sum of the parts. (Katerndahl, Burge, Ferrer, Becho, & Wood, 2010; Reed, Howe, Doyle & Bell, 2018) Health professionals work with complex problems presenting in a variety of different ways and through multiple dimensions of human experience (biological, psychological, social, spiritual). Treatment decisions are often made in the face of uncertainty as every patient responds differently to the array of therapeutic options. (Kaldjian, 2010) Furthermore, society expects health professionals to have not only knowledge of established and evolving diseases, but also interpersonal and communication skills, be able to apply ethical principles and so on. (Norcini, Anderson, Bollela, Burch, Costa, Duvivier, Galbraith, Hays, Kent, Perrott & Roberts, 2011)

McNamara was blindsided by the data, convinced the USA was winning the war despite his commanders telling him the exact opposite. (Carmody, 2019) Similarly, within clinical practice, equating 'quality' with someone who strictly adheres to guidelines or protocols, is to overlook the evidence on the more sophisticated process of expertise. (Greenhalgh, Howick & Maskrey, 2014) With regard to assessment, any

effort to prioritise and quantify one aspect of trainees' qualities, such as knowledge, will inevitably reduce emphasis on other aspects which might be deemed important. (Eva, 2015) It could actually be argued that objectivity can reduce fairness because it only measures what can be measured by a quantitative value. This is unfair to learners with broader skills and unfair to society who highly value unquantifiable competencies such as compassion, kindness and courage in their health care professionals. (O'Mahony, 2017; Wayne, Green, & Neilson, 2020) These skills, as well as other not easily quantifiable skills such as communication, collaboration and professionalism are often the ones needed within our health care systems. (Frank et al., 2010) Such reductionist approaches may also carry the risk of negatively impacting on student learning behaviour. Cilliers et al demonstrated that the effects of assessment on student learning is complex. Overreliance on 'objectivity' and quantitative results was perceived as punitive and unfair, and encouraged students' learning activities to be directed to passing assessments rather than learning to become a good clinician. (Cilliers, Schuwirth, Adendorff, Herman & van der Vleuten, 2010; Cilliers, Schuwirth, Herman, Adendorff, & van der Vleuten, 2012)

There are also further limitations to the use of objectivity in assessment. Assessment is always an evaluative process and therefore subjective. In the late 20th century, medical education assessments moved back into the authentic context of the workplace to help ease the tension between what is being measured and what should be measured. (Boulet & Durning, 2019; Rotthoff, 2018) However, in an attempt to remain true to the paradigm of objectivity, assessments such as the objective structured clinical examination (OSCE), were designed to minimise human judgement as much as possible. This was believed to improve fairness. (Gingerich, Kogan,

Yeates, Govaerts & Holmboe, 2014; Norcini, Blank, Arnold & Kimball, 1995; ten Cate & Regehr, 2019; Valentine & Schuwirth, 2019) Objectivity, from a positivist perspective that has played a prominent role in health professions education, suggests that for each item being measured, a 'true' score exists and any deviation from this true score is a measurement error. (Ten Cate & Regehr, 2019) However, even an 'objective' multiple choice examination involves a series of judgements by experts: what topics should be included, the choice of questions, specific wordings, decisions about pass scores and so on. (Norcini & Shea, 1997; ten Cate & Regehr, 2019; Valentine & Schuwirth, 2019) And, as other authors have noted, these judgements are rarely unanimous, often requiring complex negotiation between experts, which based on De Groot's definition, is far from objective. (Ten Cate & Regehr, 2019)

This can also be said of all quantitative measurement scales. Downie and colleagues stated "if the underlying purpose of questionnaires and measurement scales is to avoid the need for judgement then it does not succeed". (Downie & Macnaughton, 2013) Judgement is required in deciding what questions to ask, what numbers to assign and how to interpret final scores. Judgements can be dangerous when the professionals are unaware they are making them and believe themselves to be 'objective'. (Downie & Macnaughton, 2013) Within the health professions education literature it has been assumed that expert practitioners have a shared understanding on what competence-based assessment is and the criteria for a competent performance. (Apramian, Cristancho, Sener & Lingard, 2018) But, assessors have been shown repeatedly to have different interpretations of individual performances, (Apramian, Cristancho, Watling, Ott & Lingard, 2016a; Ten Cate & Regehr, 2019) different perceptions of whether a performance upholds competence principles,

(Apramian, Cristancho, Sener & Lingard, 2018; Bacon, Williams, Grealish & Jamieson 2015) and make different inferences about knowledge, skills and attitudes which can't be directly observed. (Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014; Kogan, Conforti, Bernabeo, lobst & Holmboe, 2011) Furthermore, it has been noted that assessors attach varying importance to different aspects of assessments or clinical procedures (Apramian, Cristancho, Watling, Ott & Lingard, 2105, 2016a; Kogan, Conforti, Bernabeo, lobst & Holmboe, 2011) as well as attach importance to factors outside of competency frameworks. (Oudkerk Pool, Govaerts, Jaarsma & Driessen, 2018) Assessors draw from multiple frames of reference (Apramian, Cristancho, Sener & Lingard, 2018; Bacon, Williams, Grealish & Jamieson 2015; Kogan, Conforti, Bernabeo, lobst & Holmboe, 2011; Kogan, Hess, Conforti & Holmboe, 2010; Yeates, O'Neill, Mann & Eva, 2013) and use variable methods to synthesise judgements into numerical ratings. (Kogan, Conforti, Bernabeo, lobst & Holmboe, 2011)

Moreover, the complexity of the task and the context of the work environment also influence assessment decisions. (Gingerich, Regehr & Eva, 2011) Returning to McNamara, The Economist observed "he was haunted by the thought that amid all the objective-setting and evaluating, the careful counting and the cost-benefit analysis, stood ordinary human beings. They behaved unpredictably." (O'Mahony, 2017) Workplace-based assessment occurs in environments where people are free to act in ways which are not predictable, and whose actions are interconnected so that one person's actions change the context for other people. (Greenhalgh & Papoutsis, 2018; Mennin, 2010; Plsek & Greenhalgh, 2001) For example, no one can predict what a patient will say in 3 minutes time and so there can be no protocol to access the learner taking the history. In their 2006 paper, Schuwirth and van der Vleuten noted "We dismiss variance between observers as error because we start from the assumption

that the universe is homogeneous, where in fact the more logical conclusion would have been that the universe is more variant". (Schuwirth & van der Vleuten, 2006)

This difficulty in obtaining agreement and a 'single truth' is not surprising because decision-making is idiosyncratic and individual. (Bacon, Williams, Grealish & Jamieson 2015; Durning, Artino & Schuwirth 2013) The psychology and cognitive science literature note that decision-making processes are highly complex, subject to multiple influences and no single theory of learning or performance can fully represent the underlying mechanisms. Furthermore, human working memory is thought to only hold approximately seven information elements at a time, and actively process no more than two to four elements at a time. (Young, Van Merriënboer, Durning & ten Cate, 2014) To overcome a limited working memory, information is rearranged and connected to pre-existing knowledge frameworks of information (schema) activated from long term memory. (Moulton, Regehr, Mylopoulos & MacRae, 2007; Young, van Merriënboer, Durning & ten Cate, 2014) These pre-existing schemas are used to judge the 'new' information being observed and influence what judgements are reached and an assessor's recall of what occurs. In assessment, assessors are active information processors who recognise, select and interpret relevant information, and integrate this information using their past experiences as well as their understandings of their social, cultural and contextual surroundings to form impressions and assign ratings. (Gingerich, Regehr & Eva, 2011; Govaerts & van der Vleuten, 2013; Govaerts, Schuwirth, van der Vleuten & Muijtjens, 2011) Previously, assessor variability has been framed primarily as a 'training issue' with the belief that the assessor is trainable. However, widespread assessor-agreement, improved test psychometrics and reduced measurement 'error' has remained elusive despite extensive efforts at faculty training.

(Downing, 2004; Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014; Gingerich, Regehr & Eva, 2011; Newble, Hoare, & Sheldrake, 1980) Delandshere and Petrosky noted: 'Judges' values, experiences, and interests are what makes them capable of interpreting complex performances, but it will never be possible to eliminate those attributes that make them different, even with extensive training and 'calibration'. (Delandshere & Petrosky, 1994)

One common attempt to overcome this variability has been to simply exclude outliers to ensure a common perspective and perceived reliability among the remaining raters. (Newble, Hoare, & Sheldrake, 1980) However, as noted by other authors, this approach does not exclude subjectivity – at best it masks subjectivity behind a constructed consensus (Ten Cate & Regehr, 2019) and perhaps eliminates the outliers who are unwilling to modify their assessment despite fear of unpleasant repercussions. (Kogan, Conforti, Bernabeo, Iobst & Holmboe, 2011) The development of the OSCE approach (Harden & Gleeson, 1979) was an illustration of reducing this assumed assessor-related 'error' through process rather than 'objectifying'. Although standardisation was initially prescribed through checklists, a sampling framework was also developed (having the candidate rotate from examiner to examiner) which accepted variability in assessors. The narrative at that time was psychometric but through the current lens, this can be seen as a procedural approach to ensuring fairness.

Moving beyond ‘bias free objectivity’ in the pursuit of fairness

Pursuing fairness in assessment through objectivity has many benefits but also has limitations. Firstly, it overlooks the complexity of clinical and educational practice and the wide variety of skills demanded of health professionals. Secondly, despite multiple efforts over several decades at both internal (such as faculty training) and external solutions (such as structured forms), researchers have not satisfactorily managed to achieve ‘bias-free objectivity’.

The call to move away from an over reliance on an objectivity paradigm has been echoed throughout the literature for several decades now. Gingerich and colleagues have suggested that perhaps the time has come to acknowledge a ‘single’ truth does not exist and consider an alternative conception of rater ‘error’ (Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014) and in 1999 Jones wrote that it is time “to acknowledge that the role of a professional vocational educator is to make educational judgements”. (Jones, 1999) These views have been echoed by other authors as they call for health professions education to have courage to acknowledge the benefits of subjectivity in assessment. (Bacon, Williams, Grealish & Jamieson 2015; Govaerts & van der Vleuten, 2013; Hodges, 2013; Rotthoff, 2018; ten Cate & Regehr, 2019)

Changing perspectives to look at fairness in assessment, rather than objectivity in assessment, allows for many different legitimate perspectives. Ten Cate and Regehr argue ‘objectivity’ might be better understood as negotiating a ‘shared subjectivity’: a convergence on an agreed-upon and socially constructed perspective. (Ten Cate &

Regehr, 2019) They note that although this convergence might achieve consensus, it isn't bias-free objectivity. (Ten Cate & Regehr, 2019) However, at the core of subjectivity is human judgment. To embrace subjectivity requires embracing human judgement and accepting that multiple different perspectives are not measurement errors to be corrected psychometrically but rather legitimate complementary perspectives of performance, much like how different angles of an object present in different but still 'correct' visions. And if we accept this premise, then we must develop new ways to determine the fairness of the judgements supervisors make.

The idea of human judgement and shared subjectivity in health professions education is widely used through competence committees and exam boards, especially since the introduction of programmatic longitudinal assessment. (Van Der Vleuten, Schuwirth, Driessen, Govaerts & Heeneman, 2015) Ten Cate and Regehr argue there are many positives in embracing multiple legitimate perspectives on a single performance, including helping the learner be alert to the fact that in real life, there are multiple ways in which behaviours can be interpreted and respond to accordingly. (Ten Cate & Regehr, 2019) Differences in opinions maybe noise in a psychometric perspective, but these differences may be very beneficial for the learning of an individual. Through the use of narrative instead of numbers, human judgement offers meaningful learning affordances through the possibility of richer feedback. Furthermore, the use of descriptive narrative in assessment has been shown in several studies to be a sensitive way of identifying at risk learners earlier, (Cohen, Blumberg, Ryan & Sullivan, 1993; Durning, Hanson, Gilliland, McManigle, Waechter & Pangaro, 2010) is a reliable way to distinguish between learners even after only a few assessment reports (Ginsburg, Eva, & Regehr, 2013; Ginsburg, van der Vleuten, & Eva, 2017) and

can predict future performance or need for remediation. (Cohen, Blumberg, Ryan & Sullivan, 1993)

Accepting and legitimising human judgement allows for context to be considered and this may be more defensible as assessors are not forced to document a context-free inference about the learner. (Ten Cate & Regehr, 2019) Furthermore, 'objectivity' in assessment can lead to many assessment forms aiming to be context independent, where assessors are forced to make judgements on a wide range of competencies not observed or in context of the clinical situation every time they complete an assessment form. Some authors have noted this can diminish the learners' trust in the assessor and process, and hides potentially credible decisions in a 'mountain of meaningless platitudes'. (McCready, 2007; Watling, 2014b) Subjectivity may help overcome this.

Human judgement also has its limitations. Multiple studies have demonstrated actuarial methods are superior to clinical prediction in many situations. (Marchese, 1992) The use of clinical guidelines has improved patient outcomes in many scenarios such as heart failure, breast cancer, atrial fibrillation, ventilator-assisted pneumonia and so on. (Murad, 2017) In both medical education and clinical medicine, objectively derived scales, guidelines and matrixes are an essential tool. However, these tools need to be used smartly. As Woolf and colleagues note it, 'clinical guidelines are useful when practitioners are unclear about appropriate practice and when scientific evidence can provide an answer. They are a poor remedy in other settings.' (Woolf, Grol, Hutchinson, Eccles & Grimshaw, 1999) Human judgement is essential in

determining when and how to apply these tools. We have tried to emphasise in this paper, that a desire to better utilise and legitimise subjective judgements in assessment, is not to dismiss the work done on objectivity in assessment over the last century. Nor does acknowledging that quantitative multiple choice tests are in fact based on a series of subjective decisions, mean they no longer have a place in modern assessment. Numerical ratings and standardised assessments are valuable elements in fair assessment. Instead, objectivity, subjectivity and human judgement are tensions which should be reconciled. (Govaerts, van der Vleuten & Holmboe, 2019) Recognising the role of human judgement in assessment, acknowledges that alongside knowledge tests there needs to be assessment of professional capabilities, and alongside debates about psychometrics there needs to be debates about the credibility and defensibility of human judgements. It isn't an either or, but rather a careful balancing of approaches in assessment programmes. (Govaerts & van der Vleuten, 2013; Govaerts, van der Vleuten & Holmboe, 2019) Changing focus from objectivity to fairness can assist with this. Widening the answer to the question: "What is fair assessment" to include human judgement can add significant value in ensuring a broader range of capabilities are being assessed to ensure learners are better equipped to be the health professionals required of the 21st century.

Conclusion

In increasingly complex clinical and educational environments, the challenge is to continue to move beyond the assumption that objectivity always leads to, and is the only way achieve fairness. Changing the focus from workplace-based assessment

being judged in terms of an objective psychometric framework to assessment being judged in terms of fairness, will help avoid falling prey to McNamara's fallacy and ensure we fulfil our social contract with society to train health professionals who are able to thrive in this ever changing environment, whilst remaining fair to the learners themselves. If we can move beyond the objective paradigm in our pursuit for fairness in assessment, we can start to explore shared subjectivity and human judgement in more depth. And a different ontological understanding of what makes human judgements and shared subjectivity fair in assessment is crucial.

There is no consensus roadmap to determine what is fair assessment conveniently published in the literature. Developing a deeper understanding of what fair human judgement looks like, how this can be defined, how it can be optimised for learning and how this can be supported is needed. What are the essential foundations of fairness and how can these be applied to judgements in complex environment of workplace-based assessment? It has been suggested fairness of human judgement can be enhanced through the use of a palette of assessor perspectives, the combination of multiple assessments, the use of narrative and paper trails in judgement decisions. (Dijkstra, Galbraith, Hodges, McAvoy, McCrorie, Southgate, van der Vleuten, Wass & Schuwirth, 2012; Dijkstra, van der Vleuten, & Schuwirth, 2010; van Der Vleuten, Schuwirth, Driessen, Govaerts & Heeneman, 2015) Some authors have also suggested looking to qualitative research strategies as an alternative to build rigour in assessment, (Driessen, van der Vleuten, Schuwirth, van Tartwijk & Vermunt, 2005; Frambach, van der Vleuten & Durning, 2013) however further research is still needed. Gipps and Stobart noted "We will never achieve fair assessment, but we can make it fairer." (Gipps & Stobart, 2009) And in 21st century

clinical practice, perhaps we can make health professions assessment fairer by looking beyond the objective paradigm.

Practice Points

- Objective approaches remain a dominant discourse in assessment.
- There are limitations to using objectivity as the “gold standard” to which assessments are judged.
- Within the literature, there is an increasing push to better utilise human judgement and subjectivity in assessment.
- Changing perspectives to focus on the fundamental underlying value of fairness in assessment may help re-set a traditional objective approach.

CHAPTER THREE: THEORETICAL PERSPECTIVES

Introduction

How we understand research and the nature of knowing influences all stages of the research process, from the conceptualisation of research questions through to interpretation and presentation of data. (Bunniss & Kelly, 2010; Rees & Monrouxe, 2010) No research is theory-free. As researchers, we need to be aware of the theoretical perspectives underpinning our research processes and declare these in our writing. (Bunniss & Kelly, 2010) Davidoff notes that “the key challenge for practitioners is not simply to base their work on theory (they always work from implicit assumptions and rationales, whether or not they do so consciously), but to make explicit the informal and formal theories they are actually using.” (Davidoff, Dixon-Woods, Leviton & Michie, 2015)

Lingard also challenges authors to be transparent in their orientation to research: “What kind of knowledge are researchers setting out to make? What are their views on knowledge, their epistemology? Are they conducting the study from an ethnographic, a critical theory, or a case study approach? These dimensions matter much more than the methodological tools, because they shape the way the research question is asked.” (Lingard, 2007)

Bunniss and Kelly support this premise, noting the quality of research is defined by the integrity and transparency of the research philosophy, and that the underlying ontological and epistemological assumptions of a study ultimately influence the nature of the knowledge claims that are constructed. (Bunniss & Kelly, 2010)

With this in mind, this chapter sets out to explicitly articulate the research paradigm and theories used in this PhD and why these were selected. Each separate study within this doctoral thesis also contains details about the research theories used in each individual study. This chapter does not intend to duplicate this information, but rather provide an overview of the overarching research paradigm and theories used in the entire PhD.

Health professions assessment

Over recent years there has been a notable shift in the way competence has been understood within health professions education. Whilst competency has never been simple or straight forward to define, competence has been previously approached from a positivist perspective. That is, competence could be subdivided into stable constructs which could be explicitly measured. (Hodges, 2013; Morcke, Dornan & Eika, 2013; Schuwirth & van der Vleuten, 2020) As a result, the health professions education literature described various different methods for measuring these

constructs (Schuwirth & van der Vleuten, 2011b) and positivist paradigms guided this inquiry.

However, health profession education understandings have evolved and changed over the last few decades. Modern education now builds on constructivist learning theories. (Mann & MacLeod, 2015; Van der Vleuten, Schuwirth, Driessen, Dijkstra, Tigelaar, Baartman & Tartwijk, 2015) In this approach, learning is seen as an active process in which learners construct the meaning of new knowledge in the light of their previous experience, knowledge, attitudes and skills. Moreover, learning is also closely intertwined with the context (or specific situation) in which it occurs. (Mann & MacLeod, 2015) In keeping with this paradigm shift, programmatic assessment has arguably become the dominant approach to modern assessment worldwide and has a constructivist paradigm at its core. (Pearce & Tavares, 2021) In programmatic assessment, assessment emphasizes context, and feedback is given to support learning, facilitate meaning making and enable remediation; all of which align with constructivist learning principles. (Van der Vleuten, Schuwirth, Driessen, Dijkstra, Tigelaar, Baartman & Tartwijk, 2015)

With regards to fairness, as I noted in the introduction, fairness is not a simple construct to define. Taking a positivist stance and seeking a 'single' truth is unlikely to be helpful as it may reduce a complex, multi-dimensional and contextual construct to a single linear, non-representative phenomenon. Therefore, the underlying theoretical paradigm of this PhD was constructivism. This aligns with the dominant world view on

learning and assessment and considers takes a pragmatic approach to how fairness is likely to be constructed.

Constructivist Paradigm

Paradigms are constellations of assumptions, values, beliefs and practices that form distinct ways of viewing the world. They can be seen as the foundational lenses through which we create or view or use theory. They play a crucial role in research, as Denzin and Lincoln described them as, the 'net that contains the researcher's epistemological, ontological and methodological premise'. (Denzin & Lincoln, 2017)

Ontology poses the question, 'what is reality?' whilst epistemology, concerned with the nature of knowledge asks 'how do we know?.' Methodology guides researchers through creating new knowledge by asking 'how can we know what can be known', and axiology explores the role of values in research. (Mann & MacLeod, 2015; Park, Konge & Artino, 2020)

Constructivism takes its roots from interpretivism and has emerged as an alternative to positivism for understanding the world. (Mann & MacLeod, 2015) Unlike positivism, the constructionist or social constructivist paradigm assumes multiple realities exist (ontology), that reality is actively and continually constructed through interactions with others (subjectivist epistemology) and uses a naturalistic (in the natural world) set of

methodological procedures, and acknowledges that research is value-based with the researcher being part of the exploration process. (Denzin & Lincoln, 2017; Mann & MacLeod, 2015)

In this chapter, the underlying assumptions of social constructivism will be described along the lines of ontology, epistemology, methodology and axiology. (Denzin & Lincoln, 2017)

Ontology

Relativism is the basic ontological premise of social constructivism. (Lincoln & Guba, 2016) In this view, social constructivism assumes that reality is relative and does not exist independently of the observer. Entities only exist in the minds of the person contemplating them, and therefore only have ontological status as an individual or group of persons grants them such status. (Lincoln & Guba, 2016)

Within our research, we acknowledged that there was no simple, universal, objective 'true' definition of fairness. Instead, we argue that fairness is a dynamic social construct shaped by shared interpretations and understandings among individuals and social groups. We acknowledged that fairness is constructed by society, it does not have a realist component, and undergoes changes over time and across cultures. The nature of this fairness can be observed, but the interpretations of these observations

are subjective and so to better understand the social world, researchers need to understand the subjective experiences of others.

Epistemology

In social constructivism, reality is seen as being constantly (re)shaped and (re)constructed through 'transactions' or interactions between individuals and their environment. (Mann & MacLeod, 2015) This perspective means that reality is not fixed but actively and continually constructed through interactions with others. It is also highly context and person specific, mediated by a person's prior experience and sociocultural factors. (Lincoln & Guba, 2016)

Fairness, and the components of fair judgement in assessment are constructed by individuals and institutions and change over time and across cultures. One person's understanding of the phenomenon may differ from another person's understanding of the phenomenon. An individual's perception of fair will also depend on context (i.e. situation) and is not only influenced by the role which the person has, ie.g. learner, supervisor, program designer, but also their experiences, their beliefs, and social and cultural background. Therefore, this PhD assumes that in order to build an understanding of fairness, the research must be aimed at exploring the meanings constructed by individuals and groups, collecting data from a multitude of perspectives, stakeholders and contexts to gain a richer and more nuanced understanding of the issue. (Varpio, Paradis, Uijtdehaage & Young, 2020)

Methodology

Lincoln notes that 'if the ontological presupposition of relativism and the epistemological presupposition of transactional subjectivism have been accepted then the methodology must involve meaning and sense making of those involved'. (Lincoln & Guba, 2016) In (social) constructivism, it is considered important for researchers to appreciate, compare and contrast the constructions individuals form to find meaning.

The methodologies appropriate for this PhD, therefore, need to be aimed at understanding how judgements are perceived and conceptualised to be fair. In this PhD, I sought multiple perspectives to gain a richer understanding of what constitutes fairness of human judgement in assessment. This understanding arises from recognising, understanding, developing and contrasting the constructions identified through dialogue. The language and narratives were explored to understand how each individual's experience was constructed within a particular context.

Axiology

In a shared and co-created reality, the values of the researcher and the values of the research participants and any other stakeholders need to be made transparent as exploration is guided and influenced by researcher position. (Lincoln & Guba, 2016)

Reflexivity was practiced throughout the research process to examine the decisions made. It is described both in this chapter and in the individual papers. Reflexivity involves the researcher reflecting critically on their role as the “human as instrument.” (Denzin & Lincoln, 2017) This process forces the researcher to reflect on the multiple identities that represent their ‘fluid self’ in the research setting. (Denzin & Lincoln, 2017)

Theory

Theories serve as conceptual tools which can aid in making sense of complex social realities. (Reeves, Albert, Kuper & Hodges, 2008) They can provide different “lenses,” allowing researcher to approach complicated problems and social issues from diverse perspectives and provide a framework for researchers to conduct their analysis. (Reeves, Albert, Kuper & Hodges, 2008) The application of theory enables maximum exploitation of learning and accumulation of knowledge, and promotes the transfer of learning from one project, context or challenge, to the next. (Davidoff, Dixon-Woods, Leviton & Michie, 2015)

Returning to the health professions education literature, it could be argued there has been a significant push for theories to be used which construct competence as a multifaceted concept rather than a single truth. (Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014; Govaerts & van der Vleuten, 2013; Govaerts, van der Vleuten,

Schuwirth & Muijtjens, 2007; Pearce & Tavares, 2021) For example, Govaerts et al has explicitly calling for a 'constructivist, social-psychological perspective' that 'integrates theories of cognition, motivation and decision making' into workplace based assessments (Govaerts, van der Vleuten, Schuwirth & Muijtjens, 2007) and a 'constructivist-interpretivist assessment framework.' (Govaerts & van der Vleuten, 2013)

In keeping with an inductive method of research, this PhD commenced with data analysis and from this abstracted a conceptual model and identified theory to inform the analysis. It acknowledged that theory is not just an abstract description, but rather resides within the researchers and frames our thinking and design choices. It is also constantly changing. As data analysis evolved, new ideas, insights and knowledge was developed. A conceptual framework was subsequently developed as our understanding of the phenomenon evolved.

As I progressed through the research process, my social constructivist view remained the same, but the nature of my ontological view shifted from linear thinking to dynamic and complex. It became increasingly clear that looking at interactions dynamically rather than solely looking for a static identification of what they were was important. Fairness, much like competence, can be viewed as the interaction with person and the situation at hand, being both idiosyncratic and contextual. Using a lens which celebrates interactions rather than dismisses these as 'noise' was needed. Therefore, a lens of complexity theory was adopted for studies 2 and 3 (chapters 6 and 7). This journey is explored in more detail in the discussion (chapter 8).

Complexity theory

Complexity theory is considered an “umbrella notion” for understanding complex phenomena. It is not a single, unified theory in itself but rather a collection of theories explaining related phenomena or characteristics from different areas such as mathematics, computer science, physics, biology, economics and so on. (Martin, McQuitty & Morgan, 2019) Complexity theory emerged as an interdisciplinary field during the 20th century as a way of examining the behaviour of systems. (Cristancho, Field & Lingard, 2019; Davis & Sumara, 1997; Martin, McQuitty & Morgan, 2019) Its origins began in the natural sciences, but has been adopted by the social sciences in more recent years. (Cleland, Patterson & Hanson, 2018) Within health professions education, complexity science application is even newer still, (Bleakley & Cleland, 2015) but it is gaining ground rapidly. There are many legitimate approaches to complexity theory and because of its transdisciplinary nature, many different terms are used such as complexity theory, complexity research, complexity science and complexity thinking. (Cristancho, Field & Lingard, 2019; Martin, McQuitty & Morgan, 2019) This diversity of terminology might be confusing but certain aspects of complex (adaptive) systems are central. The key components of a complex adaptive system are listed as below.

Some key components of complex adaptive systems

There are many components of complex adaptive systems. Some key components include:

- *Emergence*: Emergent phenomena occur in a complex system as a result of individual system elements *interacting* with one another and giving rise to diverse patterns or behaviours that were not predictable at the outset. The behaviours or outcomes can only 'emerge' as a result of the elements working together and interacting, they cannot be predetermined or directed by an external 'leader.' (Durning, Artino, Pangaro, van der Vleuten & Schuwirth, 2010; Reed, Howe, Doyle & Bell, 2018) As mentioned earlier, the emergent outcomes cannot be easily understood by simply knowing about the individual parts, but rather need to be understood at a system level because it is the interaction between the parts which produces the behaviours, not simply the parts themselves. (Lindberg, Nash & Lindberg, 2008) Understanding these individual parts will not make it possible to predict how the larger system will behave. The whole is more than the sum of the parts. (Plsek & Greenhalgh, 2001) The notion of complexity is therefore more aligned with the notion of holistic research approaches than with reductionist ones.
- *Dynamic*: Within complex systems, the dynamic agents act in parallel, constantly reacting to what the other agents are doing and in turn reverberates through the entire system and influences the behaviour of the network as a whole. (Bowe & Armstrong, 2017; Reed, Howe, Doyle & Bell, 2018) From a research perspective, using a complexity lens means the research should aim to understanding how systems evolve and how internal

and external forces drive or inhibit these systems. Focusing research on the interactions between the components may be more useful than learning about the individual agents and components themselves. (Reed, Howe, Doyle & Bell, 2018)

- *Self-organisation*: Order, innovation and progress emerge naturally from the system, they do not need to be imposed from within or from outside. (Greenhalgh & Papoutsis, 2018; Norman, 2011) Control is dispersed; the result of a huge number of decisions made by individual agents. (Van Beurden, Kia, Zask, Dietrich, & Rose, 2013) Seemingly obvious interventions can have minimal impact on system behaviour, whereas small changes can have large unintended consequences. (Bowe & Armstrong, 2017; Reed, Howe, Doyle & Bell, 2018; Van Beurden, Kia, Zask, Dietrich & Rose, 2013) So, research must not focus on strict rules, regulation and precise predictions, but rather on which internal and external forces drive the system in a specific, or even desired direction.
- *Adaption*: Agents adapt to past experience, (Fraser & Greenhalgh, 2001; Van Beurden, Kia, Zask, Dietrich & Rose, 2013) internal and external influences. However this also leads to unpredictability (Greenhalgh & Papoutsis, 2018; Mennin, 2010; Reed, Howe, Doyle & Bell, 2018) and resistance to centralised or hierarchical control. (Kurtz & Snowden, 2003) The constant adaption and interconnected nature of the agents leads to uncertainty and surprise meaning we cannot comprehend the behaviour of the system in a linear

fashion. (Plsek & Greenhalgh, 2001; Reed, Howe, Doyle & Bell, 2018) It also means that complex systems can defy intervention, and seemingly obvious interventions can have minimal impact on system behaviour, whereas small changes can have unintended consequences. (Reed, Howe, Doyle & Bell, 2018) Researching the patterns (in complexity these are typically known as fractals) which arise from complex adaptive systems is fundamental to understanding how the system works (Mennin, 2010) as they guide behaviours within it. (Reed, Howe, Doyle & Bell, 2018)

- *Fuzzy boundaries*: Complexity thinking maintains that systems can be aided by a minimal structure, such as fuzzy, ill-defined boundaries. (Fraser & Greenhalgh, 2001) These boundaries act as constraints in that they provide a stable structure within which change can occur. (Greenhalgh & Papoutsis, 2018; Mennin, 2010)
- *Embeddedness*: Individual agents and complex adaptive systems are embedded within wider complex adaptive systems. Therefore, we cannot fully understand the individual agents or systems without reference to the others. (Kurtz & Snowden, 2003; Plsek & Greenhalgh, 2001) Research should seek to understand individual's roles, backgrounds and contexts to appreciate their behaviour and perspectives within the complex adaptive system.

In this PhD, I acknowledge the diverse legitimate orientations to complexity theory without favouring one over another, and that the unifying aspect of complexity theory is that it rejects the machine-based metaphor in characterising and analysing phenomena and systems. (Davis & Sumara, 1997) Machines, although potentially very complicated, are still able to be reduced to the sum of their parts. In contrast, a complex adaptive system comprises of a collection of individual agents with freedom to act in ways that are not totally predictable. (Davis & Sumara, 1997; Plsek & Greenhalgh, 2001) Importantly, a complex adaptive system is more than just the sum of individual agents, it is a system in which the individual agents' actions are interconnected so that one agent's actions changes the context for the other agents. (Plsek & Greenhalgh, 2001) Using a clinical example, even a comprehensive understanding of the heart, brain stem and skin does not account for the emergence of complex phenomena such identity. (Davis & Sumara, 1997) Although these 'components' may contribute to such phenomena, their interrelation is too complex to study fragmentedly. (Davis & Sumara, 1997)

Components within a complex adaptive system are not fixed but constantly adapting. (Fraser & Greenhalgh, 2001; Van Beurden, Kia, Zask, Dietrich & Rose, 2013) According to complexity science, despite the unpredictable and adapting nature of complex systems, principles and patterns arise. (Davis & Sumara, 1997; Martin, McQuitty & Morgan, 2019; Reed, Howe, Doyle & Bell, 2018) Complexity theorists are interested in understanding these patterns, as this is fundamental to understanding how the system works (Mennin, 2010) as they guide behaviours within it. (Reed, Howe, Doyle & Bell, 2018) Thus complexity theorists are providing a 'rigorous

alternative to the divisive, reductionist and linear thinking that has dominated academic inquiry throughout the modern era.’ (Davis & Sumara, 1997)

Complexity theory and health professions education

The use of complexity as a lens to comprehend the nature of health professions education is not new and is increasingly encouraged. (Bowe & Armstrong, 2017; Fraser & Greenhalgh, 2001; Mennin, 2010) Health professions education research has traditionally been influenced strongly by the biomedical and physical sciences, adopting a goal of ‘evidence’ for simple, generalisable ‘truths.’ (Regehr, 2010; Rojas, 2018) It has even been suggested health professions education research tried to mimic the reference standard of quantitative randomised controlled trials by an imperative to demonstrate the effectiveness of educational interventions and their applicability across various contexts preferably with causal comparative study designs. (Rojas, 2018) Educational researchers also were influenced by biomedical sciences with Slavin maintaining the need to leverage rigorous research methods and randomised experiments to provide more effective and efficient educational programs and policies. (Slavin, 2002) However, Regehr argues that , in doing so, health professions education research failed to demonstrate the “beauty and richness of variation and context.” Furthermore, in a focus on finding a generalisable truth, opportunities were missed to develop effective approaches for representing this complexity, opting instead to dismiss it as mere interference. (Regehr, 2010) This represents an ontological and epistemological shift from ‘the nature of reality is linear

or best understood from a linear perspective' to 'the nature of reality is complex or best understood from a complexity perspective.'

More recently, health profession educational programmes have been typically perceived as taking place in complex environments. (Rojas, 2018) In particular, workplace-based assessment occurs in time pressured clinical situations with uncontrolled patient encounters where there assessment literally can 'walk through the door'.. Numerous complex relationships exist between different assessors, learners, patients, the healthcare environment and various contexts. In such a complex context, standard rules or algorithms cannot be applied to every possible situation and it is almost impossible to have control over all the vast activities and corresponding actions within these environments. As Greenhalgh notes of health services research, "the articulations, workarounds and muddling-through that keep the show on the road are not footnotes in the story but its central plot." (Greenhalgh & Papoutsi, 2018) This also is true of health professions education.

