Validity in work-based assessment: expanding our horizons

Marjan Govaerts & Cees PM van der Vleuten

CONTEXT Although work-based assessments (WBA) may come closest to assessing habitual performance, their use for summative purposes is not undisputed. Most criticism of WBA stems from approaches to validity consistent with the quantitative psychometric framework. However, there is increasing research evidence that indicates that the assumptions underlying the predictive, deterministic framework of psychometrics may no longer hold. In this discussion paper we argue that meaningfulness and appropriateness of current validity evidence can be called into question and that we need alternative strategies to assessment and validity inquiry that build on current theories of learning and performance in complex and dynamic workplace settings.

METHODS Drawing from research in various professional fields we outline key issues within the mechanisms of learning, competence and performance in the context of complex social environments and illustrate their relevance to WBA. In reviewing recent socio-cultural learning theory and research on performance and performance interpretations in work settings, we demonstrate that learning, competence (as inferred from performance) as well as performance interpretations are to be seen as inherently contextualised, and can only be understood '*in situ*'. Assessment in the context of work settings may, therefore, be more usefully viewed as a socially situated interpretive act.

DISCUSSION We propose constructivist–interpretivist approaches towards WBA in order to capture and understand contextualised learning and performance in work settings. Theoretical assumptions underlying interpretivist assessment approaches call for a validity theory that provides the theoretical framework and conceptual tools to guide the validation process in the qualitative assessment inquiry. Basic principles of rigour specific to qualitative research have been established, and they can and should be used to determine validity in interpretivist assessment approaches. If used properly, these strategies generate trustworthy evidence that is needed to develop the validity argument in WBA, allowing for in-depth and meaningful information about professional competence.

*Medical Education 2013: **47***: 1164–1174
doi:10.1111/medu.12289

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

Educational Development and Research, Maastricht University, Maastricht, the Netherlands

Correspondence: Marjan Govaerts, Educational Development and Research, Maastricht University, PO Box 616, Maastricht 6200 MD, the Netherlands. Tel: 00 31 433 885 746; E-mail: marjan.govaerts@maastrichtuniversity.nl

INTRODUCTION

Work-based assessment (WBA) is potentially the best way of assessing professional competence, i.e. the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, judgement, emotions, values and reflection in day-to-day practice.[1] Work-based assessments include assessment tools such as mini-clinical evaluation exercise, direct observation of practical skill, professionalism mini-evaluation exercise, multi-source feedback as well as in-training evaluation reports that typically require clinical assessors to convert trainee performance into a numerical score, according to predefined rules and criteria, to obtain accurate and easily communicable descriptions of a trainee's ability. However, although WBA may come closest to assessing habitual performance, research findings raise serious concerns about utility of WBA for summative assessment purposes. First, assessment tasks in the real world are unpredictable and inherently unstandardised and they will not be equivalent over different administrations. From a psychometric perspective, this poses serious threats to reliability and validity of assessment. Second, as professional judgement is inherent in WBA, serious concerns are raised about the subjectivity of assessments. Raters are generally considered to be major sources of measurement error.[2,3] Performance ratings are considered to be unacceptably biased, suffering from halo and leniency effects, and intra- and inter-rater reliability of performance ratings are often found to be substandard.[4–6] Weaknesses in the quality of measurement on top of problems in the implementation of WBA instruments have even resulted in widespread cynicism about WBA in the profession.[7]

As is apparent from a focus on quantifiable measures of assessment quality, most criticisms of WBA stem from approaches to validity and validation consistent with the quantitative framework of psychometrics. In essence, validity refers to the degree to which the proposed interpretations and the uses of assessment outcomes (e.g. performance ratings or test scores) in terms of decisions and actions are adequate and appropriate, as justified by evidence or theoretical rationales.[8,9] Validation can then be defined as 'developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use'[10] through accumulation and integration of different kinds of evidence from different sources. Or, as stated by Koch and DeLuca[11]: '..validation should be a generative process that promotes continuous inquiry into assessment practice'. What is rarely addressed explicitly, though, is that our approaches to WBA – reflected in the way we design and evaluate assessment practices – are inextricably linked to our implicit theories of learning, performance and competence. In this article, it is our intent to illustrate that an exclusive focus on traditional psychometric approaches to validity and validation in WBA may no longer be appropriate by their disregard for key issues with respect to competence development, performance and assessment in complex and dynamic workplace settings.

Within the predictive, deterministic framework of psychometrics, assessment typically aims for generalisable explanations or predictions.[9,12] Central to the psychometric discourse in current assessment are its almost exclusive focus on the inference of a true score representing true performance; its pursuit of a specified level of consistency that is assumed to be conditional on technically sound measurement (reliability) and the assumption of error (noise that needs to be eliminated) when repeated measurements fail to yield consistent results. The almost exclusive use of psychometric tools in validation of WBA, that is the way we develop the validity argument in WBA, reflects theoretical assumptions underlying our interpretations and uses of assessment outcomes that conceptualise assessment as a scientific measurement of abstract, latent and stable dispositions within individuals. In current approaches to WBA and validation of WBA, three assumptions in particular seem to stand out:

1 Learning (professional development) is a deterministic, linear process that can be identified and specified in advance; task performance and learning (as represented by assessment scores) are typically abstracted and interpreted independent of context;
2 Competence, as inferred from performance, is a fixed, permanent and decontextualised attribute, i.e. an inherent trait or ability of health care workers (or trainees), and
3 Performance can be 'objectified' and assessors, if they were only capable to do so, would be able to rate and observe some true level of performance.

