Seeing the 'black box' differently: assessor cognition from three research perspectives

Andrea Gingerich,[1] Jennifer Kogan,[2] Peter Yeates,[3] Marjan Govaerts[4] & Eric Holmboe[5]

CONTEXT Performance assessments, such as workplace-based assessments (WBAs), represent a crucial component of assessment strategy in medical education. Persistent concerns about rater variability in performance assessments have resulted in a new field of study focusing on the cognitive processes used by raters, or more inclusively, by assessors.

METHODS An international group of researchers met regularly to share and critique key findings in assessor cognition research. Through iterative discussions, they identified the prevailing approaches to assessor cognition research and noted that each of them were based on nearly disparate theoretical frameworks and literatures. This paper aims to provide a conceptual review of the different perspectives used by researchers in this field using the specific example of WBA.

RESULTS Three distinct, but not mutually exclusive, perspectives on the origins and possible solutions to variability in assessment judgements emerged from the discussions within the group of researchers: (i) the assessor as trainable: assessors vary because they do not apply assessment criteria correctly, use varied frames of reference and make unjustified inferences; (ii) the assessor as fallible: variations arise as a result of fundamental limitations in human cognition that mean assessors are readily and haphazardly influenced by their immediate context, and (iii) the assessor as meaningfully idiosyncratic: experts are capable of making sense of highly complex and nuanced scenarios through inference and contextual sensitivity, which suggests assessor differences may represent legitimate experience-based interpretations.

CONCLUSIONS Although each of the perspectives discussed in this paper advances our understanding of assessor cognition and its impact on WBA, every perspective has its limitations. Following a discussion of areas of concordance and discordance across the perspectives, we propose a coexistent view in which researchers and practitioners utilise aspects of all three perspectives with the goal of advancing assessment quality and ultimately improving patient care.

Medical Education 2014: 48: 1055–1068
doi: 10.1111/medu.12546

Discuss ideas arising from the article at "www.mededuc.com discuss"

[1]Northern Medical Program, University of Northern British Columbia, Prince George, British Columbia, Canada
[2]Faculty of Medicine, Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, USA
[3]Faculty of Medical and Human Sciences, Centre for Respiratory Medicine and Allergy, University of Manchester, Manchester, UK
[4]Faculty of Medicine, Department of Educational Development and Research, Maastricht University, Maastricht, the Netherlands
[5]Accreditation Council for Graduate Medical Education, Chicago, Illinois, USA

Correspondence: Andrea Gingerich, Northern Medical Program, University of Northern British Columbia, 3333 University Way, Prince George, British Columbia V2N 4Z9, Canada.
Tel: + 1 250 960 5432;
E-mail: gingeri@unbc.ca

INTRODUCTION

The primary goal of medical education is to produce highly competent practitioners capable of improving the health and health care of their patients and their communities.[1,2] Much has been written about the shortcomings of the current medical education system in achieving this goal,[3,4] especially around the quality of clinical performance assessment.[5] One type of performance assessment, workplace-based assessment (WBA), incorporates the assessment of complex clinical tasks within day-to-day practice through direct observation of trainees as they authentically interact with patients in real clinical settings. Direct observation provides information and data to inform judgements about trainee progress. Workplace-based assessment has become an essential component of medical education because, ultimately, clinical supervisors must be able to determine if a trainee can be entrusted with the tasks or activities critical to the profession.[6]

Despite the importance and necessity of their use, WBA and other performance assessments have measurement limitations.[7–9] These limitations, such as low inter-rater reliability, are often attributed to flaws in assessors' judgements.[10–12] In fact, when psychometrics are used to analyse performance assessments, often a greater amount of variance in ratings can be accounted for by the assessors (i.e. rater variance) than the trainees (i.e. true score variance).[13–15] Rater or assessor cognition research is a relatively new domain in medical education which focuses on the investigation of assessors' cognitive processes and their impact on assessment quality. In this paper, the term 'assessor' will be used rather than 'rater' to emphasise that assessment involves not only rating (numerical scores), but the provision of narrative comments, feedback and supervisory decisions. By better understanding the limitations and strengths of the cognitive processes used by assessors, compatible modifications in assessment practices could be made to improve the defensibility of assessment decisions and the learning value of formative feedback exchanged with trainees, and ultimately to contribute to increased public safety.

METHODS

An international group of researchers met regularly to share and critique key findings in assessor cognition research. Through iterative discussions, the group identified the prevailing approaches to assessor cognition research and noted that each of them were based on nearly disparate theoretical frameworks and literatures. This resulted in different and sometimes contrasting implications for optimising assessment practices. Given the increasing importance of performance assessment within competency-based assessment, it seemed prudent to provide a conceptual review of the different perspectives used by researchers in this field. Using the specific example of WBA, each perspective is explored individually and then jointly to further our understanding of assessor cognition. As such, this paper is not a systematic review of any of the literatures discussed and nor is it meant to be a conclusive statement on the way assessor cognition research should progress.

RESULTS

There appear to be three distinct, although not mutually exclusive, perspectives on assessor cognition within the research community. The first perspective describes potentially controllable cognitive processes invoked during assessment and draws on components of behavioural learning theory to help frame an approach to reduce unwanted variability in assessors' assessments through faculty training. The second perspective draws on social psychology research and focuses on identifying the automatic and unavoidable biases of human cognition so that assessment systems can compensate for them. A third perspective draws from socio-cultural theory and the expertise literature and proposes that variability in judgements could provide useful assessment information within a radically different assessment design. Although there is some conceptual overlap among the perspectives, there are striking differences in the fundamental assumptions being made and the theories being used. Importantly, the first two perspectives assume that any given performance exhibits a singular 'true' standard of performance; although they differ in their explanations of assessor variability, both perspectives view it as error. Conversely, the third perspective argues that variability may arise as a result of multiple legitimately different truths, which may not represent error. It seems necessary to explicitly describe these three different perspectives on assessor cognition as their differences may have challenging implications for future assessment solutions and research. In the effort to best describe the distinctiveness of each perspective, some slight oversimplification and some degree of polarisation of the perspectives were necessary.

Perspective 1. The assessor as trainable

From this perspective, inter-assessor variability in WBA is seen as the result of assessors not 'knowing' or correctly 'applying' assessment criteria. Therefore, variability in assessment judgements reflects inaccuracy in the information provided by assessors and this variability must be minimised to improve the quality of assessment information. A viable solution to reduce variability in judgements and improve measurement outcomes in assessments is the provision of targeted training for assessors. This training would aim to improve consistency in assessment judgements by providing practice using relevant guidelines, performance criteria and an agreed-upon frame of reference to assess performances.