A lens of complexity helps here because the environments in which assessment occurs are dynamic with numerous complex relationships and contexts. Outcomes and processes are the result of multiple interactions that can lead to planned or unintended events. (Rojas, 2018) Considering assessment as a complex adaptive system has substantial explanatory power and can offer an understanding to this 'muddling-through' process. Regehr suggests that the "science of education is not about creating and sharing better generalisable solutions to common problems, but about creating and sharing better ways of thinking about the problems we face."

(Regehr, 2010) This idea was echoed by Eva who noted that “the very best RCT may prove that a specific educational intervention was effective, but if it does not advance understanding of a more general phenomenon than that particular course or workshop then it will not be of substantial use to the broader community.” (Eva, 2009)

Complexity science allows researchers to go beyond looking for simple ‘truths’ and develop powerful understanding of learning and education. (Bleakley & Cleland, 2015; Martin, McQuitty & Morgan, 2019) It provides a lens which can shift thinking from a pursuit for this elusive ‘truth’, to the recognition of the dynamic, multi-faceted aspects of a clinical education including managing uncertainty and acknowledging context, which is more likely to lead to productive solutions. (Cleland, Patterson & Hanson, 2018)

Methodology

Given this social constructivist approach and using a complexity lens, I chose a qualitative approach to my methodology. I started with a hermeneutic approach to a literature review. This was a deliberate choice in line with my ontological and epistemological stance. I was not trying to describe fairness as a realist and naturally occurring phenomenon but as a phenomenon that is constructed through human interaction. A more descriptive, systematic literature review would not have been onto-epistemological aligned. Consequently, rather than the more standard thematic or other descriptive review formats, a hermeneutic approach was chosen. In this review methodology, the literature is interrogated through specific questions similar to how participants in a qualitative interview study would be questioned. In subsequent

studies, , I sought to gain insight through interviews and focus groups, from the research participants, recognising that multiple competing and even conflicting claims about fairness exist. (Mann & MacLeod, 2015) I did not seek an 'objective' truth but rather expected co-creation of understanding and meaning making with the research participants which could be used to develop a shared narrative and inform meaningful action. I was not aiming to generalise finding but rather create rich descriptions of results. (Lincoln & Guba, 2016)

The aims of my research were not to create a checklist or a numerical score to which it could be determined that fairness has been achieved, or to compare if one type of assessment was fairer than another. Instead, the focus was on how fairness is perceived and how it emerges (and is hindered from emerging) within complex, dynamic authentic workplace-based environments.

In line with my previously stated constructivist assumptions that fairness as a reality is socially constructed by multiple stakeholders, and recognising that individuals and social groups share interpretations and understandings of fairness, I collected data from multiple perspectives to gain a richer and more nuanced understanding. (Varpio, Paradis, Uijtdehaage & Young, 2020) The studies included participants from post-graduate specialities and undergraduate medical schools. The first study (chapter 5) included post-graduate trainees (known as registrars or residents) and Australian supervisors who worked with either post-graduate trainees or medical students. As is common in Australia, many supervisors had dual role in supervising both medical students and post-graduate trainees or had done so at some point in their careers.

The second and third studies included faculty members from medical schools in Australia and New Zealand (chapter 6) and the Netherlands (chapter 7).

I undertook purposeful sampling to ensure maximum variety of cases. (Lincoln & Guba, 2016) In the first study (chapter 5), this included participants from a variety of specialties (noting general practice is by far the most predominant specialty in Australia), from a variety of locations across the country, with a mix of experience and gender. In the second and third studies, I included participants with diverse roles and experience from various medical schools, both urban and regional, across their respective countries.

Further specific details of methodologies used within studies are included in each chapter and not repeated here. All studies conducted in this thesis were approved by the Flinders University Human Research Ethics Committee and each participant provided informed consent prior to participating in the studies.

Throughout this research process, as there were multiple realities which were being actively constructed, I, as a researcher, was not a passive observer of the research but rather actively involved in the research process. Researchers and participants acted together to co-create knowledge and co-create a social constructed shared reality. (Lincoln & Guba, 2016) Reflexivity was explicitly addressed within the published individual research papers, however a broader overview will be provided here as it is an ongoing process that extends throughout the entire research

endeavour. (Olmos-Vega, Stalmeijer, Varpio, & Kahlke, 2022) Olmos-Vega and colleagues note reflexivity to be “a set of continuous, collaborative, and multifaceted practices through which researchers self-consciously critique, appraise, and evaluate how their subjectivity and context influence the research processes.” (Olmos-Vega, Stalmeijer, Varpio, & Kahlke, 2022) My reflexivity statement below is described through the dimensions of ‘personal’, ‘interpersonal’, ‘methodological’ and ‘contextual’. (Olmos-Vega, Stalmeijer, Varpio, & Kahlke, 2022)

Reflexivity

I was the primary researcher for this PhD. My research ideas were influenced by my personal experiences and observations as a trainee and as a clinician, as well as through my work in clinical education. Towards the end of my PhD journey, I embarked on training in a new specialty, and so again became a trainee, some 15 years after my first post graduate training experience. In addition to my own experiences, my personal interactions with colleagues, who have informally shared stories of their assessment experiences, have played a role in shaping my research perspective as we would discuss how assessment impacted every aspect of their lives and how they often carried these experiences with them for many years after the experience. Similarly, my partner is a clinician, and his lived assessment experience in a different specialty has also influenced me as I experienced how assessment can influence not only the learner but also those near to the learner both personally and professionally. On the other side, I have been involved in clinical education and assessment implementation for 10 years. Due to this involvement in clinical education

and assessment, with a small number of the Australian and New Zealand participants, I had some pre-existing relationships of varying degrees. These were mostly not close relationships, and none of these participants had ever been my supervisors or mentors, nor had I ever been a supervisor or mentor for any of the participants. The interviews and focus groups were also held online, allowing individual participants the choice of whether or not to turn on their video camera. This approach aimed to provide participants with a sense of control over the interview process.

Positioning this research within subjectivist or social constructivist paradigms was a deliberate choice. This was because as a team of researchers, we all understood that there was no simple, universal, objective 'true' definition of fairness. Any such 'definition' would likely not be fit for purpose. Instead, we all saw fairness as an dynamic social construct that exists because individuals and social groups share interpretations and understandings of this reality. Hence, a social constructivist paradigm was used in this research. There were multiple contexts to this research. This required me to reflect on how each context shaped the perspective of the participants in the project, not with the aim of neutralising the impact of the context, but rather to add to understanding.

Conclusion

In conclusion, this chapter has outlined the philosophical underpinnings, research paradigms and theories which form the foundation of this PhD thesis. Recognising the

importance of transparency in articulating our implicit assumptions and rationales, this chapter has sought to make these explicit. This includes using a constructivist paradigm, with its multiple views of reality, dynamic nature of knowledge construction and the importance of multiple perspectives. It also includes using a complexity lens, recognising this as a tool to better understand the 'whole' of a system, more than just the components itself.

CHAPTER FOUR: FAIRNESS IN HUMAN JUDGEMENT IN ASSESSMENT: A HERMENEUTIC LITERATURE REVIEW AND CONCEPTUAL FRAMEWORK

This chapter is a published article: 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):713-38.

This article was co-authored with Professor Michael Shanahan, Professor Steven Durning and Professor Lambert Schuwirth. My contribution for this article was approximately 80% of the research design, 80% of the data collection and analysis, and 85% of the writing and editing. Specifically, for this article, I contributed to literature review, research design and conducted the literature search in accordance with the research design. I analysed the results initially independently, and later engaged in collaborative discussions with fellow authors. Furthermore, I wrote the initial draft, and incorporated edits and suggestions from the other authors. Professor Steven Durning, Professor Michael Shanahan and Professor Lambert Schuwirth equally shared the remaining percentage of the work.

Having now set the scene for this PhD research, identified my research question and made clear my theoretical positioning, I shifted my attention to undertaking a literature review. Given my constructivist paradigm and use of a complexity lens, a hermeneutic

literature review was chosen. Investigating fairness using a constructivist lens requires reviewing and compiling evidence from different disciplines and perspectives, considering unique contexts and reviewing implications for many different stakeholders. A hermeneutic approach uses a 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. 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 a coherent argument.

Throughout the review an interpretive approach was used to meaningfully synthesize and critique the existing literature. Consistent with this approach, the literature search was rigorous but flexible and iterative, and as ideas were mapped, classified and critically assessed and as the nature of the evidence became more apparent, there was further refinement of the research question. This chapter is the published literature review, presenting the literature review process and its associated findings.

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', 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, fit 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.

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 20th century health professions education research and development focussed on construct validity and reliability in assessment (Ten Cate & Regehr, 2019; Valentine & Schuwirth, 2019; van der Vleuten, Norman, & De Graaff, 1991). 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 & Billett, 2014). Competencies have been defined into professional tasks which a learner is entrusted to complete independently (Ten Cate & 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; Valentine & Schuwirth, 2019; van der Vleuten & Schuwirth, 2005). 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 (Govaerts & van der Vleuten, 2013; ten Cate & Regehr, 2019; Valentine & Schuwirth, 2019; van der Vleuten, Norman & De Graff, 1991). 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 & 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 & Schuwirth, 2019). However an exclusive focus on traditional psychometric approaches can

disregard key issues of competence, performance and assessment in complex workplace settings (Govaerts & van der Vleuten, 2013; Govaerts, van der Vleuten, Schuwirth & Muijtjens, 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 (Bacon, Williams, Grealish & Jamieson 2015; Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014; Gipps & Stobart, 2009; Govaerts & van der Vleuten, 2013; Hodges, 2013; Jones, 1999; Rotthoff, 2018; Schuwirth & van der Vleuten, 2006; ten Cate & Regehr, 2019). 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, Heifetz, Grashow, & Linsky, 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 a 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 & Cecez-Kecmanovic, 2010). The popularity of a hermeneutic review is increasing as it has value in generating insights from heterogeneous literatures which cannot be synthesised through systematic review methodology, and would otherwise produce inconclusive findings (Greenhalgh & 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 figure 2 (Boell & Cecez-Kecmanovic, 2014). Throughout the review an interpretive approach was used to meaningfully synthesise and critique the existing literature (Boell & 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 & Cecez-Kecmanovic, 2010).

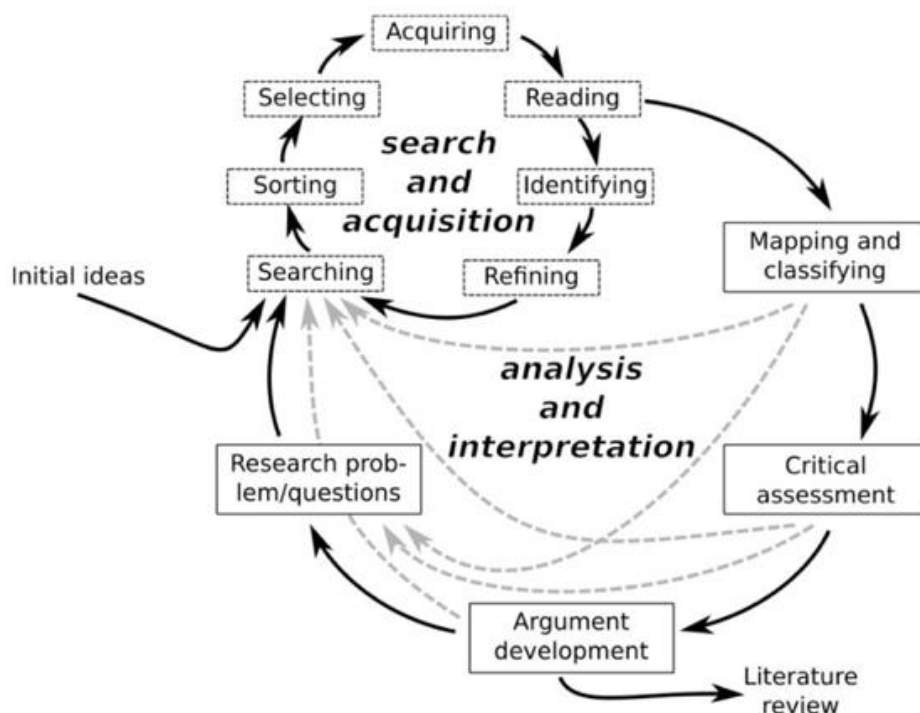


Figure 2: The hermeneutic circle as a framework for the literature review (Boell & Cecez-Kecmanovic, 2014)

Focus of the review

Following the steps outlined in figure 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:

- What are the limitations of “objectivity” in medical assessment?

- What is fair?
- Can subjective human judgement in assessment be fair?
- What is it about human judgement that makes it acceptable and defensible in clinical medicine?
- What makes an assessor's judgement in assessments legitimate?
- What are the subdimensions or components of fairness?
- 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 figure 3. 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 & Shaw, 2017) The literature searching took place over nine months to allow for subsequent

searches as new ideas emerged (Boell & 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 figure 3. There were no existing themes developed prior to starting the search. Having a less structured approach enhanced dialogical interaction between the literature and the researchers, encouraged critical assessment and supported argument development (Kusnanto, Agustian, & Hilmanto, 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.

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

1. Medical education OR
2. Education (including high school / university education) OR

3. Assess* OR
4. Post graduate OR
5. Health Professions Education OR
6. Portfolio OR
7. Learn* OR
8. Trainee*

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

1. Legal*
2. Defensib*
3. 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:

1. Value*
2. Narrative
3. Expertise
4. Holistic judgement
5. Opportunity
6. Transparency

7. Validity

AND

9. Medical education OR

10. Education (including high school / university education) OR

11. Assess* OR

12. Post graduate OR

13. Health Professions Education OR

14. Portfolio OR

15. Learn* OR

16. 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.

Figure 3: Search strategies used in the literature review

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 figure 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 (figure 4). 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 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.

Summary of included studies in the narrative review	
General background on fairness	Articles from both medical education (Harden, Lilley & Patricio 2015) and wider education literature (American Educational Research Association, American Psychological Association, National Council on Measurement in Education & Joint Committee on Standards for Educational Psychological Testing, 1999; Tierney, 2012).

<p>Values of fair human judgement in assessment:</p>	
<p>Credibility</p>	<p>Articles from the social psychology literature (Hilligoss, 2008; Lind & van den Bos, 2002; van den Bos & Miedema, 2000), the education literature (Chory, 2007; Rieh & Hilligoss, 2008; Rodabaugh, 1996) as well as perspectives and studies from the medical education literature (Ginsburg, van der Vleuten, Eva, & Lingard, 2017; Govaerts & van der Vleuten, 2013; Patterson, Zibarras, Carr, Irish, & Gregory, 2011; Telio, Regehr, & Ajjawi, 2016; Watling, 2014b; Watling, Kenyon, Zibrowski, Schulz, Goldszmidt, Singh, Maddocks & Lingard, 2008).</p>
<p>Defensibility</p>	<p>A review from the medical education (Colbert, French, Herring & Dannefer, 2017) and legal literature (Groarke; Reid, 1850; Upshur & Colak, 2003).</p>
<p>Fitness for purpose</p>	<p>Articles from the education literature (Beckett, 2008; Gipps & Stobart, 2009; Stobart, 2005), medical education literature (Duffield & Spencer, 2002; Eva, 2015; Govaerts & van der Vleuten, 2013; Viney, Rich, Needleman, Griffin & Woolf, 2017), psychology literature (Wolf, 1978), legal literature (Kaldjian, 2010; Stefan, 1993;</p>

<p>Transparency</p>	<p>Upshur & Colak, 2003) and a study from the rehabilitation literature (Ståhl, Seing, Gerdle, & Sandqvist, 2019).</p> <p>Studies, reviews and viewpoints from the medical education literature (Colbert, Dannefer & French 2017; Dijksterhuis, Voorhuis, Teunissen, Schuwirth, ten Cate, Braat & Scheele, 2009; Duffield & Spencer, 2002; Govaerts & van der Vleuten, 2013; Hays, Hamlin & Crane, 2015; Patterson, Zibarras, Carr, Irish & Gregory, 2011; Schuwirth, Southgate, Page, Paget, Lescop, Lew, Wade & Baron-Maldonado, 2002; Tavares & Eva, 2013; van der Vleuten, Schuwirth, Driessen, Govaerts & Heeneman, 2015; Watling, 2014b) and education literature (Gipps & Stobart, 2009; Rodabaugh, 1996; Tierney, 2012).</p>
<p>Components needed at an individual level:</p> <p>Narrative</p>	<p>Articles from the clinical medical literature (Greenhalgh & Hurwitz, 1999a; Greenhalgh & Hurwitz, 1999b) and the allied health education literature (Bacon, Holmes, & Palermo, 2017), perspectives and studies from the medical education literature (Cleland, Knight, Rees, Tracey, & Bond, 2008; Cohen, Blumberg, Ryan & Sullivan, 1993; Crossley & Jolly, 2012; Duffield & Spencer, 2002; Durning, Hanson, Gilliland, McManigle, Waechter & Pangaro, 2010;</p>

Ginsburg, Eva & Regehr, 2013; Ginsburg, Regehr, Lingard, & Eva, 2015; Ginsburg, van der Vleuten, Eva, & Lingard, 2016; Ginsburg, van der Vleuten, Eva & Lingard, 2017; Ginsburg, van der Vleuten & Eva, 2017; Govaerts & van der Vleuten, 2013; Kogan, Conforti, Iobst, & Holmboe, 2014; Tavares & Eva, 2013; Watling, Kenyon, Zibrowski, Schulz, Goldszmidt, Singh, Maddocks & Lingard, 2008; Weller, Misur, Nicolson, Morris, Ure, Crossley & Jolly, 2014), the education literature (Colbert, Fench, Herring & Dannefer 2017; Rodabaugh, 1996), a literature review from the nursing education literature (McCready, 2007), a legal perspective (Daniels & Sabin, 1997) and a study from the rehabilitation literature (Ståhl, Seing, Gerdle & Sandqvist, 2019).

Evidence

Articles from the clinical medicine literature (Downie & Macnaughton, 2009; Upshur & Colak, 2003), perspectives and studies from the medical education literature (Bullock, Lai, Lockspeiser, O'Sullivan, Aronowitz, Dellmore, Fung, Knight & Hauer, 2019; Duffield & Spencer, 2002; Govaerts & van der Vleuten, 2013; Southgate, Cox, David, Hatch, Howes, Johnson, Jolly, Macdonald, McAvoy, McCrorie & Turner 2001; Watling, Driessen, van der Vleuten, & Lingard, 2012; Watling, Driessen, van der Vleuten, Vanstone, & Lingard, 2013a; Watling & Ginsburg, 2019; Watling, Kenyon, Zibrowski, Schulz, Goldszmidt, Singh, Maddocks & Lingard, 2008) and papers from the allied health education literature (Bacon,

	<p>Holmes & Palermo 2017) and nursing literature (Webb, Endacott, Gray, Jasper, McMullan & Scholes, 2003).</p>
<p>Boundaries</p>	<p>Conference reports from the education literature (Houston, 2002), studies from the medical education literature (Rees & Shepherd, 2005; Watling & Ginsburg, 2019), the education literature (Rodabaugh, 1996) and the health policy literature (Kirkland, 2012).</p>
<p>Expertise</p>	<p>Studies, perspectives and a narrative review from the medical education literature (Berendonk, Stalmeijer, & Schuwirth, 2013; Govaerts, Schuwirth, van der Vleuten & Muijtjens, 2011; Govaerts & van der Vleuten, 2013; Hauer, ten Cate, Boscardian, Iobst, Holmboe, Chesluk, Baron & O'Sullivan, 2016; Jones, 1999; Telio, Regehr & Ajjawi, 2016; Watling, van der Vleuten & Lingard, 2012; Watling, Driessen, van der Vleuten, Vanstone, & Lingard, 2013b) and psychology literature (Marewski, Gaissmaier, & Gigerenzer, 2010).</p>
<p>Agility</p>	<p>Studies and perspectives from the medical education literature (Berendonk, Stalmeijer & Schuwirth, 2013; Crossley & Jolly, 2012; Flin, Youngson & Yule, 2007; Govaerts & van der Vleuten, 2013; MacRae, 1998; McCready, 2007; Watling, 2014b), papers and</p>

	<p>reviews from the clinical medical literature (Epstein, 2013; Greenhalgh, Howick & Maskrey, 2014; Kaldjian, 2010; Katerndahl, Parchman & Wood, 2010; Plsek & Greenhalgh, 2001), the education literature (Sadler, 2009), the psychology literature (lipshitz, 2001) and legal literature (Stefan, 1993).</p>
<p>Components needed at a systems level:</p> <p>Procedural Fairness</p> <p>Documentation</p>	<p>Studies, a review and perspectives from the medical education literature (Burgess, Roberts, Clark, & Mossman, 2014; Colbert, Fench, Herring & Dannefer, 2017; Hays, Hamlin & Crane, 2015; Ramani, Post, Konings, Mann, Katz & van der Vleuten 2017; van der Vleuten, Norman & De Graaff, 1991; Watling, Kenyon, Zibrowski, Schulz, Goldszmidt, Singh, Maddocks & Lingard, 2008) and studies from the psychology literature (Lind & Tyler, 1988; van den Bos, Lind, Vermunt, & Wilke, 1997; van den Bos, Wilke, & Lind, 1998).</p> <p>Papers from the medical education literature (Govaerts & van der Vleuten, 2013; Hays, Hamlin & Crane, 2015; McCready, 2007; Rees & Shepherd, 2005; Webb, Endacott, Gray, Jasper, McMullan & Scholes, 2003).</p>

Multiple
Opportunities

Papers from the clinical medical literature (Hunter, 1996), studies, a review and perspectives from the medical education literature (Boulet & Durning, 2019; Colbert, Fench, Herring & Dannefer, 2017; Dijksterhuis, Voorhuis, Teunissen, Schuwirth, ten Cate, Braat & Scheele, 2009; Eva, 2015; Govaerts & van der Vleuten, 2013; Hays, Hamlin & Crane, 2015; Schuwirth, Southgate, Page, Paget, Lescop, Lew, Wade & Baron-Maldonado, 2002; van der Vleuten & Schuwirth, 2005; Watling, Driessen, van der Vleuten, Vanstone, Lingard, 2013a; Watling, Kenyon, Zibrowski, Schulz, Goldszmidt, Singh, Maddocks & Lingard, 2008; Wycliffe-Jones, Hecker, Schipper, Topps, Robinson & Abedin et al., 2018) and papers from the education literature (Gipps & Stobart, 2009; Rodabaugh, 1996; Stobart, 2005; Tierney, 2012).

Judgements
assessed by
multiple assessors

Studies and perspectives from the medical education literature (Govaerts & van der Vleuten, 2013; Hauer, ten Cate, Boscardian, Iobst, Holmboe, Chesluk, Baron & O'Sullivan, 2016; Hauer, Chesluk, Iobst, Holmboe, Baron, Boscardian, ten Cate & O'Sullivan, 2015; Tochel, Haig, Hesketh, Cadzow, Beggs, Colhart & Peacock 2009) perspectives and a study from the allied health professions education literature (Bacon, Williams, Grealish & Jamieson 2015; Krefting, 1991; McCready, 2007; Webb, Endacott, Gray, Jasper,

Validity evidence for judgements	<p>McMullan & Scholes, 2003) and the clinical medicine literature (Ham, 1999).</p> <p>Papers from the medical education literature (Colbert, Dannefer, & French, 2015; Govaerts & van der Vleuten, 2013).</p>
----------------------------------	---

Table 1: Included papers in the literature review

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, Lilley & Patricio, 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 Educational Research Association, American Psychological Association, National Council on Measurement in Education & Joint Committee on Standards for Educational Psychological Testing, 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 breadth 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 (figure 4) 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.

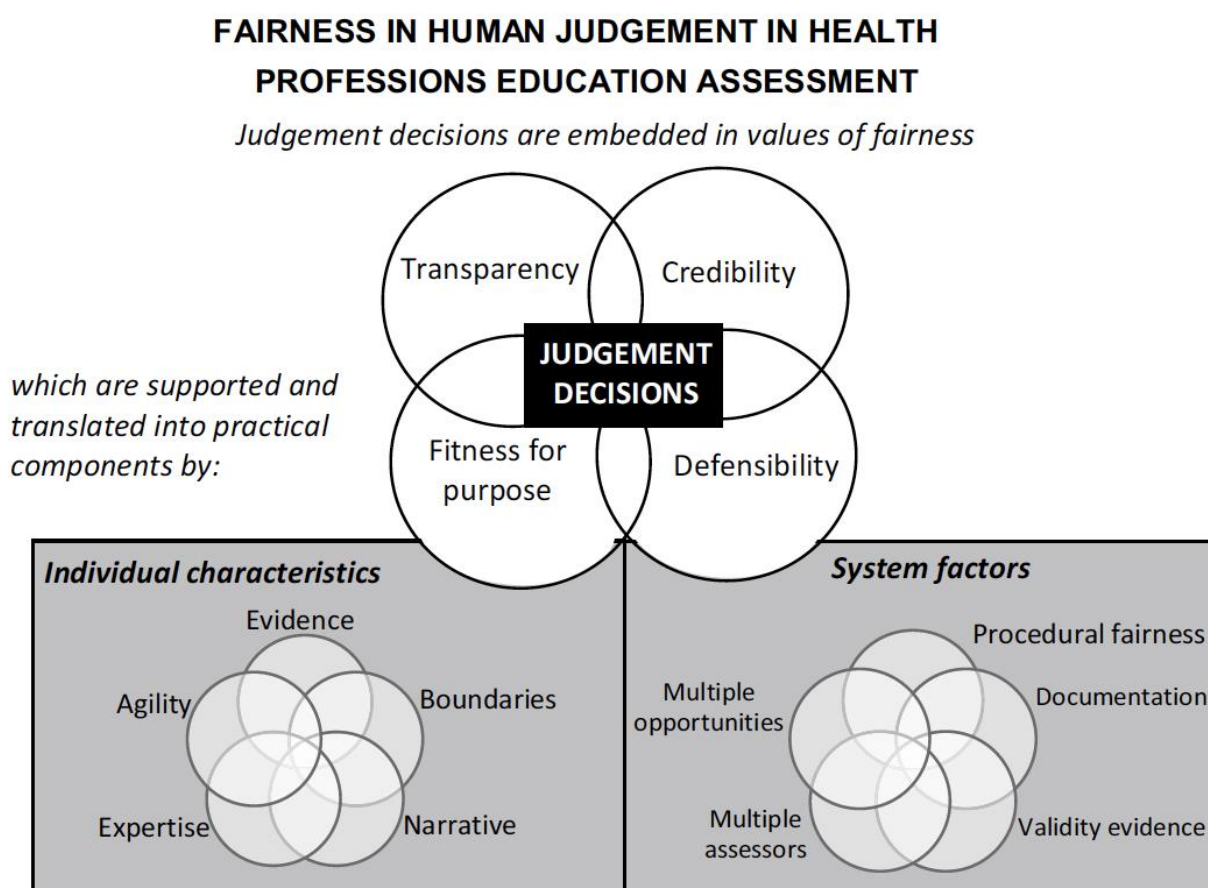


Figure 4: A conceptual framework of fairness in human judgement in assessment. The values of fairness are supported by individual characteristics and system factors.

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.

Values of fair human judgement: 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 (Lind & van den Bos, 2002; van den Bos & Miedema, 2000). There is no clear definition of credibility however an overarching view across definitions appears to be believability (Hilligoss, 2008), and confidence or trustability in the 'truthfulness' of the findings (Govaerts & van der Vleuten, 2013).

Credibility assessment is 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 & 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 & Hilligoss, 2008).

Interpersonal or interactional fairness, is an important component of credibility and fairness (Patterson, Zibarras, Carr, Irish & Gregory, 2011; Rodabaugh, 1996). 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 (Ginsburg, van der Vleuten, Eva & Lingard, 2017; Telio, Regehr & Ajjawi, 2016; Watling, Kenyon, Zibrowski, Schulz, Goldszmidt, Singh, Maddocks & Lingard, 2008). Prolonged observation, a positive learning culture, and multiple opportunities for evidence support development of this credibility judgement (Watling, 2014b; Watling, Kenyon, Zibrowski, Schulz, Goldszmidt, Singh, Maddocks & Lingard, 2008).

Values of fair human judgement: Defensibility

Judgement decisions in assessment need to be (legally) defensible as learners may seek (legal) redress with the concept of fairness often forming the basis of claims (Colbert, Fench, Herring & Dannefer, 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 (Groarke; Upshur & Colak, 2003). 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.

Values of fair human judgement: Fitness for Purpose

Many authors have argued that fairness is a social construct (Eva, 2015; Gipps & Stobart, 2009; Ståhl, Seing, Gerdle & Sandqvist, 2019; Stobart, 2005; Wolf, 1978). 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 & 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 & 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 & 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 & Spencer, 2002; Viney, Rich, Needleman, Griffin & Woolf, 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 accessibility of evidence, which is sensitive to the context of the workplace and patient situation, from which competence is inferred (Beckett, 2008).

Values of fair human judgement: Transparency

Throughout the literature, there is an emphasis on fair assessments demonstrating openness to build a shared understanding with learners (Colbert, Fench, Herring & Dannefer, 2017; Dijksterhuis, Voorhuis, Teunissen, Schuwirth, ten Cate, Braat & Scheele, 2009; Hays, Hamlin & Crane, 2015; Schuwirth, Southgate, Page, Paget, Lescop, Lew, Wade & Baron-Maldonado, 2002; van der Vleuten, Schuwirth, Driessen, Govaerts & Heeneman, 2015), with some authors arguing transparency is the best defence against unfair assessment (Gipps & 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, Zibarras, Carr, Irish, & Gregory, 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 & Stobart, 2009).

Transparency also includes a narrative which focuses on performance improvement and feedback (Colbert, Fench, Herring & Dannefer, 2017; Rodabaugh, 1996). 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 & Spencer, 2002). High quality, appropriate judgements about a performance which provide feedback build the credibility, transparency and thus fairness of a judgement decisions (Govaerts & van der Vleuten, 2013; Tavares & Eva, 2013).

However, transparency as a value can conflict with other values of fairness. (Tierney, 2012) For example, transparency provides students 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 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 (Watling, 2014b). 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 figure 4.

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 (Bacon, Holmes & Palmero, 2017; Ginsburg, Regehr, Lingard, & Eva, 2015; Govaerts & van der Vleuten, 2013), 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 & Hurwitz, 1999b). 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, Regehr, Lingard, & Eva, 2015; Govaerts & 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 & van der Vleuten, 2013; McCready, 2007). The use of descriptive narratives in assessment has been shown to identify at-risk learners earlier (Cohen, Blumberg, Ryan & Sullivan, 1993; Durning, Hanson, Gilliland, McManigle, Waechter & Pangaro, 2010; Ginsburg, Eva & Regehr, 2013; Ginsburg, van der Vleuten, & Eva, 2017) and contributes to predicting future performance or need for remediation (Cohen, Blumberg, Ryan & Sullivan, 1993). Narratives also lead assessors to more holistic judgements (Bacon, Holmes & Palermo, 2017) and allow for feedback which learners see as essential for a fair judgement (Colbert, Fench, Herring & Dannefer, 2017; Duffield & Spencer, 2002; Govaerts & van der Vleuten, 2013; Rodabaugh, 1996; Tavares & Eva, 2013; Watling, Kenyon, Zibrowski, Schulz, Goldszmidt, Singh, Maddocks & Lingard, 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, Seing, Gerdle & Sandqvist, 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, Holmes & Palmero, 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 & 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, Regehr, Lingard, & Eva, 2015; Ginsburg, van der Vleuten, Eva & Lingard, 2016, 2017). 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, van der Vleuten, Eva & Lingard, 2016, 2017), 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, Knight, Rees, Tracey, & Bond, 2008). To overcome these limitations, many have called for narratives which fit clinical practice to be used when asking assessors to make judgement (Crossley & Jolly, 2012; Kogan, Conforti, Iobst, Holmboe, 2014). Aligning rating scales to the construct of clinical independence or entrustment has been shown to improve score reliability and assessor discrimination (Crossley & Jolly, 2012; Weller, Misur, Nicolson, Morris, Ure, Crossley & Jolly, 2014). This also allows for clinical evidence to be form the basis of the narrative of the judgement which improves credibility (Watling, Driessen, van der Vleuten & Lingard, 2012). Furthermore, it also is fairer to patients, as the judgements are focused on high quality clinical care rather than rating scales. (Kogan, Conforti, Iobst, Holmboe, 2014)

Evidence

Evidence is offered as a means of supporting judgements (Upshur & Colak, 2003), and is essential for creating a validity argument (Govaerts & van der Vleuten, 2013). Without evidence, it is not a judgement but a guess (Downie & 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 & 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, Cox, David, Hatch, Howes, Johnson, Jolly, Macdonald, McAvoy, McCrorie & Turner, 2001).

Watling et al 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 (Watling, Driessen, van der Vleuten & Lingard, 2012). 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, Driessen, van der Vleuten, Vanstone, Lingard, 2013a; Watling & Ginsburg, 2019; Watling, Kenyon, Zibrowski, Schulz, Goldszmidt, Singh, Maddocks & Lingard, 2008), and this perception of the fairness is enhanced by prolonged observation (Bullock, Lai, Lockspeiser, O'Sullivan, Aronowitz, Dellmore, Fung, Knight & Hauer, 2019; Duffield & Spencer, 2002). 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 (Bacon, Holmes

& Palmero, 2017; Watling, Driessen, van der Vleuten, Vanstone, Lingard, 2013a; Webb, Endacott, Gray, Jasper, McMullan & Scholes, 2003).

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 & Shepherd, 2005). Continuous observation may mean every observation is an opportunity for learners to lose face and impact their assessment outcome (Watling & 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 (Berendonk et al., 2013; Telio, Regehr & Ajjawi, 2016; Watling, Driessen, van der Vleuten & Lingard, 2012; Watling, Driessen, van der Vleuten, Vanstone, & Lingard, 2013b). Decision making committees also value expertise, relying on faculty members' qualifications via their perceived status as expert to help ensure fairness and credibility (Hauer, ten Cate, Boscardian, Iobst, Holmboe, Chesluk, Baron & O'Sullivan, 2016).

Learners value clinical expertise over educational expertise (Watling, Driessen, van der Vleuten, Vanstone, & Lingard, 2013b). However, experts in medical education in general make more inferences on information, cluster sets of information into meaningful patterns and abstractions (Govaerts, Schuwirth, van der Vleuten & Muijtjens, 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 (Govaerts & van der Vleuten, 2013; Marewski, Gaissmaier & Gigerenzer, 2010; Watling, Driessen, van der Vleuten & Lingard, 2012). 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, Schuwirth, van der Vleuten & Muijtjens, 2011).