There is, however, increasing and compelling research evidence that challenges the assumptions underlying our approaches to WBA. For instance, findings from research in industrial and organisational psychology show that job performance lacks temporal stability, especially in highly complex

jobs.[13,14] True intra-individual variation in job performance may result from changes in the individual (e.g. due to motivation, fatigue, changing levels of competence) as well as changes in the job environment.[14] Similarly, research findings in medical education indicate that context (i.e. task environment or work environment) critically influences behaviours in practising doctors. Durning and colleagues,[15] for instance, reported that contextual factors affected clinical reasoning performance by experts (board certified internists) in ways that were very specific to the situation and were influenced by participants in the encounter (patient and doctor), their goals and the setting. So, although some aspects of job performance can be expected to be relatively stable over time (cognitive ability, perhaps), variability in performance ratings in WBA may very well reflect true performance variability within individuals. Similarly, increasing evidence from industrial and organisational psychology, as well as medical education, supports contentions that rater effects in WBA do not represent (mere) rater biases, but rather represent alternative and complementary valid perspectives on trainee performance,[16] challenging our interpretations of between-rater differences in WBA.

Recent research findings and growing understanding of learning in complex social environments therefore suggest that meaningfulness and appropriateness of current validity evidence in WBA can be called into question, and common validity theory, which is framed in psychometrics, may no longer hold: we may be operating on faulty assumptions. In the following, we will discuss changing conceptions of learning and performance in work-based settings and will present research findings to substantiate the need for expanded conceptions of validation and validity theory. Drawing from research in various professional fields we will discuss the assumptions underlying the psychometric approaches to WBA and will propose alternative strategies to assessment and validity inquiry that are embedded in qualitative research paradigms and built on current theories of workplace learning and contextual performance.

WBA AND PREDICTABILITY OF LEARNING

In medical education, perspectives originating from behaviourist, cognitivist and constructivist learning theories have long dominated developments in instruction and assessment. These learning theories have in common that they focus on individual learners, that they stress cognitive aspects of performance (i.e. thinking and reflection) and that learning is treated as a 'thing' or product located in the mind of the learner. Although these theories acknowledge that context influences quality of learning processes and thus how well learning occurs, their view is that the nature of what is learned or is to be learned, is relatively independent of context.[17] They generally treat workplace learning as a linear process, akin to formal learning, through which a learner develops from incompetent to competent, largely neglecting the role of social, cultural and organisational factors in shaping learning and performance development. During the past decades, however, more robust theories of workplace learning have emerged, expanding the limiting assumptions underlying the theories described above.

Especially the group of socio-cultural theories of workplace learning seem to offer more powerful frameworks for understanding learning in workplace settings (See Glossary Table for definition of terms used). Socio-cultural learning theories claim that learning and learning outcomes emerge through active participation in activities of a community and interaction with the complex and dynamic systems of the work environment.[18] Socio-cultural learning theories therefore consider learning and expertise development to be inextricably linked to features of the context in which the learning occurs; learning processes as well as learning outcomes change as contexts change.[17,19] What, how and why trainees learn is shaped by unique experiences and the meaning or consequences that trainees and co-participants (e.g. supervisors, assessors, co-workers and patients in a clinical context) attach to these experiences.[9] Socio-cultural learning theories, with their focus on knowledge produced by social interaction, are particularly useful for thinking about learning in clinical training and health care settings. In these settings, learning is produced by a trainee's engagement in non-standardised and unpredictable tasks of authentic health care practices and the ongoing social interaction around authentic tasks, shaped by (unique) physical, social and organisational contexts.[20] Learning in clinical work settings then inevitably becomes a dynamic, non-linear and non-deterministic process. The increasing complexity of health care as well as its ever-changing context furthermore demand that we move beyond predictability of individual learning and competence towards conceptualisations of competence as a collective, situated and dynamically produced through interaction and learning in functional clinical groups.[20]

Not only is team-based care rapidly becoming the norm in our health care systems (requiring a shift in focus from individual competence to team competence), the complex and dynamic nature of health care systems also implies that we can no longer see competence as 'a state to be achieved'. Rather, nowadays, notions of work-based learning and competence should include the ability to continuously adapt to change. Competence it is not just about acquisition of knowledge and skills, but about the ability to create new knowledge in response to changing work processes.[21] From this perspective, learning involves learning things 'that aren't there yet', through exchange and interactions in social networks and collaborative processes in communities of practice that adapt to continuously evolving circumstances.[22,23] Complex and dynamic interactive processes between the learners and their environment then 'mutually reconstruct both the learner and the environment'. Learning is 'expansive'[22] and can be conceptualised as 'an increasing (collective) capacity for acting in flexible, constructive and innovative ways appropriate to the challenges of ever changing circumstances'.[17] Learning for future practice thus implies that learning is an ongoing process without a clear endpoint; learning is never complete. This is directly opposed to traditional approaches in medical education where learning focuses on planned, formal events with well-defined and stable learning outcomes.[24] Very recent theories of workplace learning therefore explicitly question whether predictable and decidable systems of workplace learning can be designed and implemented. These theories, some of which build on complexity theory, emphasise the view that learning is an ongoing *creative* process, emergent from its context in unpredictable and unanticipated ways.[17]