This perspective is partially grounded in behavioural learning theory, which assumes that trainee learning has occurred when there are observable changes in the trainee's behaviours (or actions) which can be measured and evaluated. Learning tasks can be broken down into specific measurable behaviours[16] and, by identifying specific behavioural objectives, learners can know exactly what behaviours should be performed.[17,18] Assessment criteria are used to specify how learning will be evaluated[17,18] and rigorous standards for evaluating the educational outcome can help to ensure assessment accountability.[16] Assessment then relies on deciding if the trainee met the objectives, which can be determined by detecting the expected observable behaviour in the learner. Assessment measures (i.e. scoring rubrics) are criterion-referenced in that learners are assessed according to how well they do rather than by how well they rank among their peers.[19,20]

In WBA, in which assessors observe and assess trainees with patients, assessors must be able to identify trainees' 'desired' and 'undesired' behaviours (clinical skills). Because many core clinical skills are associated with specific criteria by which quality care can be defined,[21,22] assessors should use these criteria as they observe and assess trainees. For example, best practices have been defined for many skills related to history taking, physical examination and counselling.[23–28] If the desired endpoint of medical education is based on these definitions of clinical care quality,[29,30] then best practices for care quality should inform trainee assessment, and assessors should use these quality metrics to assess clinical skills.[31] A single stimulus, the interaction between a trainee and a patient, would then ideally result in more similar responses by assessors. However, assessors often fail to appropriately use quality metrics to assess clinical skills.

Research in WBA has revealed at least three key cognitive processes used by assessors that could adversely influence assessments. One is that assessors use variable frames of reference, or standards, against which they judge trainees' performance.[32–35] 'Unsatisfactory', 'satisfactory' and 'superior' are common anchors on many assessment tools.[36] How these anchors are interpreted is very variable. For example, some assessors use these scales normatively, defining as 'satisfactory' a performance that a trainee at a particular level of training would be expected to deliver, even if they are uncertain about what skills should be expected at a given stage of training or what constitutes competence.[32,34] Another particularly prevalent frame of reference that assessors use is themselves. While observing trainees with patients, assessors commonly use their own skills as comparators (the 'self' as the frame of reference).[32,37] This is problematic for assessment because practising physicians' clinical skills may be variable, or sometimes even deficient, in core skill domains such as history taking, physical examination and counselling.[38–41] If our goal in assessment is to determine whether care meets quality standards (the observable outcome), assessors must know and be able to accurately assess a trainee for the presence or absence of skills that define high-quality patient care. They may be less able to do this if their own clinical skills are insufficient. We know that some faculty members cannot articulate what drives their assessment and can only provide a 'gut' or 'gestalt' rating.[32] For many assessors, the criteria they use to assess trainees develop experientially and different individuals subsequently come to focus on different aspects of performance, which results in variable definitions among assessors of what determines quality.[32,33] As a consequence, it is rare for assessors to explicitly apply criteria of best practice when assessing clinical performances.[32]

A second potential source of measurement error arises when assessors make inferences during direct observation (deriving what seem to be logical conclusions from premises that are assumed to be true) rather than assessing observable behaviours.[32,42] Assessors make inferences about trainees' knowledge, skills (competence) and attitudes (work ethic, emotions, intentions, personality).[32,43] Assessors do not recognise when they are making these inferences and do not validate them for accuracy.[32] Unchecked inferences risk 'distorting' the accurate assessment of the trainee because the assessor's

inferences cannot be observed and measured; this leads to greater inter-assessor variability and ultimately faulty assessment.

A third cognitive process used by assessors that might increase assessment variability is the modifying of assessment judgements to avoid unpleasant repercussions. For example, some faculty members might artificially inflate a rating so that they are not compelled to have a conversation with a trainee about a marginal assessment, whereas others focus on their roles and responsibilities as coaches and do not shy away from giving lower ratings.[32] Some may inflate assessments in order to be perceived as popular and likable teachers, but this is not true of all assessors.[32] There is also variability in how much assessors avoid stringent assessments in order to avert conversations with institutional leaders in which they may be asked to defend their assessments.[32,44,45] There are many stimuli within the culture of WBA that may lead assessors to variably synthesise what initially may have been somewhat similar observations of trainees into different assessment judgements.

From this perspective, the aforementioned sources of error can, in part, be addressed through faculty development (i.e. the assessor is trainable) and certain principles of behavioural learning theory can be invoked to support proposed 'training solutions'. Germane to behaviourism, in a competency-based training paradigm, competencies, milestones and entrustable professional activities are articulated (with the caveat that the goals and objectives do not represent only behaviours, but also knowledge and skills) and subsequently measured.[46,47] If quality patient care is the assessment endpoint, then assessment of trainees should be based upon those competencies needed to achieve the delivery of high-quality care in unsupervised practice.[29] To accomplish this, assessors will need to learn a criterion-based approach to assessment in which trainee performance is compared with pre-specified criteria that are ideally grounded in evidence-based best practices. This may entail opportunities for assessors to refresh or acquire the clinical skills they will need to assess in WBA.[48] Preliminary research suggests that helping assessors develop a criterion-referenced shared mental model may even mitigate some of the pressures related to giving constructive feedback.

Faculty development techniques such as performance dimension training would enable assessors to break down clinical skills into agreed-upon observable behaviours and to apply assessment criteria consistently. Training might also include reflection on the biases each assessor brings to the assessment tasks, as well as learning to recognise when inferences are being used. The goal would not be to prevent inferences from being made, but to help assessors develop awareness of when their judgements may be based on inferences rather than observed behaviours. Assessors could then make better judgements about the quality of the clinical skills being performed.

The end result is that some of the cognitive processes typically used by assessors to make assessment judgements may contribute to suboptimal assessments. This situation creates problems for learners, assessors and patients. Learners receive mixed messages during assessment, as well as discrepant feedback, which can interfere with their learning because there is inconsistency in what is or is not being reinforced. An assessor, in making inaccurate assessments of the trainee, may make poorly informed supervision decisions. This, in turn, is a potential threat to patient safety and care quality.

Perspective 2. The assessor as fallible

The first perspective relies on an inherent assumption that adequately resourced assessors will function like well-tuned analytical machines, and will precisely observe and compare performance against a defined standard. Logically, any difficulties with this approach should be improved through clearer frameworks or through training in more accurate observation. Yet decades of research tell us that these approaches make comparatively little difference.[49] Why? A different body of literature challenges this 'precise analytical machine' assumption. This second perspective sees assessor variability arising from fundamental limitations in human cognition. In short, low inter-rater reliability persists despite training, not because assessors are ill prepared, but because human judgement is imperfect and will always be readily influenced.

Cognitive and social psychology assert that assessors cannot simply (passively) observe and capture performances.[50] Human working memory and processing capacity are limited.[51] Information is either lost very quickly, or must be processed and linked to a person's pre-existing knowledge structures to allow it to be retained and used.[52] As a result, there can be no such thing as 'objective' observation of performance. To retain and compare information long enough to give scores and feedback, humans

necessarily interfere with what they observe. These cognitive processes are the source of many described biases,[53] and (within this perspective) the origin of problems with judgement-based assessments.