Agility

Govaerts et al 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 (Govaerts & van der Vleuten, 2013). In contrast, many assessment forms aim to be context independent and list performance dimensions as separate distinct entities which all need to be completed regardless of the clinical situation. Although this is transparent, it is not credible or fit for purpose (McCready, 2007; Watling, 2014b) and does not recognise assessors' agility to make contextually appropriate, holistic and individualised judgement decisions (Govaerts & van der Vleuten, 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, Howick & Maskrey, 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 & 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 (Berendonk, Stalmeijer & Schuwirth, 2013; Epstein, 2013; Flin, Youngson & Yule, 2007; Kaldjian, 2010; Katerndahl, Burge, Ferrer, Becho, & Wood, 2010; lipshitz, 2001; Plsek & Greenhalgh, 2001). A continuous cycle of monitoring to assess the situation, taking appropriate actions and re-evaluating the results is required (Flin, Youngson & Yule, 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, Youngson & Yule, 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, Dannefer & French, 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, Wilke & Lind, 1998; van der Vleuten, Norman & De Graaff, 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, Lind, Vermunt, & Wilke, 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, Roberts, Clark, & Mossman, 2014; Colbert, Fench, Herring & Dannefer, 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 & Tyler, 1988), by formal, regular inclusive reviews of the judgement process, and provision of appeals process (Hays, Hamlin & Crane, 2015). Also important for procedural fairness is to ensure the learner is explicitly told of the expectations and what else is required if they did not meet the expectations (Colbert, Fench, Herring & Dannefer, 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, Lind, Vermunt, & Wilke, 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, Post, Konings, Mann, Katz & van der Vleuten, 2017; Watling, Kenyon, Zibrowski, Schulz, Goldszmidt, Singh, Maddocks & Lingard, 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 & van der Vleuten, 2013; McCready, 2007; Webb, Endacott, Gray, Jasper, McMullan & Scholes, 2003). Furthermore, procedural fairness as described above needs clear and comprehensive documentation outlining assessment policies and procedures (Hays, Hamlin & Crane, 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 & 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 & 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 & van der Vleuten, 2013; Schuwirth, Southgate, Page, Paget, Lescop, Lew, Wade & Baron-Maldonado, 2002; van der Vleuten & 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, Voorhuis, Teunissen, Schuwirth, ten Cate, Braat & Scheele, 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, Fench, Herring & Dannefer, 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, Hamlin & Crane, 2015), 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 & Stobart, 2009). Furthermore, learners value opportunities to demonstrate they have understood and incorporated feedback they have received. (Watling, Driessen, van der Vleuten, Vanstone, Lingard, 2013a; Watling, Kenyon, Zibrowski, Schulz, Goldszmidt, Singh, Maddocks & Lingard, 2008)

Fairness has often been viewed as 'equal' treatment or practice (Colbert, Fench, Herring & Dannefer, 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, Fench, Herring & Dannefer, 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, Hecker, Schipper, Topps, Robinson & Abedin 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 (Bacon, Williams, Grealish & Jamieson 2015; Govaerts & van der Vleuten, 2013; Hauer, ten Cate, Boscardian, Iobst, Holmboe, Chesluk, Baron & O'Sullivan, 2016). 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 (Bacon, Williams, Grealish & Jamieson 2015; Govaerts & van der Vleuten, 2013; Ham, 1999; Hauer, Chesluk, Iobst, Holmboe, Baron, Boscardian, ten Cate & O'Sullivan, 2015; Krefting, 1991; Tochel, Haig, Hesketh, Cadzow, Beggs, Colhart & Peacock 2009; Webb, Endacott, Gray, Jasper, McMullan & Scholes, 2003). 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, Williams, Grealish & Jamieson 2015; Govaerts & van der Vleuten, 2013; Ham, 1999; Krefting, 1991; Webb, Endacott, Gray, Jasper, McMullan & Scholes, 2003). These qualitative methods of assessing

evidence also allow for the 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, ten Cate, Boscardian, lobst, Holmboe, Chesluk, Baron & O'Sullivan, 2016). This needs to be coupled with strategies to facilitate information sharing, to overcome tendencies of the group to prioritise information known to more group members or information shared first (Hauer, ten Cate, Boscardian, lobst, Holmboe, Chesluk, Baron & O'Sullivan, 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, Chesluk, lobst, Holmboe, Baron, Boscardin, ten Cate & O'Sullivan, 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, Chesluk, lobst, Holmboe, Baron, Boscardin, ten Cate & O'Sullivan, 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, Dannefer & French, 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 & 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 & 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 & 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 Papoutsis 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 (Greenhalgh & Papoutsis, 2018). 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 fairness 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 fit 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 & van der Vleuten, 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 be 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, van der Vleuten, & Holmboe, 2018).

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 (Greenhalgh & Shaw, 2017; Whitty, 2015). 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 21st-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 & 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 our 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.

CHAPTER FIVE: MAKING IT FAIR: LEARNERS' AND ASSESSORS' PERSPECTIVES OF THE ATTRIBUTES OF FAIR JUDGEMENT

This is the peer reviewed version of the following article: Valentine N, Shanahan EM, Durning S, Schuwirth L. Making it fair: Learners' and assessors' perspectives of the attributes of fair judgement. *Med Edu.* 2021;55(9):1056-66, which has been published in final form at <https://doi.org/10.1111/medu.14574>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions. This article may not be enhanced, enriched or otherwise transformed into a derivative work, without express permission from Wiley or by statutory rights under applicable legislation.

This article was co-authored with Professor Michael Shanahan, Professor Steven Durning and Professor Lambert Schuwirth. My contribution for this article was approximately 80% of the research design, 80% of the data collection and analysis, and 85% of the writing and editing. For this article, I was involved in the research design of this project, which included designing the interview guide and vignettes. Additionally, I recruited all of the participants and subsequently conducted all of the semi-structured interviews. I undertook the analysis of the data in collaboration with my fellow authors. I wrote the original draft manuscript, and incorporated edits and revisions from the other authors. Professor Michael Shanahan, Professor Steven Durning and Professor Lambert Schuwirth were responsible for the remaining percentage of work equally.

Whilst the exhaustive literature review described in chapter 4, with its resulting conceptual model was helpful in establishing values, individual characteristics and system factors that could support fair judgement in health professions assessment, this was purely a theoretical literature derived model.

The purpose of this next study was to begin to explore the other side of the coin, and to characterise the essential building blocks of fair human judgement from the perspectives of trainees and assessors across the continuum of experience. These varying perspectives and understandings from different levels of experience could then be compared and integrated into the literature model to create a comprehensive conceptual model of fair human judgement. The aim was to create a shared understanding and a framework which could guide contextual application of determining 'what is fair human judgement'.

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)

Furthermore when addressing wicked social problems, Moore suggested 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) When considering fairness, its components and their interactions, a shared language and understanding will help create a framework to implement it. The perspectives of the assessors and learners, as well as the theoretical literature model can contribute to this.

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.

Introduction

There is broad agreement that assessment in education should be fair. (Green, Johnson, Kim & Pope, 2007) Traditionally, evidence of construct validity and reliability

have been central to defend fairness of assessment. (Ten Cate & Regehr, 2019; Valentine & Schuwirth, 2019; van der Vleuten, Norman & De Graaff, 1991) However, both the notion of validity (Kane, 2006) 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. (Desy, Coderre, Davis, Cusano, & McLaughlin, 2019; Eva & Hodges, 2012; Govaerts & van der Vleuten, 2013; Hauer & Lucey, 2019; Hodges, 2013; Rotthoff, 2018; Schuwirth & van der Vleuten, 2006; Schuwirth & van der Vleuten, 2020; ten Cate & Regehr, 2019; Valentine, Durning, Shanahan & Schuwirth, 2021) This perceived misalignment has led to an increasingly resounding push within the literature to embrace human judgement in assessment and accept its subjective nature. (Bacon, Williams, Grealish & Jamieson 2015; Boursicot, 2020; Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014; Govaerts & van der Vleuten, 2013; Hauer & Lucey, 2019; Hodges, 2013; Jones, 1999; Muller, 2020; Rotthoff, 2018; Schuwirth & van der Vleuten, 2006; ten Cate & Regehr, 2019; van der Vleuten, Norman & De Graaff, 1991) 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. (Tierney, 2013) The elusiveness of this construct makes it difficult to simply define. (Desy, Coderre, Davis, Cusano, & McLaughlin, 2019) 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, (Tierney, 2013) and is potentially not useful, then perhaps changing tack and focusing on the building blocks of fairness may be more fruitful. Better understanding the foundations of fairness can help create a shared narrative to allow 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. (Valentine, Durning, Shanahan & Schuwirth, 2021) 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, 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?

Methods

As this study focused 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. (Braun & Clarke, 2012) Thematic analysis is flexible and able to be conducted in many different ways. (Braun & Clarke, 2012) 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. (Steiner, Atzmüller, & Su, 2016) This reduces the abstract nature of the concept, in our case of fairness, but still allows for simultaneous investigation of factors and their relationships. (Steiner, Atzmüller, & Su, 2016) Three vignettes were presented during the interview (figure 5). 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 was read to ensure familiarisation with the data, and reflective memoing was used to improve immersion and engagement with the 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 was analysed to elaborate the relationships between the codes and categories, with the aim to raise the analytical level from categorical to conceptual. These themes were then reviewed and refined. It was at this point that the data was then considered in light of the existing model. We refined the conceptual model based on our study findings, examining how this 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).

Vignette 1:

A group of ten medical students were informed they had failed one of their medical school subject exams and thus had failed the year. They were not given feedback after the exam or able to see their submitted exam. The students were then given an opportunity to spend one month in remediation, followed by another exam. If they passed this exam, they would be allowed to proceed with medical school, if they failed they repeated the year. They had a practice exam a week prior to their supplementary exam. Exactly half of the group (five students) passed and half of the group failed this practice examination. The week after, the five students who had failed the practice exam passed the supplementary exam and the five students who passed the practice exam failed the supplementary exam.

Vignette 2:

A very junior consultant [attending] is working at a tertiary hospital. There are no formal assessments for this doctor as they have completed their postgraduate training. There are several complaints about the junior consultant to the head of the unit by other senior doctors. These complaints include not supporting junior doctors, not informing staff when they are unable to attend clinics, not following instructions of senior surgeons during operations and operating on patients outside their skill set. The head of the unit is unwilling to provide the feedback to the doctor because

the head of the unit is worried the doctor will resign if they receive negative feedback.

Vignette 3:

A post graduate trainee fails an end of term assessment. All six consultants in the unit provided input to the supervising assessor who collated the feedback and recommendations to complete the assessment. The assessment was discussed with the trainee and feedback for improvement provided. The assessment was given at the end of the term, the day before the trainee moved to their next placement. To be allowed to sit for final fellowship assessments, trainees need to pass 8 of 10 end of term assessments, and thus can now only fail one further end of term assessment.

Figure 5: Vignettes used in semi-structured interviews

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 was 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 (Valentine, Durning, Shanahan & Schuwirth, 2021). 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 6).

Individual Characteristics

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

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 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 subthemes 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.”

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 practicing 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.”

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 (Lombardo & Eichinger, 2000)

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 can't 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 are 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.

System Factors

Fair judgement decisions have allowed for multiple opportunities

Fair judgments about progression in training programs 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.”

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 medical 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”

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 is uncertain.

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 don't 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 is 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”

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.”

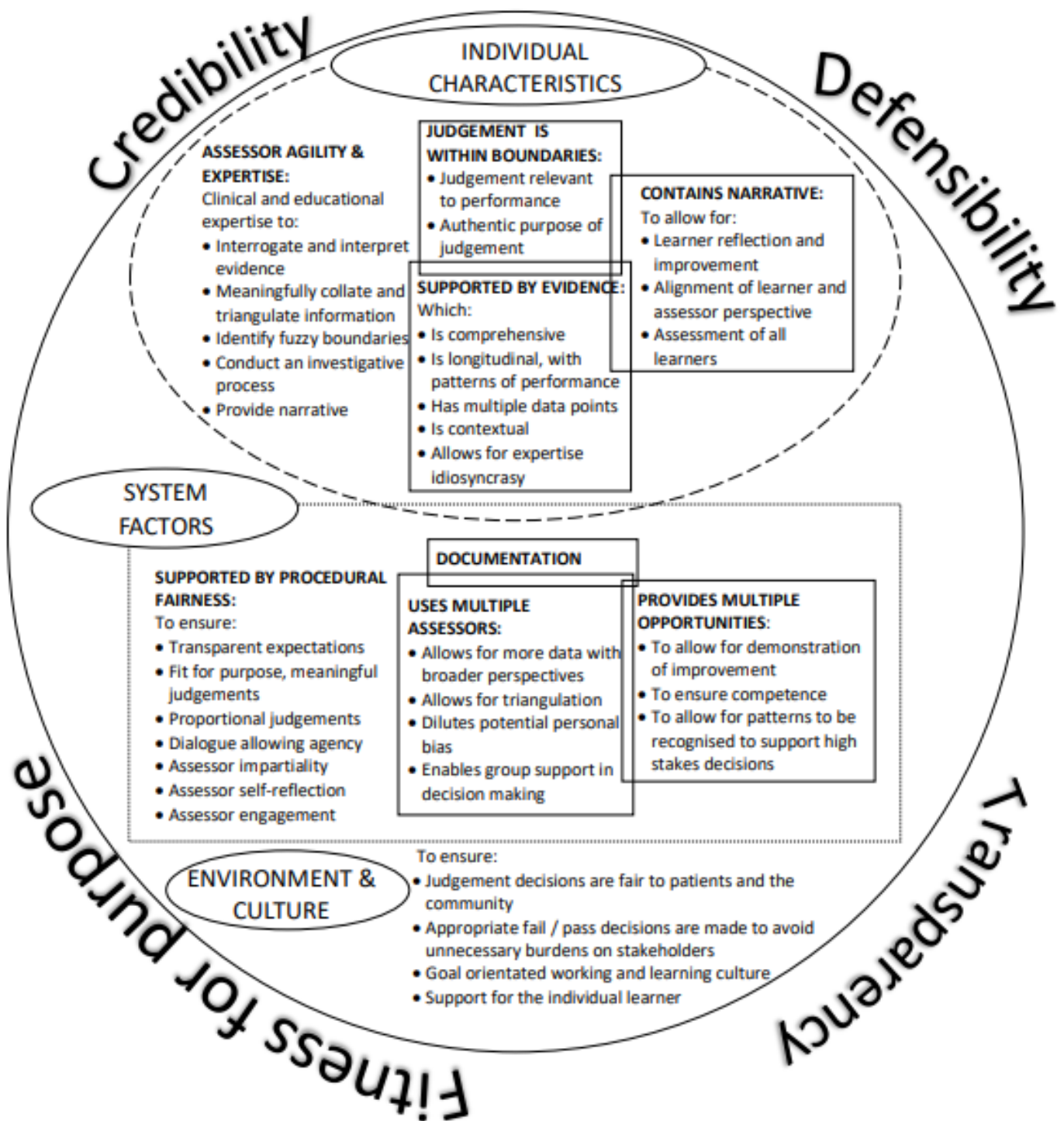


Figure 6: A conceptual model of the components of fair judgement in assessment

Discussion

The findings from this study support the conceptual model previously derived from the literature. (Valentine, Durning, Shanahan & Schuwirth, 2021) This study noted that fair judgement in assessment is multi-dimensional, complex and contextual. It highlighted there 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 was 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. (Durning, Pangaro, van der Vleuten & Schuwirth, 2010; Schuwirth & van der Vleuten, 2020) Where the system is complex, the solution likely needs to be as complex as the problem itself (Glouberman & Zimmerman, 2002) and the dynamic and unpredictable nature of complex systems logically precludes the effective use of reductionist values and methods. (Woodruff, 2021) 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. (Rosas, 2017) With prolonged observation, patterns and networks can still be revealed. (Cleland & Durning, 2015; Rosas, 2017) 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. (Van der Vleuten, Schuwirth, Driessen, Dijkstra, Tigelaar, Baartman & van Tartwijk, 2012) 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. (Dijkstra, Galbraith, Hodges, McAvoy, McCrorie, Southgate, van der Vleuten, Wass & Schuwirth, 2012; Dijkstra, van der Vleuten, & Schuwirth, 2010; Driessen, van der Vleuten, Schuwirth, van Tartwijk & Vermunt, 2005) 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 is heterogenous needs 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. (Cleland & Durning, 2015; Govaerts, van de Wiel, Schuwirth, van der Vleuten & Muijtjens, 2013; Govaerts, Schuwirth, van der Vleuten & Muijtjens, 2011; Marewski, Gaissmaier, & Gigerenzer, 2010) 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 synthesizing assessment data in a program of assessment. (Van der Vleuten, Schuwirth, Driessen, Dijkstra, Tigelaar, Baartman & van Tartwijk, 2012) 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. (Govaerts, Schuwirth, van der Vleuten & Muijtjens, 2011; Schuwirth & van der Vleuten, 2020) 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 (Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014).

The need for meaningful and actionable feedback and agreement between the assessor and learner is an important aspect in an assessment for learning philosophy. (Cantillon & Sargeant, 2008; Lockyer, Carraccio, Chan, Hart, Smee, Touchie, Holmboe, Frank & ICBME collaborators, 2017; Schuwirth & van der Vleuten, 2011b; Watling, 2014b) Lee argues that the use of specific narratives and contextual comments may be more informative for trainees than the judgement itself. (Lee, Brain,

& Martin, 2017) 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. (Turner & Harder, 2018; Young, Williamson, & Egan, 2016) This study also noted such an environment was essential for judgement decisions to be accepted as fair.

Our data suggests that embracing fair, subjective judgements can present challenges. For many institutions, this may be a cultural change (Bullock, Lai, Lockspeiser, O'Sullivan, Aronowitz, Dellmore, Fung, Knight & Hauer, 2019; McDonald, Lai, Lin, O'Sullivan & Hauer, 2021) and there may be faculty skill gaps and difficulty in making adaption to new and epistemological unfamiliar methods of assessment. (Lee, Brain, & Martin, 2017; McDonald, Lai, Lin, O'Sullivan & Hauer, 2021) 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. (Hauer, ten Cate, Boscardian, lobst, Holmboe, Chesluk, Baron & O'Sullivan, 2016) Transparent expectations, orientations, procedures and documentation are also common in most training programs. 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 (Ericsson, 2007; Eva, Bordage, Campbell, Galbraith, Ginsburg, Holmboe & Regher, 2016; Konopasek, Norcini, & Krupat, 2016; Watling, 2014a; Watling & Ginsburg, 2019). And finally, the learning environment has been gaining increasing attention in the medical education literature. (Young, Williamson & Egan, 2016) 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 programs, 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.

Conclusion

Woodruff noted that the challenge for medical education researchers is to not be distracted by 'solutions' but to look at problems more deeply. (Woodruff, 2021) 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.

CHAPTER SIX: FAIRNESS IN ASSESSMENT: IDENTIFYING A COMPLEX ADAPTIVE SYSTEM.

This chapter is a published article: Valentine N, Durning S, Shanahan EM, Schuwirth L. Fairness in assessment: identifying a complex adaptive system. *Perspect Med Educ.* 2023;12(1):315-26. This is an open access article.

This article was co-authored with Professor Steven Durning, Professor Michael Shanahan and Professor Lambert Schuwirth. My contribution for this article was approximately 80% of the research design, 80% of the data collection and analysis, and 85% of the writing and editing. Specifically for this article, I was involved in the research design of this project, undertaking tasks such as designing the focus group guide, applying for ethics approval to conduct the research and identification of potential participants. I recruited all of the participants, arranged suitable times for online focus groups and subsequently facilitated all of these focus groups. I undertook data analysis in collaboration with my fellow authors and wrote the original draft manuscript. I then incorporated the suggestions, edits and revisions from the other authors. Professor Steven Durning, Professor Michael Shanahan and Professor Lambert Schuwirth were responsible for the remaining percentage of work equally.

In this research journey exploring what constitutes fair judgement in assessment, thus far I have brought together the inferences and underpinnings from the literature to create a theoretically constructed conceptual model and undertaken a study to explore the understanding of fair human judgement from the perspectives of learners and assessors across a continuum of experiences. This first study added to the literature review, confirming the conceptual model, redefined relationships and added rich detail about the components of fairness. The resulting model assists in the development of a narrative which can be used to 'negotiate' fairness between stakeholders.

A subsequent study harnessing the perspective, skills and practical wisdom of health professions education assessment leaders and program designers was then conducted. Assessment leaders and program designers work with learners, they have a lived reality of appeals and complaints, and understand the challenges of implementing human judgement in assessment and are likely to have ideas of how to approach the issue of fair judgement. This study aimed to leverage their skills and expertise combined with their understanding of the pragmatic realities to facilitate the practical application of this conceptual model. The study allows them to be co-designers, to provide collaborative, tangible outcomes. In contrast to the previous study, focus groups were chosen to allow individual respondents to react to and build on other group members' responses, allowing for dynamic interactions. (Stalmeijer, N. McNaughton, & Van Mook, 2014)

During data analysis of this second study, we realised the same four components of fairness were present at all levels of granularity and in all contexts. We concluded we

had identified a fractal. A fractal is a manifestation of an underlying complex adaptative system (Tsoukas & Dooley, 2011) and so we began data analysis again, this time using a lens of complexity. We also then returned to the previous study's data and noted the same result. This suggested that fair judgement should be considered a complex adaptive system. This chapter is the second study, as published in Perspectives on Medical Education.

Abstract

Introduction: Assessment design in health professions education is continuously evolving. There is an increasing desire to better embrace human judgement in assessment. Thus, it is essential to understand what makes this judgement fair. This study builds upon existing literature by studying how assessment leaders conceptualise the characteristics of fair judgement.

Methods: Sixteen assessment leaders from 15 medical schools in Australia and New Zealand participated in online focus groups. Data collection and analysis occurred concurrently and iteratively. We used the constant comparison method to identify themes and build on an existing conceptual model of fair judgement in assessment.

Results: Fairness is a multi-dimensional construct with components at environment, system and individual levels. Components influencing fairness include articulated and agreed learning outcomes relating to the needs of society, a culture which allows for learner support, stakeholder agency and learning (environmental level), collection, interpretation and combination of evidence, procedural strategies (system level) and appropriate individual assessments and assessor expertise and agility (individual level).

Discussion: We observed that within the data at fractal, that is an infinite pattern repeating at different scales, could be seen suggesting fair judgement should be considered a complex adaptive system. Within complex adaptive systems, it is primarily the interaction between the entities which influences the outcome it produces, not simply the components themselves. Viewing fairness in assessment through a lens of complexity rather than as a linear, causal model has significant implications for how we design assessment programs and seek to utilise human judgement in assessment.

Introduction

Assessment design in health professions education is continuously evolving in response to new insights, ideas and research findings. Historically, assessment has been seen mainly as a measurement problem, with reliability and validity being key components of assessment. (Schuwirth & van der Vleuten, 2020) Over time, however,

evolving views about learning and rater cognition, shifting social ideals and understandings of the limitations of high stakes tests has challenged the idea that objectivity is the gold-standard of assessment. (Bacon, Williams, Grealish & Jamieson 2015; Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014; Gipps & Stobart, 2009; Govaerts & van der Vleuten, 2013; Hodges, 2013; Jones, 1999; T. Rotthoff, 2018; Schuwirth & van der Vleuten, 2006; ten Cate & Regehr, 2019)

As a result, there has been an increasing push to better utilise the role of human judgement in assessment. (Bacon, Williams, Grealish & Jamieson 2015; Dijkstra, Galbraith, Hodges, McAvoy, McCrorie, Southgate, van der Vleuten, Wass & Schuwirth, 2012; Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014; Govaerts & van der Vleuten, 2013; Hodges, 2013; Jones, 1999; T. Rotthoff, 2018; Schuwirth & van der Vleuten, 2006; ten Cate & Regehr, 2019) This was initially was under the guise of 'reliable subjectivity', utilising assessor training and large samples to ensure sufficient reliability of the assessment. (Schuwirth, Southgate, Page, Paget, Lescop, Lew, Wade & Baron-Maldonado, 2002) But more recently it has been acknowledged that rater variance may provide meaningful idiosyncrasy and should be embraced rather than controlled. (Boursicot, Kemp, Wilkinson, Finyartini, Canning, Cilliers & Fuller, 2021; Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014; Gingerich, Regehr & Eva, 2011; Govaerts & van der Vleuten, 2013; Govaerts, Schuwirth, van der Vleuten & Muijtjens, 2011)

However, assessment still needs to be fair. Subjective human judgements do not add meaningful idiosyncrasy if they are unfair to either learners or society. Nor will fair

judgements add meaning if they are part of unfair assessment systems. So, addressing what makes human judgement fair in health professions assessment is essential in legitimising subjective judgements in our assessment programs.

Fairness is often implied in assessment programs, but is not usually explicitly articulated as there is no simple definition for this complex construct (Valentine, Durning, Shanahan & Schuwirth, 2021), and fairness is dependent on cultural beliefs, social contexts and practices. (Gipps & Stobart, 2009) Despite the lack of explicit definition, the underpinnings and constituents of fairness are implied in the medical education and broader education literature. A literature review brought these inferences and underpinnings together to create a theoretically constructed conceptual model. (Valentine, Durning, Shanahan & Schuwirth, 2021) This literature review noted that the multifaceted construct of fair human judgement could be conceptualised through values, which are upheld at an individual and system level. (Valentine, Durning, Shanahan & Schuwirth, 2021) A further study exploring the understanding of residents' and supervisors' perspectives of fairness built on the theory-derived conceptual model, demonstrating that the components of fairness could be explicitly articulated whilst still embracing the complexity and contextual nature of health professions assessment. (Valentine, Shanahan, Durning, & Schuwirth, 2021) This study noted that at an individual level, contextual, longitudinally-collected evidence, which is supported by narrative, and falls within ill-defined boundaries is essential for fair judgement decisions. Assessor agility and expertise are needed to interpret and interrogate this evidence, help identify fuzzy boundaries and provide narrative feedback to ensure learners can improve. At a system level, factors such as multiple opportunities for learners to demonstrate competence and improvement, multiple

assessors to allow for different perspectives to be collected and triangulated, and documentation are all needed for fair judgement. These system features are supported through the concept of procedural fairness which provides transparent expectations, allows for fit-for-purpose, individualised, proportional judgements, and supports dialogue and engagement with the learner. Finally, the environment in which the assessment decisions are made needs to be considered for fair judgments.

(Valentine, Shanahan, Durning & Schuwirth, 2021) The resulting model can assist in developing narratives to 'negotiate' fairness between stakeholders.

Whilst this was helpful, given the fundamental nature of fairness in assessment, it is important to understand stakeholder perspectives, such as expert assessment leaders. Their insights could further help translate this concept of fairness and bring change to educational practice. In this study we, therefore, addressed the following research aims:

1. To understand what the characteristics of fair judgement are from assessment leaders' perspectives.
2. To compare and contrast these understandings with our previously reported theoretically constructed conceptual model (Valentine, Durning, Shanahan & Schuwirth, 2021; Valentine, Shanahan, Durning & Schwirth, 2021)
3. To understand how these understandings and theoretical aspects translate to practice and suggest design principles to assist in the practical application of a theory derived conceptual model.

Methods

Reflexivity

We took a subjectivist, inductive approach to this research, assuming that fairness as a reality is socially constructed, and that individuals and social groups share interpretations and understandings of the reality of fairness. (Varpio, Paradis & Uijtdehaage, 2020) The components of fair judgement in assessment are constructed by individuals and institutions, and change over time and across cultures. Therefore, we also took a constructivist stance in that the meaning of fair judgement is constructed by stakeholders, rather than the idea that there is a simple, universal true definition of fairness. Collecting data from multiple perspectives will therefore assist in gaining a richer and more nuanced understanding of this phenomenon. (Varpio, Paradis & Uijtdehaage, 2020)

Reflexivity was employed throughout the research process and is described through the dimensions of 'personal', 'interpersonal', 'methodological' and 'contextual'. (Olmos-Vega, Stalmeijer, Varpio, & Kahlke, 2022) The research team consists of experienced HPE researchers and clinicians, all familiar with the study content, having undertaken previous studies on fairness in assessment. The research team members work in diverse contexts, representing a range of specialties and HPE research environments across different continents. All team members consider themselves to be social constructivists. LS, NV & MS have been previously involved in medical education in Australia. The diversity of experiences of the research team was leveraged allowing for a range of perspectives enabling rich team discussions during data interpretation. (Varpio, Ajjawi, Monrouxe, O'Brien, & Rees, 2017) NV's interest in fairness initially

stemmed from her role in medical education and as a senior clinician. However, her perspective has shifted slightly as she has now recommenced as a trainee in a different medical specialty. NV approaches fairness from the dual perspective of both a PhD candidate and a clinician. SJD is interested in fairness as a director of academic programs spanning the continuum. SJD believes that nonlinearity and complexity often shape our interactions. SJD approached the topic and findings as both a PhD scholar and practicing physician. EMS is a full-time practicing clinician with a lifetime career committed to medical education. His interest in fairness has developed through his work as a program director and educator of students and physician trainees. LS has an interest in fairness as a researcher in assessment and with an interest of understanding assessment in a post-psychometric era. He believes that nonlinearity and complexity often shape our interactions. LS approached the topic and findings as both a research scholar and a professor of medical education. He is the first in his extended family to attend college and therefore, fairness is an important value for him.

Participants

Eligible participants were assessment leaders from the 23 medical schools in Australia and New Zealand. All 29 members of the assessment leads of the Medical Deans of Australia and New Zealand were invited to participate in 90-minute focus group conducted via Zoom. We chose focus groups to allow individuals to build on other group members' responses, allowing for dynamic interactions. (Stalmeijer, McNaughton, & Van Mook, 2014) As an aim of the study was to understand how

previously identified theoretical aspects translated to practice, participants were asked to design an assessment program for a fictional medical school utilising subjective judgements while trying to make these fair to both learners and society. Participants were instructed to employ blue sky thinking; we posed no barriers to time, money or supervisor engagement as this was not the aim of the study. A collaborative white board, Miro, was used to facilitate discussions. We provided no incentive to participate. Ethics approval was obtained (Flinders University: 4297).

Analysis

Data collection was undertaken from July to September 2021. NV conducted the focus groups and had limited familiarity with the participants. Focus group were recorded and transcribed verbatim without identifying data. Focus groups notes and the shared white board were included in data analysis. NVivo, a qualitative analysis software, was used to assist with data management.

Collection, analysis and coding of the data occurred simultaneously, each informing the other. NV initially read each transcript line-by-line to allow for familiarisation with the data. The analysis process involved discussions between researchers and comparison of different codes between and within transcripts to clarify, confirm and categorise codes. After focus groups and initial data analysis was complete, we reviewed our data in light of the previous conceptual model, examining how these

findings elaborated or contradicted the previous findings. (Valentine, Shanahan, Durning & Schuwirth, 2021)

Results

Of the 29 invited assessment leaders, 19 volunteered to participate but three withdrew prior to the focus groups. The five focus groups were attended by 12 females and four males from 15 medical schools. Fourteen medical schools were located across all six states of Australia and one was located in New Zealand. Two medical schools were located in large regional centres, 13 were in major cities. All participants had experience in assessment design and delivery at their respective medical schools. Participants' academic titles at the time of the focus groups are listed in Table 2.

Academic Assessment Lead	Academic Lead Assessment	Acting Dean
Associate dean	Associate Dean, Learning and Teaching	Associate professor (3 participants)
Chief Examiner and Head of Assessment	Director of Assessment	Director Medical School
Discipline Leader	Doctor of Medicine Program Director	Faculty Dean
Head of Assessment		

Table 2: Academic titles of focus group participants

Fair judgements are more than just the individual judgements themselves.

Judgements are not considered fair unless the environment, culture, and system in which they are made is also considered fair as demonstrated by the different sections in figure 7. So, in evaluating fairness conceptualisations and design decisions, there are many aspects of the assessment system which need to be considered in conjunction with each other.

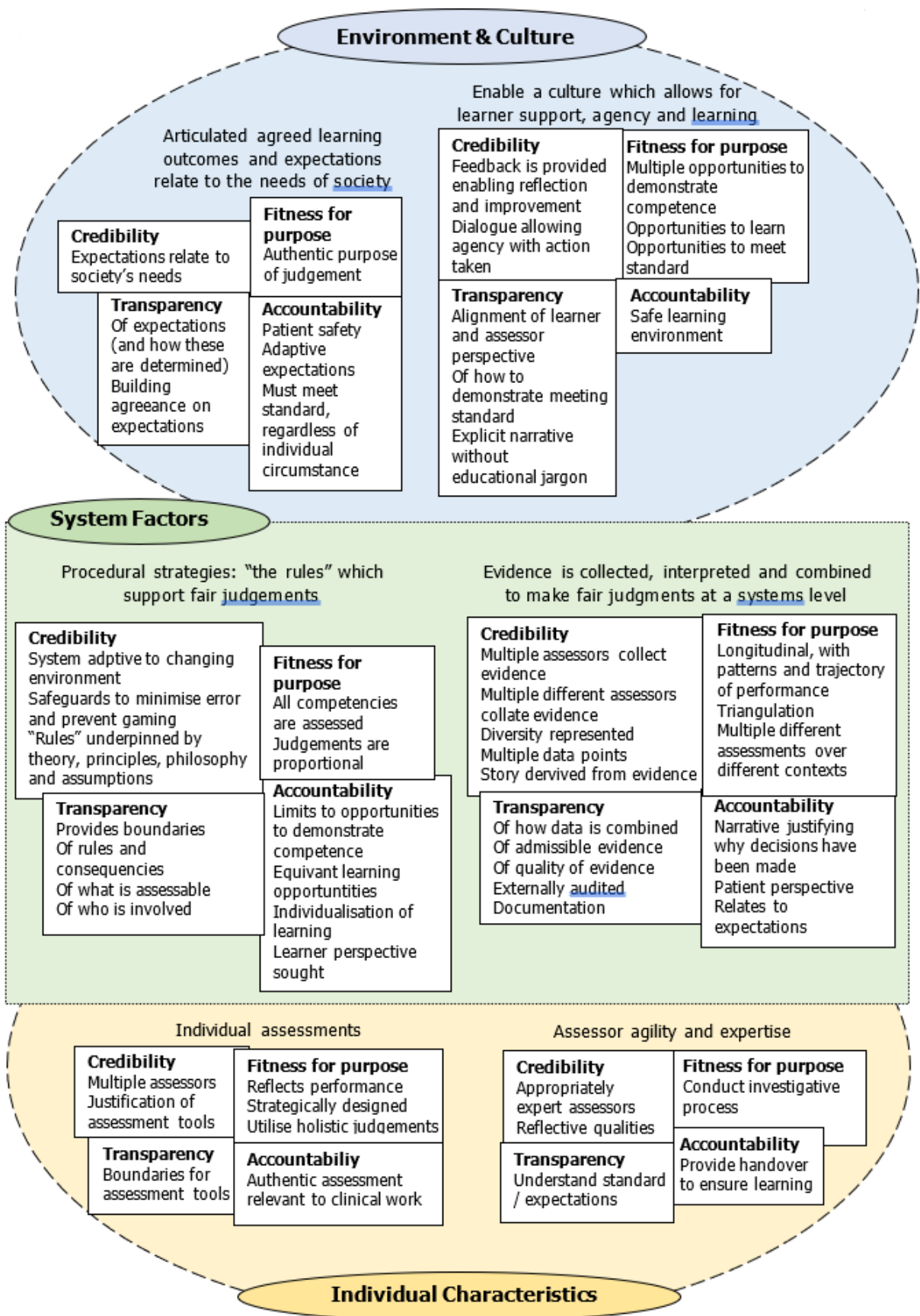


Figure 7: The components of fair judgement

Environment and Culture

Individual judgement decisions interact with their environment and culture; and so need to occur in environments and cultures which are fair to both society and learners.