Although social learning theory is increasingly being used in medical education,[19] much of current theorising still seeks to understand and explain workplace learning so that conditions that uniformly support and enhance quality learning can be identified and implemented. In fact, a lot of current efforts to improve work-based learning and assessment seem to aim for the design of clinical training that steers trainees' learning in predictable ways, through development of the 'right' theories of professional development, better analyses of task environments and the technology to model them,[12] as well as specifying standards for competent performance that have to be achieved at predefined stages in the learning process (e.g. milestones project).[25] In other words: if it would only be possible to predict what, when and how people learn, it would also be possible to design assessments using predetermined correct responses or models of performance.[12] Such (law-like) predictability is necessary to make models of assessment, learning and performance compatible with the psychometric framework. However, conceptualisations of learning as inherently situated, collaborative, transformational and expansive (i.e. focusing upon knowledge production rather than reproduction) challenge assumptions of predictability and uniformity in what is learned and what is to be learned. Assessment that focuses on predefined and specified learning outcomes then necessarily becomes an oversimplification of an arbitrary stage in the process of professional development.[26]

WBA AND COMPETENCE AS A FIXED ATTRIBUTE

Although context specificity or performance variability from one case or task to the next is a well-known phenomenon in medical education,[27] current approaches to assessment and its validation build on assumptions that there must be some level of true performance that can be 'measured': variability of an individual's performance over time or across tasks and work settings is typically viewed as measurement error. Competence is conceptualised as a stable trait, to be inferred from performance sampling within the professional domain, and expertise, once developed and established is considered to be portable and transferable from one context to another. In fact, most licensure and certification procedures seem to build on exactly this assumption.

There is an increasing body of research that challenges these conceptualisations of competence and professional performance. Within-person variation in performance is substantial and can be as large as between-person differences.[28–30] Obviously, performance of learners changes during training, as they learn and develop through participation in professional practice. Indeed, the focus of current WBA is ongoing evaluation and provision of feedback to improve performance and expertise development.[31] It would seem self-evident that conceptions of performance stability no longer hold within a context that intentionally aims for performance changes. We also readily accept that learners and professionals are not always performing at their best, and that performance varies from day to day or even within the same day. Especially in highly complex jobs, performance lacks temporal stability.[13,14] Reasons may be motivational (e.g. changes in performance goals

and effort due to conflicting tasks), physiological (e.g. fatigue) or any other unstable cause affecting individual performance, such as mood or emotional experiences.[32]

More importantly, however, there is an increasing body of research indicating that the dynamic nature of performance in work settings is caused by environmental factors, i.e. opportunities and constraints in the work setting, even in experts and talented performers. Research findings in industrial and organisational psychology and human resource management suggest that talented performance is not directly portable from one company to another, thereby challenging one of the foundational assumptions underlying human resource practices in organisations, namely that talent can be bought. In general, research findings indicate that performance is contextual and that 'talent won't transfer unless it maps to the challenges of the new environment'.[33] For instance, 'star' investment analysts on Wall Street showed significant short- and long-term performance decline after moving to another firm and the drop in performance persisted for up to 5 years.[34] Research findings suggested that specific features of the new role and work setting influenced the drop in performance. The contextual and situated nature of job performance was affirmed by findings that stars who moved with a group of colleagues performed better than those who moved solo. A study on the portability of leadership also showed that highly talented chief executive officers who were recruited by other firms did not always deliver; whether skills and experience proved valuable in the new job depended on specific characteristics of their new work environment.[33] Similarly, research on intra-individual performance variation in football players showed that a significant portion of variance could be explained by constraining actions of others, including teammates. Moreover, susceptibility to environmental constraints varied across players and job complexity, suggesting that performance is determined by the interaction between person, task and environment.[30] These findings are consistent with the notion of performance and competence being the product of cultural and social circumstances and of ongoing interaction with individuals and groups (teams) in a specific work setting.

Recent research in medical education equally challenges naïve assumptions about performance stability and generic transferability of knowledge and skilful practice. In their study on family practitioners' performance, Wenghofer and colleagues,[35] for instance, found that the doctor's work setting as well as systemic (community-related) factors significantly impacted performance, with varying effects across different performance dimensions. The study furthermore showed that, although doctor factors significantly influenced performance, they were not nearly as important as previously assumed. The critical influence of context on doctor behaviour was also illustrated in a study by Ginsburg and colleagues,[36] who reported that practising internists' approaches to professional dilemmas were malleable and dependent on individual patient characteristics, the doctor's affective response and relationship with the patient, the nature of the diagnosis as well as the doctor's relationships with co-workers in the health care system. They concluded that a doctor's performance was subject to 'multiple interdependent, idiosyncratic forces unique to each situation'.