Although numerous biases in cognition exist, some illustration is useful. To make information cognitively manageable, people activate 'schemas' or networks of related information. Thus, for example, the phrase 'heart attack' might activate a web of information that contains pathophysiological concepts, typical symptoms, likely investigation findings, and treatment algorithms.[54] It might also activate a mental image of a 'typical' heart attack patient. The notion of a 'typical' patient, or person, arises from our tendency to categorise people,[55] which leaves us open to 'representativeness bias',[56] whereby, rather than effortfully processing all available information, we tend to compare key features of a new person with those of a 'typical' example of the quantity we are interested in (i.e. a 'typical' heart attack patient or a 'typical' competent trainee). By judging the similarity between the new and 'typical' people, we judge whether it is likely that the new person is indeed having a heart attack or is a competent trainee. This saves a lot of mental effort, but means we tend to ignore important information, and this can bias our judgements. This type of bias is well illustrated by the literature on stereotypes.

Once active, stereotypes (or the tendency for impressions of a person to be influenced by his membership of a group rather than his individual features), can distort which features individuals pay attention to,[57] the judgements they reach[58] and their recall of what occurs.[59] The latter is particularly important: rather than 'objectively' recalling what they have just observed, people may unconsciously 'fill in the blanks' based on what their stereotypical beliefs suggest.[60] This is particularly important in WBA because it will distort not just scores, but also the feedback given to trainees.

Importantly, the influence of stereotypes is often not under conscious control: changes in context determine which stereotypes are activated,[61] and people are often unaware of the unconscious thoughts that influence either their cognition[62] or behaviour.[63] Emotions,[64] time pressure,[59] circadian rhythms,[65] motivation, pre-existing levels of prejudice[66] and individual cognitive preferences[67] all have bearing on the degree to which stereotypes influence individual decisions, making their influence haphazard and hard to predict. Instructions to avoid stereotyping can make their influence paradoxically worse,[68] which makes it unlikely that simple training will overcome the problem.

Although this issue has been well demonstrated in social psychology, the extent to which stereotypes influence assessment judgements in medical education is unknown. However, we do know that senior doctors possess well-developed stereotypes of the way that ethnic minority students may perform or behave[69] and that, in other aspects of education, unconscious stereotyping of ethnic minorities can be seen to account for the reduced academic achievement of these students.[70] It has previously been shown that doctors judging performances of trainees are over-confident in their judgements (they are right less often than they think).[71] Judgemental overconfidence is thought to typically arise as a result of representativeness bias,[56] which suggests that these effects may well be at work in assessment judgements.

Whereas the influence of stereotypes on assessors' judgements remains to be elucidated, the influences of other biases are clearer. Humans are known to be poor at judging or scaling absolute quantities; judgements are easily influenced by contextual information[72] through processes known as assimilation or contrast effects.[73] Recently, Yeates *et al.*[34] showed that the scores given to intermediate performances in mini-clinical examination (mini-CEX) assessments are influenced to a moderately large degree by the standard of immediately preceding performances, biasing scores away from the preceding performance. A follow-up study suggested that this effect can occur across a range of performance levels, is fairly robust and that assessors may lack insight into its operation.[33]

Other authors have theoretically considered ways that categorical thinking,[55] cognitive load[74,75] or first impressions[76] might influence assessors' judgements in medical education. Although detailed empirical investigation of these claims is awaited, initial investigation has shown that examiners in objective structured clinical examinations (OSCEs) experience higher mental workload than occurs in routine clinical work.[77] Consequently there is much reason to suggest that flaws in human cognition that have been thoroughly described in other arenas are likely to influence assessor cognition in medical education.

Having noted the often unconscious and uncontrollable nature of these limitations in human judgement,

we must face the possibility that they cannot easily be overcome through either training or more detailed assessment frameworks. In fact, as making more detailed checklists might increase the cognitive load experienced by assessors, this approach could potentially (paradoxically) worsen the very problem it hopes to improve.[75]

It would be easy, therefore, to conclude that this perspective demands a nihilistic view of judgement-based assessments: judgement is flawed and cannot be fixed. It does not. Instead, it suggests that progress may lie within a toolbox of possible cognitive interventions. For example, although (as mentioned earlier) asking people to avoid stereotyping can paradoxically worsen the influence of stereotypes, an alternative approach may be more effective. Recent research indicates that people can be induced to adopt an 'egalitarian motivation' prior to making judgements of a person.[78,79] This reduced the cognitive activation of stereotypes[78,79] and lessened the influence of stereotypes on behavioural intentions and interpersonal interactions.[79] It may therefore be that improvements in judgements can be achieved by first elucidating the effect of cognitive influences on judgements, and secondly finding corresponding cognitive tools to counteract these effects.

Needless to say, much further work is required before any claims can be made about the potential benefits of these approaches. Even if these interventions are successful, they are unlikely to completely overcome contextual influences on judgements.[74] One possible implication of this perspective would be to seek ways to replace human judgement with algorithmic measurement. Recent research has shown that a computer-learning algorithm can distinguish between novice and expert laparoscopic surgeons with reasonable accuracy by measuring and analysing their hand movements. No human judgement is involved.[80] Perhaps further developments of this sort will gradually replace human judgement. This second perspective views assessor variations as the product of fundamental limitations in human cognition. Recognising this requires the medical education community to seek a set of solutions that differ from those our traditional approaches have supplied.

Perspective 3. The assessor as meaningfully idiosyncratic

In the previous two perspectives, variability in assessors' judgements is described as being *problematic* for measuring competency, for making assessment decisions and for giving feedback. Quite reasonably then, the proposed solutions aim to increase the reliability of assessors' judgements. In the third perspective, the view of assessor variability is radically different. One of its fundamental questions concerns what happens if variability, at least in part, derives from the forming by assessors of relevant and legitimate but different, and sometimes conflicting, interpretations. This perspective examines potential sources of idiosyncrasy within assessor cognition that could provide meaningful assessment information, but also lead to variability, assessor disagreement and low inter-rater reliability.