Articulated agreed learning outcomes and expectations relate to the needs of society

One way of linking individual judgements with the environment is through transparency of expectations, typically through establishing agreement with relevant stakeholders. That way, fairness is ensured through allowing assessors and society opportunity to provide relevant perspective on competence and its practical usefulness.

“That common standard, even if it’s subjectively deployed, could be understood in some written words to be where everyone was aiming for.” (Participant 1)

Patient safety is a central concept in society’s needs and thus essential for judgements to be fair to patients. This includes ensuring that learners meet agreed expectations at certain points in time, regardless of their individual circumstances.

“if they’ve got all these extenuating circumstances et cetera but the other side of that is the duty to assure to the public that the student is competent and safe.” (Participant 9)

But these expectations are not static, and so fairness also includes adapting learning to meet society’s ongoing evolving needs.

Enable a culture which allows for learner support, agency, and learning

Turning to the learners themselves, fair judgements need to be accompanied by meaningful feedback that enables reflection and learning. If assessments are just summative hurdles to clear, they unfairly limit opportunities for students to learn and grow in their journey to becoming health care professionals.

Fairness also requires that the assessment and feedback are a dialogue and enable the learner to share their perspective on the judgement and take agency over their own learning, with action taken following this.

“you have to show that you’ve heard the student’s story.” (Participant 6)

Safe, fit-for-purpose learning and assessment environments are essential for fairness to the learner themselves, and are in the best interest of patient care as they allow

learners to continually improve. A learner who feels safe enough to recognise their weaknesses and to focus on continual improvement is more likely to be educated as a lifelong learner, even after graduation.

“Competency isn’t a do it once pass fail. It’s, didn’t do well so have another go. Didn’t do well, have another go, more feedback, have another go, more feedback.” (Participant 8)

Transparency within the learning environment allows for reflection and learning and thus is essential for fairness. The narrative used in both expectations and feedback therefore needs to be clear, explicit and without educational jargon to allow for this learning.

“Our university in its infinite wisdom has stopped us using those sort of descriptors and they’ve told us we have to give feedback to students on their university scale ... it is causing all sorts of issues” (Participant 5)

Institutions must also be transparent about how learners can demonstrate that they are meeting the expected standard and what they need to do if not. In addition, transparency ensuring the learner is aware of how they are performing against expectations is required as it allows for alignment of learner and assessor

perspectives. A surprise judgement is considered unfair as it denies the learner the opportunity to improve.

“it’s the no unexpected news at the end, because they’ve been forewarned as they’ve gone through.” (Participant 1)

System Factors

Procedural strategies: ‘the rules’ which support fair judgements

Procedural strategies provide boundaries at a system level. These provide clarity for assessors and protection for the learners thus allowing for development of a partnership. Procedural strategies includes ensuring transparency of rules and consequences, of what is assessable, who is involved in the assessment process and provision of safeguards to minimise sources of error and prevent gaming.

“So, what’s fair game for not being assessed, like asking a stupid question, let’s say. We’re not going to judge on that.” (Participant 6)

Procedural strategies can also facilitate fairness through ensuring appropriate proportionally is assigned to judgements. This proportionality might be aligned with the

stakes of the assessment, the richness of the information from the assessment or the number of opportunities to pass an assessment.

To meet the agreed learning outcomes and society's expectations of a practicing professional, multiple competencies are needed. Procedural strategies were suggested to ensure judgements could only be fair at a system level if all competences were likely to be assessed.

"...looking at the domains across different assessments" (Participant 2)

It was acknowledged that learners develop competence at different paces, and if the aim is to develop competent health practitioners, then it is fair to allow individual variation.

"We're allowing some to go slower and some to go faster" (Participant 7)

However, this allowance for individualisation of learners needs to be balanced with the need to ensure fairness to society through placing limits on the opportunities provided to demonstrated competence. Furthermore, failure to fail is also unfair to learners as this may deny them the opportunity to learn and undertake remediation.

“...is not fair is that the poorer students are given many more opportunities to scrape through the course rather than [fail]” (Participant 1)

Due to the unpredictable nature of workplace-based learning and assessment, not all learners will have exactly the same experiences, but they are all entitled to the same quality of learning experiences and assessment. So, equity may be more important than standardisation.

Fairness can be supported by ensuring the “rules” are underpinned by theory, principles, philosophy and assumptions, providing a framework for the fuzzy boundaries of the procedural strategies, and a guide for future scenarios. As the environment and situations change, the system will need to be reviewed, evaluated and adapted to ensure it remains fit-for-purpose and fair to both learner and society.

“It probably is principles-based. It's probably influenced by theory. It probably is conceptually based as well. It's probably strategically designed. It's probably purposeful.” (Participant 3)

Given these fuzzy boundaries, student perspective and trust in the system are essential. Trust is important in all aspects of the system, from expectations, to process, to decision making. There are several ways this trust can be built, but without

it, judgements are not considered fair, as learners do not believe their interests are considered.

“I think that was one of the best things that we did [having students on competency committees] for the student body because no matter how much work we'd done and how much communication and consultation, the thing that's convinced them that it genuinely was low stakes was their own peers going out, going oh no, it's true, when it goes to the panel they really look at everything together and holistically” (Participant 4)

Evidence is collected, interpreted and combined to make fair judgments at a systems level

At a systems level, collection, interpretation, and collation of evidence is required. To ensure fairness, a richness of data is needed to build a picture about a learner's progress which is reassuringly comprehensive enough to make high-stakes decisions. This includes longitudinal data from multiple different assessments, over many different contexts which allows for triangulation of data and identification of patterns of performance.

“If you're gathering narrative from a wide range of people, you'll often start to see patterns of behaviour or a consensus appearing. That can make it more fair” (Participant 7)

Multiple longitudinal data points also allows for a trajectory to be considered reducing uncertainty about a student's learning journey and adding to the richness of the picture.

"...not only have you not got to the point we want you to get to, but you're showing no inclining of making any progress either. Different story from, you haven't quite got there but boy we've been really encouraged by how much progress you've made over the last three months and we think if you had another three months you probably will get there." (Participant 6)

Multiple different assessors collecting evidence adds to the credibility of the picture of evidence being collated about the learner. Diversity of opinions also adds to the richness of the picture rather than creating unreliability.

"if they know ... there's going to be multiple judgements made by multiple clinicians, that multiple perspectives ... then they're much more confident in the fairness of the assessment." (Participant 9)

Having a different group of assessors meaningfully collate and weigh up multiple pieces of evidence at a system level adds a safety net for learners and 'on-the-ground'

supervisors alike. It ensures a second 'check' for learners and allows support for supervisors in decisions making, particularly for difficult decision making.

"We always find reassuring to our supervisors that actually it's the [university name] Board of Examiners who makes the decision. Their job is just to tell us what they saw and be as frank as they can be about the student's performance ... but we'll take the decision-making on our shoulders, not theirs and that does help" (Participant 7)

As evidence is collated, a story is created. This story provides meaningfulness and credibility to the judgement which makes it fair. It connects the evidence with previous knowledge and experience about the learner and provides justification for the judgement, both of which help make the judgement fair. Furthermore, from the story, areas for improvement can be identified which is also essential for fair judgement.

"So I wonder if that narrative and the pattern equals a story. ... Because if you just got a six out of 10 and a seven out of 10 and a B minus, that's not telling you a story. But a narrative – and a narrative doesn't tell you a decision, but it contributes to a story and then once you've got the story, you can make a decision." (Participant 6)

Transparency at a systems level may involve considering what evidence could be considered in fair judgements and how the quality of the evidence would be determined. Additionally, understanding how data is combined was also seen as important as it helps define assessment boundaries for the learner.

“I think you need to be clear on how you’re adding things up too. If you’re going to say, they have lots and lots of direct observations, but actually we’re going to look at all of them at the end of the rotation and make some sort of narrative judgement based on all the feedback provided in those and you need to make that very clear for students.” (Participant 8)

External auditing of judgements can ensure accountability to learners and society and thus fairness. This requires documentation of how and why the judgement is made including the ‘story of evidence’ behind the judgement decision. This may also involve discussing the result with the learner.

“If you’ve got independent verification of the judgment, then that makes it a fair assessment in the student’s eyes type thing.” (Participant 1)

For judgements decisions to be fair, they needed to have an authentic purpose, that is to meet the needs of society, and relate back to the agreed outcomes for the learner.

This provides accountability to society but also makes the judgement credible, transparent and fit-for-purpose for the learner.

“...there is a tangible outcome at the end of this which is basically work readiness” (Participant 15)

As judgement decisions relate to the needs of society, ensuring the patients' perspective is represented is important for accountability. This may be through patient representation on committees or allowing patients opportunity to provide feedback.

“...it's also important to give – allow some sort of patient voice in assessment as well” (Participant 15)

Individual factors

Individual assessments

To be accountable to learners, judgements are only fair if they add to the rich picture of a learners' performance, progress, and possibly prognosis. Assessors pushing their own agenda or making judgement decisions which are irrelevant to the outcome of assisting learners to become competent healthcare providers were seen as unfair.

“the lack of relevance.. did just go off on the examiner's flight of fantasy”

(Participant 4)

This means fair judgements must only consider factors relevant to the outcome of assisting learners to become competent healthcare professionals. Any other factor, such as reputation is outside of the boundaries of fair judgement, does not add to the meaningfulness of the judgement and so therefore is unfair.

“..is this assessment an accurate ... reflection of the learning outcomes or are there issues causing irrelevant ease or irrelevant difficulty to subsets of the group that we're assessing.” (Participant 3)

To support this, individual assessments should also be transparent with boundaries for each assessment tool to ensure fair judgements.

“...asking the right questions of the right people in the right way. Fit-for-purpose tools, these are all things that help guide and direct and support both your trainee or learner and your assessor, so they don't go off on tangents and they know what it's about.” (Participant 3)

In addition to multiple assessors being used at a system level to make high stakes decisions, multiple assessors can also be used at an individual assessment level. This is not from an inter-rater reliability point of view but rather to ensure that different perspectives are combined and the whole picture is seen.

“...some types of assessments actually require that type of triangulation like multisource feedback or sometimes supervised supports where you actually have to draw on the whole team” (Participant 10)

There also needs to be justification to ensure the correct assessment tool has been selected for the right situation. There is no one-size-fits-all medical school program, and credibility of the tools needs to be demonstrated to ensure the resultant judgement is also fair and fit-for-purpose. If the combination of the collected evidence is not relevant or does not add to the whole picture it leads to the perception of unfairness as it denies the learner the opportunity to be genuinely judged and provided with feedback. It also is unfair to society as the learner is denied the opportunity for improvement.

“...you need some sort of credibility with the tools. So you probably need to show that you have got the right tools out of the toolbox” (Participant 6)

Assessor agility and expertise

Assessors need both agility and expertise to make fair judgements. Agility is required because assessment judgements typically involve interactive processes between assessors and learners. Assessors also need to understand the outcomes and the standard to which they are assessing. Whilst diversity of perspectives adds to the richness and completeness of the picture of the learner's progress, prejudiced perspectives due to sociocultural factors such as racism creates unfairness. Similarly, irrelevant perspectives which do not relate to the task of being a health professional also creates unfairness.

“qualities of the decision maker. What I meant by the ability to see multiple perspectives is the awareness of one's own biases and positioning” (Participant 11)

Assessors may be required to search for extra information to make fair judgements. This is needed for saturation of information and to ensure the complete picture of a learner's progress is known.

“I think sometimes you actually have to go back and get some additional information about some particular aspects of individual's capacity” (Participant 14)

As previously mentioned, fair judgments necessitate meaningful feedback to be given to enable learning. This requires clinical and educational expertise and agility of assessors to ensure this is credible and fit-for-purpose. In addition, assessors can ensure their judgements demonstrate accountability to learners through providing a 'handover' to other assessors help facilitate future learning.

“Because if the purpose of assessment is to help medical students be good future doctors, then we would be passing on information about their strengths, and particularly about their weaknesses as they're progressing through the course with a view of helping them, and our future patients, so that they get better doctors.” (Participant 3)

Discussion

This study has highlighted there is no simple definition or formula for fair judgements, but rather fair judgement is multi-dimensional and context dependent. It supports the previous contextual model (Valentine, Shanahan, Durning & Schuwirth, 2021) demonstrating there are multiple layers to fair judgment; with significant overlap between these layers. The components of fairness noted in the previous study with residents and supervisors (Valentine, Shanahan, Durning & Schuwirth, 2021) were again found in this study, however, with different emphasises as this group has a different perspective, and work in different contexts.

However, perhaps more significantly, during data analysis we realised the same four components of fairness were occurring at all levels of granularity and in all contexts. We concluded we had identified a fractal. A fractal is a shape or concept, which remains the same at different scales. (Lipsitz & Goldberger, 1992) An infinite number of repeating patterns at different sizes are combined together to give a fractal its shape. Their defining feature, is their 'self-similarity', that is the same shape is found regardless of whether you zoom in or out. (Holbrook, 2003; Lipsitz & Goldberger, 1992) Our fractal pattern or 'shape' was made up of four components: credibility, fitness for purpose, transparency, and accountability.

Whilst our data has been presented as categories and themes, the fractals can still be seen. During our data analysis, we noted that when participants spoke about what is required for fair judgements, underlying all they said were these four elements. This occurred whether they were speaking about judgements at a 'corridor consult' level, at a workplace-based assessment level all the way through a competency committee meeting level. There were different emphases on these four components in different contexts and at different levels, but all four were always present. When we compared this with our previous research, these components were also noted. (Valentine, Durning, Shanahan & Schwirth 2021; Valentine, Shanahan, Durning & Schwirth, 2021)

Features of Complex Adaptive Systems	An example of how this relates to fair judgement
<p><u>COMPLEX</u></p> <p>CAS consist of individual agents (Plsek & Greenhalgh, 2001) who make independent choices about their actions. (Fraser & Greenhalgh, 2001) Each individual agent reacts to what the other agents are doing. (Bowe & Armstrong, 2017; Reed, Howe, Doyle & Bell, 2018) This interaction between the agents directs the CAS and influences the outcomes it produces. (Durning, Artino, Pangaro, van der Vleuten, & Schuwirth, 2010; Reed, Howe, Doyle & Bell, 2018) The principle of connectivity is that a system's behaviour relies less on the nature of the individual agents than on the quantity and quality of connections between them. Therefore learning how things are interconnected is often more</p>	<p>Medical schools need to determine if students meet the standard expected to graduate.</p> <p>Judgement decisions are made by a diverse group of individuals or committees considering multiple different assessments and evidence.</p> <p>Assessors are independent experts allowing them to make independent judgement decisions depending on their interaction with the data and other individuals. It is not possible to create specific rules for how judgement decisions are made. Each judgement decisions will involve different data, with different circumstances and will be perceived in different ways. Furthermore, the determination of the outcome is more than simply including more measurement points in the model. Although further data may improve judgement decisions, the</p>

useful than learning about the pieces.
(Fraser & Greenhalgh, 2001)

Despite the unpredictable and adapting nature of complex systems, principles and patterns arise. (Reed, Howe, Doyle & Bell, 2018) Understanding these patterns is fundamental to understanding how the system works (Mennin, 2010) as they guide behaviours within it. (Reed, Howe, Doyle & Bell, 2018)

interactions between these factors also needs to be considered.

Expert assessors recognise that a **multitude of factors** should be considered in assessment and can perceive **information from multiple interactions** simultaneously process this information to **identify patterns**. Making meaning of these **relationships** is encouraged.

ADAPTIVE

The efficacy and effectiveness of CAS is mainly due to the adaptability of the system. Agents adapt to past experience, (Fraser & Greenhalgh, 2001; van Beurden, Kia, Zask, Dietrich & Rose, 2013) internal and external influences. However this also leads to unpredictability, (Greenhalgh & Papoutsis, 2018; Mennin, 2010; Reed, Howe, Doyle & Bell, 2018) and resistance to centralised control. (Kurtz

The assessors and the system of assessment are **adaptive**. Previous experience, new information, a different assessment method or a change in expectations causes the agents and thus the system to change. **Adaption** is often enhanced in crisis, this may be seen in the case of a struggling trainee, making decisions with incomplete data or changing environments such as pandemics.

<p>& Snowden, 2003) Control is dispersed; the result of a huge number of decisions made by individual agents. (Van Beurden, Kia, Zask, Dietrich & Rose, 2013)</p> <p>Work arounds and muddling through are central to CAS. (Fraser & Greenhalgh, 2001; Greenhalgh & Papoutsi, 2018) Tensions and paradox do not necessarily need to be resolved. (Plsek & Greenhalgh, 2001) Order, innovation and progress emerge naturally from the system, they do not need to be imposed from within or from outside. (Greenhalgh & Papoutsi, 2018; Norman, 2011) Seemingly obvious interventions can have minimal impact on system behaviour, whereas small changes can have large unintended consequences. (Bowe & Armstrong, 2017; Reed, Howe, Doyle & Bell, 2018; van Beurden, Kia, Zask, Dietrich & Rose, 2013)</p>	<p>Agents self-organise to consciously improve the interactions between patients, learners, the environment and the university to ensure judgements are fair. The desire is often to apply more rules, however these rules alone are less likely to influence judgement decisions.</p> <p>If a judgement is not obvious, the system is still able to move forward and judgement decisions made. Effective judgements can emerge, even from minimum initial data.</p> <p>There will always be tensions when making judgement decisions. For example between what is fair for the patient and what is fair for the individual student, or balancing learning with assessment.</p>
---	---

<u>SYSTEMS</u>	Within assessment, boundaries, ground
<p>Complexity thinking maintains that systems can be aided by a minimal structure, such as fuzzy, ill-defined boundaries. (Fraser & Greenhalgh, 2001) These boundaries act as constraints in that they provide a stable structure within which change can occur. (Greenhalgh & Papoutsis, 2018; Mennin, 2010)</p>	<p>rules and processes, can provide assessors with security and confidence to make judgement decisions.</p> <p>To ensure fair judgement, sufficient organisational structure is needed to keep stakeholders focused on the task, without limiting flexibility, initiative and commitment to overall improvement.</p>
<p>Individual agents and CAS are embedded within wider CAS. Therefore, we cannot fully understand the individual agents or systems without reference to the others. (Kurtz & Snowden, 2003; Plsek & Greenhalgh, 2001)</p>	<p>Humans are not limited to one identity, but are also members of clinical workplaces, families and social groups which are embedded within cultural environments and wider society. These external memberships influence how agents behave and the perspectives they bring to judgement decisions.</p>

Table 3: Fair judgement demonstrated as a complex adaptive system

A fractal is a manifestation of an underlying complex adaptive system (CAS). (Tsoukas & Dooley, 2011) CAS are systems with collections of individual agents which are interconnected so that each individual agent reacts to and influences what the

other agents are doing. (Mennin, 2010; Plsek & Greenhalgh, 2001) It is these interactions that influence the system and the emergent phenomena it produces. (Durning, Artino, Pangaro, van der Vleuten, & Schuwirth, 2010; Reed, Howe, Doyle & Bell, 2018) Reed illustrates it as, 'life is more than molecules and atoms – it is the complex patterns of organisation that emerge between them'. (Reed, Howe, Doyle & Bell 2018) How fair judgement can be perceived as a CAS is demonstrated in Table 3.

These findings provide a new perspective of how fair judgement can be conceptualised in assessment. Whilst there has been an increasing push over recent years to view assessment as a system (Boursicot, Kemp, Wilkinson, Finyartini, Canning, Cilliers & Fuller, 2021; Schuwirth & van der Vleuten, 2020), recommendations can theoretically still be viewed from a linear, causal perspective with less consideration given to the interactions within the system, and how the system responds to these many interactions. (Boursicot, Kemp, Wilkinson, Finyartini, Canning, Cilliers & Fuller, 2021)

Implications of viewing fairness through a complexity lens

We must acknowledge though that the use of complexity science to comprehend the complex nature of medical education is not new and is indeed encouraged. (Bowe & Armstrong, 2017; Fraser & Greenhalgh, 2001; Greenhalgh & Papoutsi, 2018; Mennin, 2010) Switching focus, and taking the view of assessment as a system one step further could have significant implications. The first implication is that it is people who

create the components of the fractal and their interactions, and thus it is people who create fairness. Fairness emerges from how people use and combine credibility, accountability, fitness for purpose and transparency within our assessment systems. These interactions are mediated by strategies or effectivities. Expert and agile assessors, armed with situational and contextual awareness as well as a broad repertoire of strategies navigate these components and the interactions. For example, in making a fair judgement for an end-of-term assessment, an assessor may ask other staff about a learner, obtaining multiple pieces evidence collected over time. The assessor will interact with other stakeholders, the evidence, the context and the 'pattern' of fair judgement. They will potentially ask other assessors to help self-calibration, and will discuss with the learner, obtaining their perspective on the assessment. Based on these interactions they will combine information in a credible way, which is accountable, transparent and fit-for-purpose to create judgement. After giving the learner the judgement, they may then adapt, perhaps by providing more targeted feedback to help the learner improve by identifying where they are not meeting expectations.

Therefore, based on our findings, fairness cannot be reduced to a linear checklist exercise, where reductionist algorithms or 'objective' values and methods can be used to ensure fair judgement in assessment. (Valentine, Shanahan, Durning & Schwirth, 2021) Just as putting all of the components of a human body in a bucket does not make life, neither does simply ensuring all four fractal components of fair judgements are ticked off build fairness in assessment. In complexity, the system behaviour relies less on the nature of the individual people and strategies but more on the strength and nature of the connections between them. (Fraser & Greenhalgh, 2001) For example,

the notion of programmatic assessment contends that individual data points are insufficient to provide a fair judgement about a learner's performance. Instead, a fair judgement requires analysing combined data, identifying factors and contexts which may influence the learner's performance, collecting evidence to support the judgement and provide feedback for improvement. (Roberts, Khanna, Lane, Reimann, & Schuwirth, 2021; Schuwirth & van der Vleuten, 2011b) Complexity thinking allows for the explicit articulation of both the components and dynamic interactions of fair judgement. Both are needed to create fairness. This has implications for the way assessments are designed and implemented.

Complexity also challenges the idea of prediction and control. In complex systems, people need sufficient freedom to interact with one another independently. (Fraser & Greenhalgh, 2001; Plsek & Greenhalgh, 2001) Strict rules or policies restrict the agility and freedom of people to interact with each other and if agents do not interact, fairness cannot emerge. (Woodruff, 2019) For managers and institutions, understanding how people, patterns or fractal and strategies interact is key to making changes to the direction of the CAS (Wakefield, 2013), which counterintuitively may include reducing the rules.

Despite these implications there are many unanswered questions from this research. For example, who decides what is fair and unfair and who negotiates disagreements? What happens when disagreements cannot be resolved? What happens when fairness cannot be achieved? Shared decision making with a shared narrative to negotiate fairness rather than creating rules and regulations from a top-down

approach would be preferable to allow for fairness to emerge through these interactions. However, learners, assessors and intuitions may be unfair in their interactions and prevent negotiation on fairness. An external stakeholder may need to be involved in this situation to negotiate fairness. These questions highlight future areas of research. This study also focused only on the stakeholder perspective of expert assessment leaders of medical schools and did not consider the perspectives of medical students themselves. Future research should include their valuable perspective. Furthermore, given our findings, further research should now be done considering fair judgement as a CAS. For example, researchers could consider what prevents fairness from emerging, what is the influence of other systems, external powers and pressures on the dynamics of the CAS.

There are limitations to this study. Fairness is not 'a-cultural' and the sociocultural context in which assessment occurs is relevant. (Gipps & Stobart, 2009) Indeed what is fit for purpose, credible, accountable and transparent will be determined by the local context and culture. This study was done in a Western orientated cultural context. It is therefore plausible the findings are limited in their generalisability. In line with our ontological and epistemological views, we do not define generalisability as replicability but rather as the extent to which we have been able to incorporate sufficient different perspectives on fairness. As demonstrated by the roles held, the participants in this research were heterogeneous with different expertise and responsibility. This diversity is likely to influence their understanding of fairness.

Conclusion

So, whilst the individual components identified in our results are not unique; approaching fairness from an ontological viewpoint of complexity is perhaps the most significant insight from this study. Within CAS, it is primarily the interaction between the entities which influences the outcome it produces, not simply the components themselves. Our study supported this premise by noting that fairness is created by people through how they use and combine the different fractal components of fairness within the assessment system. Fractal patterns can assist in enabling sense making in complex systems. Understanding fair judgement not as a linear process with a predictable trajectory but rather as a dynamic CAS may lead to purposeful, meaningful changes in our assessment systems which supports the use of fair judgement in assessment.

CHAPTER SEVEN: WHAT STOPS FAIRNESS FROM EMERGING IN ASSESSMENT? THE FORCES ON A COMPLEX ADAPTIVE SYSTEM

This chapter is a published article: Valentine N, Durning S, Shanahan EM, Schuwirth L. What stops fairness from emerging in assessment? The forces on a complex adaptive system. *Perspect Med Educ.* 2023; 12(1):338-347. This is an open access article.

This article was co-authored with Professor Steven Durning, Professor Michael Shanahan and Professor Lambert Schuwirth. My contribution for this article was approximately 80% of the research design, 80% of the data collection and analysis, and 85% of the writing and editing. My role in the research design of this project included collaborating with my fellow authors in the research design, developing a video to play to research participants, applying for ethics approval to undertake the research, recruiting research participants and coordinating suitable times for focus groups. Subsequently, I facilitated all of the focus groups. I also undertook data analysis in collaboration with my fellow authors. Finally, I wrote the original draft manuscript before incorporating suggested edits and revisions from the other authors. I was also responsible for the submission process, ensure the article adhered to the publication guidelines. Professor Steven Durning, Professor Michael Shanahan and Professor Lambert Schuwirth shared the remaining percentage of work equally.

The demonstration of fair judgement in assessment as complex phenomenon rather than a linear process in the previous study changed the focus of this PhD slightly. It ensured this next study was different from the previous two. Whilst it was still important to explore the meanings of fairness constructed by many different stakeholders in accordance with a constructivist paradigm, this study specifically sought to understand more about fairness as a complex adaptative system.

This final study aimed to canvas the expert thoughts and conceptualisations from another a distinct group of program designers and health professions education experts outside of Australia and New Zealand. Specifically, it sought to engage in discussions regarding the external forces on the complex adaptive system which could potentially disrupt fairness from emerging. This chapter is the third study, as published in *Perspectives on Medical Education*.

Abstract

Introduction: Workplace-based assessment occurs in authentic, dynamic clinical environments where reproducible, measurement-based assessments can often not be implemented. In these environments, research approaches that respect these multiple dynamic interactions, such as complexity perspectives, are encouraged. Previous research has shown that fairness in assessment is a nonlinear phenomenon that emerges from interactions between its components and behaves like a complex

adaptive system. The aim of this study was to understand the external forces on the complex adaptive system which may disrupt fairness from emerging.

Methods: We conducted online focus groups with a purposeful sample of nineteen academic leaders in the Netherlands. We used an iterative approach to collection, analysis and coding of the data and interpreted the results using a lens of complexity, focusing on how individual elements of fairness work in concert to create systems with complex behaviour.

Results: We identified three themes of forces which can disrupt fairness: forces impairing interactivity, forces impairing adaptation and forces impairing embeddedness. Within each of these themes, we identified subthemes: assessor and student forces, tool forces and system forces.

Discussion: Consistent with complexity theory, this study suggests there are multiple forces which can hamper the emergence of fairness. Whilst complexity thinking does not reduce the scale of the challenge, viewing forces through this lens provides insight into why and how these forces are disrupting fairness. This allows for more purposeful, meaningful changes to support the use of fair judgement in assessment in dynamic authentic clinical workplaces.

Introduction

Workplace-based assessment affords learners the opportunity to experience and overcome the real-life challenges clinicians face in delivering patient care. However,

this level of 'authenticity' can create significant challenges for assessment. For example, time pressured clinical situations, uncontrolled encounters and the prioritisation of patient care all make standardised workplace assessment challenging. Different strategies have been used to attempt to overcome these problems such as the use of programmes of assessment (Van der Vleuten & Schuwirth, 2005), improved understanding of sampling and validity evidence (Schuwirth & van der Vleuten, 2020), the use of entrustment decisions (Ten Cate, 2005) and narrative approaches (Valentine & Schuwirth, 2019). However, the realities of the healthcare environment means that despite the best intentions, assessment may occur in an unpredictable manner. Despite these challenges, assessment and learning still occur. Medical students become future health care professionals. As Greenhalgh notes of health services research, "the articulations, workarounds and muddling-through that keep the show on the road are not footnotes in the story but its central plot." (Greenhalgh & C. Papoutsis, 2018) This also is true of medical education.

A lens of complexity has been encouraged to comprehend the nature of health professions education (Bowe & Armstrong, 2017; Cristancho, Field & Lingard 2019; Fraser & Greenhalgh, 2001; Long, McDermott, & Meadows, 2018; Mennin, 2010) because the environments in which assessment occurs are dynamic with numerous complex relationships and contexts. Linear rules or algorithms cannot be applied to every possible situation. Thus, to offer an understanding of this "muddling-through" (Greenhalgh & Papoutsis, 2018), considering assessment as a complex adaptive system (CAS), which has significant explanatory power, is warranted. A CAS is a collection of individual agents with freedom to act in ways that are not always predictable, and whose actions are interconnected so that one agent's actions change

the context for the other agents. (Plsek & Greenhalgh, 2001) The key features of a CAS are described in Table 4.

Independence:	CAS consist of individual agents (Plsek & Greenhalgh, 2001) who make independent choices about their actions. (Fraser & Greenhalgh, 2001)
Adaptability:	Each agent adapts to changes in the context, past experience and to each other’s behaviour. (Bowe & Armstrong, 2017; Fraser & Greenhalgh, 2001; Long, McDermott & Meadows, 2018; Reed, Howe, Doyle & Bell 2018)
Unpredictability:	The independence and adaptability of the agents leads to non-linearity and unpredictability. (Greenhalgh & Papoutsi, 2018; Long, McDermott & Meadows, 2018; Mennin, 2010; Reed, Howe, Doyle & Bell 2018)
Emergence:	Interactions between agents create outcomes that are greater than the sum of the individual agent behaviours. (Long, McDermott & Meadows, 2018) The system’s behaviour relies less on the nature of the individual agents than on the quantity and quality of connections between them. (Fraser & Greenhalgh, 2001)
Patterns:	Despite the unpredictability, principles and patterns arise. (Reed, Howe, Doyle & Bell 2018) These patterns provide understanding to how the system works (Mennin, 2010) as they guide behaviours within it. (Reed, Howe, Doyle & Bell 2018)

Distributed control:	Control is dispersed as a result of a huge number of decisions made by individual agents (Van Beurden, Kia, Zask, Dietrich & Rose, 2013) making the system is resistance to centralised control. (Kurtz & Snowden, 2003; Long, McDermott & Meadows, 2018)
Self-organisation:	Order, innovation and progress naturally arise from within the system. (Greenhalgh & Papoutsi, 2018; Norman, 2011) Work arounds and muddling through are central to CAS. (Fraser & Greenhalgh, 2001; Greenhalgh & Papoutsi, 2018) Tensions and paradoxes do not necessarily need to be resolved. (Plsek & Greenhalgh, 2001)
Embeddedness:	Agents and CASs are embedded within wider other CASs. (Long, McDermott & Meadows, 2018) Therefore, agents or systems cannot be understood without reference to the other systems. (Kurtz & Snowden, 2003; Plsek & Greenhalgh, 2001)
Fuzzy, ill-defined boundaries:	The system boundaries are permeable and hard to define. (Long, McDermott & Meadows, 2018)

Table 4: Key features of a complex adaptive system (CAS)

One specific area in which a complexity perspective may provide plausible insights is understanding the nature of fair judgements in assessment. Traditionally, fairness has been seen as synonymous with objectivity, however, measurement-based assessment has struggled to adequately evaluate the wide variety of competencies demanded of today's health professionals. This has led to an increasing push to

embrace human judgement in assessment and accept its subjective nature. (Eva & Hodges, 2012; Govaerts & van der Vleuten, 2013; Hauer & Lucey, 2019; Hodges, 2013; Rotthoff, 2018; Schuwirth & van der Vleuten, 2006; Schuwirth & van der Vleuten, 2020; ten Cate & Regehr, 2019; Valentine, Durning, Shanahan, van der Vleuten, & Schuwirth, 2022) But to do this, we need to understand ‘What makes human judgement fair?.’

Unfortunately, there is no simple definition of fairness. It has been noted that fairness is a multi-faceted construct, with many different interacting components. (Valentine, Durning, Shanahan & Schuwirth, 2021; Valentine, Durning, Shanahan & Schuwirth, 2023; Valentine, Shanahan, Durning & Schuwirth, 2021) Recently, researchers identified a fractal suggesting fairness behaves as a CAS rather than a linear process, and that it emerges from numerous dynamic interactions between its components, which can be influenced internally. (Valentine, Durning, Shanahan & Schuwirth, 2023) However, CAS are also nested within other systems and may be impacted by the external context in which it exists. The aim of this study was, therefore, to understand how external forces on the CAS disrupt fair judgement emerging.

Methods

Theoretical underpinnings

We employed reflexivity throughout the research process and it is described through the dimensions of ‘personal’, ‘interpersonal’, ‘methodological’ and ‘contextual’ (Olmos-Vega, Stalmeijer, Varpio, & Kahlke, 2022). Our research team consists of experienced

health professions education researchers and clinicians, all familiar with the study content, having undertaken previous studies on fairness in assessment. As a research team we work in diverse contexts, representing a range of specialties and medical education research environments across different continents. LS has previously been involved in education in the Netherlands, however NV (who conducted the focus groups) was not previously known to the participants. This diversity of experiences was leveraged through allowing for a range of perspectives and enabling rich team discussions during data interpretation. (Varpio, Ajjawi, Monrouxe, O'Brien, & Rees, 2017)

Barriers and enablers often, have a real and realist aspect. In this study, however, we wanted to focus on limiting and enabling forces as partially constructed by participants in line with our constructionist paradigm. (Bergman, Feijter, Frambach, Godefrooij, Slootweg, Stalmeijer & van der Zwet, 2012) For the rest of the paper, for readability, barriers and enablers will be referred to as forces. Specifically, we used principles of complexity theory as a 'lens' to identify how external forces may facilitate or disrupt fairness in assessment. (Bleakley & Cleland, 2015) We recognise that complexity science is not singular, but rather has multiplicity of legitimate orientations.