Despite powerful research evidence, however, the notion that performance genuinely fluctuates over (short) periods of time and cannot be defined independently of its context has not really affected assessment researchers yet. If we want to capture the complex and multifaceted construct of professional competence we need to focus on aspects that go beyond the technical and context-free aspects of performance. On the contrary, unique and continually changing work contexts in modern health care systems demand that we assess our learners' and doctors' ability to adapt and to flexibly apply and develop knowledge and skills in the face of evolving circumstances. In line with this approach, performance variability resulting from interaction with contextual factors should not be dismissed as 'measurement error', but considered as potentially valuable and meaningful information in the appreciation of an individual's professional competence.[37]

WBA AND OBJECTIFICATION OF PERFORMANCE

From a socio-cultural perspective, performance is socially constructed and determined by each person's perception of and interaction with situational characteristics of the task at hand. When this framework is applied to the assessment of performance in work settings, a picture emerges of performance that can never be 'objective', but is always conceptualised and constructed according to the perspectives and values of an individual assessor, influenced by his or her unique experiences and the social structures in the assessment task and its context.[38]

In fact, research findings in industrial and organisational psychology indicate that assessors' judgements of performance in work settings can only be understood *in situ*: assessor behaviours are framed within the context in which assessment takes place. In WBA, assessors are engaged in complex and unpredictable tasks, more often than not in a context of time pressures and conflicting as well as ill-defined goals.[39,40] Assessors' behaviours and assessment outcomes are furthermore influenced by a broad range of other factors in the work context, such as interpersonal relationships (with the learner as well as with co-workers), political, emotional and cultural factors.[41,42] Central to constructivist, socio-cultural approaches to assessment is the view that assessors can no longer be seen as passive measurement instruments, but as active information processors who interpret and construct their own personal reality of the assessment context. Or, as stated by Delandshere and Petrosky[43]: '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".' This implies that there can be honest disagreement within and across communities of practice: a specific supervisor–assessor's conception of appropriate performance in, for instance, a patient encounter may be different from that of co-workers, the trainee or the patient. Differences in an assessor's interpretation and scoring of performance-related behaviours may then be viewed as 'distinct views of a common individual's job performance that may be equally valid'[44] or 'meaningful differences in….. behavior across sources, especially when each source rates… behavior in different situations'.[16]

Recent research in medical education[45,46] confirms findings from industrial and organisational psychology. A study by Govaerts *et al.*[46] for instance, explored the use of performance theories by experienced and trained assessor–supervisors in general practice. Findings showed that, when observing and evaluating trainee performance, assessors interactively used general as well as task-specific performance theory and person schemas to arrive at judgements and decisions about performance effectiveness. Between-assessor differences in the performance dimensions used in the assessment of performance were substantial, though, reflecting assessor idiosyncrasy in the interpretation of task performance as a result of differing personal experiences, beliefs and professional values. These findings provide support for socio-cultural approaches to WBA, in which assessors are to be seen as 'social perceivers' who construct and reconstruct their own performance theories and conceptualisations of competence through training, socialisation and task experience. Consequently, assessors in work settings are inherently idiosyncratic, and multiple assessors will have multiple constructed realities. Assessment that is framed in socio-cultural, constructivist theories thus challenges the assumption, underlying psychometric assessment theory, of the existence of a single true score.

IMPLICATIONS FOR WBA AND VALIDATION

What emerges from learning theories as described above and research evidence about performance and performance interpretations being inherently contextualised is the need to reconsider assumptions underlying common WBA practices.

On the basis of the research and insights presented in this paper, we want to argue that assessment in work settings is a socially situated interpretive act, which is inherently value laden. It reflects the experiences, the meanings, intentions and interpretations of individuals involved in the assessment process ('the interpretive community').[47] Conceptions of learning and performance based in socio-cultural theory call for assessment that does not just focus on learning outcomes, but also (and perhaps even more so) on the processes underlying learning, performance and performance interpretations in dynamic, complex workplace settings. This implies that the purpose of assessment is not to 'objectively' and 'accurately' quantify learning or learning outcomes, but to understand what, how and why trainees and doctors are learning. This entails understanding and explicating context, i.e. the relationship between learners, the learning environment and the larger social systems within which learning is occurring.[9] Assessment questions need to address learners' experiences, the activities that they are engaged in as well as the social, cultural and ethical issues that shape learning, learning outcomes and performance interpretations.[12] Assessment questions, in other words, need to be grounded in inquiry traditions that offer rich, situated accounts of contextualised learning, performance and assessor judgements in order to capture, understand and evaluate multiple, diverse instances and interpretations of learning and performance in complex social systems. Inquiry systems that are situated within qualitative research paradigms (e.g. constructivist-interpretive) seem to be well suited for this task.