In the non-standardised reality of WBA, variance attributable to the idiosyncrasies of assessors is only outmatched by variance attributable to context specificity.[81–83] From a psychometric measurement standpoint, neither of these sources of variance reveal anything about the trainee's competence and are generally assumed to contribute to measurement error. Viewed from situated cognition theory and socio-cultural (learning) theories, however, context-specific variation is not 'error'. According to these theories, context is not an inert or interchangeable detail separate from a trainee's performance, but instead is viewed as enabling and constraining the trainee's ability to perform any intended or required skills.[84–86] This is because context is understood to encompass all the dynamic interactions between everyone and everything within an environment, and is not just a label for the physical location.[84,85,87,88] Based on this understanding of context, trainees will not have full control over the events within a clinical encounter and their competence will instead be shaped by, revealed within, and linked to that unique context.[89,90]

Viewpoints such as these make it more difficult to think of context as something to be disregarded or averaged across. They also call into question the idea of competence as something that resides solely within each trainee and remains stable across different places, patients and time.[91] On the contrary, competence has been described as being socially constructed and needing to be demonstrated and perceived by others.[92–94] The idea of perceiving others' competence is especially important for WBA because many of the key constructs that must be assessed are not directly observable.[95] Instead, constructs such as patient-centredness, professionalism, humanism and many others must be *inferred* from observable demonstrations.[89,93] Inferences are also required for making judgements of responsibility,

praise and blame[96–99] that are essential to clinical supervision decisions. If we assume that trainees' clinical performance is constructed through dynamic interactions with contexts, then there is a need for contextualised interpretations of those performances. Consequently, a WBA designed to accommodate this would require an instrument with the sensitivity to detect unpredictable changes in performance across contexts. The instrument would also need sufficient specificity to pick out key features of a performance amidst a barrage of potentially useful data. In addition, it would need the wisdom to make useful inferences and extrapolations from observed events. Fortuitously or not, such an instrument is already being used. Expert assessors could perform these tasks by making social inferences and may be essential for high-quality WBA designs.

With this in mind, it may be informative to refer to discussions about expert judgement in clinical diagnosis. Research increasingly suggests that assessor expertise resembles diagnostic expertise in the clinical domain to a remarkable extent.[43,100,101] Experienced clinicians use rapid, automatic pattern recognition to form diagnostic impressions; they very rapidly cluster sets of information into meaningful patterns, enabling fast and accurate diagnostic reasoning.[102] They do not use detailed checklists with signs and symptoms based on textbook knowledge as novices would do, and more than that, they use information reflecting (subtle) variations in the context of the patient encounter.[103] The cognitive processing used by experts is heavily reliant on the identification and interpretation of relevant contextual cues. Hofstadter writes: 'In fact, that's the main business of human brains – to take a complex situation and to put one's finger on *what matters* in it, to distil from an initial welter of sensations and ideas what a situation really is all about. To spot the gist.'[104] In addition, experts can recognise anomalies that violate expectancies, note the significance of the situation beyond the immediate events, identify what events have already taken place based on the current situation, and form expectations of events that are likely to happen also based on the current situation.[105–107] Human cognition is superb at filtering through unlimited bits of incoming data to discern relevant cues and make sense of situations.

In WBA, research findings indicate that experienced assessors are similarly able to note situation-specific cues in the assessment task, link task-specific cues to task-specific performance requirements and performance assessment, explicitly link specific aspects of trainee performance to patient behaviours and the outcome of the consultation, and form more comprehensive interpretations of performance.[42,43] Even when experienced clinical assessors are engaged in complex tasks, often under time pressures and with conflicting as well as ill-defined goals, they seem to be capable of identifying cues in trainees' performances that correlate with future performances.[100] They spot the gist.

Using humans as the assessment instrument adds additional complexity, however. Assessor expertise, as with any professional expertise, develops through immersion within specific contexts.[108] As each assessor's expertise will have been influenced by different contexts and shaped by unique experiences, different mental models of general performance, task-specific performance and person schemas might be expected, with each assessor inevitably developing a unique cognitive filter.[42,43] When interpreting performance in context, assessors will give meaning to their observations by using their past experiences and understandings of their social, cultural and contextual surroundings. Consequently, assessors may spot different 'gists' or underlying concepts within a complex performance and construct different interpretations of them.[89,109] As Delandshere and Petrosky write: '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".'[110] Variations in assessor judgements may very well represent variations in the way performance can be understood, experienced and interpreted.

From this perspective, differences in assessor judgements are not something to eliminate. This perspective does not deny that expert reasoning in performance judgements may be flawed in a manner comparable with errors in experts' diagnostic reasoning.[111] However, rather than reflecting suboptimal judgements, inconsistencies among assessors' interpretations may very well reflect the complexity of the performance and the inherently 'subjective' interpretation of that performance filtered through the assessor's understanding. If differences in assessment judgements were to come from differences in the way the trainee's performance can be perceived and experienced by others, then the inconsistencies among assessors' interpretations might be complementary and equally valid. Assessor disagreement may look less like error, for example, if many

interpretations are collected and considered as pieces of a composition that thoroughly describes the trainee's perceived competence.[112] There could be significant value in the aggregated information if it could reveal specific, context-dependent patterns of performance and performance interpretations.[93] Even contradictory judgements might be informative if judgements were collected purposefully until some type of information saturation was reached.[113] A key benefit of using saturation, rather than reliability, to analyse assessors' judgements is that it provides the power to capture pockets of repeated interpretations that may differ from the majority interpretation yet represent important variants of how that resident's behaviour can be perceived.

There are certainly other implications for WBA design. If experienced assessors are viewed as potentially important assessment instruments for WBA, then it will be important to cultivate expertise in assessors through the provision of ongoing feedback and deliberate practice in making assessment judgements. Solutions that aim to minimise assessor variability, such as checklists and the reduction of tasks into observable subcomponents, would be best avoided as they may interfere with assessors making expert judgements.[91,114,115] Assessors would probably need to provide some form of narrative assessment information to help reveal the context-dependence of their interpretations. As for trainees, because they may receive conflicting assessment information from assessors, guided reflection may help them to reconcile how others can derive an interpretation of their behaviour that differs from how it was intended. These conversations could be incorporated into an assessment culture that focuses on deliberate practice for the continual improvement of patient care and outcomes.

In the third perspective, and similarly to the second perspective, ideas of why it is unreasonable to expect different assessors to interpret the same trainee's performance in exactly the same way are shared. By contrast with the second perspective, variability has been described as a potentially useful source of assessment information that stems from assessors differently developing expertise and using expert judgement. Workplace-based assessment is filled with unpredictable assessment situations in which assessors are continuously challenged to identify critical features of context-dependent performances. Experts may be well suited for this task, but the inferences and extrapolations they use to interpret the performance may also introduce variability into their judgements. To harness their insights, radically new assessment analyses, designs and culture may be needed. Even if it were possible for assessors to be objective, this perspective would argue that it is not desirable to eliminate these differences because in compilation they may contribute to a more comprehensive understanding of the trainee's abilities.

DISCUSSION

It is important to recognise that, for the foreseeable future, WBA will be highly dependent on the judgement of humans. Few would deny that the primary goal of medical education is to produce a highly competent health care workforce to care for patients and populations and that WBA is a critical component of clinical training. The three perspectives on assessor cognition discussed above not only highlight a number of difficulties with WBA, but, more importantly, challenge some of our preconceptions of assessment and cognition. When considered separately, each proposes a reasonable and logical view of assessor cognition. However, when considered simultaneously, the three perspectives may seem initially to be irreconcilable. Instead of summarising each perspective, we will take the opportunity to highlight important commonalities that were not previously discussed before noting some points of divergence. We recognise this represents a synthesis developed through literature review and an iterative group process. Accordingly, it is not meant to cover all possible perspectives or to serve as a systematic review of the literature. Despite the challenges highlighted in this paper, we believe that WBA can be improved by integrating the areas of concordance and discordance amongst the three perspectives.

