Advances in Health Sciences Education
https://doi.org/10.1007/s10459-020-10003-0

INVITED PAPER

A history of assessment in medical education

Lambert W. T. Schuwirth[1,2] · **Cees P. M. van der Vleuten**[1,2]

Received: 28 July 2020 / Accepted: 19 October 2020

Abstract
The way quality of assessment has been perceived and assured has changed considerably in the recent 5 decades. Originally, assessment was mainly seen as a measurement problem with the aim to tell people apart, the competent from the not competent. Logically, reproducibility or reliability and construct validity were seen as necessary and sufficient for assessment quality and the role of human judgement was minimised. Later, assessment moved back into the authentic workplace with various workplace-based assessment (WBA) methods. Although originally approached from the same measurement framework, WBA and other assessments gradually became assessment processes that included or embraced human judgement but based on good support and assessment expertise. Currently, assessment is treated as a whole system problem in which competence is evaluated from an integrated rather than a reductionist perspective. Current research therefore focuses on how to support and improve human judgement, how to triangulate assessment information meaningfully and how to construct fairness, credibility and defensibility from a systems perspective. But, given the rapid changes in society, education and healthcare, yet another evolution in our thinking about good assessment is likely to lurk around the corner.

Keywords Assessment · History · Programmatic assessment · Workplace based assessment

Introduction

This special issue provides a perfect opportunity to reflect on where we are at the moment in health professions education and where we have come from. We would not be exaggerating by claiming that this Journal has played an important role in this history. Right from the start, under the leadership of its founding editor-in-chief, it has contributed to the development of strong research approaches in health professions education research. We must acknowledge as well that although Geoff Norman, as founding editor-in-chief,

✉ Lambert W. T. Schuwirth
lambert.schuwirth@flinders.edu.au

[1] FHMRI: Prideaux Research in Health Professions Education, College of Medicine and Public Health, Flinders University, Sturt Road, Bedford Park, South Australia, 5042, GPO Box 2100, Adelaide, SA 5001, Australia

[2] Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands

came from a quantitative, experimental research tradition, there has always been room in the journal for breadth and research of different ontological, epistemological, theoretical and methodological backgrounds. This breadth, but with the requirement of scientific rigour, has made the journal one of the important ones in the field.

In this paper we want to describe our perspective on the history of assessment in medical education, and it has been an interesting one. It has been marked by both evolutionary and revolutionary changes. Current views on what constitutes good assessment in medical education differ vastly from, for example, 50 years ago. Some wonder whether this really means that the current state is better or that we are just following new fads. It may come as no surprise that we are convinced that assessment has evolved and is better now. We would also contend that this is due to a logical sequence of developments, where each one built and improved upon insights of the previous. Therefore, in this paper we want to describe history of developments in assessment of medical competence from the 1960s to the current time.

In doing so, we realise that in every description of history, choices have to be made as to what to include and what not. For instance, we want to declare here that whenever we speak about 'assessment' in this paper we pertain to assessment in medical education. This is perhaps a limitation because there are many health professions education disciplines that have made important contributions the to the developments in assessment and perhaps even earlier or better, but we may not be across that vast body of literature well enough. We also had to choose a certain narrative and aggregation in our description and we have made those choices trying to gauge what we think is most meaningful to most readers. In this article we will describe the developments in three phases: 'assessment as measurement', 'assessment as judgement', and 'assessment as system'. We do not want to suggest there to be a sharp delineation in time between these three phases; they did overlap considerably and also informed each other in an iterative way.

Assessment as measurement

Assessment research and development in medical education in the 1960s aimed at producing more structured, standardised and 'objective' assessment, because of dissatisfaction with prevailing practice, which was often seen as subjective, unreliable and biased. Much was learnt and copied from test psychology. Test psychology as a discipline already had a well-developed measurement paradigm focussing on measuring personality characteristics with standardised methods, for example intelligence, motivation or extraversion/introversion. The most widely known examples of such personality trait tests are the Wechsler Adult Intelligence Scale (WAIS) or the Minnesota Multiphasic Personality Inventory (MMPI). This had several implications for our views in assessment research and development.

The first and most obvious implication was the view that competence could and even should be captured purely quantitatively and that it could be expressed as a (single) score. In this view, assessment design was mainly a psychometric measurement problem. So, unsurprisingly, the hallmarks of assessment quality were construct validity and reliability.