(Cristancho, Field & Lingard, 2019) Our focus was on how individual elements work in concert to create systems with complex behaviour. This approach to complexity lends itself to social phenomena and is commonly used in medical education. Whilst there are many legitimate approaches to and principles of a CAS, we have used three of the key principles, interaction, adaption and embeddedness to identify how external forces may impact fairness. (Fraser & Greenhalgh, 2001; Long, McDermott & Meadows, 2018; Reed, Howe, Doyle, Bell, 2018) The first two were chosen because it is only through interaction and adaption of individual elements that diverse behaviour or

outcomes can emerge from the CAS. (Durning, Artino, Pangaro, van der Vleuten & Schuwirth, 2010; Reed, Howe, Doyle, Bell, 2018) In other words, the whole is more than the sum of the parts. Embeddedness was chosen because the individual agents and systems also sit within other systems. Medical schools, for example are embedded within clinical workplaces, which are embedded within society. CASs cannot be fully understood without reference to these other systems. (Kurtz & Snowden, 2003; Plsek & Greenhalgh, 2001) In addition, we have defined three subcategories of forces (assessor and student, tool and system forces) to follow the three components described by Schuwirth and van der Vleuten in their description of the history assessment. (Schuwirth & van der Vleuten, 2020) In this paper, assessment was described as a measurement (tool), judgement (assessors and students involved in the judgement) and system. Specifically, a 'tool' is information, assessment or any strategy which would normally support interactions to facilitate the emergence of fairness.

Setting and participants

This study was conducted online via Zoom with a purposeful sample of academic leaders from eight universities across the Netherlands. All universities involved in this research were either utilising or transitioning to a programmatic assessment framework. The Netherlands also has a thriving, collaborative medical education community and participants would be well informed to discuss this topic with a good understanding of the literature. Medical training within the Netherlands involves a three-year Bachelor of Medicine programme followed by a three-year Masters of Medicine programme. (van der Vleuten & Schuwirth, 2005) All eight medical schools

were invited to participate through the Dutch Association for Medical Education. Each medical school was able to nominate appropriate individual members to participate. Ethics approval was obtained (Flinders University: 4297).

We invited participants via email to participate in focus groups. Focus groups were chosen to allow individuals to build on other group members' responses, allowing for dynamic interactions. (Stalmeijer, N. McNaughton, & Van Mook, 2014) As we wished to understand the external forces on the CAS which may impact fair judgement emerging, participants were shown a video explaining the findings of our related series of studies. In the focus groups, we asked them to discuss their perspectives on considering fairness as a complex adaptative system, as well as the external systems or factors which could influence the CAS. We provided no further incentive to participate.

Data analysis

The focus groups were recorded and transcribed verbatim without any identifying data. Reflective notes collected during focus groups and the shared white boards were included in data analysis. We used NVivo (Denver, Colorado) qualitative software to assist with data management. Although there are elements of abductive analysis, most of the first and second order themes were already preconceived and so this methodology is best described as thematic analysis. As data collection progressed, we developed codes, and refined and revised them in an iterative matter. The analysis process involved development of a coding book, comparison of different codes between and within transcripts to clarify, confirm and categorise codes. Themes were

then developed and illustrative quotes were used to bring the participant experiences to light. All authors were involved in the discussions during the coding process. The data collected was considered to offer a sufficient understanding (drawn from Dey's notion of theoretical sufficiency) to answer the research question. (Varpio, Ajjawi, Monrouxe, O'Brien, & Rees, 2017)

Results

Four focus groups were held between February and March 2022 lasting between 70 and 95 minutes. Nineteen individuals from six medical schools participated. As described above, there are forces which can impair interactivity, adaptability and embeddedness and restrict fairness from emerging. Within these themes, the forces have been subcategorised as assessor and students forces, tool forces and system forces. (Table 5)

“COMPLEX”: FORCES IMPAIRING INTERACTIVITY

Assessor and student forces

- Assessors' enthusiasm and engagement in the judgements process
- Assessor self-doubt and lack of confidence in their own judgement
- Student not empowered to interact with the complex adaptive system
- Student chooses not to engage
- Lack of situational awareness

Tool forces

- Not using evidence or tools to mediate interactions
- Use of convenience not purposeful sampling to support interactions
- Lacking information to support meaningful interactions
- Lack of access to information

System forces

- System barriers, hierarchical systems and cultural norms can inhibit opportunities for interactions between stakeholders

“ADAPTIVE”: FORCES IMPAIRING ADAPTABILITY

Assessor and student forces

- Assessor inexperience which impacts their ability to adapt in response to their interactions
- Assessors not adapting due to fear of change and uncertainty, or of doing wrong
- Assessors not appreciating need to adapt (I know best) or not wanting to adapt (easier not to)
- Learners are unaware of how to interact and adapt with the judgement
- Learners unwillingness to adapt following negative feedback
- Learners inappropriately adapt their behaviour towards those assessing them to receive a desired outcome

Tool forces

- Articulation of judgement to facilitate adaption
- Willingness of assessors to give and receive feedback to each other

System forces

- System which does not allow for feedback and adaption

“SYSTEM”: FORCES IMPAIRING EMBEDDEDNESS

Assessor and student forces

- Unsafe for a learner to be vulnerable to judgements
- Vulnerability of assessors as they have ultimate responsibility for their patients
- Lack of support for assessor to make a judgement

Tool forces

- High stakes nature impacts perception of fair

System forces

- Judgements influenced by bias, such as gender bias, harmful discrimination or specific prejudices which are outside agreed fuzzy boundaries
- Conflict in the purpose of judgement for the individual: is it a progression judgement or feedback?
- Conflict in the purpose of judgement for the system: it is distinguishing between learners, ie ranking or determining if meeting a standard?
- Fear of an external force which will disrupt the system
- University regulations limit the freedom of assessors to make judgement decisions
- System ensures some judgements are more intensive than others, ie fail judgements
- University limitations (ie high student numbers, money, assessor time, inefficient technology) impact how assessors interact with the system

Table 5: The forces preventing fairness emerging from the complex adaptive system

Forces impairing interactivity

A fundamental characteristic of a CAS is that the system's behaviour relies less on the nature of the individual agents than on the quantity and quality of the interactions between them. Barriers to these interactions can cause significant disruption the output of the system.

Assessor and student forces

Assessors can self-limit their interactions with the components of fairness, for example as a result of their self-doubt in making a judgement, limiting the emergence of fair judgement. Their self-doubt may be in their own abilities, or it may be due to concern there is a lack of information to form comprehensive picture of a learner's progress.

“Confidence in their own judgement ... People doubt whether they really have all the reason(s) to give this nice person the judgement you're not good enough at this moment.” (Participant 1)

On the other hand, enthusiasm and engagement of assessors, or perception of engagement, in the assessment judgement process impacts the quantity and quality of these interactions. This may be through permitting or not empowering the learner to be an active participant in the learning process, or because the student chooses not to engage. Either way, the level of engagement may impact on the intent of wanting students to learn and improve, which directly influences the perception of fairness from both an assessor and learner perspective.

“These often are also students that don’t have ownership of their portfolio. They don’t own their learning path” (Participant 6)

Provision of feedback to the student requires situational awareness and meaning-making of the situation in the here and now. Unless this meaning-making, situational awareness and agile adjustment of behaviour occurs, interactions will be limited. Typically, the process is likely to be perceived as ritualistic and going through the motions.

“She [student] said, in the one internship I got the command that I wasn’t assertive enough, I didn’t speak up enough, so I tried to change that in the next internship, and then they told me I did too much. I spoke too much, and spoke up too much. I was too dominant, I was - so, actually, I don’t know any more what I should do.” (Participant 11)

This is also an example of gender bias which is described later. If an assessor’s judgement is influenced by any factor other than the student’s performance, then this is outside of the agreed fuzzy boundaries and a pressure on the CAS. This includes biases, stigma or harmful discrimination.

Tool forces

Having sufficient information about the learner facilitates interactions. So, logically, barriers to the provision and availability of information limits interactions. Examples of this include assessors not making the effort to collect sufficient information to support interactions,

“...many teachers [sic] do not look in the portfolio” (Participant 18)

or only using evidence that is convenient to find,

“...not use like convenience sampling, like the patients that are just coming along, and also use like more purposeful sampling that you say okay, we're missing your data on your ability to handle such kind of patients” (Participant 2)

or not being able to obtain the meaningful information, that is needed to support these interactions.

“the system doesn't provide this type of feedback because they did a multiple-choice test” (Participant 7)

System forces

Hierarchical systems and cultural norms may limit interactions between assessors and students. For example, learners with different backgrounds and cultural differences may face difficulties in adopting to assessment within their new system. If their cultural norms do not align with assessment process this can limit their ability to engage with the system. Traditional beliefs about the value proposition or role of assessment in education can also discourage dialogue about assessment or feedback. This can further be complicated by system forces such as scheduling which can also limit the opportunities for interactions to occur both between assessor and student, and also between assessors.

“It’s a conservative, hierarchic system and a lot of surgeons, especially in our region, they’re not raised there, they didn’t have their education there, ... so they’re not used to this kind of assessing, and they just think you shut up and you do your job.” (Participant 7)

Forces impairing adaptability

Assessors in fair assessment processes agilely adapt to past experience, internal and external influences and feedback. This makes a CAS efficient and effective because assessors use their expertise and situational awareness to quickly adjust their behaviour and interactions when necessary. Learners also contribute to this process of adaption. Any barrier to the adaptive processes will impact the behaviour of the system.

Assessor and student forces

Assessors may lack the expertise and situational awareness required to adapt to the incoming feedback and changing contexts, for example as a result of insufficient staff development. This increases the likelihood of a fear of change or intolerance of uncertainty, leading assessors to want to stick with ‘what they have always done’. But this comes with a lack of perspective as to the need to adapt, or not wanting to adapt, or even the belief that everything was better in the past. Such a misalignment between what the assessor has to offer and what the situation at hand requires, easily leads to a perception of unfairness.

“We know that teachers have certain conceptions of learning, which have been formed by their own experience, and those conceptions of learning – and assessment – are deeply rooted, and it’s very hard to change them...They’re often also they’re formed by their personal experience, and perhaps even related to their identity as a teacher.” (Participant 12)

Assessors may also fear they are not acting in the best interest of both students and society, so do not adapt to avoid doing the ‘wrong’ thing; or that formative and summative assessment or assessment of and assessment for learning are zero-sum games.

“...they [assessors] want to do the best thing, and they’re afraid that with the new way of assessing, these programmatic assessments, they are afraid that they are not doing the right thing” (Participant 13)

Assessors may fear students too may be uncertain about how to interact and adapt to the judgments and that they see it as a zero-sum game as well. This, again, may impact on the adaptability of stakeholders in the CAS and limit the emergence of fairness.

“...in a summative system, we sort of educate learners to ignore feedback”
(Participant 12)

This may be due to many reasons such as cultural values, previous experience with assessment, or expectations placed on doctors within society to not show weakness and thus not need to adapt.

“...one of my major goals in life is if I can achieve that doctors consider it normal to be vulnerable, I would think that we have gained a lot. But there are still many around who feel themselves or think they are still on this pedestal, and they can’t make any mistakes.” (Participant 17)

Students may be unwilling to adapt due to the stigma or embarrassment of receiving negative feedback which may hinder future interactions in the system. The student may identify reasons for receiving this negative feedback, including being a surprise result, or it being the fault of an unfamiliar way of testing, but either way an unwillingness to adapt and learn from the negative feedback remains a barrier to the CAS.

“What I often see is that students who fail the test will say that the test was subjective. So, it was not their fault, it was the test’s fault. It’s used as a mechanism to not be open to learn” (Participant 3)

Students may inappropriately adapt their behaviour towards those assessing them to receive a desired outcome. Both the assessor and student are behaving in a way in which means they are complying with the system but are self-limiting the quality of interactions because they don’t want to play the ‘real’ game of vulnerable, in-the-moment authentic learning and feedback. This may hinder the quality of further interactions.

Tool forces

Students and assessors are only able to adapt to enable fairness to occur if they receive information from stakeholder interactions. If information is not provided either because it is difficult to articulate or because assessors are unwilling to share, then purposeful adaptation cannot occur.

“...a student can tick the boxes but you have the feeling that it’s not going to be a good doctor and how do you make that visual, visualise that to other teachers.” (Participant 4)

This includes interactions both between student and assessor and between assessors.

“...the openness to feedback and to give feedback to your fellow assessors”
(Participant 2)

System forces

A system which makes stakeholders feel unsafe in providing feedback and engaging in interactions will not support interactions and adaptations.

“Someone [whistle-blower] who’s leaking information about certain circumstances, but they are the people who don’t feel they can be – they are not free to be honest” (Participant 13)

Forces impairing embeddedness

Individual agents and systems are embedded within wider systems. Each individual or system cannot be fully understood without reference to their roles within wider systems.

Assessor and student forces

The diversity of roles of both the student and assessor can create barriers with the CAS. Students are not just learners, they are also future doctors attempting to obtain a grade, a residency training position or impress a future colleague. This variety of future roles within a variety of systems can make it unsafe for them to be vulnerable, engage in quality interactions and adapt appropriately to the judgement decisions.

“They all feel like they can’t show what they have in themselves, they can’t show their talents and they’re very worried that they will not get the job that they want so much and so on. It’s really a lot of tension actually.” (Participant 15)

Similarly, assessors are not just assessors. They are also clinicians with responsibility for patients or in private practice, even business owners. Assessors may feel vulnerable and unable to trust learners with patient care. A limited perception of this entrustment hinders interactions in the workplace. When, for example, a learner feels that they could have been entrusted more than the assessor, this creates the perception of unfairness.

“I think they also feel vulnerable, they just have to let their student go to their very ill patients and ... they don't get feedback themselves of the capability of the student and it's very difficult to let loose of that control.” (Participant 15)

Assessors also work with the learners themselves, and need support to make judgement decisions, especially difficult judgement decisions.

Tool forces

Judgement decisions usually have wide-reaching effects. Some high stakes decisions can have significant financial, social or motivation consequences, especially if the judgement is that the learner is unsatisfactory or not ready to progress to a next phase. Tools that do not provide sufficient information to form such high-stakes decisions and require decisions that are not proportional with the richness of information available will be perceived as not fair and may evoke volatile emotional responses. These, in turn, will impact on future interaction with others and the system, such as leading to leniency bias or retreatism.

“But the two or three per cent of the students which will get an unsatisfactory grade, they say it's not fair. It's always the same [problem].” (Participant 7)

System forces

Complexity thinking maintains that systems are not synonymous with complete chaos and that they can be maintained by fuzzy, ill-defined boundaries. If a judgement is influenced by any factor other than the student's performance, then this was seen as

being outside of these agreed boundaries and a pressure on the CAS. This includes biases, stigma or harmful discrimination.

“...sometimes students are discussed in the staff meeting or someone will tell you about a student and I think that also influences your judgement. That can be particularly detrimental for what we call non-traditional students, so for people with migration backgrounds with lower socio-economic status.”

(Participant 2)

As the system is not isolated but rather overlaps with other systems, there may be conflicts between the various purposes of the judgements. This can put pressure on the system and its ability to adapt and future interactions.

“Am I giving feedback or I’m also giving a judgment?” (Participant 14)

An external force, such as the COVID pandemic, or fear of an external force such as litigation is also likely to put pressure on the system. University level regulations tend to limit the freedom of assessors to interact, adapt and make judgement decisions and can make some judgement decisions more time consuming than other decisions.

“it also depends on the system, because I still can have the courage but I still can't do it because of the system” (Participant 2)

Institutions also make decisions about how to spend finite resources which impact how assessors are able to interact with the system.

“You want to have a very individual, personal relationship, like a mentor, but for the start of the study, the bachelor part, numbers are too high. It’s very difficult.”

(Participant 13)

Discussion

This study adds a different perspective on the external forces which may impact the emergence of fairness. Often barriers are described in realist terms; for example, lack of time or resourcing. However, this study uses the lens complexity to describe how forces prevent fairness from emerging from the CAS. Viewing forces in this way provides insight into why and how these forces are disrupting fairness from emerging. This is not trivial. When barriers and enablers are described in realist or objectivist terms it carries the connotation that they are relatively established. On the other hand, when barriers and enablers are explored from a subjectivist and complexity lens, it allows us to critically examine the factors which are contributing the creation and persistence of these forces, and agilely adapt or create more levers to influence the impact these forces have on the CAS and the emergence of fairness.

Typically, workplace-based assessment occurs in unpredictable clinical environments where implementation of replicable and standardised, measurement-based assessments is largely impossible. However, if we look beyond linear, objective thinking and a complexity lens is applied, then these challenges can be reconsidered. For example, they make us reconsider the value of equality versus the value of equity with respect to fairness. Equality, as in standardisation and structuring assessment

may seem fair because everybody receives the same process of assessment. Our concept of fairness is one from an equity lens. Everybody receives the same quality of assessment, but the assessment process is bespoke; it recognises that people have different strengths and weaknesses, and that the assessment process needs to be bespoke to cater to those. Complexity thinking does not reduce the scale of the challenge, nor does it provide simple fixes to tensions in assessment. (Greenhalgh & Papoutsis, 2018) But, it does provide a different perspective to approach them. It also presents forces as interactional problems which can be modified allowing institutions more agency over the situation.

Consistent with complexity theory, this study suggests there is no single force or factor which needs to be addressed for fairness to emerge. The study highlights an almost overwhelming number of potential forces to address. However, viewing these forces with a systems mindset has at least two important implications. Firstly, a systems mindset shifts responsibility from away the individual. Forces need to be addressed at a system level because forces arise from changing interrelationships or adaptations (or lack of) between parts of the systems. (Cristancho & Taylor, 2019) Secondly, addressing this as a system, allows for a framework to allow the researcher and educator to better identify, explore and address the force and related potential forces.

The forces described in this study are not exhaustive; there are likely many others. Similarly, the generalisability of the forces identified is limited by the nature of this study. However, the intent of this inquiry was not to identify an exhaustive list, nor was it to design solutions as these too are likely to be context specific. The aim of this study was to understand the external forces which may impact the emergence of fairness using a lens of complexity. Considering fairness as a CAS changes our views

about how we can improve assessment and legitimise human judgement in our assessment programs. Because in CAS, the interactions between the entities are most important, meaning strict regulatory frameworks and tick box approaches to managing fairness are counterproductive because they limit the interactions between components.

Embracing complexity in fair judgement also means understanding that managers or policies cannot control the judgements assessors make, or that linear, causal thinking cannot predict behaviour of the individuals in the system. (Kurtz & Snowden, 2003; Van Beurden, Kia, Zask, Dietrich & Rose, 2013; Woodruff, 2021) Instead, systems designs and management practices which encourage interactions, develop expertise, enable access to all necessary information, facilitate self-organisation and individual responsibility can contribute to better outcomes. (Bowe & Armstrong, 2017; Holden, 2005) Providing a variety of strategies to enable assessors to adapt to the situation in the here and now rather than enforcing one 'gold standard' strategy is also likely to enhance system behaviour. (Fraser & Greenhalgh, 2001; Woodruff, 2019)

Our previous research into the components of fairness noted a fractal which consisted of credibility, fitness for purpose, transparency and accountability. (Valentine, Durning, Shanahan & Schuwirth, 2023) Fractals are shapes or concepts which exhibit "self-similarity" at different scales, meaning they remain the same regardless of whether you zoom in or out. (Holbrook, 2003; Lipsitz & Goldberger, 1992) A fractal is a manifestation of an underlying complex adaptative system (CAS). (Tsoukas & Dooley, 2011) Understanding fractal patterns can enable sense making in complex systems and guide rational changes in the system and influence the agent's behaviour. (Mennin, 2010; Reed, Howe, Doyle, Bell, 2018) Fractals can provide structure and

fuzzy boundaries to help CAS remain in stable equilibrium. (Fraser & Greenhalgh, 2001; Holbrook, 2003; Mennin, 2010; Woodruff, 2021) Because of the organised, adaptive nature of CAS, if any of the fractal elements are missing, the system becomes unstable and may breakdown. (Golberger, 1996) This has implications for the way we design assessments.

There are limitations to this study. As already mentioned, the forces identified are not exhaustive; there are likely many others. Similarly, given the Western orientated cultural context, there may be additional relevant meaningful perspectives to be found in other contexts. Furthermore, whilst we specifically sought to look at forces from a constructionist perspective, researching forces from a realist perspective could complement this view, and may enable a more comprehensive approach to future system changes.

As 21st century health professions education moves to embrace human judgement in its assessment programs (Boursicot, Kemp, Wilkinson, Finyartini, Canning, Cilliers & Fuller, 2021) understanding what makes this judgement fair beyond an objective framework is essential. Understanding and thus modifying the forces which prevent fairness emerging in light of a CAS system can lead to more purposeful, meaningful changes to support the use of fair judgement in assessment in the authentic clinical workplaces.

Appendix to this published article: Focus Group Video Script and Question Guide:

Most people agree that assessment needs to be fair. Traditionally, objectivity was seen as the main way to ensure fairness in assessment. But more recently, views have changed, and it is now generally accepted that subjective human judgement plays an key role in comprehensive assessment programs. However, in embracing subjective judgement an important question has arisen, what makes human judgement in assessment fair?

That is what we have been looking at with a series of studies.

SLIDE TRANSITION

And what we've found is that fair judgement in assessment is complex. It can actually be considered to be a complex adaptive system. As such, there are many interacting and sometimes conflicting factors to consider and understand.

SLIDE TRANSITION

To help explain what we've found, let's use an analogy. Consider a pine tree. A pine tree is composed of branches which are composed of smaller branches which in turn are composed of even smaller branches and so on. Branches on pine trees have an interesting feature: no matter where you look, or how much you zoom in or zoom out, the shape or pattern remains the roughly the same. From the largest branch to the smallest branch the pattern seems to repeat, over and over again at different scales. This is called a fractal. Fractals can also be produced mathematically. The equation behind repeating fractals is actually quite simple, but it produces an incredibly complex shape which repeats for infinity.

SLIDE TRANSITION

We think fair judgement in assessment is a little like this. It is complex and seems to be different in different circumstances. But if you look more closely, our research has demonstrated that there is a recurrent and repetitive shape to fair judgement. We've conducted a literature review, spoken with learners, teachers and education designers and managers. And what we found was that underlying everything they said were the same four components of fair judgement: transparency, accountability, fitness for purpose and credibility. This is the basic "shape" of fair judgement.

SLIDE TRANSITION

Just like the equations in fractals, these four components of fair judgement are reasonably straightforward in themselves. However, these components are not enough to create fairness in judgements on their own. Fair judgement 'emerges' from the purposeful and meaningful interactions between these four components. And it is these interactions which makes fair judgement complex. To use another analogy, when you take all of the components of the human body and put them into a bucket that does not create life. Life only exists when all of those body systems work together and interact with one another.

SLIDE TRANSITION

And there are many layers or sizes of pine branches. In fact, there is an almost infinite number of sizes that this same complex shape can be. The same is true of fair judgement. There are an infinite number of layers of judgement, for example, an individual utterance of the learner during the assessment, whether they were

able to take a history from a patient, right the way through to is a learner ready to graduate? It doesn't make a difference if you look up closely, ie as an on the ground supervisor, or take a step back as a program coordinator, in all of the layers the same four components of fair judgement can be seen.

SLIDE TRANSITION

There are also forces which influence the development of the complex shapes. Going back to our tree analogy, the growth and size of the pine tree is influenced by the sunlight, or soil quality or water. And if someone builds a great big building next to the pine tree, the shape is going to be altered too.

SLIDE TRANSITION

Similarly in fair judgement, there are forces which influence these four key components. These forces include being able to have multiple assessors, whether longitudinal data collection is possible, having a narrative or vocabulary to support the judgment and so on. These are demonstrated in the diagram provided. It is these forces which influence the interactions and linkages between the four components of fair judgement.

Our model demonstrates how we see the complex adaptive system of fairness in assessment. There might be other components but this is a framework to help understand and construct fair judgement in different contexts.

SLIDE TRANSITION

We do know though that know that fairness does not operate in a vacuum. It is impacted by various other systems and forces, for example university regulations, patient demands or power imbalances between assessors and

learners. It is these other systems and forces, and how they impact on fair judgement that we are interested in for this study.

We'd love to know 3 things:

- What do you think of the model? We will go through this at the beginning of the focus group.
- What external systems or factors could influence our model?
- How do these external systems or factors influence the interactions between the elements of our model?

CHAPTER EIGHT: DISCUSSION AND CONCLUSIONS

Introduction to discussion

This PhD describes a program of research exploring fairness in assessment. As described in chapter 1, objectivity has previously been seen as the predominant way to ensure fairness in assessment. (Hodges, 2013; ten Cate & Regehr, 2019) This perspective existed because of competence being viewed as something which could and should be captured quantitatively and expressed as a numerical value. (Schuwirth & van der Vleuten, 2020) From a positivist perspective, the dominant assumption was that competence is a single independently existing reality or combination of independently existing realities which can be identified and measured. (Park, Konge & Artino, 2020) In a desire to seek this 'single correct judgement' or 'reality', it seemed logical to therefore ensure that fair assessment was objective and free from any personal biases. This objectivity was constructed as the hallmark of high-quality assessment and used to justify the fairness of assessment and the discrimination and differentiation between learners with its subsequent consequences. (Govaerts & van der Vleuten, 2013; McGuire, 1993; ten Cate & Regehr, 2019; Valentine & Schuwirth, 2019)

However as highlighted in chapter 1, this approach to assessment has many limitations and the health professions community has been advocating for a change in direction for some time. Indeed, the views on assessment of medical competence

have changed substantially, leading to important implications for how we understand what constitutes fair assessment. (Bacon, Williams, Grealish & Jamieson, 2015a; Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014; Govaerts, van de Wiel, Schuwirth, van der Vleuten & Muijtjen 2013; Hodges, 2013; Jones, 1999; Rotthoff, 2018; Schuwirth & van der Vleuten, 2006; ten Cate & Regehr, 2019) However, specific research into the nature of fairness has been lacking and is needed to help further inform future developments in assessment.

In contrast to the previously described positivist paradigm, I took a social constructivist approach as described in chapter 3. Through my research, I sought to explore the meanings constructed by individuals and groups, collecting data from different stakeholders and contexts.

The results of this body of research confirmed that there is no simple definition or formula for fairness. Fairness is multi-dimensional and context dependent. Most importantly, this body of research offers a different perspective of fairness in assessment by approaching it from an ontological viewpoint of complexity. Through the identification of a fractal in our data from multiple studies, we have come to understand fairness as a complex phenomenon that emerges from the dynamical interaction between components. In line with complexity theoretical notions, the same four elements of fairness (transparency, fitness for purpose, accountability and credibility) occurred throughout the data. These same four elements interacted with each other at all levels in the assessment program and therefore behaved like a fractal. This has important implications for our understanding of fairness, because

within a complex adaptive system, a system's behaviour relies less on the mere presence of the individual components but more on the dynamic strength and nature of the interactions between them. In line with this, people seek to create fairness through managing the interplay between fitness for purpose, credibility, transparency and accountability when interacting with others rather than using them as a tick box list.

In this discussion chapter I will initially explore the process of how I came to change from a linear perspective to viewing fairness as a complex adaptive system during this program of research. Subsequently, I will discuss the implications of seeing fairness as a complex adaptive system. I will then turn to the strategies identified during this research which support fairness emerging by mediating the interactions between the components of fairness. Unfortunately, there are forces which can hinder the emergence of fairness as identified in chapter 7. The implications of this will be discussed briefly. Finally, I will also explore the limitations of this research and propose suggestions for future research.

Changing from a linear perspective to seeing fairness as a complex adaptive system

The concept of complexity was recognised very early in this program of research but the importance and implications for this research were not understood until midway

through the program. To illustrate this learning process, I have drawn a parallel to the changes that have occurred in the field of assessment over the last few decades.

In 2020, Schuwirth and Van der Vleuten highlighted the evolution of assessment. Initially, assessment was predominantly viewed as a measurement problem, approached with a linear, static, reductionist perspective. (Schuwirth & van der Vleuten, 2020) Later, assessment transitioned into being a judgement problem, initially still viewed within a measurement framework, where it was approached from a reductionist perspective. It eventually evolved to see judgement as a narrative that embraces diverse, complementary views on competence. (Schuwirth & van der Vleuten, 2020) Even though this was a significant change, this was still a linear approach to assessment. It is only more recently that there has been a shift towards considering assessment as a system where narrative and diverse views of competence are still valued but the thinking has become more integrated and dynamic, taking a system perspective. (Schuwirth & van der Vleuten, 2020)

With this analogy in mind, when I commenced my PhD, I approached it from the perspective of 'judgement as a narrative, embracing diverse, complementary views of competence', aligning with my social constructivist ontology. I was not seeking a 'single truth' but rather explore the meanings constructed by individuals and groups. (Varpio, Paradis & Uijtdehaage, 2020) However my thinking remained linear and static as I sought to understanding the different 'components' of fairness.

Despite this linear and static thinking, I observed indications of complexity early which influenced the choices I made. For example, a hermeneutic literature review chosen partly because of its cyclical nature. This process is rigorous but it is also flexible and iterative, allowing for refinement of the research question as data analysis progresses. (Boell & Cecez-Kecmanovic, 2010)

The initial findings from the literature review (chapter 4) and the interview study of learners and assessors (chapter 5) demonstrated that fairness is a multi-faceted and contextual construct. These research programs identified values of fairness which were supported by 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 decisions. Assessor agility and expertise are needed to interpret and interrogate this evidence, help identify the fuzzy boundaries and provide narrative feedback to ensure learners can improve. At a system level, factors such as multiple opportunities for learners to demonstrate competence and improvement, multiple assessors to allow for different perspectives to be collected and triangulated, and documentation are all needed for fair judgement. These system features are supported through the concept of procedural fairness which provides transparent expectations, allows for fit-for-purpose, individualised, proportional judgements, and supports dialogue and engagement with the learner. Finally, the environment in which the assessment decisions are made needs to be considered for fair judgments. In line with my social constructivist approach, it was acknowledged that these components should not be seen as a mere 'tick box' list or check list to ensure fairness as there is no simple recipe for fairness. Furthermore, one of the benefits of identifying these components

was the opportunity to provide a narrative to support dialogue between stakeholders to help create a shared understanding. With the benefit of hindsight, perhaps it could be suggested that creating a shared narrative supports interactions between stakeholders which is essential in complexity. However, at the time of doing this research I did not appreciate the importance of the interactions between the components I was identifying.

It was during the data analysis of the second study (chapter 6) that complexity became clearer as the data demonstrated a fractal pattern. As already highlighted in this chapter, the fractal pattern or 'shape' was made up of four components: credibility, fitness for purpose, transparency and accountability. We noted that when assessment leaders spoke about what is required for fair judgements, underlying all they said were these four elements.

A fractal is a shape that remains the same at different scales. (Lipsitz & Goldberger, 1992) The defining feature of fractals is their "self-similarity". (Holbrook, 2003) Fractals are shapes made by the same basic repeating pattern, so that the same shape is found regardless of whether you zoom in or out. (Lipsitz & Goldberger, 1992) An infinite number of repeating patterns at different sizes are combined together to give a fractal its shape. (Lipsitz & Goldberger, 1992)

A fractal is a typical manifestation of complexity and thus it was thought that applying a complexity lens may assist in our understanding of the phenomenon of fairness. On review of previous paper (chapter 5), complexity was seen.

As discussed in chapter 3 the use of complexity as lens to comprehend health professions education not new and is indeed encouraged. (Bowe & Armstrong, 2017; Fraser & Greenhalgh, 2001; Mennin, 2010) The key features of a complex adaptive system have been highlighted in chapter 3 and in chapter 6 and so will not be repeated again in this chapter. Instead, the implications of these key features and using a complexity lens will be discussed.

If fairness is considered through a complexity lens, then consistent with the idea of emergent phenomena, fairness can only emerge through interactions between components. Fairness emerges from how people use and combine credibility, accountability, fitness for purpose and transparency within our assessment systems and thus it is people who create fairness.

Consider the example of an end of term assessment. In making a judgement, an assessor may gather multiple firsthand observations from patients and healthcare staff who have interacted with the learner, obtaining multiple pieces of evidence over time, and combining these with their own observations. The assessor will interact with other stakeholders, the evidence, the context and the 'pattern' of fair judgement. They will potentially ask other assessors to help with self-calibration, and will discuss with the

learner, obtaining their perspective on the assessment. Based on these interactions they will combine information in a credible way, which is accountable, transparent and fit for purpose to create the judgement. After giving the learner the judgement, they may then adapt, perhaps by providing more targeted feedback to help the learner improve by identifying where they are not meeting expectations.

It is people who create fairness. In line with complexity, fairness cannot be directed by an external person or pre-determined. (Durning, Artino, Pangaro, van der Vleuten, Schuwirth, 2010; Reed, Howe, Doyle, Bell, 2018) Nor can it be reduced to a linear checklist exercise, where reductionist algorithms or 'objective' values and methods can be used to ensure fair judgement in assessment. Just as putting all of the components of the body in a bucket does not make life, neither does simply ensuring all four fractal components of fair judgements are ticked off build fairness in assessment. Just as it is interactions between the organs of the body which makes life, it is the interactions between the components from which fairness emerges.

A short video which demonstrates this idea is available at:

<https://youtu.be/6HZo8kpt3g8>.

I wish to pause at this point prior to discussing the implications of seeing fairness from a complexity lens and speak to one of the individual components of fairness. In the literature review, one of the four components of fairness was noted as defensibility, whereas in a later study this was noted to be accountability.

The literature review highlighted an emphasis on ensuring judgement was a legally defensible assessment of a trainee's learning. As data analysis progressed throughout the research program, there was a shift towards dual accountability; an accountability that extended not only to the learners but also to society at large. This included a commitment to assessment *for* learning which would enable learners to develop and improve which was fair to both themselves and wider society. As a result, extensive discussions occurred regarding these concepts and terms, and how this data should be interpreted. It was decided that substituting "defensible" with "accountability" in the framework would be a more appropriate reflection of this this evolving perspective.

Practical implications of fairness as a complex adaptive system

There are many practical implications of seeing fairness as a complex adaptive system. I will describe some of these below.

Assessment needs to be adaptable and agile with a focus on connections

As mentioned previously, clinical assessment often occurs in the workplace-based environments. And whilst workplace-based assessment offers learners the chance to gain real-life experience in delivering patient care, the reality of the clinical

environment also means that assessment is both unpredictable - as it 'walks through the door' - and time pressured as it competes with the time required to meet patient needs. As a result, the implementation of standardised, reproducible measurement-based assessments is often neither fair nor feasible. A complexity lens suggests that having individuals, both learners and assessors, who are able to agilely apply a variety of solutions to different circumstances, rather than relying on one standardised solution, is more likely result in the emergence of the desired fairness. (Woodruff, 2019)

When we acknowledge that learning takes place in a complex context, we recognise the need for assessment to also occur in a complex context. Consequently, adopting a lens of complexity to our assessment processes helps us to better understand fairness.

However, whilst complexity thinking can be considered at odds with standardised assessment it shouldn't be seen as an either-or approach. All forms of assessment are required to be fair. Given that assessment is now commonly approached from the level of the whole program, (Schuwirth & van der Vleuten, 2020) complete programs of assessment may still include standardised and structured tests as well as judgement based assessment, just as patient care still includes standardised lab testing in addition to expert judgement narratives from history taking, radiology reports and so on. However, the triangulation of information from standardised, numerical sources and narrative sources should be individualised. Modern assessment requires multiple pieces of information to be triangulated to be used for learning or for decision-

making. (Boursicot, Kemp, Wilkinson, Finyartini, Canning, Cilliers & Fuller, 2021) A complexity approach to fairness states this should be an ongoing interactive process between learners, assessors and their environment, and based on combining quantitative and qualitative information. It would make sense therefore, at a program level, that a complexity lens is more appropriate than a linear causal one.