Areas of concordance

Several areas of concordance deserve elaboration. Firstly, all three perspectives require assessors to actually observe trainees interacting with patients and all recognise that the current quantity and frequency of observation-based assessment of undergraduate and postgraduate medical trainees is less than ideal. This is a serious deficiency in assessment programmes, which requires immediate attention.[36,116–124] Regardless of the perspective on assessor cognition, institutions must create clinical and medical education systems that permit, promote and sustain direct observation of trainees. Hence, the first step to improving WBA requires institutions to provide support and to ensure that faculty staff actually do it.

A second area of concordance among the three perspectives concerns the need for faculty members to achieve and maintain their own clinical competence, while concomitantly developing expertise as assessors. An impediment to assessing the quality of specific skills performed by a trainee is an assessor's lack of awareness of the specific skills required to competently perform that task. Therefore, faculty development for assessors may need to include training that refers to their own clinical skills development in addition to training in how to assess those skills.

Finally, there are two mechanisms common to each perspective that may help to maximise the strengths and minimise the weaknesses of assessor cognition. One concerns the robust sampling of tasks performed by each trainee and assessed by an equally robust sample of assessors and is intended to improve the reliability, validity, quality and defensibility of assessment decisions. The other is facilitated group discussions among assessors and assessment decision makers that provide opportunities to synthesise all available assessment data to create a clearer composite picture of a trainee's overall performance.[125] Group discussions allow both consistent and variable judgements to be explored and better understood.[126]

Areas of discordance

There are also areas of discordance, or incompatibilities, among the three perspectives that cannot be ignored. For example, whether there exists one or multiple 'truths', the goals of faculty development, the utility of making inferences and the pursuit of reliability have been previously discussed. Rather than trying to overcome the discordances and fully integrate the different perspectives into a unified theory, it may be useful to identify circumstances in which the strengths of a particular perspective may be especially advantageous.

A simple football (soccer) analogy might help to illustrate how different perspectives on assessor cognition could be purposefully matched to fundamentally different assessment situations to improve WBA. A football player must place the ball into the net in order to score a goal and anything outside the boundary of the net is a miss. The delivery of health care is similarly bounded; there are not limitless ways for trainees to provide safe, effective patient-centred care. Some clinical tasks have tighter boundaries, or a smaller 'net'. For example, the insertion of central venous catheters and the management of mechanical ventilators to prevent pneumonia should be performed within the boundaries specified by the latest evidence-based medicine or procedural checklists. Variance from the standards in these cases should be limited. Correspondingly, it would be advantageous for assessor judgements of these performances to have less variability. However, there are situations in which determining the quality of the trainee's performance depends on a larger number of contextual factors such as the patient's medical condition, needs and culture, and system factors. For example, although there are guidelines for delivering bad news (e.g. the SPIKES[127] framework), an appropriate performance under a specific combination of factors may not be appropriate in a different combination of factors. In other words, the boundary zone (i.e. the size of the net) is wider for breaking bad news than it is for central venous catheter insertion, but neither is infinite. For clinical encounters that can be highly influenced by contextual factors, an assessment system that can accommodate variability and expertise in assessors' judgements may be appropriate and valuable.

Moving forward

As the emerging field of 'assessor cognition' research grows, these perspectives will help to align and situate research, signposting ways to discuss discordant (even contradictory) empirical findings to inform and develop assessment practice. Our goal should not be to accept inherent limitations in assessor cognition as an excuse to avoid improving assessment design. Instead, we should critically reflect on and strategically incorporate both the concordant and discordant views presented by each of these perspectives to enhance the quality of our assessments.

All three perspectives will also need to account for rapidly changing clinical care delivery models, a critical contextual variable, that will substantially impact medical education. Both learning and clinical care now occur increasingly in the context of the interprofessional team and this will affect how we think about the assessment of individuals. Patients entrust faculty staff and education programmes to perform supervision and assessment in a manner that effectively meets their needs in this new context. Furthermore, it is likely that judgements of competence will be made through a group process, meaning that groups will make inferences based on others' observations and ratings. In the end, regardless of which perspective of

assessor cognition is emphasised or utilised, and of how that perspective is used, the ultimate outcome must be the same: the delivery of safe and effective patient-centred care.

Contributors: all authors were involved in developing the conceptual framework for this study and made substantive contributions to the writing and editing of the entire paper. Each author took the lead on different sections of the paper: JK for 'The assessor as trainable'; PY for 'The assessor as fallible'; AG and MG for 'The assessor as meaningfully idiosyncratic', and EH for the discussion. All authors approved the final manuscript for submission. All authors are accountable for all aspects of the work and for ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.
Acknowledgements: the authors wish to thank Lisa Conforti from the ABIM for assisting with the logistics required for a successful international collaboration; and Glenn Regehr for providing helpful comments on a previous draft.
Funding: the authors are grateful for the in-kind support provided by the American Board of Internal Medicine used to cover the costs of conference calls and other collaboration expenses.
Conflicts of interest: EH receives royalties from Mosby-Elsevier for a textbook on assessment.
Ethical approval: previous presentations: The authors discussed ideas from this paper in symposiums presented at the AAMC RIME conference in Philadelphia, PA (November 2013) and the Ottawa conference in Ottawa, ON (April 2014).