During the past decades, 'interpretivist approaches' to assessment have been proposed, in line with social-constructivist and socio-cultural theories of learning and performance.[9,11,12,48,49] A central feature of these approaches is that performance assessments are seen as social constructions or interpretations, rather than absolute, objective truths[49]; there is no single 'true' score or 'objective' rating of performance. Rather, 'truth' is a matter of consensus among assessors who have to arrive at judgements on performance that are as informed and sophisticated as can be at a particular point in time. Various methodological approaches in interpretivist assessment have been described. Kuper *et al.*[50] for instance, suggested an ethnographic approach and use of interviews and focus groups to capture a broad range of interpersonal behaviours in specific contexts and to generate rich, meaningful assessments of doctor competence. In the setting of teacher education, case study approaches have been adopted to develop an assessment scheme for the purpose of teacher certification.[43] Although each approach has its own origin and nuances, key characteristics of interpretivist assessment approaches could be summarised as follows[43,48,49,51]:

1 In WBA assessment, tasks are not interchangeable, but make unique contributions to learning and assessment. As assessments in work settings are 'socially constructed' between assessors and the person who is being assessed, learners typically prepare a paper or portfolio documenting their learning and assessment activities to capture situated assessment processes. Assessment asks learners to describe the contexts in which they work (and learn), to document their learning experiences, learning goals and learning plans as well as assessment activities (work sampling, for instance) and performance evaluations. Knowing how a learner perceives the demands of any particular assessment task is considered critical information in performance interpretations. Therefore, the learner's point of view is typically incorporated in the assessment process, as are intermittent feedback cycles with critical analyses and reflection on learning and task performance;
2 Assessments rely on narratives rather than numerical scores: assessments seek to purposefully generate elaborate, written evaluative statements about performance by expert judges – those who are most knowledgeable about the context in which assessment occurs, intentionally capturing and accounting for context-specific aspects of performance. As scores have little intrinsic meaning, assessment instruments challenge assessors to provide narrative comments that are useful in guiding the learner's competence development as well as meaningful in decision making about competence achievement;
3 All stakeholders in the assessment process are thus continuously challenged and required to document their performance interpretations as well as to articulate underlying values and assumptions;
4 Written performance evaluations are collected across a broad range of tasks, contexts and assessors, in order to gain in-depth understanding of a person's performance repertoire and capability to adapt to various task requirements, and
5 Inferences about professional competence are based on critical review of all available performance evidence, through open deliberative and critical dialogue among stakeholders in the assessment process. An interpretive approach does not imply that interpretations are bound to single assessment occasions or to single performance documentations. Meaningful interpretations can, and should be, constructed across assessment occasions and performance evaluations. Data from multiple sources are to be triangulated, reviewed and discussed to identify patterns of performance across tasks and contexts as well as any outlying aspects of performance. Interpretations are repeatedly tested against all available evidence, until a coherent interpretation or an integrative judgement on an overall level of performance can be accounted for[43,48]. If necessary, decisions involve inquiry strategies for additional information gathering about specific aspects of performance. This does not mean that 'anything goes'; essentially, final decision making requires professional judgements that should be corroborated, motivated and substantiated in such a way that the judgement is defensible and credible. To guide the performance evaluation, interpretive categories or dimensions can be developed through collective discussion of values and standards. The critical review of the evidence, the questioning of the different interpretations and assumptions as well as the documentation of the decision-making process are all essential and contribute to the validity and fairness of the final decision. Part of the strength of interpretive approaches to assessment is its traceability, through documentation of rich, meaningful information and

> articulation of values and standards. External evaluators may then assume an auditing role to ensure that the process is equitable, reflects professional standards and is sufficiently rigorous to protect the public from incompetent professionals. In this respect, interpretive assessment may be more trustworthy than assessments relying on a set of scores that mask assessors' thinking.[51]

These views on assessment are fundamentally different from prevailing psychometric-based, reductionist (positivist-oriented) approaches to assessment. What both the psychometric-based and constructivist-interpretivist assessment approaches have in common, though, is that inferences about professional competence need to be credible and defensible, based on trustworthy evidence. Within both frameworks, assessment validation comprises the 'development of a series of inferences and assumptions leading from the observed performances to conclusions and decisions…' and 'evaluation of the plausibility of these inferences and assumptions ….. using appropriate evidence'.[52] Clearly, traditional notions of reliability and validity related to quantitative evaluation of assessment practices have limited usefulness in the evaluation of situated performance interpretations. The theoretical assumptions underlying interpretivist assessment approaches, as described above, call for validity theory that provides the theoretical framework and the conceptual tools to guide the validation process in qualitative assessment inquiry. Although we acknowledge that there is considerable debate about the value and legitimacy of alternative sets of criteria and standards to assess qualitative inquiry, basic principles of rigour specific to qualitative inquiry have been put forward over the past decades, and we argue that they can and should be used to determine 'validity' (i.e., trustworthiness, credibility and defensibility) of the qualitative inquiry in interpretivist assessment approaches. Criteria and standards that can be used to judge the adequacy of constructivist-interpretivist assessment have been suggested by Lincoln and Guba[53,54] in their classical work on evaluation. They suggest the use of criteria such as trustworthiness (i.e. credibility, transferability, dependability and confirmability) and authenticity (i.e. fairness, openness, negotiation and shared understanding) to evaluate assessment quality. They furthermore propose the use of various techniques or methodological strategies to bring rigour to the qualitative inquiry. These techniques include: prolonged engagement in the assessment process; peer debriefing; analysis of disconfirming evidence (i.e. actively seeking counterexamples that challenge emerging interpretations), member checks and progressive subjectivity (to achieve credibility) as well as thick, rich description (to achieve transferability) and the audit trail, external audit and documentation of the assessment decision processes (to achieve dependability and confirmability). Some strategies need to be addressed in the assessment design stage, whereas others are applied during data collection and interpretation or after interpretation of performance data (similar to the application of techniques and strategies to ensure validity in standardised assessments).[55] Examples of these approaches to assessment validation have been described in typically context-bound assessments of portfolios.[49,56–58] If used properly, methodological approaches as described above generate trustworthy evidence that is needed to develop the validity argument in interpretivist assessment approaches. In conclusion, similar to the positivist approach to validation, interpretivist assessment has the intent to construct generalising interpretations about a learner and his performance. However, the strategies to arrive at these interpretations and to provide evidence on the strength of these generalisations rest on different approaches.