Reliability was not defined in the everyday meaning of the word, such as "the quality of being able to be trusted or believed because it is working or behaving well" but merely as the extent to which scores would be reproducible across items, cases, examiners, etc. or as internal consistency. At that time there was general agreement on the notion

and importance of reliability. Validity in educational assessment, on the other hand, was a more disputed concept. This was exemplified in the early 1980s by an interesting polemic between Robert Ebel and Lee Cronbach (Cronbach 1983; Ebel 1983). Cronbach argued, in line with his landmark publication about construct validity, that an assessment can only be valid if its scores 'behaved' in alignment with the assumptions about the construct (Cronbach and Meehl 1955). As a simple example, if an assumption is that expert clinicians are better medical problem solvers than lesser experts, a test for clinical problem solving should lead to higher scores for experts than for lesser experts. If our instrument finds that candidates of intermediate expertise outperform expert—a finding from the patient management literature—this argues against the construct validity of the instrument. But it can also be the other way around, if we assume that there is one best way of clinical reasoning for each medical problem and we find that an assessment instrument shows dissent amongst experts rather than consensus, this may challenge our theoretical assumptions about the construct (Young 2019). Ebel on the other hand, argued that educational assessments were not psychological tests and therefore, validity has to be built into the test, for instance by careful blueprinting and item writing. In short, the former view sees each item only as meaningful to the extent to which it contributes numerically to the total score and the latter sees each item as intrinsically meaningful and the score as a summary statement (Ebel 1983).

Another implication from mimicking assessment design on test psychology was to define medical competence as a combination of personality traits; typically, these were, 'knowledge', 'skills', 'attitudes and 'problem-solving ability'. And, like assumptions in test psychology, these individual attributes were assumed to be generic and independent. A popular view at the time was that each of these could be measured independently of the others. For example, it was held that problem-solving ability could be measured independently of knowledge, or that an assessment of skills—such as the OSCE—should not include knowledge aspects (Van der Vleuten and Swanson 1990).

When assessment is seen as a measurement of competence it is only logical to also strive to make it objective. Therefore, much of the assessment design aimed at minimising the role of human judgement, and structuring and standardisation were seen as important ways to increase reliability of the assessment.

Another consequence of using psychological testing as the basis for assessment design pertains to the definition of its purpose, namely, to tell people apart. Psychological tests are typically designed to tell people apart based on their personality traits; high extraversion-low extraversion, high and low intelligence, etc. so it was almost inevitable that assessments of that time were also designed to tell people apart: high competence and low competence. This way of thinking is still dominant in widely used item parameters such as Discrimination Index or Item-Total correlations. Although telling people apart may be one of the purposes of assessment in some contexts—especially in assessment *of* learning- in the early era of test development it was generally seen as the only one: students were categorised into 'sufficiently competent' and 'not sufficiently competent'. Incompetent or not-yet-competent students cannot progress to the next phase and would have to either resit the exam at some point in time to be allowed to progress. This was common practice under the assumption it would automatically lead to graduating only highly competent students.

In itself, the thinking of this era was not incoherent, but research findings and new ways of thinking gave rise to some critical concerns. Research, for example, showed that subjectivity is not the main source of unreliability, but poor sampling strategies are (Swanson 1987; Swanson and Norcini 1989). Poor sampling mainly leads to lack of reliability because of domain specificity (Swanson and Norcini 1989; Eva et al. 1998; Eva 2003); the

way a candidate solves a problem or item on a test is a poor predictor of how they would solve any other problem, and consequently high numbers of cases or items are needed to produce a sufficiently generalisable or reliable result. Moreover, the notion of objectivity was challenged (Norman et al. 1991; Van der Vleuten et al. 1991). Increasingly, it was acknowledged that assessment is always a process of collecting information about a learner's achievement and progress and *valuing* it. This 'valuing' always incorporates human judgement. Even the most structured multiple-choice test is preceded by a process that includes a series of human judgments: blueprinting, standard setting, relevance of items to include, wording of items and so on.

Another important finding was that traits could not be measured as independently of each other with different forms of assessment as previously thought (Norman et al. 1985; Norman 1988; Van der Vleuten et al. 1988). The 'holy grails' in assessment in medical education, clinical reasoning and problem solving, were found to be highly reliant on background knowledge, and so logically, performance does not generalise well across content (Swanson et al. 1987). Counterintuitively though, performance does generalise well across assessment formats (Norman et al. 1985). If for example, similar content was asked using open ended questions and multiple-choice questions, correlations were extremely high (Ward 1982; Schuwirth et al. 1996). Even when students' performance on a written test on clinical skills was compared to an actual OSCE, performance generalised surprisingly well (Van der Vleuten et al. 1988).