REFERENCES

1 McGaghie WC, Lipson L. *Competency-based Curriculum Development in Medical Education: An Introduction*. Geneva: World Health Organization 1978.
2 Frenk J, Chen L, Bhutta ZA *et al*. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. *Lancet* 2010;**376** (9756):1923–58.
3 Hodges BD. A tea-steeping or i-doc model for medical education? *Acad Med* 2010;**85** (9) (Suppl): 34–44.
4 Irby DM, Cooke M, O'Brien BC. Calls for reform of medical education by the Carnegie Foundation for the advancement of teaching: 1910 and 2010. *Acad Med* 2010;**85** (2):220–7.
5 van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ* 2005;**39**:309–17.
6 ten Cate TJO, Snell L, Carraccio C. Medical competence: the interplay between individual ability and the health care environment. *Med Teach* 2010;**32** (8):669–75.
7 Lurie SJ, Mooney CJ, Lyness JM. Measurement of the general competencies of the Accreditation Council for Graduate Medical Education: a systematic review. *Acad Med* 2009;**84** (3):301–9.
8 Albanese M. Rating education quality: factors in the erosion of professional standards. *Acad Med* 1999;**74** (6):652–8.
9 Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med* 2003;**15** (4): 270–92.
10 Elliot DL, Hickam DH. Evaluation of physical examination skills. Reliability of faculty observers and patient instructors. *JAMA* 1987;**258** (23):3405–8.
11 Noel GL, Herbers JE, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? *Ann Intern Med* 1992;**117** (9): 757–65.
12 Herbers JE Jr, Noel GL, Cooper GS, Harvey J, Pangaro LN, Weaver MJ. How accurate are faculty evaluations of clinical competence? *J Gen Intern Med* 1989;**4** (3):202–8.
13 Cook DA, Beckman TJ, Mandrekar JN, Pankratz VS. Internal structure of mini-CEX scores for internal medicine residents: factor analysis and generalisability. *Adv Health Sci Educ Theory Pract* 2010;**15** (5):633–45.
14 Margolis MJ, Clauser BE, Cuddy MM, Ciccone A, Mee J, Harik P, Hawkins RE. Use of the mini-clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: a validity study. *Acad Med* 2006;**81** (10 Suppl):56–60.
15 Hill F, Kendall K, Galbraith K, Crossley J. Implementing the undergraduate mini-CEX: a tailored approach at Southampton University. *Med Educ* 2009;**43**:326–34.
16 Tyler R. *Basic Principles of Curriculum and Instruction*. Chicago, IL: University of Chicago Press 1949.
17 Ertmer PA, Newby TJ. Behaviourism, cognitivism, constructivism: comparing critical features from a design perspective. *Perform Improve Q* 1993;**6** (4):50–72.
18 Saettler P. *The Evolution of American Educational Technology*. Englewood, CO: Libraries Unlimited 1990.
19 Smith P, Ragan T. *Instructional Design*, 2nd edn. New York, NY: John Wiley & Sons 1999.
20 Torre DM, Daley BJ, Sebastian JL, Elnicki DM. Overview of current learning theories for medical educators. *Am J Med* 2006;**119**:903–7.
21 Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: Institute of Medicine 2001.
22 World Health Organization. *Quality of Care: A Process for Making Strategic Choices in Health Systems*. 2006. http://www.who.int/management/quality/assurance/QualityCare_B.Def.pdf. [Accessed 22 November 2013.]

23 Lyles JS, Dwamena FC, Lein C, Smith RC. Evidence-based patient-centred interviewing. *J Clin Outcomes Manage* 2001;**8** (7):28–34.
24 Braddock CH III, Edwards KA, Hasenberg NM, Laidley TL, Levinson W. Informed decision making in outpatient practice: time to get back to basics. *JAMA* 1999;**282** (24):2313–20.
25 Murray CJ, Frenk J. Health metrics and evaluation: strengthening the science. *Lancet* 2008;**371** (9619): 1191–9.
26 Ptacek JT, Eberhardt TL. Breaking bad news. A review of the literature. *JAMA* 1996;**276** (6):496–502.
27 Searight R. Realistic approaches to counselling in the office setting. *Am Fam Physician* 2009;**79** (4):277–84.
28 McGee S. *Evidence-Based Physical Diagnosis*. St Louis, MO: Saunders Elsevier 2007.
29 Sklar DP, Lee R. Commentary: what if high-quality care drove medical education? A multiattribute approach *Acad Med* 2010;**85** (9):1401–4.
30 Asch DA, Nicholson S, Srinivas S, Herrin J, Epstein AJ. Evaluating obstetrical residency programmes using patient outcomes. *JAMA* 2009;**302** (12):1277–83.
31 Haan CK, Edwards FH, Poole B, Godley M, Genuardi FJ, Zenni EA. A model to begin to use clinical outcomes in medical education. *Acad Med* 2008;**83** (6):574–80.
32 Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: a conceptual model. *Med Educ* 2011;**45**:1048–60.
33 Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly observed performance assessments. *Adv Health Sci Educ Theory Pract* 2013a;**18** (3):325–41.
34 Yeates P, O'Neill P, Mann K, Eva KW. Effect of exposure to good vs poor medical trainee performance on attending physician ratings of subsequent performances. *JAMA* 2012;**308** (21):2226–32.
35 Yeates P, O'Neill P, Mann K, Eva KW. 'You're certainly relatively competent': assessor bias due to recent experiences. *Med Educ* 2013;**47**:910–22.
36 Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA* 2009;**302** (12):1316–26.
37 Kogan JR, Hess BJ, Conforti LN, Holmboe ES. What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. *Acad Med* 2010; **85** (10 Suppl):25–8.
38 Vukanovic-Criley JM, Criley S, Warde CM, Broker JR, Guevara-Mathews L, Churchill WH, Nelson WP, Criley JM. Competency in cardiac examination skills in medical students, trainees, physicians, and faculty: a multicentre study. *Arch Intern Med* 2006;**166** (6): 610–6.
39 Paauw DS, Wenrich MD, Curtis JR, Carline JD, Ramsey PG. Ability of primary care physicians to recognise physical findings associated with HIV infection. *JAMA* 1995;**274** (17):1380–2.
40 Ramsey PG, Wenrich MD. Use of peer ratings to evaluate physician performance. *JAMA* 1993;**269** (13): 1655–60.
41 Braddock CH III, Fihn SD, Levinson W, Jonsen AR, Pearlman RA. How doctors and patients discuss routine clinical decisions. Informed decision making in the outpatient setting. *J Gen Intern Med* 1997;**12** (6):339–45.
42 Govaerts MJB, van de Wiel MWJ, van der Vleuten CPM. Quality of feedback following performance assessments: does assessor expertise matter? *Eur J Train Dev* 2013a;**37** (1):105–25.
43 Govaerts MJB, Schuwirth L, van der Vleuten CP, Muijtjens AMM. Workplace-based assessment: effects of rater expertise. *Adv Health Sci Educ Theory Pract* 2011;**16** (2):151–65.
44 Dudek NL, Marks MB, Regehr G. Failure to fail: the perspectives of clinical supervisors. *Acad Med* 2005;**80** (10) (Suppl):84–7.
45 Cleland JA, Knight LV, Rees CE, Tracey S, Bond CM. Is it me or is it them? Factors that influence the passing of underperforming students. *Med Educ* 2008; **42**:800–9.
46 Nasca TJ, Brigham T, Philibert I, Flynn TC. The next GME accreditation system – rationale and benefits. *New Engl J Med* 2012;**366** (11):1051–6.
47 Carraccio CL, Englander R. From Flexner to competencies: reflections on a decade and the journey ahead. *Acad Med* 2013;**88** (8):1067–73.
48 Frankel RM, Eddins-Folensbee F, Inui TS. Crossing the patient-centred divide: transforming health care quality through enhanced faculty development. *Acad Med* 2011;**86** (4):445–52.
49 Landy FJ, Farr JL. Performance rating. *Psychol Bull* 1980;**87** (1):72–107.
50 Ilgen DR, Barnes-Farrell JL, McKellin DB. Performance appraisal process research in the 1980s: what has it contributed to appraisals in use? *Organ Behav Hum Dec* 1993;**54** (3):321–68.
51 Baddeley A. The magical number seven: still magical after all these years? *Psychol Rev* 1994;**101** (2):353.
52 van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies. *Med Educ* 2010;**44**:85–93.
53 Macrae CN, Bodenhausen GV. Social cognition: categorical person perception. *Br J Psychol* 2001;**92** (Pt 1):239–55.
54 Schmidt HG. Foundations of problem-based learning: some explanatory notes. *Med Educ* 1993; **27**:422–32.
55 Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgements: rethinking the aetiology of rater errors. *Acad Med* 2011;**86** (10) (Suppl):1–7.