CONCLUDING REMARKS

Based on contemporary learning theories and research evidence illuminating the context specificity of performance and performance interpretations, we argue that we need to expand our approaches to assessment inquiry in work settings and validity theory underlying validation processes.

We do not want to claim that contextualised perspectives on assessment can only be covered by the constructivist-interpretivist assessment framework. Alternative frameworks, such as Brunswik's Probabilistic Functionalism and Lens Model, also describe ecological perspectives on judgement and decision making.[59] Our argument, however, is that when building on specific frameworks in (evaluation of) assessments, one has to be very clear about assumptions underlying its use. On the basis of socio-cultural learning theories we propose approaches towards WBA that are grounded in qualitative (constructivist-interpretivist) research paradigms, to generate in-depth understanding of and meaningful information about critical aspects of professional competence. Rich, narrative evaluations of performance as well as articulation of underlying performance theories and values not only enhance the formative function of the assessment system to

maximise learning,[58] but are indispensible for trustworthy decision making in summative assessments. Our constructivist-interpretivist approach to WBA seems to cater to the growing awareness in the literature that an exclusive focus on the psychometric discourse may no longer be helpful in facing assessment challenges in modern health care practices and education.[60,61]

We do not want to pretend that approaches as described in this paper provide solutions to all problems in WBA. Nor do we want to build an argument against the use of quantitative performance data in assessment of professional competence. Numerical ratings as well as standardised assessments are valuable elements in programmatic approaches to competence assessment.[62] Rather, we should aim for careful balancing of quantitative and qualitative approaches in our assessment programmes, justifying our choices on the basis of assessment purposes as well as conceptualisations of learning and performance/competence.

Implications of interpretivist approaches to WBA include a shift from numbers to words in performance assessment as well as assessors who are willing and able to create an 'interpretive community'. This means that assessors must be able to demonstrate commitment to articulation of their own values and assumptions underlying judgements; they must be willing to engage in critical dialogue and meaningful negotiation, offer criticisms to others and be open for change in the light of the negotiation. The biggest challenge may very well be to make the necessary commitments of time and energy that are required to achieve trustworthiness in the assessment process. However, we feel that expanding our assessment repertoire with constructivist-interpretivist approaches may support new and much-needed directions in assessment and professional accountability. Engagement in discussion about performance values by communities of practice may furthermore fuel the debate about what constitutes excellence in professional competence and how assessment systems may contribute to improving the quality of patient care.

Finally, we think that conceptualisations of assessment and validity as described in this paper apply to all kinds of unstandardised assessments – in a range of (school-based) educational contexts. Changes in assessment towards assessment for learning, as well as acknowledgement that current measurement practices in educational assessment are not in line with current theories of learning and cognition, increasingly call for reconsideration of conventional notions of assessment and assessment validity. In medical education, research into questions raised by interpretivist assessment approaches is badly needed.

Contributors: MG and CvdV worked collaboratively to develop the primary content of this paper. MG wrote the initial draft of the manuscript. Both MG and CvdV contributed to revisions of the initial draft for intellectual content and clarity. Both authors approved the final manuscript for publication.
Acknowledgements: The authors want to thank Kevin Eva, Tim Wilkinson, Cathy Haigh and an anonymous reviewer for their valuable comments, which helped improve the contents of this paper.
Funding: None.
Conflicts of interest: None.
Ethical approval: Not applicable.