Assessment as judgement

A notable change in thinking about assessment took place in the 1990s. Discontent with the dominance of the measurement 'paradigm' grew, mainly because in this paradigm only certain, limited aspects of competence can be captured. A paper by Boud et al. illustrates this clearly by arguing that assessment should also promote independence, thoughtfulness and critical thinking and that when assessment focuses purely on measurement, it runs contrary to achieving these aims (Boud 1990). It was further argued that assessment could only promote these values if the students were included as active and responsible stakeholders in the assessment process and were provided with meaningful feedback (Boud 1995). This may be more commonplace now but at the time this view was quite new. Up until then, the main ways through which assessment impacted on learning was by behaviourist mechanisms, through reinforcement and punishment. Of course, the notions of formative assessment and feedback existed, but in a system in which the summative aspects were aimed at telling people apart in a mainly quantitative way, the impact of formative aspects was often negligible (Harrison et al. 2015; Harrison et al. 2016).

How assessment drives learning is more complex than simply by punishment and reward, however. It is highly influenced by the way students construct meaning from the assessment (Cilliers et al. 2010, 2012). Three changes in thinking took place. First, the notion of competence was redefined as competencies rather than as personality traits (Hager and Gonczi 1996; Canmeds 2005). Until today, the notion of competencies is not undisputed and there are many definitions and uses. (Albanese et al. 2008; Govaerts 2008) However, in general, competencies are an attempt to define the outcomes of medical education more meaningfully than traits. This is important because that opens up possibilities to also provide more meaningful feedback to the learner, and thus foster their learning (Ericsson et al. 1993). Second, because objectivity and standardisation

are not as essential to reliability as good sampling is, assessment could be allowed to move back into the authentic context (Norcini et al. 1995). This enabled the inclusion of more facets, such as critical thinking, professionalism, reflection and self-regulation in the assessment. Finally, there was a reappraisal of the role of human judgement in the assessment process (Epstein and Hundert 2002). This was not a return to the traditional ad-hoc and unreliable assessment practice of before; the ensuing workplace-based assessments (WBA) were developed using better knowledge and understanding around sampling, validity and reliability from previous research.

One of the perceived advantages of WBA over previous structured assessment methods such as the OSCE, is its ability to assess candidates in a real authentic setting. Authenticity has advantages in that it allows for the assessment of aspects which cannot be tested with an OSCE, such as management under pressure, agile interaction with patients and navigating boundary conditions of healthcare systems. It must be kept in mind though, that authenticity is not automatically the same as validity (Cronbach and Meehl 1955; Swanson et al. 1987; Kane 2006).

As said before, validity is the extent to which the assessment assesses what it purports to assess. In the sense of ensuring validity, direct observation-based assessment or WBA is fundamentally different to standardised testing. In standardised testing, validity can be built into the method. For example, a multiple-choice test has its own validity and reliability built into it, and it can even be administered by computers. This is not the case in WBA, where human observation and interpretation are essential. In current validity theory (Kane) observation and interpretation by the examiner are essential for the first inference in the validity chain, and without it, validity cannot be established (Kane 2006).

Logically, the role of the examiner became more central with respect to validity and for this, examiners need to have sufficient expertise with regard to the clinical content of the WBA—or any other form of direct observation-based assessment—*but also* with regard to the assessment aspects, what to look for, how to interpret, where to draw the line between satisfactory and unsatisfactory performance, et cetera. We see this as a fundamental change from the previous phase with respect to assessment design. Instead of designing assessment such that it removes the human judgement component—'objective assessment'—it now had to be designed to embrace human judgement. But this so-called assessment literacy (Popham 2009) was, and often still is, a challenge in WBA context (Berendonk et al. 2013). The importance of such assessment literacy was further demonstrated by Govaerts et al. (2011, 2012), who showed similarities between the cognitive processes in diagnosing disease and 'diagnosing competence'.

Initially, indicators for quality of WBA approaches were also borrowed from test psychology. For example, most WBA instruments still try to capture the complex observed performance in a single numerical outcome, studies look at reliability/generalisability of scores and scores are the summative part and feedback the formative part of WBA (Moonen-van Loon et al. 2013).

But gradually, different views emerged. A notable development was the realisation that standard psychometric quality criteria—construct validity and reliability—as the only hallmarks of assessment utility had their limitations. This was highlighted in the very first issue of this Journal, when Van der Vleuten introduced a broader view on the utility of assessment than reliability and construct validity alone (Van der Vleuten 1996). After that, several publications emerged which raised awareness for a broader conceptual understanding of quality in assessment. Schuwirth and Van der Vleuten made a plea for an extension of the 'toolkit' in psychometrics to provide more versatile

modelling to cater to the increasing variety in assessment (Schuwirth and Van der Vleuten 2006), and later with respect to programmatic assessment (Schuwirth and Van der Vleuten 2012). Important and perhaps even more foundational work in this area was done by Hodges and colleagues. cf. (Hodges and Lingard 2012; Hodges 2013).