56 Tversky A, Kahneman D. Judgement under uncertainty: heuristics and biases. *Science* 1974;**185** (4157):1124–31.
57 Bodenhausen GV, Wyer JRS. Effects of stereotypes on decision making and information-processing strategies. *J Pers Soc Psychol* 1985;**48** (2):267–82.
58 Bodenhausen GV. Stereotypic biases in social decision making and memory: testing process models of stereotype use. *J Pers Soc Psychol* 1988; **55** (5):726–37.
59 Dijksterhuis A, van Knippenberg A. Memory for stereotype-consistent and stereotype-inconsistent information as a function of processing pace. *Eur J Soc Psychol* 1995;**25** (6):689–93.
60 MacRae CN, Schloerscheidt AM, Bodenhausen GV, Milne AB. Creating memory illusions: expectancy-based processing and the generation of false memories. *Memory* 2002;**10** (1):63–80.
61 Macrae CN, Bodenhausen GV, Milne AB. The dissection of selection in person perception: inhibitory processes in social stereotyping. *J Pers Soc Psychol* 1995;**69** (3):397–407.
62 Nisbett RE, Wilson TD. Telling more than we can know: verbal reports on mental processes. *Psychol Rev* 1977;**84** (3):231–59.
63 Bargh JA, Chartrand TL. The unbearable automaticity of being. *Am Psychol* 1999;**54** (7):462–79.
64 Bodenhausen GV, Sheppard LA, Kramer GP. Negative affect and social judgement: the differential impact of anger and sadness. *Eur J Soc Psychol* 1994; **24** (1):45–62.
65 Bodenhausen GV. Stereotypes as judgemental heuristics: evidence of circadian variations in discrimination. *Psychol Sci* 1990;**1** (5):319–22.
66 Kunda Z, Spencer SJ. When do stereotypes come to mind and when do they colour judgement? A goal-based theoretical framework for stereotype activation and application. *Psychol Bull* 2003;**129** (4):522–44.
67 Crawford MT, Skowronski JJ. When motivated thought leads to heightened bias: high need for cognition can enhance the impact of stereotypes on memory. *Pers Soc Psychol Bull* 1998;**24** (10):1075–88.
68 Macrae CN, Bodenhausen GV, Milne AB, Jetten J. Out of mind but back in sight: stereotypes on the rebound. *J Pers Soc Psychol* 1994;**67** (5):808–17.
69 Woolf K, Cave J, Greenhalgh T, Dacre J. Ethnic stereotypes and the underachievement of UK medical students from ethnic minorities: qualitative study. *BMJ* 2008;**337** (7670):611–5.
70 van den Bergh L, Denessen E, Hornstra L, Voeten M, Holland RW. The implicit prejudiced attitudes of teachers: relations to teacher expectations and the ethnic achievement gap. *Am Educ Res J* 2010;**47** (2): 497–527.
71 Tweed M, Ingham C. Observed consultation: confidence and accuracy of assessors. *Adv Health Sci Educ Theory Pract* 2010;**15** (1):31–43.
72 Stewart N, Brown GDA, Chater N. Absolute identification by relative judgement. *Psychol Rev* 2005; **112** (4):881–911.
73 Mussweiler T. Comparison processes in social judgement: mechanisms and consequences. *Psychol Rev* 2003;**110** (3):472–89.
74 Wood TJ. Mental workload as a tool for understanding dual processes in rater-based assessments. *Adv Health Sci Educ Theory Pract* 2013;**18** (3):523–5.
75 Tavares W, Eva KW. Exploring the impact of mental workload on rater-based assessments. *Adv Health Sci Educ Theory Pract* 2013;**18** (2):291–303.
76 Wood TJ. Exploring the role of first impressions in rater-based assessments. *Adv Health Sci Educ Theory Pract* 2014;**19** (3):409–27.
77 Byrne A, Tweed N, Halligan C. A pilot study of the mental workload of objective structured clinical examination examiners. *Med Educ* 2014;**48**:262–7.
78 Moskowitz GB, Li P. Egalitarian goals trigger stereotype inhibition: a proactive form of stereotype control. *J Exp Soc Psychol* 2011;**47** (1):103–16.
79 Todd AR, Bodenhausen GV, Richeson JA, Galinsky AD. Perspective taking combats automatic expressions of racial bias. *J Pers Soc Psychol* 2011;**100** (6):1027.
80 Watson RA. Use of a machine learning algorithm to classify expertise: analysis of hand motion patterns during a simulated surgical task. *Acad Med* 2014; **89** (8):1163–1167.
81 Colliver JA. Educational theory and medical education practice: a cautionary note for medical school faculty. *Acad Med* 2002;**77** (12):1217–20.
82 Baig LA, Violato C, Crutcher RA. Assessing clinical communication skills in physicians: are the skills context specific or generalisable. *BMC Med Educ* 2009;**9**:22.
83 Keller L, Mazor KM, Swaminathan H, Pugnaire MP. An investigation of the impacts of different generalisability study designs on estimates of variance components and generalisability coefficients. *Acad Med* 2000;**75** (10) (Suppl):21–4.
84 Durning SJ, Artino AR Jr, Pangaro LN, van der Vleuten C, Schuwirth L. Redefining context in the clinical encounter: implications for research and training in medical education. *Acad Med* 2010;**85** (5):894–901.
85 Durning SJ, Artino AR. Situativity theory: a perspective on how participants and the environment can interact: AMEE Guide no. 52. *Med Teach* 2011;**33** (3):188–99.
86 Richter Lagha RA, Boscardin CK, May W, Fung C-C. A comparison of two standard-setting approaches in high-stakes clinical performance assessment using generalisability theory. *Acad Med* 2012;**87** (8):1077–82.
87 Hager P, ed. *Theories of Workplace Learning*. Los Angeles, CA: Sage Publications 2011.