Fuzzy boundaries contain unpredictability

Complex systems are to a certain extent unpredictable. (Greenhalgh & Papoutsi, 2018; Mennin, 2010; Reed, Howe, Doyle & Bell, 2018) People are required to navigate interactions between components, and in doing so they adapt based on their own past experiences. (Fraser & Greenhalgh, 2001; Van Beurden, Kia, Zask, Dietrich & Rose, 2013) In addition, in complexity, people will also need to agilely adapt to internal and external influences that cannot always be predicted or controlled. (Kurtz & Snowden, 2003)

Although a certain level of unpredictability is a feature of a complex situation, there are still some fuzzy boundaries. For example, there are certain actions or utterances which are deemed acceptable and some which clearly are not, but the boundaries are not sharp and often situational. Taking a history with a patient can serve as an illustration. What exactly will be said during the consult is unpredictable, but there are still certain boundaries about what is appropriate to say. To assist in the navigation between these boundaries, health professionals are required to have a repertoire of

strategies and the agility to adapt to what is occurring during the consult. This repertoire of strategies and the agility to adapt is the difference between an expert taking history and providing the patient with a predefined questionnaire to complete. The expert is able to expertly navigate the unpredictability of the consult armed with the appropriate strategies to do so.

Fractal patterns can also help. (Mennin, 2010; Reed, Howe, Doyle & Bell, 2018)
Explicitly communicating desired expectations of patients and health professions can help in recognising these fuzzy boundaries. Similarly, being clear about expected outcomes of assessment programs can assist in defining boundaries.

The focus moves from solving problems to identifying patterns

Complex problems require complex solutions. Because many of our cause to effect experiences involve direct relationships, for example eating relieves hunger, exercise improves fitness and so on, it seems logical to think in terms of a linear chain of events. But linear causal thinking cannot predict the behaviour of individuals or the system.

A common told anecdotal tale is the one of the cobra effect. This term was said to have originated during the time of the British colonial rule in Delhi. In an effort to reduce the number of deadly snakes in Delhi, the British government instituted an

incentive program in which a bounty was provided for each dead cobra presented by a citizen. This program was initially successful with a large number of snakes killed for reward. However, the numbers of cobra in the city did not continue to drop as expected. Instead, citizens began breeding cobras for the lucrative bounty. When the government became aware of the breeding program, the incentive program was scrapped. As a result, the cobra breeders set their now-worthless snakes free, which in turn increased the wild cobra population in Delhi. Thus, the solution intended to address the issue actually ended up making the problem worse through an unforeseen consequence.

This is also another example of Goodheart's law which I described in chapter 1. Strathern generalised this law as 'When a measure becomes a target, it ceases to be a good measure'. (Strathern, 1997) When the focus of a policy is set on only one measure, people, such as the cobra breeders, are able to optimise or manipulate that measure to meet a target. These overly simplistic rules do not allow for agility to respond to the unpredictable nature of the interactions occurring within the system. They are like trying to use a questionnaire to take a history, a protocol to individualise a treatment plan as the measures are pre-defined, or a tick list to control fairness.

Whilst linear thinking often feels intuitive, understanding system dynamics is important because assessment occurs in complex systems. Complex adaptive systems are not in constant equilibrium meaning there is a continual change and response to changes in the system. (Holden, 2005) Any pressure on one part of the system will be reflected

elsewhere. Seemingly small static changes on one part of the system can have large and often unexpected impacts as demonstrated by the cobra effect.

In contrast to trying to enforce one solution to solve a problem, complexity theorists are interested in understanding and recognising patterns. Studying the unpredictable patterns, principles and models which arise may help decode a hidden order to the system (Gleick, 2008; Golberger, 1996; Mennin, 2010; Newell, 2008; Storey & Butler, 2013) and enable sense making and rational choice amid 'turbulence'. (Wakefield, 2013) This typically requires reflection on action; retrospection to understand what the emerging patterns are. Only through studying these patterns can we understand the emergent behaviours at a larger scale and influence the system towards fairness. The data from this program of research suggests that at every level of assessment design and implementation, from corridor conversations to licencing decisions, the components of fitness for purpose, accountability, credibility and transparency need to be considered and included.

This research has also not specifically considered the impact of technology on fairness. However, over the course of this PhD (2018 – 2023) technology has disrupted assessment more than most could have foreseen. Artificial intelligence has advanced rapidly due to increasing data and computing power. (Lee, Wu, Li & Kulasegaram, 2021) There is a tension between the acknowledgement of how this may transform how health care is delivered, how best to teach learners how to optimally utilise artificial intelligence (Lee, Wu, Li & Kulasegaram, 2021) and the threat this poses to traditional assessment. Chat Generative Pre-trained Transformer (Chat

GPT) has been shown to correctly pass high stakes examinations and can be compared with the performance of medical students in the second half of their studies. (Friederichs, Friederichs & März, 2023) Educators have had to negotiate not only new technical challenges but also moral and pedagogical challenges. (Fawns & Schaepkens, 2022)

Fawns and Schuwirth have suggested that “our response to GenAI [General Artificial Intelligence] should align with our value proposition and not purely react to the threat or challenge we face.” (Fawns & Schuwirth, 2023) In this sense, returning to the underlying concept of fairness may be of value. What is the value proposition of assessment within our particular assessment system and how does this align with the values of fairness? Does this align with the value proposition of learning and is it fair for wider society?

In addition to considering the value proposition of assessment, a lens of complexity may be helpful. A review of an online proctored high stakes exam during COVID, found that while projecting objectivity, this process actually compromised objectivity through rigid scripts and inflexibility and inadequate adaptability in interpretation of these scripts. Furthermore, these proctorial services, driven by a perceived need to combat cheating and legitimise changing assessment to the online format, inadvertently exacerbated tensions between agendas of commercialisation, accountability and the education of trustworthy professionals. (Fawns & Schaepkens, 2022) Instead of focusing on trying to solve the problem of ‘cheating’ with a simple solution, returning to identifying fractal patterns which embrace values of fairness,

supporting interactions between stakeholders and technology, and allowing for agility of assessment and assessors may be more fit for purpose.

Simple rules can assist in creating an environment in which fairness can emerge

Complexity theorists state that 'simple rules' often provide a way of understanding and potentially managing the emergent behaviour of complex systems. (Reed, Howe, Doyle & Bell, 2018) Simple rules typically assist with direction pointing, recognition of boundaries, and permissions to help create an environment in which the outcome can emerge. (Plsek & Wilson, 2001) Having these few, but flexible and simple rules prevents unnecessarily limiting self-organisation and innovation which are naturally embedded in an organisation. (Plsek & Wilson, 2001) These simple rules however are not simple solutions or designed to limit freedom but rather are guiding principles to assist in understanding of the system.

Rich behaviour comes from collaborating and competing agents, and therefore environments and culture needs to support interactions between stakeholders

Thinking in complexity implies shifting from meticulously crafted regulations or solutions to developing systems, environments and conditions that support interactions and enable to emergence of many solutions. Without interactions, desired

outcomes cannot emerge. Therefore, any rules or policies that restrict people's agility and freedom to interact with each other and their environment will limit the likelihood of the desired outcome to emerge. (Plsek & Wilson, 2001; Woodruff, 2019)

In medical education, although time constraints are common, pre-emptively taking control of the interactions may also limit the opportunity for informative patterns to emerge. Setting the stage, cultivating creative emergent behaviour and allowing for order, and self-organisation will support desirable fractal patterns to emerge and enhance the efficiency of the system. (Holden, 2005) Even in crisis, when the urge to apply formal structures, detailed guidelines, protocols and regulations may be strong, emergence is often seen with agents rising to the occasion, organising and adapting to the demands of the hour. (Holden, 2005) This explains why the current views on fairness in assessment are no longer entirely build on the notion of standardisation and structuring.

This also requires institutions to place trust in and to foster quality relationships with learners, on-the-ground assessors, faculty, assessment leaders and other stakeholders. Collectively, these agents can positively impact the system through the decisions they make. (Van Beurden, Kia, Zask, Dietrich & Rose, 2013) They are highly autonomous, skilled and often with significant opinions about practices or policies which have been implemented at a managerial level. Leadership should therefore refrain as much as possible from making new rules based on incidental poor behaviour but use it to further capacity build staff. This is not easy in many educational organisations as it requires management to understand that it does not have total

control over the system or the judgements the agents make, they can only influence agents to change the emergent behaviour of the systems. (Bowe & Armstrong, 2017; Kurtz & Snowden, 2003; Van Beurden, Kia, Zask, Dietrich & Rose, 2013)

Leadership inspired by complexity theory recognises that change occurs naturally within the systems and individuals engage in change for a variety of reasons. (Plsek & Wilson, 2001) The leader's role is to create systems that disseminate rich information about better practices, allowing others to adapt to those practices in ways that are most meaningful to them. (Plsek & Wilson, 2001)

Furthermore, standardising the assessment or structuring the rating forms does not typically add to fairness as these are not the source of validity and reliability evidence of the assessment. This is another fundamental shift in thinking about the origins of fairness. In the traditional, measurement-oriented view, the reliability and validity – the quality – was built into the method, e.g., the test paper. (Schuwirth & van der Vleuten, 2020) But in modern assessment these qualities are seen as a result of the interaction between the assessment method or program and the users. So, instead of structuring and standardisation, training of stakeholders – both in giving and receiving feedback for instance -, using rubrics to support judgement rather than to replace judgement, and supporting an organisational culture focused on fairness is more effective.

'Seeing' a complex system is hard. No one individual is capable of knowing all parts of a system. Seeking multiple perspectives to understand the patterns within system and

how they interact within one another is necessary. Supervisors, assessors, learners, assessment leads all hold tacit and explicit knowledge about problems at a local level and how to overcome them. (Reed, Howe, Doyle, & Bell, 2018) Acknowledging and allowing for this knowledge to be shared through interactions is important in allowing for emergence. So, whilst in testing, reliability and validity evidence are an individual feature of the test, fairness is always the result of a collaborative effort. This is one of the challenges of an educational change from a testing program to programmatic assessment, and requires strong knowledge brokers in the process. (Torre, Schuwirth, van der Vleuten & Heeneman, 2022)

Individuals learn through deliberate practice and adapting to prior experience.

A final practical implication of complexity is learning by action. Individuals adapt to past experiences, (Fraser & Greenhalgh, 2001; Van Beurden, Kia, Zask, Dietrich & Rose, 2013) and so, development of assessment expertise through deliberate practice and adaptation is essential for that learning to occur. This is more than just the standard simple online courses. Rather, it requires learning through different narratives and lived experiences, through collaborative learning or communities of practice. This requires a further change in faculty and assessor development, for example by focusing on deliberate practice including on becoming an expert in nonlinear dynamic practice and systems thinking. The roles of a coach or mentor to improve performance makes more sense from a complexity perspective than from a linear perspective. (Durning, Artino, Pangaro, van der Vleuten & Schuwirth, 2010)

If we consider again the history of assessment, the importance of developing assessor expertise can be demonstrated. Initial developments in workplace-based assessment involved seeking to find the most appropriate instruments for each competence trait to be assessed. Research was undertaken to determine how best to manipulate such assessment instruments, i.e. were open ended or multiple choice questions the most appropriate, how many points should be included in each scale and so on. (Schuwirth & van der Vleuten, 2011a) Much of this research was trying to 'solve' the problem of improving the validity of the test. However modern theories and conceptualisations of validity are shifting away from validity as a test characteristic and instead can be seen as an argument-based approach. (Boursicot, Kemp, Wilkinson, Finyartini, Canning, Cilliers & Fuller, 2021; Schuwirth & van der Vleuten, 2011a) Evidence has suggested that the development of expertise has an impact on the validity of workplace-based assessment. Expert assessors are better able to take a broader, more holistic review in interpreting learner behaviour and integrating different aspects of performance. (Govaerts, Schuwirth, van der Vleuten & Muijtjens, 2011) Empowering the assessor by giving them language which fitted their expertise also has a positive impact. For example, Weller et al. demonstrated that by changing the wording on an assessment from what could be considered as education jargon to what is now known as entrustable professional activities had a dramatic impact on the assessment. (Weller, Misur, Nicolson, Morris, Ure, Crossley & Jolly, 2014) This expertise and agility to deal with whatever situation arises makes more sense from a complexity lens than from a linear one.

In supporting assessors and learners within a complex adaptive system it important to acknowledge that individuals occupy multiple roles within various interconnected

systems. (Kurtz & Snowden, 2003) Systems are not isolated entities but are embedded within other systems. (Plsek & Greenhalgh, 2001) For example, medical schools are embedded within clinical workplaces, and both of these systems are intricately interwoven within broader society. Within these interconnected complex adaptive systems individuals assume many different roles, for example, as a student, a parent, a patient, a trainee aiming to be accepted into training and so on. Consequently, individuals will behave differently depending on their role and context. This dynamic shifting of roles can occur both individually and collectively such as a group of assessors or trainees. To appreciate an individual's behaviour, it is important to understand the diverse roles which they fulfill within these different systems. Recognising the external pressures and influences that come from these roles is needed to assist in the interpretation of their behaviour. (Kurtz & Snowden, 2003)

These implications which have been mentioned have all been summarised in the figure 8 below.

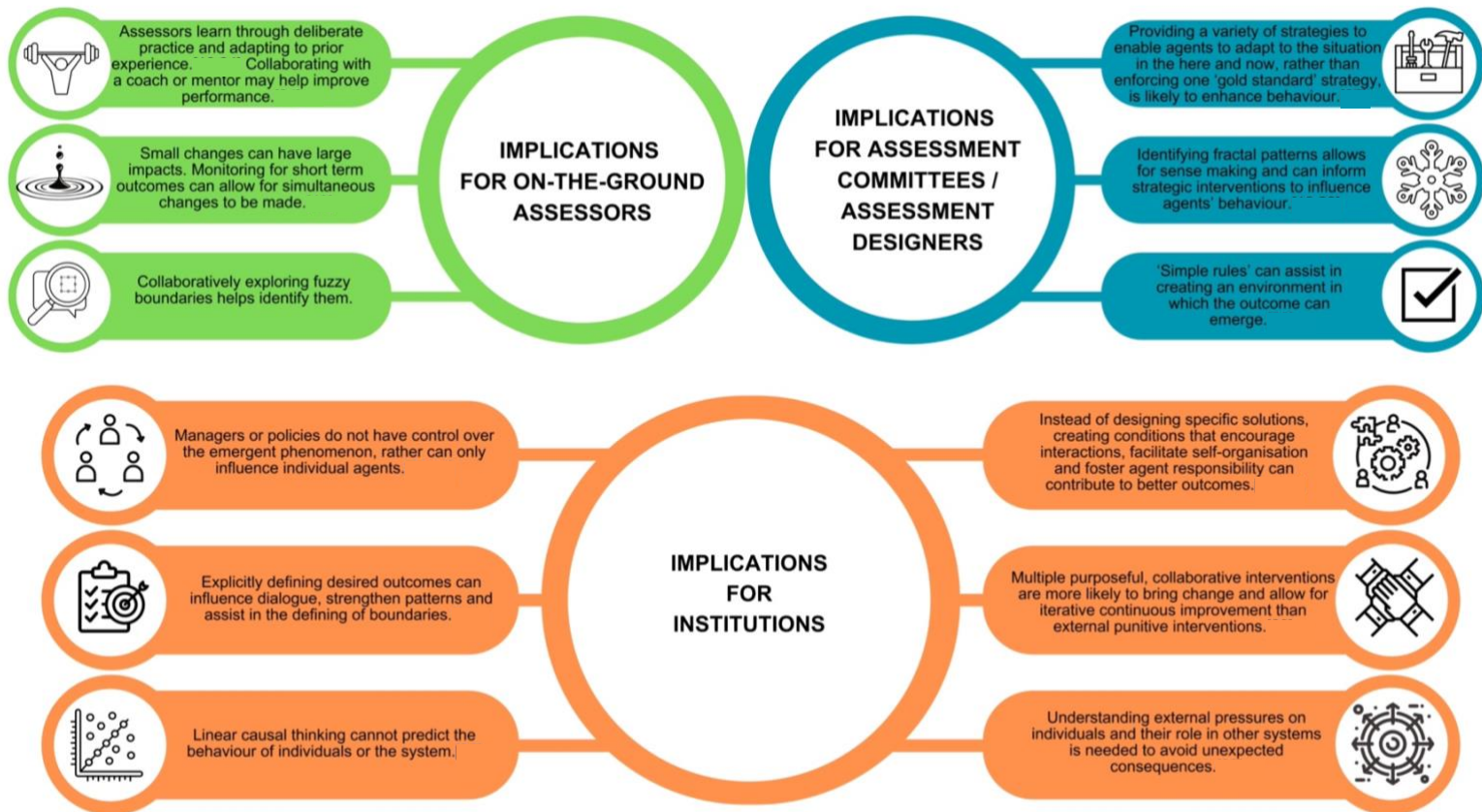


Figure 8: Implications of considering fairness with a complexity lens

Strategies to support fairness emerging in practice

Within our research, we identified strategies which can be used to support fairness emerging through strengthening and mediating the interactions between the components of fairness. It is important, however, that in line with complexity, these strategies are not used as a tick box list or a reductionist algorithm but rather as strategies and narratives to use as appropriate, simply because there is no standard recipe to create fairness, nor a one-size-fits-all solution.

These strategies have been summarised in figure 9 below.

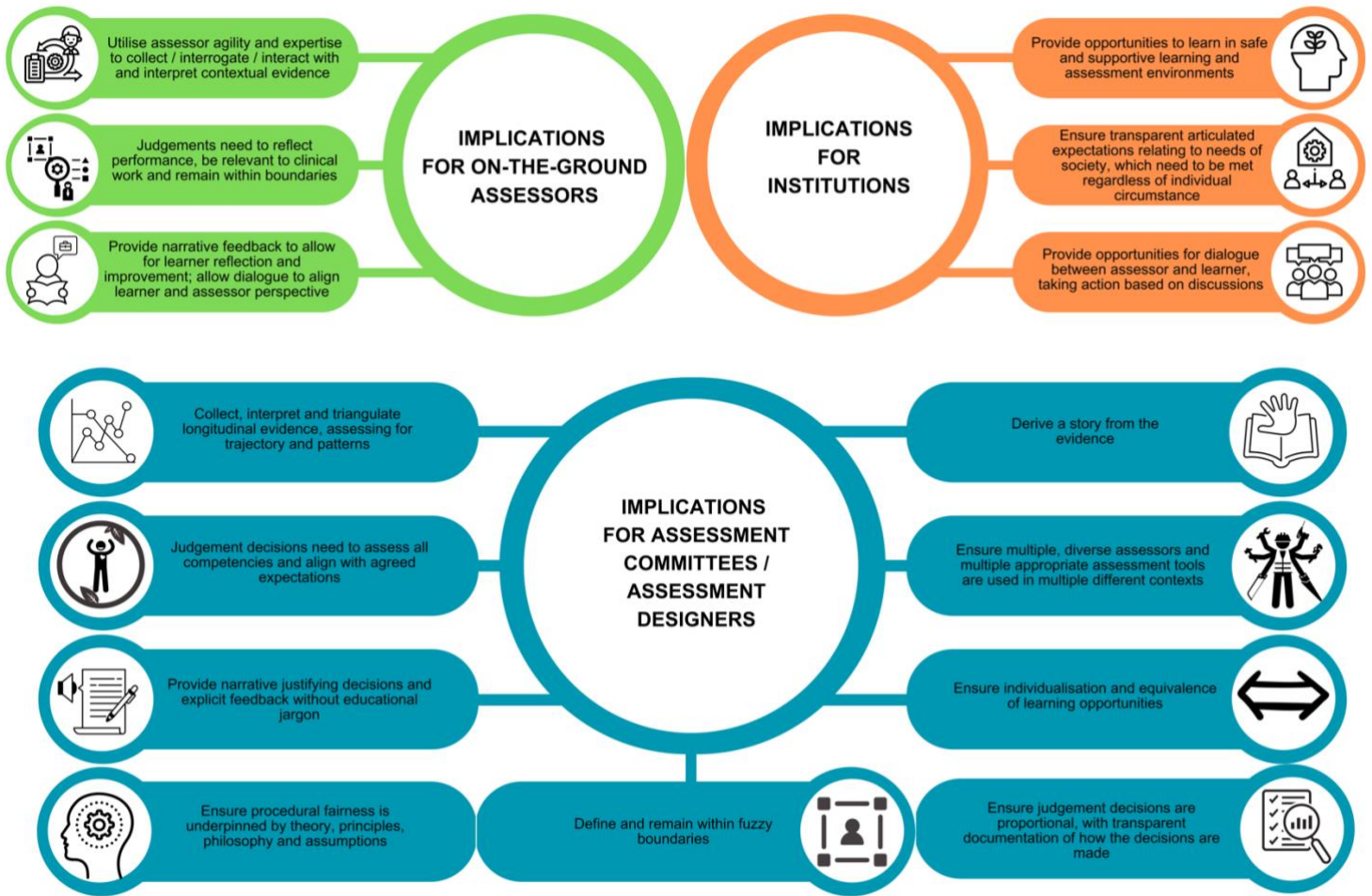


Figure 9: Strategies to facilitate the emergence of fairness through supporting the interaction of its components

Many of these strategies will be familiar. For example, programmatic assessment principles include the use of multiple data sources, meaningful triangulation of data, informed decision making and meaningful feedback to learners. (Van der Vleuten, Schuwirth, Driessen, Dijkstra, Tigelaar, Baartman & van Tartwijk, 2012) Competency frameworks explicitly articulate the dimensions or domains in which learners are guided to develop and provide a shared narrative common to all assessment programs. These expectations are used to inform judgement decisions.

Learners are regarded as active acquirers of their own knowledge, skills and competencies (Heeneman, Oudkerk Pool, Schuwirth, van der Vleuten & Driessen, 2015) and are required to harness the learning potential from their assessment. (Van der Vleuten, Schuwirth, Driessen, Dijkstra, Tigelaar, Baartman & van Tartwijk, 2012) However to accomplish this, there are several things which are required. Firstly, learners must receive sufficient meaningful information on their performance. (Heeneman, Oudkerk Pool, Schuwirth, van der Vleuten & Driessen, 2015) This will require effective communication and explicit narrative feedback, free from educational jargon. It may require a narrative justifying the decisions made or storytelling of the how decisions were made derived from the evidence collected.

In addition to receiving sufficient meaningful information on their performance, learners need to perceive assessment as a learning opportunity, and feel safe to be vulnerable and to fail. (Cilliers, Schuwirth, Adendorff, Herman & van der Vleuten, 2010; Harrison, Konings, Dannefer, Schuwirth, Wass & van der Vleuten, 2016; Watling & Ginsburg, 2019) But failing in an assessment can be distressing and dangerous (Watling & Ginsburg, 2019) and, thus, learners may not feel supported to learn or explore the boundaries of their knowledge. (Schuwirth & Ash, 2013) Creating a culture of safe learning environments is required for assessment programs to be fair to both learners and society, as it allows learners to grow and develop into future health professionals. Changing from a linear, behaviourist perspective checking only if the learner learnt enough information to a complexity perspective focussing on whether the learner has the right repertoire of strategies supports this explorative mindset in learners. This should not be confused as being lenient. Complexity or fairness, for

example, is not about letting all learners pass. This would not be fair to either society or the learner.

Procedural fairness is a multifaceted concept with various dimensions. Strategies to promote and uphold procedural fairness may include establishing a sound theoretical basis for assessment design which may include considerations of equity, equivalence of learning opportunities and individualisation of learning. Inequitable assessment is increasingly being recognised as having negative effects on learners. (Teherani, Hauer, Fernandez, King & Lucey, 2018) Linear reductionist approaches to assessment are limited in their ability to consider inequities and the unique needs of learners, and the impact this can have on society, for example through workforce distribution. A complexity lens allows assessment to be agile, fit for purpose and individualised to ensure it is equitable. A complexity approach also allows for growth and development of trainees as they have the flexibility to tailor their learning to align with their own specific needs. Using a clinical analogy, health professionals do not discharge every patient from their care with exactly the same treatment plan, instead, it they tailor treatment to meet the individual needs of each patient. However, in assessment, it is still not uncommon for students to complete identical learning and assessment tasks regardless of their unique strengths and weaknesses.

Forces which can limit fairness emerging

Unfortunately, as the study in chapter 7 noted, there are forces can limit fairness from emerging. As highlighted in this study, often barriers can be described in realist terms, for example lack of time or resourcing. Looking at these barriers through a lens of complexity, allows us to critically examine the factors which are contributing the creation and persistence of these barriers, and agilely adapt or create more levers to influence the impact these barriers have on the complex adaptive system and the emergence of fairness.

Limitations of this research

This research has identified four fractal components of fairness within the context of the global North. It is important to note that different contexts may identify additional fractal components, given this research was conducted from a social constructivist perspective and thus the intent was not to exhaustively discover all fractal components. Further research in diverse contexts should be undertaken to identify and understand these components.

Following on from this, fairness is not 'a-cultural', it is influenced by society, context and culture, (Gipps & Stobart, 2009) and so these components which have been identified are likely to be contextually and socio-culturally specific. The components of fairness are dependent on the underlying narrative and general discourse. For

example, in some cultures or situations credibility may be expected to be supported by a clear and convincing rationale, whereas in others rely more on authority or expert opinion. Institutions and organisations should, therefore, collaborate with stakeholders and allow fractal patterns to emerge which are relevant to their context and situations. Research should be undertaken to better understand how fairness emerges in these differing cultures, for example, how does it change if fairness to patients and society is prioritised over fairness towards an individual learner? Or how does it change in cultures with a predominant religious focus? Exploring these variations in different cultures could provide valuable additional insights into fairness.

Directions for future research

Complexity could be considered a threshold concept. A threshold concept is a transformed way of understanding, or interpreting, or viewing something without which a learner cannot progress. (Barradell, 2013) Once a threshold concept is understood, it changes the way that a person thinks about that topic. Once assessment has been seen through the lens of complexity, it cannot be unseen. Whilst this program of research focused on assessment, a complexity perspective could be used in other areas of health professions education allowing for other fractal patterns to be identified. Some of these have already been highlighted, for example, Cleland and colleagues argue that considering selection and widening access with a complexity lens allows for genuine reframing and consideration of different responses than previously when an elusive objective truth has been sought. (Cleland, Patterson & Hanson, 2018) There are multiple other situations, such as when new education

reforms are implemented, where a complexity perspective may be more appropriate than a linear reductionist perspective as it allows for exploration of interconnected components, adaptive behaviour, and emergent phenomena.

With regards to fairness and complexity, further research could be done around what conditions are necessary for emergence? Or understanding relationships between stakeholders such as tutors, students or social networks. What impacts interactions? How does this impact outcomes? Or feedback loops and adaptations? What impact does this have on diversity, inclusion and equity? Interdisciplinary teams? Personal adaptation within teams?

Research into understanding relevance would also be of assistance in furthering our understanding of fairness. Relevance can help define fuzzy boundaries by assisting in defining what is fit for purpose and transparent assessment. Relevance may also help in defining what is accountable to society and learners, and how an assessor determines relevance may provide information about credibility.

Whilst complexity does provide significant explanatory power to how we view fairness in assessment, there are still many unanswered questions. For example, does fairness only exist if assessment is perceived to be fair by everyone? And if so, it is reasonable to assume that there will be times in which fairness will never be achieved? So, in the absence of being able to achieve agreement on fairness, what external forces or power will determine the course of action? How will this impact on

the complex adaptive system? What consequences does this have? These are just some of the questions future research should consider.

Conclusion

In the first chapter of this thesis, I outlined some of the limitations of using an objectivity lens to view fairness. As a result of these limitations, I also highlighted some changes that have occurred in how the health professions community view assessment of competence as a result of these limitations. Looking to the future, the use of a complexity lens has significant explanatory power to enhance our understanding of fairness and assists us change how we approach the assessment of competence, as well as how we define the quality of assessment of competence.

If we view assessment as idiosyncratic, non-linear and constructed in the here and now, the complexity lens aligns with this perspective. Similarly, it aligns with the lived reality of workplaces in which medical education occurs characterised by systems which are open, in continuous evolution and encompass individuals who continuously construct reality and meaning in their lives. (Woodruff, 2021) Imperfect situations arise as a result of unpredictable circumstances, and workarounds and improvisations occur to ensure the continuity of assessment. Complexity offers an approach to these tensions. Woodruff notes 'success in complexity cannot rest on pre-planned compliance alone'. (Woodruff, 2021) Just as clinicians develop capability to handle the unknown, unpredictable and emergent; as assessors, learners and researchers we

need to do the same. (Greenhalgh & Papoutsis, 2018; Schuwirth, van der Vleuten & Durning, 2017) Engaging pragmatically with these uncertainties, rather than trying to solve them, and using a complexity lens can ensure that fairness still emerges.

Complexity thinking does not offer the promise of simple fixes or tick box lists to ensure fair assessment programs. However, it does have implications for the way we view what is quality assessment and how we design assessment programs. Replacing linear causal views with recognition and articulation of recognising and adapting to patterns will enable better understanding of what is fair assessment. This may not be a straightforward, as it lacks the familiarity of a direct line of sight from assessment development to concrete solution focused action with standard evaluation. However, a lens of complexity may lead to more purposeful, meaningful changes to assessment systems which are more aligned with 21st century assessment.

STUDENT PUBLICATIONS DURING HIGHER DEGREE RESEARCH CANDIDATURE

Valentine N, Durning S, Shanahan EM, Schuwirth L. What stops fairness from emerging in assessment? The forces on a complex adaptive system. *Perspect Med Educ.* 2023;12(1):338-347

Valentine N, Shanahan EM, Durning S, Schuwirth L. Fairness in assessment: identifying a complex adaptive system. *Perspect Med Educ.* 2023;12(1):315-26.

Valentine N, Schuwirth L. Using fairness to reconcile tensions between coaching and assessment. *Med Edu.* 2023;57(3):213-6.

Valentine N, Durning S, Shanahan EM, Van der Vleuten CM, Schuwirth L. The pursuit of fairness in assessment: Looking beyond the objective. *Med Teach.* 2022;44(4):353-9.

Valentine N, Shanahan EM, Durning S, Schuwirth L. Making it fair: Learners' and assessors' perspectives of the attributes of fair judgement. *Med Edu.* 2021;55(9):1056-66.

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):713-38.

ADDITIONAL PUBLICATION: USING FAIRNESS TO RECONCILE TENSIONS BETWEEN COACHING AND ASSESSMENT

This was an invited commentary which was completed during my higher degree research candidature.

This is the peer reviewed version of the following article: Valentine N, Schuwirth L. Using fairness to reconcile tensions between coaching and assessment. *Med Edu.* 2023;57(3):213-6 which has been published in final form at <https://doi.org/10.1111/medu.14968>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions. This article may not be enhanced, enriched or otherwise transformed into a derivative work, without express permission from Wiley or by statutory rights under applicable legislation.

This article was co-authored with Professor Lambert Schuwirth. My contribution to this publication was 70%. After receiving an invitation to contribute to write this article, I met with the journal's editor and collaborated with Professor Schuwirth regarding the content and design of the manuscript. I wrote the first draft of the manuscript and incorporated edits into the final version. Following the publication of the article I met with the editor again to record a podcast discussing the published article.

Manuscript

Coaching is dedicated to supporting learners' personal and professional development to assist them reach their potential. (Atkinson, Watling, & Brand, 2022) In contrast, traditionally, assessment has been focused on achievement or selecting out the 'bad apples'. From this perspective, coaching and assessment may seem competing tensions but actually both are essential partners of each other. Coaching without assessment can be directionless and closing the feedback loop often requires some form of assessment of whether goals have been attained. Assessment without coaching, on the other hand means the only driver for learning is behaviourist and reductionist via grades.

But for coaching and assessment to be successful partners in learning, there needs to be mutual engagement, interaction, and partnership between coach and learner. (Watling & LaDonna, 2019) There are varied methods of coaching, (Stoddard & Borges, 2016) but a core component includes coach and learner collaborating on setting individual goals based on assessment and feedback. (Lovell, 2018) The research study "*Goal co-construction and dialogue in an internal medicine longitudinal coaching program*", Farrell et al. 2022 focuses on how goal developments unfolded between coach and learner. (Farrell, Cuncic, Hartford, Hatala, & Ajjawi, 2023) This research followed eight coach-resident dyads over a twelve-month period and noted co-construction mainly occurred in how to meet goals, rather than prioritizing of goals or co-constructing new goals. This appears to be a clash between an assessment **of** and an assessment **for** learning purpose in the coaching context. On the one hand, the coach seeks to support the learner but in order to fulfil that role, they also have to

form a judgement as to the learners' progress, and strengths and weaknesses. This judgement can easily be perceived as an assessment of learning which may hamper the uptake of the feedback. (Harrison, Konings, Dannefer, Schuwirth, Wass & van der Vleuten, 2016) Navigating this dilemma requires a coaching situation to be created in which both coach and learner see the process and judgement as fair. Without this perception of fairness, coaching is likely to be ineffective.

Fairness is a fundamental quality of (health professions) education. It is often implied in assessment programs but is not explicitly articulated as there is no simple definition. (Valentine, Durning, Shanahan & Schuwirth, 2021) Just as there is no set formula for coaching, there is no set formula which can be used for fairness. Previous research into fair judgements in assessment programs showed that fairness has four key components: credibility, transparency, fitness for purpose and accountability, (Valentine, During, Shanahan, & Schuwirth, 2023) and the relevance in the context of coaching is plausible. Credibility is related not only to the judgement itself but also to the person making the judgement; (Chory, 2007) learners are more receptive to feedback coming from sources they perceive as credible. (Atkinson, Watling & Brand, 2022) There is no recipe for a credible coach, but Lovell notes coaches are expected to have expertise and experience within the relevant field. (Lovell, 2018) Assessor engagement has also been noted to be important in the learner's credibility judgements. (Valentine, Durning, Shanahan & Schuwirth, 2021)

Transparency in coaching relies on the provision of meaningful and useful feedback, enabling a shared understanding with the learner. Transparency can include a

narrative which focuses on performance improvement (Colbert, Fench, Herring & Dannefer, 2017) to ensure learners do not continue to make the same mistakes.

Coaching allows for individualisation of learning goals. Learning in the workplace is produced by engagement with authentic clinical care and shaped by individual physical, social and organisational contexts. (Govaerts & van der Vleuten, 2013)

Therefore, what is fit for purpose and fair to that individual must be determined by the coach and learner specifically to the individual contexts.

Finally, coaches have accountability to both learners and patients. Providing a culture within the coaching relationship which allows for learner agency and an opportunity to learn demonstrates accountability to learners. In addition, learners will become future health care professionals and need to the needs of the community. By ensuring coaching focuses genuinely on developing the learner to be the best professional they can be, this accountability can work both ways.