This conceptual change of views was needed because the realisation grew that competence and competencies are not simple, straightforward phenomena which can be captured and sufficiently summarised in a single numerical outcome. Instead, they are complex and multifaceted. For example, where formerly increasing quality in WBA was pursued by minimising variability between assessors, Gingerich et al. (2015, 2017) explored the nature of assessor variability from a different standpoint. They argue that different expert assessors may differ because they observe different aspects of a multifaceted phenomenon such as competence. So, rather than seeing them only as dissenting, they were now seen as potentially complementary. In a further paper she and co-authors explored the nature of any variability from three different perspectives: an error based perspective in which different assessors use different frames of reference or apply criteria incorrectly, assessor fallibility and cognitive biases as a result of cognitive load restrictions, but also as meaningful idiosyncrasy (Gingerich et al. 2014).

Obviously, the latter can be seen as contributing to the complementary nature of examiner variability, but the former two are logically seen as limiting the validity of direct observation-based assessment.

There may be other threats to validity as well. One may be leniency bias or unwillingness to express concerns with a candidate in an attempt to avoid negative consequences (Berendonk et al. 2013; Shanahan et al. 2019), or another may occur in any situation where the candidate is allowed to choose the case or their examiner.

The problems of incorrect frames of reference and application of criteria, cognitive load restrictions and leniency bias can be typically counteracted by improving the assessment literacy of examiners through staff development. This is firstly, because expertise is always associated with efficiency (Chi and Rees 1982; Norman 1988; Boreham 1994; Norman 2009), and efficiency is associated with reduction of cognitive load (Van Merrienboer and Sweller 2005), and the same is likely to hold for assessment literacy (Govaerts et al. 2011, 2012). Secondly, having a fit-for-purpose vocabulary to support and defend one's judgement plausibly improves agency and empowers the assessor, and reduces the likelihood of differences between so-called private and public judgement, and leniency (Berendonk et al. 2013; Valentine and Schuwirth 2019). Thirdly, because increased assessment literacy involves the development of a so-called shared subjectivity and shared narrative, reducing the likelihood of incorrect frames of reference or interpretation of criteria (Ginsburg et al. 2015, 2017; Cook et al. 2016). Although this is still an area subject to much research, there are examples in the literature that show impressive effects from relatively minor adaptations. One example is the introduction of entrustable professional activities (Ten Cate Th 2005, Ten Cate Th and Scheele 2007) (EPA) as a form of narrative for judgement. How this works is best illustrated by a study by Weller et al. (2014). By introducing an EPA-based scale in WBA the judgements supervisors were asked to make, mimicked more the high-stakes judgements they were used to be making about their registrars, so basically changing the rubric to one that better supported the supervisors' existing expertise. This dramatically improved the psychometric properties of the assessment. The usefulness and validity of using EPAs in WBA contexts has been demonstrated for example, in general practice (Valentine et al. 2019), but more research in this area is definitely needed.

Assessment as a system

Gradually, the realisation grew that education, competence and assessment are more complex phenomena than originally thought (Durning et al. 2010). A new narrative emerged in which words such as 'complexity', 'systems' and 'non-linear dynamics' arose. These words have a longer history in other scientific domains, such as meteorology and physics, and for medical education they were not meant as one-on-one equivalents and, as Norman argued, nor should they (Norman 2011). Instead, they were indicators of a fundamental rethink about the ontological and epistemological foundations of 'education', 'competence' and 'assessment', using the basis of systems theory (Checkland 1985; Ulrich 2001). In general, the main implications of this thinking were:

- Education is a problem solving process which at any point in time may have multiple equally acceptable solution pathways (i.e. educational problem-solving processes like clinical reasoning are idiosyncratic processes)
- Yet, there are more or less fuzzy boundaries between acceptable and unacceptable solutions and not it is not a matter of 'just everything goes'
- At any point in time, the stakeholders need to be able to change tack if a solution pathway is not optimal and for this, they need situational awareness, a repertoire of strategies and the agility to change

Obviously, this also involved a rethink in assessment, from a methods-oriented approach to whole-systems approach. This is quite a fundamental change because until that time assessment typically operated by deconstructing competence into discrete, individually assessable units. However, that still left us with the huge challenge of reconstituting the complex phenomenon of competence from only few discrete elements. For example, even when an assessment programme contains 10 individual tests, each of those tests will only generate a binary result (pass/fail). That way, the reconstitution of competence will have to be done with those 10 binary data points. Using grades and weighting may only mitigate this problem slightly. Unfortunately, early uses of competencies did not seem to solve this problem either and they too used a reductionist approach with organisations often defining competencies, sub-competencies and even sub-sub-competencies, ad infinitum.