GLOSSARY TABLE

Social/socio-cultural learning theories emphasise learning through active participation in social (authentic, professional activities). Learners develop by actively engaging in ongoing processes of workplaces. The learning processes as well as learning outcomes (performance) are determined by social, organisational, cultural and other contextual factors. However, socio-cultural learning theories also reject the idea that the individual learner should be the exclusive focus of analysis: learning can be either individual or social (collective).[17]
Constructivist-interpretivist assessment approaches view assessment to be value laden and socially constructed. Assessors are social beings who construct the assessment according to their own values, beliefs and perceptions. Performance can therefore never be objective. The interpretive approach focuses on participants' own perspectives in conceptualising and reconstructing their experiences, expectations, interpretations and assumptions.[38]
Trustworthiness of qualitative assessment inquiry is important to evaluate its worth. Trustworthiness involves establishing[55]:
Credibility, or confidence in the 'truth' of the findings;
Transferability, or showing that findings have applicability in other contexts*;*
Dependability, or showing that findings are consistent and could be repeated*;*
Confirmability, or the degree of 'neutrality' (findings not shaped by investigator bias, motivation or interest).
Specific strategies can be used for establishing each of these criteria in qualitative assessment inquiry.[58]

REFERENCES

1 Epstein RM, Hundert EM. Defining and assessing professional competence. *JAMA* 2002;**287** (2):226–35.

2 Albanese MA. Challenges in using rater judgments in medical education. *J Eval Clin Pract* 2000;**6** (3):305–19.
3 Downing SM. Threats to the validity of clinical teaching assessments: what about rater error? *Med Educ* 2005;**39**:353–5.
4 Kreiter CD, Ferguson KJ. Examining the generalizability of ratings across clerkships using a clinical evaluation form. *Eval Health Prof* 2001;**24**:36–46.
5 van Barneveld C. The dependability of medical students' performance ratings as documented on in-training evaluations. *Acad Med* 2005;**80** (3):309–12.
6 Cook DA, Beckman TJ, Mandrekar JN, Pankratz VS. Internal structure of mini-CEX scores for internal medicine residents: factor analysis and generalizability. *Adv Health Sci Educ Theory Pract* 2010;**15** (5):633–45.
7 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** (1):28–37.
8 Kane M. Validation. In: Brennan RL, ed. *Educational Measurement*. Westport, CT: American Council on Education/Praeger 2006; 621–94.
9 Moss PA, Girard BJ, Haniford LC. Validity in educational measurement. *Rev Res Educ* 2006;**30**: 109–62.
10 American Educational Research Association, American Psychological Association & National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association 1999.
11 Koch MJ, DeLuca C. Rethinking validation in complex high-stakes assessment contexts. *Assess Educ Princ Pol Pract* 2012;**19** (1):99–116.
12 Delandshere G. Assessment as inquiry. *Teach Coll Rec* 2002;**104** (7):1461–84.
13 Fisher CD. What if we took within-person performance variability seriously? *Ind Organ Psychol* 2008;**1**:185–9.
14 Sturman MC, Cashen LH, Cheramie RA. The impact of job complexity and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. *J Appl Psychol* 2005;**90** (2):269–83.
15 Durning SJ, Artino AR, Boulet JR, Dorrance K, Van der Vleuten CPM, Schuwirth LWT. 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.
16 Lance CE, Hoffman BJ, Gentry WA, Baranik LE. Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Hum Resource Manag Rev* 2008;**18**:223–32.
17 Hager P. Theories of workplace learning. In: Malloch M, Cairns L, Evans K, O'Connor BN, eds. *The SAGE Handbook of Workplace Learning*. Los Angeles, CA: Sage Publications 2011; 17–32.
18 Bleakley A. Broadening conceptions of learning in medical education: the message from teamworking. *Med Educ* 2006;**40**:150–7.
19 Mann KV. Theoretical perspectives in medical education: past experience and future possibilities. *Med Educ* 2011;**45**:60–8.
20 Lingard L. Rethinking competence in the context of teamwork. In: Hodges BD, Lingard L, eds. *The Question of Competence: Reconsidering Medical Education in the Twenty-first Century*. London: Cornell University Press 2012; 42–70.
21 Fraser SW, Greenhalgh T. Coping with complexity: educating for capability. *BMJ* 2001;**323**:799–803.
22 Engeström Y, Sannino A. Studies of expansive learning: foundations, findings and future challenges. *Educ Res Rev* 2010;**5** (1):1–24.
23 Mennin S. Self-organisation, integration and curriculum in the complex world of medical education. *Med Educ* 2010;**44**:20–30.
24 Bleakley A. Blunting Occam's razor: aligning medical education with studies of complexity. *J Eval Clin Pract* 2010;**16**:849–55.
25 Ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad Med* 2007;**82**:542–7.
26 Hager P, Hodkinson P. Moving beyond the metaphor of transfer of learning. *Br Educ Res J* 2009;**35** (4): 619–38.
27 Eva KW. On the generality of specificity. *Med Educ* 2003;**37** (7):587–8.
28 Fisher CD, Noble CS. A within-person examination of correlates of performance and emotions while working. *Hum Perf* 2004;**17**:145–68.
29 Deadrick D, Bennett N, Russell C. Using hierarchical linear modeling to examine dynamic performance criteria over time. *J Manage* 1997;**23**:745–57.
30 Stewart GL, Nandkeolyar AK. Exploring how constraints created by other people influence intraindividual variation in objective performance measures. *J Appl Psychol* 2007;**92** (4):1149–58.
31 Norcini J, Burch V. Workplace-based assessment as an educational tool. AMEE Guide no. 31. *Med Teach* 2007;**29** (9):855–71.
32 Beal DJ, Weiss HM, Barros E, MacDermid SM. An episodic process model of affective influences on performance. *J Appl Psychol* 2005;**90** (6):1054–68.
33 Groysberg B, McLean AN, Nohria N. Are leaders portable? *Harv Bus Rev* 2006;**84** (5):92–101.
34 Groysberg B, Lee L-E. Hiring stars and their colleagues: exploration and exploitation in professional service firms. *Organ Sci* 2009;**20** (4):740–58.
35 Wenghofer E, Williams AP, Klass DJ. Factors affecting physician performance: implications for performance improvement and governance. *Health Policy* 2009;**5** (2):e141–60.
36 Ginsburg S, Bernabeo E, Ross KM, Holmboe ES. "It depends": results of a qualitative study investigating how practicing internists approach professional dilemmas. *Acad Med* 2012;**87** (12):1–9.