These components of fairness are not simply a tick box list. At times, these components may appear to be in tension with one another. For example, a structured form forcing assessors to make judgements in a reductionist way may seem transparent but it is not credible or fit for purpose. It may actually diminish a learner's trust in the assessor and process. (Watling, 2014b)

Like the clinical setting in which learning, assessment and coaching occur, fairness is a complex phenomenon. Using a complexity perspective is plausible and indeed encouraged within health professions education because clinical and learning environments are dynamic with numerous complex relationships and contexts. (Fraser & Greenhalgh, 2001; Mennin, 2010)

In the coaching situation, considering fair judgment as a complex adaptive system has strong explanatory power and can offer a better understanding of these tensions than linear or reductionist perspectives. Complexity holds that interactions and adaption of different components of the system are needed for an outcome to emerge.

(Greenhalgh & Papoutsi, 2018) Fairness is, therefore, created from the interactions between its components (Valentine, During, Shanahan, & Schuwirth, 2023) so there is no standard recipe to fairness, nor a one-size-fits-all solution. Expert and agile coaches have a repertoire of different strategies to support the interactions between credibility, transparency, fitness for purpose and accountability. In an assessment context, research demonstrated the types of strategies used by assessors to facilitate the interactions between the components of fairness include utilising narrative, aggregating evidence from multiple sources, procedural strategies, enabling a culture allowing for learner agency with a focus on their learning, articulating reasonable expectations of learners and ensuring a sound theoretic basis of assessment design. (Valentine, During, Shanahan, & Schuwirth, 2023) These strategies may be different in the coaching scenario and an extension of the aforementioned research study could be to review the existing 12 months of data to consider how fairness was created by the coaches in this study.

Coaching and assessment are not irreconcilable but rather partners in a learning journey, with fair judgments being the essential linchpin necessary to ensure mutual engagement and interaction between coach and learner. Counterintuitively, overly strict regulatory frameworks and tick box approaches to managing this fairness may be appealing, but they would not do justice to the complexity of the real-world clinical and learning situation. Fairness can only be created through the interactions of its different components, facilitated by different strategies.

Just as clinicians develop capability to handle the unknown, unpredictable and emergent; as coaches, learners and researchers we need to do the same.

(Greenhalgh & Papoutsi, 2018) So whilst complexity thinking does not provide simple fixes, it does have implications for coach and learner training and the way coaching programs are designed.

BIBLIOGRAPHY

Apramian, T., Cristancho, S., Sener, A., & Lingard, L. (2018). How Do Thresholds of Principle and Preference Influence Surgeon Assessments of Learner Performance? *Annals of Surgery, 268*(2), 385-390.

<https://doi.org/10.1097/SLA.0000000000002284>

Apramian, T., Cristancho, S., Watling, C., Ott, M., & Lingard, L. (2015). Thresholds of Principle and Preference: Exploring Procedural Variation in Postgraduate Surgical Education. *Academic Medicine, 90*(11, Suppl.), S70-S76.

<https://doi.org/10.1097/ACM.0000000000000909>

Apramian, T., Cristancho, S., Watling, C., Ott, M., & Lingard, L. (2016a). "Staying in the Game": How Procedural Variation Shapes Competence Judgments in Surgical Education. *Academic Medicine, 91*(11, Suppl.), S37-S43.

<https://doi.org/10.1097/ACM.0000000000001364>

Apramian, T., Cristancho, S., Watling, C., Ott, M., & Lingard, L. (2016b). "They Have to Adapt to Learn": Surgeons' Perspectives on the Role of Procedural Variation in Surgical Education. *Journal of Surgical Education, 73*(2), 339-347.

<https://doi.org/10.1016/j.jsurg.2015.10.016>

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. American Educational Research Association.

- Atkinson, A., Watling, C. J., & Brand, P. L. P. (2022). Feedback and coaching. *European Journal of Pediatrics*, 181(2), 441-446. <https://doi.org/10.1007/s00431-021-04118-8>
- Bacon, R., Holmes, K., & Palermo, C. (2016). Exploring subjectivity in competency-based assessment judgements of assessors. *Nutrition & Dietetics*, 74(4), 357-364. <https://doi.org/10.1111/1747-0080.12326>
- 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), 57-72. <https://doi.org/10.11157/fohpe.v16i3.86>
- Bacon, R., Williams, L. T., Grealish, L., & Jamieson, M. (2015). Competency-based assessment for clinical supervisors: design-based research on a web-delivered program. *JMIR Research Protocols*, 4(1), Article e26. <https://doi.org/10.2196/resprot.3893>
- Barradell, S. (2013). The identification of threshold concepts: A review of theoretical complexities and methodological challenges. *Higher Education*, 65, 265-276. <https://doi.org/10.1007/s10734-012-9542-3>
- Beckett, D. (2008). Holistic competence: Putting judgements first. *Asia Pacific Education Review*, 9(1), 21-30. <https://doi.org/10.1007/BF03025822>
- Berendonk, C., Stalmeijer, R. E., & Schuwirth, L. W. (2013). Expertise in performance assessment: assessors' perspectives. *Advances in Health Sciences Education*, 18(4), 559-571. <https://doi.org/10.1007/s10459-012-9392-x>

- Bergman, E., de Feijter, J., Frambach, J., Godefrooij, M., Slootweg, I., Stalmeijer, R., & van der Zwet, J. (2012). AM Last Page: A Guide to Research Paradigms Relevant to Medical Education. *Academic Medicine*, 87(4), 545.
<https://doi.org/10.1097/ACM.0b013e31824fbc8a>
- Bleakley, A., & Cleland, J. (2015). Sticking with messy realities: how 'thinking with complexity' can inform healthcare education research. In J. Cleland, & S. J. Durning (Eds.), *Researching Medical Education* (pp.81-92). The Association for the Study of Medical Education. <https://doi.org/10.1002/9781118838983.ch8>
- Boell, S. K., & Cecez-Kecmanovic, D. (2010). Literature Reviews and the Hermeneutic Circle. *Australian Academic & Research Libraries*, 41(2), 129-144.
<https://doi.org/10.1080/00048623.2010.10721450>
- Boell, S. K., & Cecez-Kecmanovic, D. (2014). A hermeneutic approach for conducting literature reviews and literature searches. *Communications of the Association for Information Systems*, 34(12), 257-286. <https://doi.org/10.17705/1CAIS.03412>
- Booth, C. M., & Eisenhauer, E. A. (2012). Progression-free survival: meaningful or simply measurable? *Journal of Clinical Oncology*, 30(10), 1030-1033.
<https://doi.org/10.1200/JCO.2011.38.7571>
- Boreham, N. C. (1994). The dangerous practice of thinking. *Medical Education*, 28(3), 172-179. <https://doi.org/10.1111/j.1365-2923.1994.tb02695.x>
- Boshuizen, H. P., & Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cognitive Science*, 16(2), 153-184. https://doi.org/10.1207/s15516709cog1602_1

Boud, D. (1990). Assessment and the promotion of academic values. *Studies in Higher Education*, 15(1), 101-111.

<https://doi.org/10.1080/03075079012331377621>

Boulet, J. R., & Durning, S. J. (2019). What we measure ... and what we should measure in medical education. *Medical Education*, 53(1), 86-94.

<https://doi.org/10.1111/medu.13652>

Boursicot, K. (2020). *Consensus Statement Reports: Performance Assessment*. Paper presented at the Ottawa 2020, Kuala Lumpur, Malaysia.

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

<https://doi.org/10.1080/0142159X.2020.1830052>

Bowe, C. M., & Armstrong, E. (2017). Assessment for Systems Learning: A Holistic Assessment Framework to Support Decision Making Across the Medical Education Continuum. *Academic Medicine*, 92(5), 585-592.

<https://doi.org/10.1097/ACM.0000000000001321>

Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol. 2. Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 57–71). American Psychological Association.

<https://doi.org/10.1037/13620-004>

Bullock, J. L., Lai, C. J., Lockspeiser, T., O'Sullivan, P. S., Aronowitz, P., Dellmore, D., Cha-Chi, F., Knight, C., Hauer, K. E. (2019). In pursuit of Honors: a multi-

institutional study of students' perceptions of clerkship evaluation and grading. *Academic Medicine*, 94(11, Suppl.), S48-S56. <https://doi.org/10.1097/ACM.0000000000002905>

Bunniss, S., & Kelly, D. R. (2010). Research paradigms in medical education research. *Medical Education*, 44(4), 358-366. <https://doi.org/10.1111/j.1365-2923.2009.03611.x>

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. <https://doi.org/10.1186/s12909-014-0261-6>

Cantillon, P., & Sargeant, J. (2008). Giving feedback in clinical settings. *British Medical Journal*, 337(1). <https://doi.org/10.1136/bmj.a1961>

Carmody, J. B. (2019). On Residency Selection and the Quantitative Fallacy. *Journal of Graduate Medical Education*, 11(4), 420-421. <https://doi.org/10.4300/JGME-D-19-00453.1>

Chory, R. M. (2007). Enhancing Student Perceptions of Fairness: The relationship between Instructor Credibility and Classroom Justice. *Communication Education*, 56(1), 89-105. <https://doi.org/10.1080/03634520600994300>

Cilliers, F. J., Schuwirth, L. W., Adendorff, H. J., Herman, N., & Van der Vleuten, C. P. (2010). The mechanism of impact of summative assessment on medical students' learning. *Advances in Health Sciences Education*, 15, 695-715. <https://doi.org/10.1007/s10459-010-9232-9>

Cilliers, F. J., Schuwirth, L. W., Herman, N., Adendorff, H. J., & van der Vleuten, C. P. (2012a). A model of the pre-assessment learning effects of summative

assessment in medical education. *Advances in Health Sciences Education*, 17, 39-53. <https://doi.org/10.1007/s10459-011-9292-5>

Cleland, J., & Durning S. J. (2015). *Researching Medical Education*. The Association for the Study of Medical Education. <https://doi.org/10.1002/9781118838983.ch8>

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 Education*, 42(8), 800-809. <https://doi.org/10.1111/j.1365-2923.2008.03113.x>

Cleland, J. A., Patterson, F., & Hanson, M. D. (2018). Thinking of selection and widening access as complex and wicked problems. *Medical Education*, 52(12), 1228-1239. <https://doi.org/10.1111/medu.13670>

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*, 5(1), 10-15. <https://doi.org/10.1080/10401339309539580>

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. <https://doi.org/10.4300/JGME-D-14-00448.1>

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. <https://doi.org/10.1007/s40037-017-0359-8>

- General Medical Council. (2017). *Designing and maintaining postgraduate assessment programmes*. General Medical Council. https://www.gmc-uk.org/-/media/documents/designing-and-maintaining-postgraduate-assessment-programmes-2109_pdf-70434370.pdf
- Cristancho, S., Field, E., & Lingard, L. (2019). What is the state of complexity science in medical education research? *Medical Education*, *53*(1), 95-104. <https://doi.org/10.1111/medu.13651>
- Cristancho, S. M., & Taylor, T. (2019). The agility of ants: lessons for grappling with complexity in health care teamwork. *Medical Education*, *53*(9), 855-857. <https://doi.org/10.1111/medu.13937>
- 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. <https://doi.org/10.1111/j.1365-2923.2011.04166.x>
- 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. <https://doi.org/10.1111/j.1088-4963.1997.tb00082.x>
- Dauphinee, W. D. (1995). Assessing clinical performance: where do we stand and what might we expect? *The Journal of the American Medical Association*, *274*(9), 741-743. <https://doi.org/10.1001/jama.1995.03530090073025>
- Davidoff, F., Dixon-Woods, M., Leviton, L., & Michie, S. (2015). Demystifying theory and its use in improvement. *BMJ Quality & Safety*, *24*(3), 228-238. <https://doi.org/10.1136/bmjqs-2014-003627>

Davis, B., & Sumara, D. J. (1997). Cognition, complexity, and teacher education. *Harvard Educational Review*, 67(1), 105-125.
<https://www.proquest.com/scholarly-journals/cognition-complexity-teacher-education/docview/212252964/se-2>

Delandshere, G., & Petrosky, A. R. (1994). Capturing Teachers' Knowledge: Performance Assessment: a) and Post-Structuralist Epistemology b) From a Post-Structuralist Perspective c) and Post-Structuralism d) None of the Above. *Educational Researcher*, 23(5), 11-18.
<https://doi.org/10.3102/0013189X023005011>

Denzin, N. K., & Lincoln, Y. S. (2017). *The SAGE Handbook of Qualitative Research* (5th ed). Sage.

Desy, J., Coderre, S., Davis, M., Cusano, R., & McLaughlin, K. (2019). How can we reduce bias during an academic assessment reappraisal? *Medical Teacher*, 41(11), 1315-1318. <https://doi.org/10.1080/0142159X.2019.1638503>

Dijksterhuis, M. G. K., Voorhuis, M., Teunissen, P. W., Schuwirth, L. W. T., ten Cate, O. T. J., Braat, D. D. M., & Scheele, F. (2009). Assessment of competence and progressive independence in postgraduate clinical training. *Medical Education*, 43(12), 1156-1165. <https://doi.org/10.1111/j.1365-2923.2009.03509.x>

Dijkstra, J., Galbraith, R., Hodges, B. D., McAvoy, P. A., McCrorie, P., Southgate, L. J., van der Vleuten, C. P., Wass, V., Schuwirth, L. W. (2012). Expert validation of fit-for-purpose guidelines for designing programmes of assessment. *BMC Medical Education*, 12(20)<https://doi.org/10.1186/1472-6920-12-20>

- Dijkstra, J., van der Vleuten, C. P., & Schuwirth, L. W. (2010). A new framework for designing programmes of assessment. *Advances in Health Sciences Education*, 15(3), 379-393. <https://doi.org/10.1007/s10459-009-9205-z>
- Downie, R., & Macnaughton, J. (2009). In defence of professional judgement. *Advances in psychiatric treatment*, 15(5), 322-327. *Advances in psychiatric treatment: the Royal College of Psychiatrists' journal of continuing professional development*, 15(5), 322-327.
- Downie, R., & Macnaughton, J. (2013). In defence of professional judgement. *Advances in psychiatric treatment: the Royal College of Psychiatrists' journal of continuing professional development*, 15(5), 322-327.
- Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Medical Education*, 38(9), 1006-1012. <https://doi.org/10.1111/j.1365-2929.2004.01932.x>
- Driessen, E., van der Vleuten, C., Schuwirth, L., van Tartwijk, J., & Vermunt, J. (2005). The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. *Medical Education*, 39(2), 214-220. <https://doi.org/10.1111/j.1365-2929.2004.02059.x>
- Duffield, K., & Spencer, J. (2002). A survey of medical students' views about the purposes and fairness of assessment. *Medical Education*, 36(9), 879-886. <https://doi.org/10.1046/j.1365-2923.2002.01291.x>
- Durning, S. J., Artino, A. R., Jr., Pangaro, L. N., van der Vleuten, C., & Schuwirth, L. (2010). Perspective: redefining context in the clinical encounter: implications for research and training in medical education. *Academic Medicine*, 85(5), 894-901. <https://doi.org/10.1097/ACM.0b013e3181d7427c>

Durning, S. J., Artino, A. R., Jr., Schuwirth, L., & van der Vleuten, C. (2013). Clarifying assumptions to enhance our understanding and assessment of clinical reasoning. *Academic Medicine*, 88(4), 442-448.

<https://doi.org/10.1097/ACM.0b013e3182851b5b>

Durning, S. J., Hanson, J., Gilliland, W., McManigle, J. M., Waechter, D., & Pangaro, L. N. (2010). Using qualitative data from a program director's evaluation form as an outcome measurement for medical school. *Military Medicine*, 175(6), 448-452.

<https://doi.org/10.7205/MILMED-D-09-00044>

Epstein, R. M. (2013). Whole mind and shared mind in clinical decision-making.

Patient Education and Counselling, 90(2), 200-206.

<https://doi.org/10.1016/j.pec.2012.06.035>

Ericsson, K. A. (2007). An expert-performance perspective of research on medical expertise: the study of clinical performance. *Medical Education*, 41(12), 1124-1130. <https://doi.org/10.1111/j.1365-2923.2007.02946.x>

Eva, K. W. (2003). On the generality of specificity. *Medical Education*, 37(7), 587-588.

<https://doi.org/10.1046/j.1365-2923.2003.01563.x>

Eva, K. W. (2008). On the limits of systematicity. *Medical Education*, 42(9), 852-853.

<https://doi.org/10.1111/j.1365-2923.2008.03140.x>

Eva, K. W. (2009). Broadening the debate about quality in medical education research. *Medical Education*, 43(4), 294-296. <https://doi.org/10.1111/j.1365-2923.2009.03342.x>

Eva, K. W. (2015). Moving beyond childish notions of fair and equitable. *Medical Education*, 49(1), 1-3. <https://doi.org/10.1111/medu.12640>

- Eva, K. W., Bordage, G., Campbell, C., Galbraith, R., Ginsburg, S., Holmboe, E., & Regehr, G. (2016). Towards a program of assessment for health professionals: from training into practice. *Advances in Health Sciences Education, 21*(4), 897-913. <https://doi.org/10.1007/s10459-015-9653-6>
- Eva, K. W., & Hodges, B. D. (2012). Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? *Medical Education, 46*(9), 914-919. <https://doi.org/10.1111/j.1365-2923.2012.04310.x>
- Farrell, L., Cuncic, C., Hartford, W., Hatala, R., & Ajjawi, R. (2023). Goal co-construction and dialogue in an internal medicine longitudinal coaching programme. *Medical Education, 57*(3), 265-271. <https://doi.org/10.1111/medu.14942>
- Fawns, T., & Schaepkens, S. (2022). A Matter of Trust: Online Proctored Exams and the Integration of Technologies of Assessment in Medical Education. *Teaching and Learning in Medicine, 34*(4), 444-453. <https://doi.org/10.1080/10401334.2022.2048832>
- Fawns, T., & Schuwirth, L. (2023). Rethinking the value proposition of assessment at a time of rapid development in generative artificial intelligence. *Medical Education*, Online ahead of print. <https://doi.org/10.1111/medu.15259>
- Flin, R., Youngson, G., & Yule, S. (2007). How do surgeons make intraoperative decisions? *Quality & Safety in Health Care, 16*(3), 235-239. <https://doi.org/10.1136/qshc.2006.020743>

Frambach, J. M., van der Vleuten, C. P., & Durning, S. J. (2013). AM last page.

Quality criteria in qualitative and quantitative research. *Academic Medicine*, 88(4), 552. <https://doi.org/10.1097/ACM.0b013e31828abf7f>

Frank, J. R., Snell, L. S., ten Cate, O., Holmboe, E. S., Carraccio, C., Swing, S. R., Harris, P., Glasgow, N. J., Campbell, C., Dath, D., Harden, R. M., Lobst, W., Long, D. M., Mungroo, R., Richardson, D. L., Sherbino, K., Silver, I., Taber, S., Talbot, M. & Harris, K. A. (2010). Competency-based medical education: theory to practice. *Medical Teacher*, 32(8), 638-645. <https://doi.org/10.3109/0142159X.2010.501190>

Fraser, S. W., & Greenhalgh, T. (2001). Coping with complexity: educating for capability. *British Medical Journal*, 323(7316), 799-803. <https://doi.org/10.1136/bmj.323.7316.799>

Frenk, J., Chen, L., Bhutta, Z. A., Cohen, J., Crisp, N., Evans, T., Fineberg, H., Garcia, P., Ke, Y., Kelley, P., Kistnasamy, B., Meleis, A., Naylor, D., Pablos-Mendez, A., Reddy, S., Scrimshaw, S., Sepulveda, J., Serwadda, D. & Zarayk, H. (2010). Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. *The Lancet*, 376(9756), 1923-1958. [https://doi.org/10.1016/S0140-6736\(10\)61854-5](https://doi.org/10.1016/S0140-6736(10)61854-5)

Friederichs, H., Friederichs, W. J., & März, M. (2023). ChatGPT in medical school: how successful is AI in progress testing? *Medical Education Online*, 28(1), <https://doi.org/10.1080/10872981.2023.2220920>

- Gingerich, A. (2015). *Questioning the rater idiosyncrasy explanation for error variance by searching for multiple signals within the noise*. [Doctoral dissertation, University of Northern British Columbia].
- 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. <https://doi.org/10.1111/medu.12546>
- Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-based assessments as social judgments: rethinking the etiology of rater errors. *Academic Medicine*, 86(10, Suppl.), S1-7. <https://doi.org/10.1097/ACM.0b013e31822a6cf8>
- 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. <https://doi.org/10.1097/ACM.0b013e3182a36c3d>
- 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. <https://doi.org/10.1111/medu.12637>
- 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 Sciences Education*, 21(1), 175-188. <https://doi.org/10.1007/s10459-015-9622-0>
- Ginsburg, S., van der Vleuten, C. P., Eva, K. W., & Lingard, L. (2017). Cracking the code: residents' interpretations of written assessment comments. *Medical Education*, 51(4), 401-410. <https://doi.org/10.1111/medu.13158>

Ginsburg, S., van der Vleuten, C. P. M., & Eva, K. W. (2017). The Hidden Value of Narrative Comments for Assessment: A Quantitative Reliability Analysis of Qualitative Data. *Academic Medicine*, 92(11), 1617-1621.

<https://doi.org/10.1097/ACM.0000000000001669>

Gipps, C., & Stobart, G. (2009). Fairness in Assessment. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational Assessment in the 21st Century: Connecting Theory and Practice* (pp. 105-118). Dordrecht: Springer Netherlands.

Gleick, J. (2008). *Chaos: Making a new science*. Penguin.

Glouberman, S., & Zimmerman, B. (2002). *Complicated and Complex Systems: What Would Successful Reform of Medicare Look Like? Commission on the Future of Healthcare in Canada: Discussion Paper No. 8*. Commission on the Health Care in Canada. <https://publications.gc.ca/site/eng/235920/publication.html>

Golberger, A. L. (1996). Non-linear dynamics for clinicians: chaos theory, fractals, and complexity at the bedside. *The Lancet*, 347(9011), 1312-1314.

[https://doi.org/10.1016/S0140-6736\(96\)90948-4](https://doi.org/10.1016/S0140-6736(96)90948-4)

Govaerts, M., van de Wiel, M., Schuwirth, L., van der Vleuten, C., & Muijtjens, A. (2013). Workplace-based assessment: raters' performance theories and constructs. *Advances in Health Sciences Education*, 18(3), 375-396.

<https://doi.org/10.1007/s10459-012-9376-x>

Govaerts, M., & van der Vleuten, C. P. (2013). Validity in work-based assessment: expanding our horizons. *Medical Education*, 47(12), 1164-1174.

<https://doi.org/10.1111/medu.12289>

- 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 Sciences Education, 16*(2), 151-165. <https://doi.org/10.1007/s10459-010-9250-7>
- 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 Sciences Education, 18*(3), 375-396. <https://doi.org/10.1007/s10459-012-9376-x>
- 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 Sciences Education, 12*(2), 239-260. <https://doi.org/10.1007/s10459-006-9043-1>
- 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. <https://doi.org/10.1111/medu.13656>
- Grant, J. (1999). The Incapacitating Effects of Competence: A Critique. *Advances in Health Sciences Education, 4*(3), 271-277. <https://doi.org/10.1023/A:1009845202352>
- Green, S. K., Johnson, R. L., Kim, D.-H., & Pope, N. S. (2007). Ethics in classroom assessment practices: Issues and attitudes. *Teaching and Teacher Education, 23*(7), 999-1011. <https://doi.org/10.1016/j.tate.2006.04.042>
- Greenhalgh, T., Howick, J., & Maskrey, N. (2014). Evidence based medicine: a movement in crisis? *British Medical Journal, 348*, (g3725). <https://doi.org/10.1136/bmj.g3725>

Greenhalgh, T., & Hurwitz, B. (1999a). Why study narrative? *The Western Journal of Medicine*, 170(6), 367-369.

Greenhalgh, T., & Hurwitz, B. (1999b). Narrative based medicine: why study narrative? *British Medical Journal*, 318(7175), 48-50.
<https://doi.org/10.1136/bmj.318.7175.48>

Greenhalgh, T., & Papoutsis, C. (2018). Studying complexity in health services research: desperately seeking an overdue paradigm shift. *BMC Medicine*, 16(1), 95. <https://doi.org/10.1186/s12916-018-1089-4>

Greenhalgh, T., A'Court, C., & Shaw, S. (2017). Understanding heart failure; explaining telehealth – a hermeneutic systematic review. *BMC Cardiovascular Disorders*, 17(1), 156. <https://doi.org/10.1186/s12872-017-0594-2>

Groarke, L. (2019, Summer). Informal Logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford Metaphysics Research Lab.
<https://plato.stanford.edu/archives/sum2019/entries/logic-informal/>.

Ham, C. (1999). Tragic choices in health care: lessons from the Child B case. *British Medical Journal*, 319(7219), 1258-1261.
<https://doi.org/10.1136/bmj.319.7219.1258>

Harden, R. M., & Gleeson, F. (1979). Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education*, 13(1), 41-54.

Harden, R. M., Lilley, P., & Patricio, M. (2015). *The Definitive Guide to the OSCE: The Objective Structured Clinical Examination as a performance assessment*. Elsevier Health Sciences.

- Harrison, C. J., Konings, K. D., Dannefer, E. F., Schuwirth, L. W., Wass, V., & van der Vleuten, C. P. (2016). Factors influencing students' receptivity to formative feedback emerging from different assessment cultures. *Perspectives on Medical Education*, 5(5), 276-284. <https://doi.org/10.1007/s40037-016-0297-x>.
- Harrison, C. J., Könings, K. D., Schuwirth, L., Wass, V., & van der Vleuten, C. (2015). Barriers to the uptake and use of feedback in the context of summative assessment. *Advances in Health Sciences Education*, 20(1), 229-245. <https://doi.org/10.1007/s10459-014-9524-6>
- Hauer, K. E., ten Cate, O., Boscardin, C. K., Iobst, W., Holmboe, E. S., Chesluk, B., Baron, R. B., O'Sullivan, P. S. (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. <https://doi.org/10.4300/JGME-D-15-00144.1>
- Hauer, K. E., Chesluk, B., Iobst, W., Holmboe, E., Baron, R. B., Boscardin, C. K., ten Cate, O., O'Sullivan, P. S. (2015). Reviewing residents' competence: a qualitative study of the role of clinical competency committees in performance assessment. *Academic Medicine*, 90(8), 1084-1092. <https://doi.org/10.1097/acm.0000000000000736>
- Hauer, K. E., & Lucey, C. R. (2019). Core Clerkship Grading: The Illusion of Objectivity. *Academic Medicine*, 94(4), 469-472. <https://doi.org/10.1097/ACM.0000000000002413>

Hays, R. B., Hamlin, G., & Crane, L. (2015). Twelve tips for increasing the defensibility of assessment decisions. *Medical Teacher*, 37(5), 433-436.

<https://doi.org/10.3109/0142159X.2014.943711>

Heeneman, S., Oudkerk Pool, A., Schuwirth, L. W. T., van der Vleuten, C. P. M., & Driessen, E. W. (2015). The impact of programmatic assessment on student learning: theory versus practice. *Medical Education*, 49(5), 487-498.

<https://doi.org/10.1111/medu.12645>

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*. Harvard Business Press.

Hilligoss, B., Young Rich, S. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics and interaction in context. *Information Processing & Management*, 44(4), 1467-1484.

<https://doi.org/10.1016/j.ipm.2007.10.001>

Hodges, B. (2013). Assessment in the post-psychometric era: learning to love the subjective and collective. *Medical Teacher*, 35(7), 564-568.

<https://doi.org/10.3109/0142159X.2013.789134>

Hodges, B. D., Ginsburg, S., Cruess, R., Cruess, S., Delport, R., Hafferty, F., Ho, M.-J., Holmboe, E., Holtman, M., Ohbu, S., Rees, C., ten Cate, O., Tsugawa, Y., Mook, W. V., Wilkinson, T., Wade, W. (2011). Assessment of professionalism: recommendations from the Ottawa 2010 Conference. *Medical Teacher*, 33(5), 354-363. <https://doi.org/10.3109/0142159X.2011.577300>

- Holbrook, M. B. (2003). Adventures in complexity: An essay on dynamic open complex adaptive systems, butterfly effects, self-organizing order, coevolution, the ecological perspective, fitness landscapes, market spaces, emergent beauty at the edge of chaos, and all that jazz. *Academy of Marketing Science Review*, 6(1), 1-184.
- Holden, L. M. (2005). Complex adaptive systems: concept analysis. *Journal of Advanced Nursing*, 52(6), 651-657. <https://doi.org/10.1111/j.1365-2648.2005.03638.x>
- Houston, D. (2002). Quality and the University: Stakeholders, boundary judgements and systems. Paper presented at the 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. *Theoretical Medicine*, 17(3), 225-241. <https://doi.org/10.1007/bf00489447>
- Jones, A. (1999). The place of judgement in competency-based assessment. *Journal of Vocational Education & Training*, 51(1), 145-160. <https://doi:10.1080/13636829900200073>
- 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. <https://doi.org/10.1136/jme.2009.035295>
- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport: ACE/Praeger.

- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319-342. <https://doi.org/10.1111/j.1745-3984.2001.tb01130.x>
- 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. <https://doi.org/10.1111/j.1365-2753.2010.01532.x>
- Katerndahl, D. A., Burge, S. K., Ferrer, R. L., Becho, J., & Wood, R. (2010). Complex dynamics in intimate partner violence: a time series study of 16 women. *The Primary Care Companion for CNS Disorders*, 12(4).
<https://doi.org/10.4088/PCC.09m00859whi>
- 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.
<https://doi.org/10.1215/03616878-1496020>
- Kogan, J. R., Conforti, L., Bernabeo, E., Lobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: a conceptual model. *Medical Education*, 45(10), 1048-1060. <https://doi.org/10.1111/j.1365-2923.2011.04025.x>
- Kogan, J. R., Conforti, L. N., Lobst, W. F., & Holmboe, E. S. (2014). Reconceptualizing variable rater assessments as both an educational and clinical care problem. *Academic Medicine*, 89(5), 721-727.
<https://doi.org/10.1097/ACM.0000000000000221>
- Kogan, J. R., Hess, B. J., Conforti, L. N., & Holmboe, E. S. (2010). What Drives Faculty Ratings of Residents' Clinical Skills? The Impact of Faculty's Own

Clinical Skills. *Academic Medicine*, 85(10, Suppl.), S25-28.

<https://doi.org/10.1097/ACM.0b013e3181ed1aa3>

Konopasek, L., Norcini, J., & Krupat, E. (2016). Focusing on the formative: building an assessment system aimed at student growth and development. *Academic Medicine*, 91(11), 1492-1497. <https://doi.org/10.1097/ACM.0000000000001171>

Krefting, L. (1991). Rigor in Qualitative Research: The Assessment of Trustworthiness. *American Journal of Occupational Therapy*, 45(3), 214-222. <https://doi.org/10.5014/ajot.45.3.214>

Kuper, A., Reeves, S., Albert, M., & Hodges, B. D. (2007). Assessment: do we need to broaden our methodological horizons? *Medical Education*, 41(12), 1121-1123. <https://doi.org/10.1111/j.1365-2923.2007.02945.x>

Kurtz, C. F., & Snowden, D. J. (2003). The new dynamics of strategy: Sense-making in a complex and complicated world. *IBM Systems Journal*, 42(3), 462-483. <https://doi.org/10.1147/sj.423.0462>

Kusnanto, H., Agustian, D., & Hilmanto, D. (2018). Biopsychosocial model of illnesses in primary care: A hermeneutic literature review. *Journal of Family Medicine and Primary Care*, 7(3), 497-500. https://doi.org/10.4103/jfmpc.jfmpc_145_17

Lazcano-Ponce, E., Angeles-Llerenas, A., Rodríguez-Valentín, R., Salvador-Carulla, L., Domínguez-Esponda, R., Astudillo-García, C. I., León, E. M., Katz, G. (2020). Communication patterns in the doctor–patient relationship: evaluating determinants associated with low paternalism in Mexico. *BMC Medical Ethics*, 21(125). <https://doi.org/10.1186/s12910-020-00566-3>

- Lee, J., Wu, A. S., Li, D., & Kulasegaram, K. (2021). Artificial Intelligence in Undergraduate Medical Education: A Scoping Review. *Academic Medicine*, 96(11, Suppl.), S62-S70. <https://doi.org/10.1097/ACM.0000000000004291>
- Lee, V., Brain, K., & Martin, J. (2017). Factors Influencing Mini-CEX Rater Judgments and Their Practical Implications: A Systematic Literature Review. *Academic Medicine*, 92(6), 880-887. <https://doi.org/10.1097/acm.0000000000001537>
- Lincoln, Y. S., & Guba, E. G. (2016). *The Constructivist Credo*. Routledge. <https://doi.org/10.4324/9781315418810>
- Lind, E., & Tyler, T. (1988). *The Social Psychology of Procedural Justice*. Springer. <https://doi.org/10.1007/978-1-4899-2115-4>
- Lind, E. A., & Van den Bos, K. (2002). When fairness works: Toward a general theory of uncertainty management. *Research in Organizational Behavior*, 24(1), 181-223. [https://doi.org/10.1016/S0191-3085\(02\)24006-X](https://doi.org/10.1016/S0191-3085(02)24006-X)
- Lindberg, C., Nash, S., & Lindberg, C. (2008). *On the Edge: Nursing in the Age of Complexity*. CreateSpace.
- Lingard, L. (2007). Qualitative Research in the RIME Community: Critical Reflections and Future Directions. *Academic Medicine*, 82(10, Suppl.), S129-S130. <https://doi.org/10.1097/ACM.0b013e318140593e>
- 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. <https://doi.org/10.1002/bdm.381>

- Lipsitz, L. A., & Goldberger, A. L. (1992). Loss of 'complexity' and aging. Potential applications of fractals and chaos theory to senescence. *The Journal of the American Medical Association*, 267(13), 1806-1809.
- Lockyer, J., Carraccio, C., Chan, M.-K., Hart, D., Smee, S., Touchie, C., Holmboe, E. S., Frank, J. R., ICBME Collaborators. (2017). Core principles of assessment in competency-based medical education. *Medical Teacher*, 39(6), 609-616.
<https://doi.org/10.1080/0142159X.2017.1315082>
- Lombardo, M. M., & Eichinger, R. W. (2000). High potentials as high learners. *Human Resource Management*, 39(4), 321-329. [https://doi.org/10.1002/1099-050X\(200024\)39:4<321::AID-HRM4>3.0.CO;2-1](https://doi.org/10.1002/1099-050X(200024)39:4<321::AID-HRM4>3.0.CO;2-1)
- Long, K. M., McDermott, F., & Meadows, G. N. (2018). Being pragmatic about healthcare complexity: our experiences applying complexity theory and pragmatism to health services research. *BMC Medicine*, 16(1), 94.
<https://doi.org/10.1186/s12916-018-1087-6>
- Lovell, B. (2018). What do we know about coaching in medical education? A literature review. *Medical Education*, 52(4), 376-390. <https://doi.org/10.1111/medu.13482>
- Lucey, C., & Souba, W. (2010). Perspective: The Problem With the Problem of Professionalism. *Academic Medicine*, 85(6), 1018-1024.
<https://doi.org/10.1097/ACM.0b013e3181dbe51f>
- Regher, G., MacRae, H., Reznick, R. K., Szalay, D. (1998). Comparing The Psychometric Properties of Checklists and Global Rating Scales for Assessing Performance on an OSCE-format Examination. *Academic Medicine*, 73(9), 993-997. <https://doi.org/10.1097/00001888-199809000-00020>

- Mann, K., & MacLeod, A. (2015). Constructivism: learning theories and approaches to research. In J. Cleland and S.J. Durning (Eds.), *Researching Medical Education*, (pp. 49-66). Wiley-Blackwell. <https://doi.org/10.1002/9781118838983.ch6>
- Marchese, M. C. (1992). Clinical versus actuarial prediction: A review of the literature. *Perceptual and Motor Skills*, 75(2), 583-594.
- Marewski, J. N., Gaissmaier, W., & Gigerenzer, G. (2010). Good judgments do not require complex cognition. *Cognitive Processing*, 11(2), 103-121.
<https://doi.org/10.1007/s10339-009-0337-0>
- Martin, S. D., McQuitty, V., & Morgan, D. N. (2019). Complexity Theory and Teacher Education. *Oxford Research Encyclopedia of Education*.
<https://doi.org/10.1093/acrefore/9780190264093.013.479>
- McCready, T. (2007). Portfolios and the assessment of competence in nursing: A literature review. *International Journal of Nursing Studies*, 44(1), 143-151.
<https://doi.org/10.1016/j.ijnurstu.2006.01.013>
- McDonald, J. A., Lai, C. J., Lin, M. Y. C., O'Sullivan, P. S., & Hauer, K. E. (2021). "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. *Academic Medicine*, 96(2), 263-270.
<https://doi.org/10.1097/acm.0000000000003730>
- McGuire, C. (1993). Perspectives in assessment. *Academic Medicine*, 68(2, Suppl.), S3-S8. <https://doi.org/10.1097/00001888-199302000-00022>
- McNamara, R. (2017). *In Retrospect: The Tragedy and Lessons of Vietnam*. Vintage.