From an assessment point of view, programmatic assessment—or 'making the whole course count' as one of its similar developments in general education is named (Cooper et al. 2010)—has attempted to combine the complexity views with the need to keep the assessment integrated and holistic. It is based on students and their teachers/supervisors constructing a meaningful holistic narrative rather than a set of individual measurements. One of the reasons why this change in thinking was deemed necessary is because in the earlier years of assessment the reconstitution of the 'whole' from the individual measurement outcomes required hugely arbitrary decisions, such as 'This assessment counts for 40%' or 'The pass fail score is 55%'. Of course, approaching assessment as a system issue does not negate the need to make 'ready to progress'/'not ready to progress' decisions at some phases in the educational continuum. But these decisions must be made on the basis of meaningful triangulation of information from various sources, longitudinal data collection, meaningful feedback with targeted learning activities and proportional decision making (Van der Vleuten and Schuwirth 2005; Van der Vleuten et al. 2012, 2015), always requiring a clear and transparent rationale behind each high-stakes decision.

This change of approach has had significant implications for our conceptualisations of quality of assessment. For example, the process of triangulating assessment information across methods on similar content, rather than solely within method is different compared with traditional practice. Traditionally, assessment information was combined because it was of the same format. An OSCE station on knee examination and on abdominal examination are of the same format and that is why, traditionally, they were combined; poor performance on the one can be compensated for by good performance on the other. This practice is contrary to most evidence about generalisation though. Numerous studies have demonstrated that competence generalises better across formats than across content, whether it is with open-ended and multiple-choice tests (Ward 1982; Norman et al. 1987) or even comparing written and practice based tests (Van der Vleuten et al. 1988). But, triangulating information across formats requires a narrative rather than a numerical process, and historically numbers are often seen a more 'objective' and 'reliable' than words.

Others may argue that triangulation of information can be done reliably, and that assessment practice would be best served by following the information collection and collation principles in clinical health care provision (Schuwirth et al. 2017). But this is purely rhetorical. More recent research has therefore, focussed on the quality of narratives and how they can be used in the context of assessment. For instance, Ginsburg et al. published important work on how language is used in the support of forming judgements and communicating them. She and her co-workers explored the language used by consultants to conceptualise the performance or their registrars (Ginsburg et al. 2010, 2015, 2017) and how stakeholders use and interpret these judgments and feedback. Cook and co-workers explored how narratives can become valid parts of an assessment, drawing on qualitative research methodological rigour (Cook et al. 2016). Valentine et al. studied the narratives expert assessors use when assessing clinical case write ups and how these are used to inform their judgements and feedback, as a sort of 'symptomatology' of competence (Valentine and Schuwirth 2019). Finally, Driessen et al. showed how concepts of rigour in qualitative research can be applied to ensure rigour in the interpretation and decision making process in assessment at the organisational level (Driessen et al. 2005).

So, in summary, current research seeks to improve our understanding of the building blocks of judgement in assessment and how the so-called private judgement is formed and substantiated. Research also explores how stakeholders conceptualise competence, communicate their judgements and feedback, and how they interpret it. Or, how validity of non-numerical outcomes or judgements can be ensured, and how this can be done at a programme level. Additionally, in a study involving a large number of the world's top assessment experts, Dijkstra et al. explored the issue of quality of assessment as a system and developed a framework for the quality of a programme of assessment (Dijkstra et al. 2010). By using a Delphi technique with a large group of international assessment experts, consensus was reached for a comprehensive set of design guidelines (Dijkstra et al. 2012).

Although the concepts of assessment as a system or programmatic assessment become more widely accepted, the implementation is far from easy. Because its fundamental philosophy is so different to tradition, it runs contrary to that of many prevailing organisational cultures (Watling et al. 2013; Harrison et al. 2017), and it requires a rethink about the nature of fairness of an assessment system that does not require reductionist and/or purely quantitative approach (Valentine et al. accepted for publication).

The future of assessment

So, "where might all this be heading?", would be an important question to answer. Making predictions is not easy though, and often with hindsight, predictions of the past are mostly silly. If there had been any prediction modelling done in the mid-1800s it would probably have been that he quantity of horse manure was going to be the main issue to deal with in traffic. Yet, in medical education—or better, health professions education as a whole—considering future scenarios is a must as we educate healthcare professionals for the future. There are numerous technological and ensuing societal changes taking place that are likely to impact on health professions education and assessment. The most notable are the increasing availability of freely accessible information—not always knowledge though—through open access journals and cognitive surplus; such as freely available online instruction videos and resources (Shirky 2010). The emergence of distributed trust systems and peer economy models (Botsman 2017) are other examples. These will undoubtedly have an impact on what students expect from their education and assessment and how universities will have to design their education and curricula; from a knowledge and solutions-oriented perspective to a curation of problems perspective. These developments will also have an impact on assessment. Where the focus of much assessment at the moment is still on whether the student possesses sufficient knowledge, skills, competencies and is able to apply them, there will inevitably be a shift toward the assessment of the extent to which a student is able to use all ICT affordances, incorporate them meaningfully in their development of competence and is able to balance ICT derived 'competence' with their organic brain competence in a complex practical environment. What we mean by this, is that modern students, though their continual access to ICT, have the affordances of communicating with multiple communities and collaboratives almost simultaneously. They also have modes of creation of artifacts of their learning and achievement far beyond paper and pencil—such as videos, podcasts, animated presentations, complex evolving diagrams, etc. (Friedman and Friedman 2008). The change in our thinking about assessment, assessment quality and assessment practice will form a solid foundation for future health professions education developers and researchers to adapt to these changes, and we are sure that publications in Advances of Health Sciences Education will continue to play an important role in this future 'history'.