37 Schuwirth LWT, van der Vleuten CPM. A plea for new psychometric models in educational assessment. *Med Educ* 2006;**40**:296–300.
38 Gipps C. Socio-cultural aspects of assessment. *Rev Res Educ* 1999;**24**:355–92.
39 Murphy KR, Cleveland JN. *Understanding Performance Appraisal. Social, Organizational and Goal-based Perspectives*. Thousand Oaks, CA: Sage Publications 1995.
40 Levy PE, Williams JR. The social context of performance appraisal: a review and framework for the future. *J Manag* 2004;**30**:881–905.
41 Tziner A, Murphy KR, Cleveland JN. Contextual and rater factors affecting rating behaviour. *Group Organ Manag* 2005;**30** (1):89–98.
42 Ferris GR, Munyon TP, Basik K, Buckley MR. The performance evaluation context: social, emotional, cognitive, political and relationship components. *Hum Resource Manag Rev* 2008;**18**:146–63.
43 Delandshere G, Petrosky A. Capturing teachers' knowledge: performance assessment and post-structuralism. *Educ Res* 1994;**23** (5):11–8.
44 Landy FJ, Farr JL. Performance rating. *Psychol Bull* 1980;**87** (1):72–107.
45 Kogan JR, Conforti L, Bernabeo E, Lobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: a conceptual model. *Med Educ* 2011;**45**:1048–60.
46 Govaerts MJB, Van de Wiel MWJ, Schuwirth LWT, Van der Vleuten CPM, Muijtjens AMM. Workplace-based assessment: raters' performance theories and constructs. *Adv Health Sci Educ Theory Pract* 2013;**18** (3):375–96.
47 Shay SB. The assessment of complex performance: a socially situated interpretive act. *Harv Educ Rev* 2004;**74** (3):307–29.
48 Moss PA. Enlarging the dialogue in educational measurement: voices from interpretive research traditions. *Educ Res* 1996;**25** (1):20–8. 43.
49 Johnston B. Summative assessment of portfolios: an examination of different approaches to agreement over outcomes. *Stud High Educ* 2004;**29** (3): 395–412.
50 Kuper A, Reeves S, Albert M, Hodges BD. Assessment: do we need to broaden our methodological horizons? *Med Educ* 2007;**41**:1121–3.
51 Delandshere G, Petrosky AR. Assessment of complex performances: limitations of key measurement assumptions. *Educ Res* 1998;**27** (2):14–24.
52 Kane MT. Terminology, emphasis, and utility in validation. *Educ Res* 2008;**37** (2):76–82.
53 Guba E, Lincoln Y. *Fourth Generation Evaluation*. London: Sage Publications 1989.
54 Lincoln YS, Guba EG. *Naturalistic Inquiry*. Newbury Park, CA: Sage Publications 1985.
55 Krefting L. Rigor in qualitative research: the assessment of trustworthiness. *Am J Occup Ther* 1991;**45** (3):214–22.
56 Driessen E, van der Vleuten CPM, Schuwirth L, van Tartwijk J, Vermunt J. The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. *Med Educ* 2005;**39**:214–20.
57 Tigelaar DEH, Dolmans DHJM, Wolfhagen IHAP, van der Vleuten CPM. Quality issues in judging portfolios: implications for organizing teaching portfolio assessment procedures. *Stud High Educ* 2005;**30** (5):595–610.
58 van der Vleuten CPM, Schuwirth LWT, Scheele F, Driessen EW, Hodges B. The assessment of professional competence: building blocks for theory development. *Best Pract Res Cl Ob* 2010;**24** (6):703–19.
59 Wigton RS. What do the theories of Egon Brunswik have to say to medical education? *Adv Health Sci Educ Theory Pract* 2008;**13**:109–21.
60 Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach* 2013;**35**(7):564–8.
61 Schuwirth L, Ash J. Assessing tomorrow's learners: in competency-based education only a radically different holistic method of assessment will work. *Six things we could forget. Med Teach* 2013;**35**(7):555–9.
62 Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach* 2011;**33**:478–85.

Received 6 April 2013; editorial comments to author 8 May 2013; accepted for publication 14 June 2013