- Mennin, S. (2010). Self-organisation, integration and curriculum in the complex world of medical education. *Medical Education*, 44(1), 20-30.
<https://doi.org/10.1111/j.1365-2923.2009.03548.x>
- 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.
- Morcke, A. M., Dornan, T., & Eika, B. (2013). Outcome (competency) based education: an exploration of its origins, theoretical basis, and empirical evidence. *Advances in Health Sciences Education*, 18(4), 851-863.
<https://doi.org/10.1007/s10459-012-9405-9>
- Moulton, C. A., Regehr, G., Mylopoulos, M., & MacRae, H. M. (2007). Slowing Down When You Should: A New Model of Expert Judgment. *Academic Medicine*, 82(10, Suppl.), S109-116. <https://doi.org/10.1097/ACM.0b013e3181405a76>
- Muller, J. (2020). The Tyranny of Metrics: On the Use and Misuse of Metrics in Medicine and Education. Paper presented at the AMEE Conference, Online.
- Murad, M. H. (2017). Clinical Practice Guidelines: A Primer on Development and Dissemination. *Mayo Clinic Proceedings*, 92(3), 423-433.
<https://doi.org/10.1016/j.mayocp.2017.01.001>
- Newble, D., Hoare, J., & Sheldrake, P. (1980). The selection and training of examiners for clinical examinations. *Medical Education*, 14(5), 345-349.
<https://doi.org/10.1111/j.1365-2923.1980.tb02379.x>

- Newell, C. (2008). The class as a learning entity (complex adaptive system): An idea from complexity science and educational research. *SFU Educational Review*, 2(1), 5-17. <https://doi.org/10.21810/sfuer.v2i.335>
- Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., Galbraith, R., Hays, R., Kent, A., Perrot, V., Roberts, T. (2011). Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, 33(3), 206-214. <https://doi.org/10.3109/0142159X.2011.551559>
- Norcini, J., & Shea, J. (1997). The Credibility and Comparability of Standards. *Applied Measurement in Education*, 10(1), 39-59. https://doi.org/10.1207/s15324818ame1001_3
- Norcini, J. J., Blank, L. L., Arnold, G. K., & Kimball, H. R. (1995). The Mini-CEX (Clinical Evaluation Exercise): A Preliminary Investigation. *Annals of Internal Medicine*, 123(10), 795-799. <https://doi.org/10.7326/0003-4819-123-10-199511150-00008>
- Norman, G. (2011). Chaos, complexity and complicatedness: lessons from rocket science. *Medical Education*, 45(6), 549-559. <https://doi.org/10.7326/0003-4819-123-10-199511150-00008>
- Norman, G., Tugwell, P., Feightner, J., Muzzin, L. J., & Jacoby, L. (1985). Knowledge and clinical problem-solving. *Medical Education*, 19(5), 344-356. <https://doi.org/10.1111/j.1365-2923.1985.tb01336.x>

O'Mahony, S. (2017). Medicine and the McNamara fallacy. *Journal of the Royal College of Physicians of Edinburgh*, 47(3), 281-287.
<https://doi.org/10.4997/jrcpe.2017.315>

Olmos-Vega, F. M., Stalmeijer, R. E., Varpio, L., & Kahlke, R. (2023). A practical guide to reflexivity in qualitative research: AMEE Guide No. 149. *Medical Teacher*, 45(3), 241-251. <https://doi.org/10.1080/0142159X.2022.2057287>

Oudkerk Pool, A., Govaerts, M. J. B., Jaarsma, D., & Driessen, E. W. (2018). From aggregation to interpretation: how assessors judge complex data in a competency-based portfolio. *Advances in Health Sciences Education*, 23(2), 275-287. <https://doi.org/10.1007/s10459-017-9793-y>

Park, Y. S., Konge, L., & Artino, A. R. (2020). *The positivism paradigm of research. Academic Medicine*, 95(5), 690-694.
<https://doi.org/10.1097/ACM.0000000000003093>

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. <https://doi.org/10.1111/j.1365-2923.2010.03808.x>

Pearce, J., & Tavares, W. (2021). A philosophical history of programmatic assessment: tracing shifting configurations. *Advances in Health Sciences Education*, 26(4), 1291-1310. <https://doi.org/10.1007/s10459-021-10050-1>

Periyakoil, V. S. (2008). Using Metaphors in Medicine. *Journal of Palliative Medicine*, 11(6), 842-844. <https://doi.org/10.1089/jpm.2008.9885>

- Plsek, P. E., & Greenhalgh, T. (2001). Complexity science: The challenge of complexity in health care. *British Medical Journal*, 323(625), 625-628.
<https://doi.org/10.1136/bmj.323.7313.625>
- Plsek, P. E., & Wilson, T. (2001). Complexity, leadership, and management in healthcare organisations. *British Medical Journal*, 323(746), 746-749.
<https://doi.org/10.1136/bmj.323.7315.746>
- Ramani, S., Post, S. E., Konings, K., Mann, K., Katz, J. T., & van der Vleuten, C. (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.
<https://doi.org/10.1080/10401334.2016.1244014>
- Reed, J. E., Howe, C., Doyle, C., & Bell, D. (2018). Simple rules for evidence translation in complex systems: A qualitative study. *BMC Medicine*, 16(1), 92.
<https://doi.org/10.1186/s12916-018-1076-9>
- 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. <https://doi.org/10.1111/j.1365-2929.2004.02032.x>
- Rees, C. E., & Monrouxe, L. V. (2010). Theory in medical education research: how do we get there? *Medical Education*, 44(4), 334-339. <https://doi.org/10.1111/j.1365-2923.2009.03615.x>

- Reeves, S., Albert, M., Kuper, A., & Hodges, B. D. (2008). Why use theories in qualitative research? *British Medical Journal*, 337(1), Article a949.
<https://doi.org/10.1136/bmj.a949>
- Regehr, G. (2010). It's NOT rocket science: rethinking our metaphors for research in health professions education. *Medical Education*, 44(1), 31-39.
<https://doi.org/10.1111/j.1365-2923.2009.03418.x>
- Reid, T. (1850). *Essays on the intellectual powers of man*. J. Bartlett.
- Rieh, S. Y., & Hilligoss, B. (2008). College students' credibility judgments in the information-seeking process. *Digital Media, Youth, and Credibility*, 49-72.
- Roberts, C., Khanna, P., Lane, A. S., Reimann, P., & Schuwirth, L. (2021). Exploring complexities in the reform of assessment practice: a critical realist perspective. *Advances in Health Sciences Education*, 26(5), 1641-1657.
<https://doi.org/10.1007/s10459-021-10065-8>
- 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.
<https://doi.org/10.1002/tl.37219966608>
- Rojas, D. (2018). Operationalising complexity in health professions education. *Medical Education*, 52(12), 1216-1217. <https://doi.org/10.1111/medu.13765>

Rosas, S. R. (2017). Systems thinking and complexity: considerations for health promoting schools. *Health promotion international*, 32(2), 301-311.
<https://doi.org/10.1093/heapro/dav109>

Rotthoff, T. (2018). Standing up for Subjectivity in the Assessment of Competencies. *GMS Journal of Medical Education*, 35(3), Doc29.
<https://doi.org/10.3205/zma001175>

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, 34(2), 159-179.
<https://doi.org/10.1080/02602930801956059>

Schmidt, H. G., & Boshuizen, H. P. (1993). On acquiring expertise in medicine. *Educational psychology review*, 5(3), 205-221.
<https://doi.org/10.1007/BF01323044>

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

Schuwirth, L., Southgate, L., Page, G., Paget, N., Lescop, J., Lew, S., Wade, W. B., Baron-Maldonado, M. (2002). When enough is enough: a conceptual basis for fair and defensible practice performance assessment. *Medical education*, 36(10), 925-930. <https://doi.org/10.1046/j.1365-2923.2002.01313.x>

Schuwirth, L., van der Vleuten, C., & Durning, S. J. (2017). What programmatic assessment in medical education can learn from healthcare. *Perspectives on Medical Education*, 6(4), 211-215. <https://doi.org/10.1007/s40037-017-0345-1>

- Schuwirth, L. W., & van der Vleuten, C. P. (2006). A plea for new psychometric models in educational assessment. *Medical Education*, 40(4), 296-300.
<https://doi.org/10.1111/j.1365-2929.2006.02405.x>
- Schuwirth, L. W., & van der Vleuten, C. P. (2011a). General overview of the theories used in assessment: AMEE Guide No. 57. *Medical Teacher*, 33(10), 783-797.
<https://doi.org/10.3109/0142159X.2011.611022>
- Schuwirth, L. W., & van der Vleuten, C. P. (2011b). Programmatic assessment: from assessment of learning to assessment for learning. *Medical Teacher*, 33(6), 478-485. <https://doi.org/10.3109/0142159X.2011.565828>
- Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2020). A history of assessment in medical education. *Advances in Health Sciences Education*, 25(5), 1045-1056.
<https://doi.org/10.1007/s10459-020-10003-0>
- Shaw, S., & Nisbet, I. (2021). Attitudes to Fair Assessment in the Light of COVID-19. *Research Matters: A Cambridge Assessment publication*, 31, 6-21.
- Slavin, R. E. (2002). Evidence-Based Education Policies: Transforming Educational Practice and Research. *Educational Researcher*, 31(7), 15-21.
<https://doi.org/10.3102/0013189X031007015>
- Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, 10(2), 176-199.
<https://doi.org/10.1177/1745691615569000>
- Southgate, L., Cox, J., David, T., Hatch, D., Howes, A., Johnson, N., Jolly, B., Macdonald, E., McAvoy, P., McCrorie, P., Turner, J. (2001). The General Medical Council's Performance Procedures: peer review of performance in the

workplace. *Medical education*, 35 Suppl 1, 9-19. <https://doi.org/10.1046/j.1365-2923.2001.0350s1009.x>

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. <https://doi.org/10.1080/09638288.2017.1401675>

Stalmeijer, R. E., McNaughton, N., & Van Mook, W. N. K. A. (2014). Using focus groups in medical education research: AMEE Guide No. 91. *Medical Teacher*, 36(11), 923-939. <https://doi.org/10.3109/0142159X.2014.917165>

Stefan, S. (1993). What constitutes departure from professional judgment? *Mental and Physical Disability Law Reporter*, 17(2), 207-213.

Steiner, P. M., Atzmüller, C., & Su, D. (2016). Designing valid and reliable vignette experiments for survey research: A case study on the fair gender income gap. *Journal of Methods and Measurement in the Social Sciences*, 7(2), 52-94. <https://doi.org/https://doi.org/10.2458/v7i2.20321>

Stobart, G. (2005). Fairness in multicultural assessment systems. *Assessment in Education: Principles, Policy & Practice*, 12(3), 275-287. <https://doi.org/10.1080/09695940500337249>

Stoddard, H. A., & Borges, N. J. (2016). A typology of teaching roles and relationships for medical education. *Medical Teacher*, 38(3), 280-285. <https://doi.org/10.3109/0142159X.2015.1045848>

Storey, B., & Butler, J. (2013). Complexity thinking in PE: game-centred approaches, games as complex adaptive systems, and ecological values. *Physical Education*

and Sport Pedagogy, 18(2), 133-149.

<https://doi.org/10.1080/17408989.2011.649721>

Strathern, M. (1997). 'Improving ratings': audit in the British University system.

European review, 5(3), 305-321. doi: [https://doi.org/10.1002/\(SICI\)1234-981X\(199707\)5:3<305::AID-EURO184>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4)

Swanson, D. B., & Norcini, J. J. (1989). Factors influencing reproducibility of tests

using standardized patients. *Teaching and Learning in Medicine: An International Journal*, 1(3), 158-166. <https://doi.org/10.1080/10401338909539401>

Tavares, W., & Eva, K. W. (2013). Exploring the impact of mental workload on rater-

based assessments. *Advances in Health Sciences Education*, 18(2), 291-303. <https://doi.org/10.1007/s10459-012-9370-3>

Teherani, A., Hauer, K. E., Fernandez, A., King, T. E., Jr., & Lucey, C. (2018). How

Small Differences in Assessed Clinical Performance Amplify to Large Differences in Grades and Awards: A Cascade With Serious Consequences for Students Underrepresented in Medicine. *Academic Medicine*, 93(9), 1286-1292. <https://doi.org/10.1097/acm.0000000000002323>

Telio, S., Regehr, G., & Ajjawi, R. (2016). Feedback and the educational alliance:

examining credibility judgements and their consequences. *Medical Education*, 50(9), 933-942. <https://doi.org/10.1111/medu.13063>

Ten Cate, O. (2005). Entrustability of professional activities and competency-based

training. *Medical Education*, 39(12), 1176-1177. <https://doi.org/10.1111/j.1365-2929.2005.02341.x>

- Ten Cate, O. (2006). Trust, competence, and the supervisor's role in postgraduate training. *British Medical Journal*, 333(7571), 748-751.
<https://doi.org/10.1136/bmj.38938.407569.94>
- Ten Cate, O. (2017). Competency-Based Postgraduate Medical Education: Past, Present and Future. *GMS Journal of Medical Education*, 34(5), Doc69.
<https://doi.org/10.3205/zma001146>
- Ten Cate, O., & Billett, S. (2014). Competency-based medical education: origins, perspectives and potentialities. *Medical education*, 48(3), 325-332.
<https://doi.org/10.1111/medu.12355>
- Ten Cate, O., & Regehr, G. (2019). The Power of Subjectivity in the Assessment of Medical Trainees. *Academic Medicine*, 94(3), 333-337.
<https://doi.org/10.1097/ACM.0000000000002495>
- Ten Cate, O., & Scheele, F. (2007). Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Academic Medicine*, 82(6), 542-547. <https://doi.org/10.1097/ACM.0b013e31805559c7>
- Tierney, R. D. (2012). Fairness in classroom assessment. Sage. In J. H. McMillian (Ed.), *SAGE Handbook of Research on Classroom Assessment*, (pp. 125). Sage Publications.
- Tierney, R. D. (2013). Fairness in Classroom Assessment. In J. H. McMillian (Ed.), *SAGE Handbook of Research on Classroom Assessment*, (pp. 125). Sage Publications.

- Tierney, R. D. (2014). Fairness as a multifaceted quality in classroom assessment. *Studies in Educational Evaluation, 43*, 55-69.
<https://doi.org/10.1016/J.STUEDUC.2013.12.003>
- Tochel, C., Haig, A., Hesketh, A., Cadzow, A., Beggs, K., Colthart, I., & Peacock, H. (2009). The effectiveness of portfolios for post-graduate assessment and education: BEME Guide No 12. *Medical Teacher, 31*(4), 299-318.
<https://doi.org/10.1080/01421590902883056>
- Torre, D., Schuwirth, L., van der Vleuten, C., & Heeneman, S. (2022). An international study on the implementation of programmatic assessment: Understanding challenges and exploring solutions. *Medical Teacher, 44*(8), 928-937.
<https://doi.org/10.1080/0142159X.2022.2083487>
- Tsoukas, H., & Dooley, K. J. (2011). Introduction to the special issue: Towards the ecological style: Embracing complexity in organizational research. *Organization Studies 32*(6), 729-735 <https://doi.org/10.1177/0170840611410805>
- Turner, S., & Harder, N. (2018). Psychological Safe Environment: A Concept Analysis. *Clinical Simulation in Nursing, 18*, 47-55.
<https://doi.org/10.1016/j.ecns.2018.02.004>
- Upshur, R. E., & Colak, E. (2003). Argumentation and evidence. *Theoretical Medicine and Bioethics, 24*(4), 283-299. <https://doi.org/10.1023/A:1026006801902>
- Valentine, N., Durning, S., Shanahan, E. M., & Schuwirth, L. (2021). Fairness in human judgement in assessment: a hermeneutic literature review and conceptual framework. *Advances in Health Sciences Education, 26*(2), 713-738.
<https://doi.org/10.1007/s10459-020-10002-1>

- Valentine, N., Durning, S. J., Shanahan, E. M., & Schuwirth, L. (2023). Fairness in Assessment: Identifying a Complex Adaptive System. *Perspectives on Medical Education*, 12(1), 315-326. <https://doi.org/10.5334/pme.993>
- Valentine, N., Durning, S. J., Shanahan, E. M., van der Vleuten, C., & Schuwirth, L. (2022). The pursuit of fairness in assessment: Looking beyond the objective. *Medical Teacher*, 44(4), 353-359. <https://doi.org/10.1080/0142159X.2022.2031943>
- Valentine, N., & Schuwirth, L. (2019). Identifying the narrative used by educators in articulating judgement of performance. *Perspectives on Medical Education*, 8(2), 83-89. <https://doi.org/10.1007/s40037-019-0500-y>
- Valentine, N., Shanahan, E. M., Durning, S. J., & Schuwirth, L. (2021). Making it fair: Learners' and assessors' perspectives of the attributes of fair judgement. *Medical Education*, 55(9), 1056-1066. <https://doi.org/10.1111/medu.14574>
- Valentine, N., Durning, S., Shanahan, M., Schuwirth L. (2023). Fairness in assessment: identifying a complex adaptive system. *Perspectives on Medical Education*, 12(1):315-326. <https://doi.org/10.5334/pme.993>.
- Van Beurden, E. K., Kia, A. M., Zask, A., Dietrich, U., & Rose, L. (2013). Making sense in a complex landscape: how the Cynefin Framework from Complex Adaptive Systems Theory can inform health promotion practice. *Health Promotion International*, 28(1), 73-83. <https://doi.org/10.1093/heapro/dar089>
- 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.

<https://doi.org/10.1037//0022-3514.72.5.1034>

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. [https://doi.org/10.1037/0022-](https://doi.org/10.1037/0022-3514.79.3.355)

[3514.79.3.355](https://doi.org/10.1037/0022-3514.79.3.355)

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), 1449-1458. <https://doi.org/10.1037/0022-3514.75.6.1449>

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.

<https://doi.org/10.1111/j.1365-2923.1991.tb00036.x>

van der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional

competence: from methods to programmes. *Medical Education*, 39(3), 309-317.

<https://doi.org/10.1111/j.1365-2929.2005.02094.x>

van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar,

D., Baartman, L. K. J., & van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher*, 34(3), 205-214.

<https://doi.org/10.3109/0142159X.2012.652239>

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. <https://doi.org/10.3109/0142159X.2014.973388>

- Van Merriënboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*, 147-177. <https://doi.org/10.1007/s10648-005-3951-0>
- Van Merriënboer, J. J., & Sweller, J. (2010). Cognitive load theory in health professional education: design principles and strategies. *Medical Education, 44*(1), 85-93. <https://doi.org/10.1111/j.1365-2923.2009.03498.x>
- Varpio, L., Ajjawi, R., Monrouxe, L. V., O'Brien, B. C., & Rees, C. E. (2017). Shedding the cobra effect: problematising thematic emergence, triangulation, saturation and member checking. *Medical Education, 51*(1), 40-50. <https://doi.org/10.1111/medu.13124>
- Varpio, L., Paradis, E., Uijtdehaage, S., & Young, M. (2020). The distinctions between theory, theoretical framework, and conceptual framework. *Academic Medicine, 95*(7), 989-994. <https://doi.org/10.1097/ACM.0000000000003075>
- 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. <https://doi.org/10.1177/0141076817690713>
- Wakefield, T. H. (2013). An ontology of storytelling systemicity: Management, fractals and the Waldo Canyon fire. [Doctoral dissertation, Colorado Technical University]
- Watling, C. (2014a). Cognition, culture, and credibility: deconstructing feedback in medical education. *Perspectives on Medical Education, 3*(2), 124-128. <https://doi.org/10.1007/s40037-014-0115-2>

- Watling, C. (2014b). Unfulfilled promise, untapped potential: feedback at the crossroads. *Medical Teacher*, 36(8), 692-697.
<https://doi.org/10.3109/0142159x.2014.889812>
- 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. <https://doi.org/10.1111/j.1365-2923.2011.04126.x>
- 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. <https://doi.org/10.1111/medu.12150>
- Watling, C., Driessen, E., van der Vleuten, C. P., 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.
<https://doi.org/10.1111/medu.12235>
- 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., & LaDonna, K. A. (2019). Where philosophy meets culture: exploring how coaches conceptualise their roles. *Medical Education*, 53(5), 467-476.
<https://doi.org/10.1111/medu.13799>
- Watling, C. J., Kenyon, C. F., Zibrowski, E. M., Schulz, V., Goldszmidt, M. A., Singh, I., Maddocks, H. L., Lingard, L. (2008). Rules of engagement: residents' perceptions of the in-training evaluation process. *Academic medicine*, 83(10 Suppl), S97-100. <https://doi.org/10.1097/acm.0b013e318183e78c>

Wayne, D. B., Green, M., & Neilson, E. G. (2020). Medical education in the time of COVID-19. *Science Advances*, 6(31), pp. eabc7110
<https://doi.org/10.1126/sciadv.abc7110>

Webb, C., Endacott, R., Gray, M. A., Jasper, M. A., McMullan, M., & Scholes, J. (2003). Evaluating portfolio assessment systems: what are the appropriate criteria? *Nurse Education Today*, 23(8), 600-609. [https://doi.org/10.1016/S0260-6917\(03\)00098-4](https://doi.org/10.1016/S0260-6917(03)00098-4)

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

Whitty, C. J. (2015). What makes an academic paper useful for health policy? In: *BMC Medicine*, 13, 301. <https://doi.org/10.1186/s12916-015-0544-8>

William, D. (2001). An overview of the relationship between assessment and the curriculum. In D. Scott (Ed.) *Curriculum and Assessment*, (pp. 165-181) JAI Press

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. <https://doi.org/10.1901/jaba.1978.11-203>

Woodruff, J. N. (2019). Accounting for complexity in medical education: a model of adaptive behaviour in medicine. *Medical Education*, 53(9), 861-873.
<https://doi.org/10.1111/medu.13905>

- Woodruff, J. N. (2021). Solutionism: A study of rigour in complex systems. *Medical Education*, 55(1), 12-15. <https://doi.org/10.1111/medu.14377>
- Woolf, S. H., Grol, R., Hutchinson, A., Eccles, M., & Grimshaw, J. (1999). Clinical guidelines: potential benefits, limitations, and harms of clinical guidelines. *British Medical Journal*, 318(7182), 527-530. <https://doi.org/10.1136/bmj.318.7182.527>
- Wycliffe-Jones, K., Hecker, K. G., Schipper, S., Topps, M., Robinson, J., & Abedin, T. (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.
- Yankelovich, D. (1972). *Corporate priorities: A continuing study of the new demands on business*. Yankelovich Inc., Stamford, CT.
- Yeates, P., O'Neill, P., Mann, K., & Eva, K. (2013). Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. *Advances in Health Sciences Education*, 18(3), 325-341. <https://doi.org/10.1007/s10459-012-9372-1>
- Young, J., Williamson, M., & Egan, T. (2016). Students' reflections on the relationships between safe learning environments, learning challenge and positive experiences of learning in a simulated GP clinic. *Advances in Health Sciences Education*, 21(1), 63-77. <https://doi.org/10.1007/s10459-015-9611-3>
- Young, J. Q., van Merriënboer, J., Durning, S., & ten Cate, O. (2014). Cognitive Load Theory: implications for medical education: AMEE Guide No. 86. *Medical Teacher*, 36(5), 371-384. <https://doi.org/10.3109/0142159X.2014.889290>

APPENDIX 1: CO AUTHORSHIP APPROVALS FOR HIGHER DEGREE RESEARCH THESIS EXAMINATION



Office of Graduate Research
Room 003, Registry Building
Bedford Park, SA 5042
GPO Box 2100, Adelaide 5001 Australia
Email: hdrexams@flinders.edu.au
Phone: (08) 8201 5961
Website: <https://students.flinders.edu.au/my-course/hdr>
CRICOS Provider: 00114A

CO-AUTHORSHIP APPROVALS FOR HDR THESIS FOR EXAMINATIONS

In accordance with Clause 5, 7 and 8 in the [HDR Thesis Rules](#), a student must sign a declaration that the thesis does not contain any material previously published or written by another person except where due reference is made in the text or footnotes. There can be no exception to this rule.

- a. Publications or significant sections of publications (whether accepted, submitted or in manuscript form) arising out of work conducted during candidature may be included in the body of the thesis, or submitted as additional evidence as an appendix, on the following conditions:
 - I. they contribute to the overall theme of the work, are conceptually linked to the chapters before and after, and follow a logical sequence
 - II. they are formatted in the same way as the other chapters (i.e. not presented as reprints unless as an appendix), whether included as separate chapters or integrated into chapters
 - III. they are in the same typeface as the rest of the thesis (except for reprints included as an appendix)
 - IV. published and unpublished sections of a chapter are clearly differentiated with appropriate referencing or footnotes, and
 - V. unnecessary repetition in the general introduction and conclusion, and the introductions and conclusions of each published chapter, is avoided.
- b. Multi-author papers may be included within a thesis, provided:
 - I. the student is the primary author
 - II. there is a clear statement in prose for each publication at the front of each chapter, recording the percentage contribution of each author to the paper, from conceptualisation to realisation and documentation.
 - III. The publication adheres to Flinders [Authorship of Research Output Procedures](#), and
 - IV. each of the other authors provides permission for use of their work to be included in the thesis on the [Submission of Thesis Form](#) below.
- c. Papers where the student is not the primary author may be included within a thesis if a clear justification for the paper's inclusion is provided, including the circumstances relating to production of the paper and the student's position in the list of authors. However, it is preferable to include such papers as appendices, rather than in the main body of the thesis.

STUDENT DETAILS

Student Name	Nyoli Valentine
Student ID	2064689
College	College of Medicine and Public Health
Degree	Doctor of Philosophy
Title of Thesis	What is fairness in assessment?

CO-AUTHORSHIP APPROVALS FOR HDR THESIS EXAMINATION

PUBLICATION 1

This section is to be completed by the student and co-authors. If there are more than four co-authors (student plus 3 others), only the three co-authors with the most significant contributions are required to sign below.

Please note: A copy of this page will be provided to the Examiners.

Full Publication Details

Valentine N, Durning S, Shanahan EM, Van der Vleuten CM, Schuwirth L. The pursuit of fairness in assessment: Looking beyond the objective. Med Teach. 2022;44(4):353-9.

Section of thesis where publication is referred to

Chapter two

Student's contribution to the publication

80	%	Research design
N/A	%	Data collection and analysis
80	%	Writing and editing

Outline your (the student's) contribution to the publication:

I formulated and refined the ideas and perspectives presented in this article by engaging in collaborative discussions with my fellow authors. Additionally, I was responsible for writing the initial draft manuscript, and incorporating the suggested edits and revisions provided by the other authors. I was also responsible for the submission process, ensure the article adhered to the publication guidelines.

APPROVALS

By signing the section below, you confirm that the details above are an accurate record of the students contribution to the work.

Name of Co-Author 1 Michael Shanahan Signed  Date 22/09/2023

Name of Co-Author 2 Steven Durning Signed  Date 22-9-23

Name of Co-Author 3 Lambert Schuwirth Signed  Date 22-9-23

PUBLICATION 2

This section is to be completed by the student and co-authors. If there are more than four co-authors (student plus 3 others), only the three co-authors with the most significant contributions are required to sign below.

Please note: A copy of this page will be provided to the Examiners.

Full Publication Details

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):713-38.

Section of thesis where publication is referred to

Chapter four

Student's contribution to the publication

80	%	Research design
80	%	Data collection and analysis
85	%	Writing and editing

Outline your (the student's) contribution to the publication:

I contributed to literature review research design and conducted the literature search in accordance with the research design. I analysed the results initially independently and later engaged in collaborative discussions with fellow authors. Furthermore, I wrote the initial draft, and then incorporated edits and suggestions from the other authors.

APPROVALS

By signing the section below, you confirm that the details above are an accurate record of the students contribution to the work.

Name of Co-Author 1 Steven Durning Signed  Date 22.9.23

Name of Co-Author 2 Michael Shanahan Signed  Date 22/09/2023

Name of Co-Author 3 Lambert Schuwirth Signed  Date 22-9-23

PUBLICATION 3

This section is to be completed by the student and co-authors. If there are more than four co-authors (student plus 3 others), only the three co-authors with the most significant contributions are required to sign below.

Please note: A copy of this page will be provided to the Examiners.

Full Publication Details

Valentine N, Shanahan EM, Durning S, Schuwirth L. Making it fair: Learners' and assessors' perspectives of the attributes of fair judgement. Med Edu. 2021;55(9):1056-66

Section of thesis where publication is referred to

Chapter 5

Student's contribution to the publication

80	%	Research design
80	%	Data collection and analysis
85	%	Writing and editing

Outline your (the student's) contribution to the publication:

I was involved in the research design of this project, which included designing the interview guide and vignettes. Additionally, I recruited all of the participants and subsequently conducted all of the semi-structured interviews. I undertook the analysis of the data in collaboration with my fellow authors. I wrote the original draft manuscript, and incorporated edits and revisions from the other authors.

APPROVALS

By signing the section below, you confirm that the details above are an accurate record of the students contribution to the work.

Name of Co-Author 1	<u>Michael Shanahan</u>	Signed		Date	<u>22/09/2023</u>
Name of Co-Author 2	<u>Steven Durning</u>	Signed		Date	<u>22.9.23</u>
Name of Co-Author 3	<u>Lambert Schuwirth</u>	Signed		Date	<u>22-9-23</u>

CO-AUTHORSHIP APPROVALS FOR HDR THESIS EXAMINATION

PUBLICATION 4

This section is to be completed by the student and co-authors. If there are more than four co-authors (student plus 3 others), only the three co-authors with the most significant contributions are required to sign below.

Please note: A copy of this page will be provided to the Examiners.

Full Publication Details

Valentine N, Durning S, Shanahan EM, Schuwirth L. Fairness in assessment: identifying a complex adaptive system. *Perspect Med Educ.* 2023;12(1):315-26.

Section of thesis where publication is referred to

Chapter 6

Student's contribution to the publication

80	%	Research design
80	%	Data collection and analysis
85	%	Writing and editing

Outline your (the student's) contribution to the publication:

I was involved in the research design of this project, undertaking tasks such as designing the focus group guide, applying for ethics approval to conduct the research and identification of potential participants. I recruited all of the participants, arranged suitable times for online focus groups and subsequently facilitated all of these focus groups. I undertook data analysis in collaboration with my fellow authors and wrote the original draft manuscript. I then incorporated the suggestions, edits and revisions from the other authors.

APPROVALS

By signing the section below, you confirm that the details above are an accurate record of the students contribution to the work.

Name of Co-Author

1

Steven Durning

Signed



Date 22.9.23

Name of Co-Author 2

Michael Shanahan

Signed



Date 22/09/2023.

Name of Co-Author 3

Lambert Schuwirth

Signed



Date 22.9.23

CO-AUTHORSHIP APPROVALS FOR HDR THESIS EXAMINATION

PUBLICATION 5

This section is to be completed by the student and co-authors. If there are more than four co-authors (student plus 3 others), only the three co-authors with the most significant contributions are required to sign below.

Please note: A copy of this page will be provided to the Examiners.

Full Publication Details

Valentine N, Durning S, Shanahan EM, Schuwirth L. What stops fairness from emerging in assessment? The forces on a complex adaptive system. *Perspect Med Educ.* 2023; 12(1):338-347

Section of thesis where publication is referred to

Chapter 7

Student's contribution to the publication

80	%	Research design
80	%	Data collection and analysis
85	%	Writing and editing

Outline your (the student's) contribution to the publication:

My role in the research design of this project included collaborating with my fellow authors in the research design, developing a video to play to research participants, applying for ethics approval to undertake the research, recruiting research participants and coordinating suitable times for focus groups. Subsequently, I facilitated all of the focus groups. I also undertook data analysis in collaboration with my fellow authors. Finally, I wrote the original draft manuscript before incorporating suggested edits and revisions from the other authors. I was also responsible for the submission process, ensure the article adhered to the publication guidelines.

APPROVALS

By signing the section below, you confirm that the details above are an accurate record of the students contribution to the work.

Name of Co-Author

1

Steven Durning

Signed

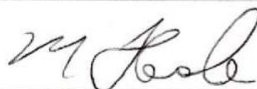


Date 22.9.23

Name of Co-Author 2

Michael Shanahan

Signed



Date 22/09/2023

Name of Co-Author 3

Lambert Schuwirth

Signed



Date 22-9-23

PUBLICATION 6

This section is to be completed by the student and co-authors. If there are more than four co-authors (student plus 3 others), only the three co-authors with the most significant contributions are required to sign below.

Please note: A copy of this page will be provided to the Examiners.

Full Publication Details

Valentine N, Schuwirth L. Using fairness to reconcile tensions between coaching and assessment. Med Edu. 2023;57(3):213-6.

Section of thesis where publication is referred to

Appendix

Student's contribution to the publication


60%	%	Research design
N/A	%	Data collection and analysis
80%	%	Writing and editing

Outline your (the student's) contribution to the publication:

After receiving an invitation to contribute to write this article, I met with the journal's editor and collaborated with Professor Schuwirth regarding the content and design of the manuscript. I wrote the first draft of the manuscript and incorporated edits into the final version. Following the publication of the article I met with the editor again to record a podcast discussing the published article.

APPROVALS

By signing the section below, you confirm that the details above are an accurate record of the students contribution to the work.

Name of Co-Author 1 Lambert Schuwirth Signed  Date 8 December 2023

Name of Co-Author 2 _____ Signed _____ Date _____

Name of Co-Author 3 _____ Signed _____ Date _____