Epilogue

The search for the perfect assessment approach continues and is probably never finished. This is logical, medical education and medical practice are currently facing possible disruptive changes. There have been multiple symposia about the impact of modern technology on education and future healthcare at which future scenarios have been considered. The scenario that was considered most plausible by far is one in which healthcare providers are increasingly technology supported and or even technology substituted for those tasks that are traditionally the doctor's main added value; to diagnose and determine therapeutic management. Such future scenarios will require health professionals to have different skills, abilities and competencies, most likely in the humanistic domain. Patients will probably still need someone who can partner with and enable them in navigating their illness,

and who can help them make meaning of their situation. Obviously, this would require yet a new rethink of assessment.

References

Albanese, M. A., Mejicano, G., Mullan, P., Kokotailo, P., & Gruppen, L. (2008). Defining characteristics of educational competencies. *Medical Education, 42*(3), 248–255.

Berendonk, C., Stalmeijer, R. E., & Schuwirth, L. W. T. (2013). Expertise in performance assessment: Assessors' perspectives. *Advances in Health Sciences Education, 18*(4), 559–571.

Boreham, N. C. (1994). The dangerous practice of thinking. *Medical Education, 28,* 172–179.

Botsman, R. (2017). *Who can you trust?: How technology brought us together and why it might drive us apart*. New York: Hachette.

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

Boud, D. (1995). Assessment and learning: Contradictory or complementary. In P. Knight (Ed.), *Assessment for learning in higher education* (pp. 35–48). London: Kogan.

Canmeds. (2005). Retrieved 26 July, April 2017 from, http://www.royalcollege.ca/portal/page/portal/rc/canmeds.

Checkland, P. (1985). From optimizing to learning: A development of systems thinking for the 1990s. *The Journal of the Operational Research Society, 36*(9), 757–767.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 7–76). Hillsdale: Lawrence Erlbaum Associates.

Cilliers, F. J., Schuwirth, L. W. T., Adendorff, H. J., Herman, N., & Van der Vleuten, C. P. M. (2010). The mechanisms of impact of summative assessment on medical students' learning. *Advances in Health Sciences Education, 15,* 695–715.

Cilliers, F. J., Schuwirth, L. W. T., Herman, N., Adendorff, H. J., & Van der Vleuten, C. P. M. (2012). A model of the pre-assessment learning effects of summative assessment in medical education. *Advances in Health Sciences Education, 17,* 39–53.

Cook, D. A., Kuper, A., Hatala, R., & Ginsburg, S. (2016). When assessment data are words: Validity evidence for qualitative educational assessments. *Academic Medicine, 91*(10), 1359–1369.

Cooper, L., Orrell, J., & Bowden, M. (2010). *Work integrated learning: A fuide to effective practice*. Milton Park: Routledge.

Cronbach, L. J. (1983). What price simplicity? *Educational Measurement: Issues and Practice, 2*(2), 11–12.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

Dijkstra, J., Galbraith, R., Hodges, B., McAvoy, P., McCrorie, P., Southgate, L., et al. (2012). Expert validation of fit-for-purpose guidelines for designing programmes of assessment. *BMC Medical Education, 12*(20), 1–8.

Dijkstra, J., Van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2010). A new framework for designing programmes of assessment. *Advances in Health Sciences Education, 15,* 379–393.

Driessen, E., Van der Vleuten, C. P. M., Schuwirth, L. W. T., Van Tartwijk, J., & Vermunt, J. (2005). The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: A case study. *Medical Education, 39*(2), 214–220.

Durning, S. J., Artino, A., Pangaro, L., Van der Vleuten, C., & Schuwirth, L. (2010). Redefining context in the clinical encounter: implications for research and training in medical education. *Academic Medicine, 85*(5), 894–901.

Ebel, R. L. (1983). The practical validation of tests of ability. *Educational Measurement: Issues and Practice, 2*(2), 7–10.