88 Engeström Y, Sannino A. Studies of expansive learning: foundations, findings and future challenges. *Educ Res Rev* 2010;**5** (1):1–24.
89 Kuper A, Reeves S, Albert M, Hodges BD. Assessment: do we need to broaden our methodological horizons? *Med Educ* 2007;**41**:1121–3.
90 Ginsburg S, Bernabeo E, Ross KM, Holmboe ES. 'It depends': results of a qualitative study investigating how practising internists approach professional dilemmas. *Acad Med* 2012;**87** (12):1685–93.
91 Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach* 2013;**35** (7):564–8.
92 Lingard L. What we see and don't see when we look at 'competence': notes on a god term. *Adv Health Sci Educ Theory Pract* 2009;**14** (5):625–8.
93 Delandshere G, Petrosky AR. Assessment of complex performances: limitations of key measurement assumptions. *Educ Res* 1998;**27** (2):14–24.
94 Hodges B. Medical education and the maintenance of incompetence. *Med Teach* 2006;**28** (8):690–6.
95 Pangaro L, ten Cate O. Frameworks for learner assessment in medicine: AMEE Guide no. 78. *Med Teach* 2013;**35** (6):e1197–210.
96 Read SJ, Jones DK, Miller LC. Traits as goal-based categories: the importance of goals in the coherence of dispositional categories. *J Pers Soc Psychol* 1990;**58** (6):1048–61.
97 Reeder GD, Kumar S, Hesson-McInnis M, Trafimow D. Inferences about the morality of an aggressor: the role of perceived motive. *J Pers Soc Psychol* 2002;**83** (4): 789–803.
98 Malle BF, Pearce GE. Attention to behavioural events during interaction: two actor–observer gaps and three attempts to close them. *J Pers Soc Psychol* 2001; **81** (2):278–94.
99 Weiner B. Inferences of responsibility and social motivation. In: Zanna MP, ed. *Advances in Experimental Social Psychology*, vol **27**. San Diego, CA: Academic Press 1995;1–47.
100 Berendonk C, Stalmeijer RE, Schuwirth LWT. Expertise in performance assessment: assessors' perspectives. *Adv Health Sci Educ Theory Pract* 2013;**18** (4):559–71.
101 Govaerts MJB, Wiel MWJ, Schuwirth LWT, van der Vleuten CPM, Muijtjens AMM. Workplace-based assessment: raters' performance theories and constructs. *Adv Health Sci Educ Theory Pract* 2013b;**18** (3):375–96.
102 Gruppen LD, Frohna AZ. Clinical reasoning. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education*. Dordrecht: Kluwer Academic Publishers 2002;205–30.
103 Durning SJ, Artino AR, Boulet JR, Dorrance K, van der Vleuten C, Schuwirth L. The impact of selected contextual factors on experts' clinical reasoning performance (does context impact clinical reasoning performance in experts?). *Adv Health Sci Educ Theory Pract* 2012;**17** (1):65–79.
104 Hofstadter D. *I am a Strange Loop*. New York, NY: Basic Books 2007.
105 Klein G. *Streetlights and Shadows: Searching for the Keys to Adaptive Decision Making*. Cambridge, MA: MIT Press 2009.
106 Chi MTH, Feltovich PJ, Glaser R. Categorisation and representation of physics problems by experts and novices. *Cognit Sci* 1981;**5** (2):121–52.
107 Norman GR, Tugwell P, Feightner JW, Muzzin LJ, Jacoby LL. Knowledge and clinical problem-solving. *Med Educ* 1985;**19**:344–56.
108 Webster-Wright A. Reframing professional development through understanding authentic professional learning. *Rev Educ Res* 2009;**79** (2):702–39.
109 Gipps C. Socio-cultural aspects of assessment. *Rev Res Educ* 1999;**24**:355–92.
110 Delandshere G, Petrosky AR. 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. *Educ Res* 1994;**23** (5):11–8.
111 Elstein AS, Schwarz A. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ* 2002;**324** (7339):729–32.
112 Schuwirth LW, der van Vleuten CP. Assessing competence. In: Hodges BD, Lingard LA, eds. *The Question of Competence: Reconsidering Medical Education in the Twenty-First Century*. Ithaca, NY; London: ILR Press, Cornell University Press 2012;113–30.
113 Ross S, Poth CN, Donoff M, Humphries P, Steiner I, Schipper S, Janke F, Nichols D. Competency-based achievement system: using formative feedback to teach and assess family medicine residents' skills. *Can Fam Physician* 2011;**57** (9):e323–30.
114 Kahneman D, Klein G. Strategic decisions: when can you trust your gut? [Interview]. McKinsey Quartley & Co. Inc. 2010. http://www.mckinsey.com/insights/strategy/strategic_decisions_when_can_you_trust_your_gut. [Accessed 28 October 2013.]
115 Crossley J, Jolly B. Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. *Med Educ* 2012;**46**:28–37.
116 Nutter D, Whitcomb ME. The AAMC project on the clinical education of medical students. 2008. https://www.aamc.org/download/68522/data/clinicalskillsnutter.pdf. [Accessed 16 December 2013.]
117 Daelmans HEM, Hoogenboom RJI, Donker AJM, Scherpbier AJJA, Stehouwer CDA, van der Vleuten CPM. Effectiveness of clinical rotations as a learning environment for achieving competences. *Med Teach* 2004;**26** (4):305–12.

118 Holmboe ES. Faculty and the observation of trainees' clinical skills: problems and opportunities. *Acad Med* 2004;**79** (1):16–22.
119 Bindal N, Goodyear H, Bindal T, Wall D. DOPS assessment: a study to evaluate the experience and opinions of trainees and assessors. *Med Teach* 2013; **35** (6):e1230–4.
120 Bindal T, Wall D, Goodyear HM. Trainee doctors' views on workplace-based assessments: are they just a tick box exercise? *Med Teach* 2011;**33** (11):919–27.
121 Weston PSJ, Smith CA. The use of mini-CEX in UK Foundation Training six years following its introduction: lessons still to be learned and the benefit of formal teaching regarding its utility. *Med Teach* 2014;**36** (2):155–63.
122 McKavanagh P, Smyth A, Carragher A. Hospital consultants and workplace based assessments: how foundation doctors view these educational interactions? *Postgrad Med J* 1037;**2012** (88):119–24.
123 Sabey A, Harris M. Training in hospitals: what do GP specialist trainees think of workplace-based assessments? *Educ Primary Care* 2011;**22** (2):90–9.
124 Tokode OM, Dennick R. A qualitative study of foundation doctors' experiences with mini-CEX in the UK. *Int J Med Educ* 2013;**4**:83–92.
125 Pangaro L. A new vocabulary and other innovations for improving descriptive in-training evaluations. *Acad Med* 1999;**74** (11):1203–7.
126 Surowiecki J. *The Wisdom of Crowds*. New York, NY: Random House Digital 2005.
127 Baile WF, Buckman R, Lenzi R, Glober G, Beale EA, Kudelka AP. SPIKES – a six-step protocol for delivering bad news: application to the patient with cancer. *Oncologist* 2000;**5** (4):302–11.

Received 10 April 2014; editorial comments to author 4 June 2014; accepted for publication 11 July 2014