Epstein, R. M., & Hundert, E. M. (2002). Defining and assessing professional competence. *The Journal of the American Medical Association, 287*(2), 226–235.

Ericsson, K. A., Krampe, R. T., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363–406.

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

Eva, K. W., Neville, A. J., & Norman, G. R. (1998). Exploring the etiology of content specificity: Factors influencing analogic transfer and problem solving. *Academic Medicine, 73*(10), s1–s5.

Friedman, L. W., & Friedman, H. H. (2008). The new media technologies: Overview and research framework. *Available at SSRN 1116771*.
Gingerich, A. (2015). *Questioning the rater idiosyncrasy explanation for error variance by searching for mulitple signals within the noise*. Maastricht: Maastricht University.
Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014). Seeing the 'black box'differently: Assessor cognition from three research perspectives. *Medical Education, 48*(11), 1055–1068.
Gingerich, A., Ramlo, S. E., Van der Vleuten, C. P. M., Eva, K. W., & Regehr, G. (2017). Inter-rater variability as mutual disagreement: Identifying raters' divergent points of view. *Advances in Health Sciences Education, 22*(4), 819–838.
Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K., & Regehr, G. (2010). Toward authentic clinical evaluation: Pitfalls in the pursuit of competency. *Academic Medicine, 85*(5), 780–786.
Ginsburg, S., Regehr, G., Lingard, L., & Eva, K. (2015). Reading between the lines: Faculty interpretations narrative evaluation comments. *Medical Education, 49,* 296–306.
Ginsburg, S., Vleuten, C. P. M., Eva, K. W., & Lingard, L. (2017). Cracking the code: Residents' interpretations of written assessment comment. *Medical Education, 51,* 401–410.
Govaerts, M. (2008). Educational competencies or education for professional competence? *Medical Education, 42*(3), 234–236.
Govaerts, M. J. B., Schuwirth, L. W. T., Van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2011). Workplace-based assessment: Effects of rater expertise. *Advances in Health Sciences Education, 16*(2), 151–165.
Govaerts, M. J. B., Wiel, M. W. J., Schuwirth, L. W. T., Vleuten, C. P. M., & Muijtjens, A. M. M. (2012). Workplace-based assessment: Raters' performance theories and constructs. *Advances in Health Sciences Education, 18,* 1–22.
Hager, P., & Gonczi, A. (1996). What is competence? *Medical Teacher, 18*(1), 15–18.
Harrison, C. J., Könings, K. D., Dannefer, E. F., Schuwirth, L. W. T., Wass, V., & Van der Vleuten, C. P. M. (2016). Factors influencing students' receptivity to formative feedback emerging from different assessment cultures. *Perspectives on Medical Education, 5,* 276–284.
Harrison, C. J., Könings, K. D., Schuwirth, L., Wass, V., & Van der Vleuten, C. (2015). Barriers to the uptake and use of feedback in the context of summative assessment. *Advances in Health Sciences Education, 20*(1), 229–245.
Harrison, C. J., Könings, K. D., Schuwirth, L. W. T., Wass, V., & Van der Vleuten, C. P. M. (2017). Changing the culture of assessment: the dominance of the summative assessment paradigm. *BMC Medical Education, 17*(1), 73.
Hodges, B. (2013). Assessment in the post-psychometric era: Learning to love the subjective and collective. *Medical Teacher, 35*(7), 564–568.
Hodges, B., & Lingard, L. (2012). *The question of competence: Reconsidering medical education in the twenty-first century*. Ithaka New York: Cornell University Press.
Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (Vol. 1, pp. 17–64). ACE/Praeger: Westport.
Moonen-van Loon, J. M. W., Overeem, K., Donkers, H. H. L. M., Van der Vleuten, C. P. M., & Driessen, E. W. (2013). Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. *Advances in Health Sciences Education, 18*(5), 1087–1102.
Norcini, J., Blank, L. L., Arnold, G. K., & Kimball, H. R. (1995). The mini-CEX (clinical evaluation exercise);a preliminary investigation. *Annals of Internal Medicine, 123*(10), 795–799.
Norman, G. (2009). Dual processing and diagnostic errors. *Advances in Health Sciences Education, 14,* 37–49.
Norman, G. (2011). Chaos, complexity and complicatedness: Lessons from rocket science. *Medical Education, 45,* 549–559.
Norman, G., Tugwell, P., Feightner, J., Muzzin, L., & Jacoby, L. (1985). Knowledge and clinical problem-solving. *Medical Education, 19,* 344–356.
Norman, G. R. (1988). Problem-solving skills, solving problems and problem-based learning. *Medical Education, 22,* 270–286.
Norman, G. R., Smith, E. K. M., Powles, A. C., Rooney, P. J., Henry, N. L., & Dodd, P. E. (1987). Factors underlying performance on written tests of knowledge. *Medical Education, 21,* 297–304.
Norman, G. R., Van der Vleuten, C. P. M., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: Issues of validity, efficiency and acceptability. *Medical Education, 25*(2), 119–126.
Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice, 48,* 4–11.
Schuwirth, L. W. T., & Van der Vleuten, C. P. M. (2006). A plea for new psychometrical models in educational assessment. *Medical Education, 40*(4), 296–300.

Schuwirth, L. W. T., & Van der Vleuten, C. P. M. (2012). Programmatic assessment and Kane's validity perspective. *Medical Education, 46*(1), 38–48.
Schuwirth, L. W. T., Van der Vleuten, C. P. M., & Donkers, H. H. L. M. (1996). A closer look at cueing effects in multiple-choice questions. *Medical Education, 30,* 44–49.
Schuwirth, L. W. T., Vleuten, C. P. M., & Durning, S. J. (2017). What programmatic assessment in medical education can learn from healthcare. *Perspectives on Medical Education, 6,* 1–5.
Shanahan, M. E., Van der Vleuten, C., & Schuwirth, L. (2019). Conflict between clinician teachers and their students: the clinician perspective. *Advances in Health Sciences Education, 25,* 401–414.
Shirky, C. (2010). *Cognitive surplus: Creativity and generosity in a connected age*. London: Penguin.
Swanson, D. B. (1987). A measurement framework for performance-based tests. In I. Hart & R. Harden (Eds.), *Further developments in Assessing Clinical Competence* (pp. 13–45). Montreal: Can-Heal Publications.
Swanson, D. B., & Norcini, J. J. (1989). Factors influencing reproducibility of tests using standardized patients. *Teaching and Learning in Medicine, 1*(3), 158–166.
Swanson, D. B., Norcini, J. J., & Grosso, L. J. (1987). Assessment of clinical competence: Written and computer-based simulations. *Assessment and Evaluation in Higher Education, 12*(3), 220–246.
Ten Cate, Th J. (2005). Entrustability of professional activities and competency-based training. *Medical Education, 39,* 1176–1177.
Ten Cate, Th J, & Scheele, F. (2007). Competency-based postgraduate training: Can we bridge the gap between theory and clinical practice. *Academic Medicine, 82,* 542–547.
Ulrich, W. (2001). The quest for competence in systemic research and practice. *SystemsResearch and Behavioral Science, 18,* 3–28.
Valentine, N., Durnig, S. J., Shanahan, E. M. & Schuwirth, L. W. T. (accepted for publication). Fairness in human judgement in assessment: A hermeneutic literature review and conceptual framework. *Advances in Health Sciences Education*.
Valentine, N., & Schuwirth, L. W. T. (2019). Identifying the narrative used by educators in articulating judgement of performance. *J Perspectives on Medical Education, 8*(2), 1–7.
Valentine, N., Wignes, J., Benson, J., Clota, S., & Schuwirth, L. W. T. (2019). Entrustable professional activities for workplace assessment of general practice trainees. *Medical Journal of Australia, 210*(8), 354–359.
Van der Vleuten, C. P. M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Science Education, 1*(1), 41–67.
Van der Vleuten, C. P. M., Norman, G. R., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: Issues of reliability. *Medical Education, 25,* 110–118.
Van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*(3), 309–317.
Van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., et al. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher, 34,* 205–214.
Van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Govaerts, M. J. B., & Heeneman, S. (2015). 12 Tips for programmatic assessment. *Medical Teacher, 37*(7), 641–646.
Van der Vleuten, C. P. M., & Swanson, D. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine, 2*(2), 58–76.
Van der Vleuten, C. P. M., Van Luyk, S. J., & Beckers, H. J. M. (1988). A written test as an alternative to performance testing. *Medical Education, 22,* 97–107.
Van Merrienboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*(2), 147–177.
Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement, 6*(1), 1–11.
Watling, C., Driessen, E., Van der Vleuten, C. P. M., Vanstone, M., & Lingard, L. (2013). Beyond individualism: Professional culture and its influence on feedback. *Medical Education, 47*(6), 585–594.
Weller, J. M., Misur, M., Nicolson, S., Morris, J., Ure, S., & Jolly, B. (2014). Can I leave the theatre? A key to more reliable workplace-based assessment. *British Journal of Anaesthesia, 112*(6), 1083–1091.
Young, M., Thomas, A., Gordon, D., Gruppen, L., Lubarsky, S., Rencic, J., et al. (2019). The terminology of clinical reasoning in health professions education: Implications and considerations. *Medical Teacher, 41,* 1–8.