Best Practice & Research Clinical Obstetrics and Gynaecology 24 (2010) 703–719

Contents lists available at ScienceDirect

Best Practice & Research Clinical Obstetrics and Gynaecology

journal homepage: www.elsevier.com/locate/bpobgyn

2

The assessment of professional competence: building blocks for theory development

C.P.M. van der Vleuten, PhD, Professor of Education [a,*], L.W.T. Schuwirth, MD, PhD, Professor for Innovative Assessment [a,d], F. Scheele, MD, PhD, Gynaecologist and Professor of Medical Education [b,e], E.W. Driessen, PhD, Senior Lecturer in Education [a,d], B. Hodges, PhD, Psychiatrist, Richard and Elizabeth Currie Chair in Health Professions Education Research [c,f]

[a] *Department of Educational Development and Research, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, The Netherlands*
[b] *Department of Obstetrics and Gynaecology, Saint Lucas Andreas Hospital, Jan Tooropstraat 164, 1016 AE Amsterdam, The Netherlands*
[c] *The Wilson Centre for Research in Education, 200 Elizabeth Street, 1ES 565, Toronto, Ontario, Canada, M5G 2C4*

Keywords:
assessment
professional competence
assessment of clinical performance
principles of assessment (programmes)

This article presents lessons learnt from experiences with assessment of professional competence. Based on Miller's pyramid, a distinction is made between established assessment technology for assessing 'knows', 'knowing how' and 'showing how' and more recent developments in the assessment of (clinical) performance at the 'does' level. Some general lessons are derived from research of and experiences with the established assessment technology. Here, many paradoxes are revealed and empirical outcomes are often counterintuitive. Instruments for assessing the 'does' level are classified and described, and additional general lessons for this area of performance assessment are derived. These lessons can also be read as general principles of assessment (programmes) and may provide theoretical building blocks to underpin appropriate and state-of-the-art assessment practices.

* Corresponding author: C. van der Vleuten, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Tel.: +31433885726; Fax: +31433885779.
E-mail addresses: c.vandervleuten@educ.unimaas.nl (C.P.M. van der Vleuten), L.Schuwirth@educ.unimaas.nl (L.W.T. Schuwirth), f.scheele@slaz.nl (F. Scheele), Driessen@educ.unimaas.nl (E.W. Driessen), brian.hodges@utoronto.ca (B. Hodges).
[d] Tel: +31433885726, Fax: +31433885779.
[e] Tel: +31205108911, Fax: +31 206853879.
[f] Tel.: +416 340 3079, Fax: +416 340 3792.

1521-6934/$ – see front matter
doi:10.1016/j.bpobgyn.2010.04.001

Assessment of professional competence is one area in medical education where significant progress has been made. Many developments and innovations have inspired good research, which has taught valuable lessons and prompted steps leading to further innovations. In this article, the most salient general lessons are presented, which differs from our previous reviews of assessment of competence.[1,2] Previously, the research around certain instruments to arrive at general conclusions was examined, but, in this article, this order is deliberately reversed to provide a different perspective.

Miller's pyramid is used by the authors as a convenient framework[3] to organise this review of assessment (Fig. 1).

When the literature on assessment of medical competence is surveyed from a historical perspective, what is striking is that, over the past decades, the research appears to have been steadily 'climbing' Miller's pyramid. Current developments are concentrated at the top: the 'does' level, while assessment at the lower layers, directed at (factual) knowledge, application of knowledge and demonstration of skills, has a longer history and might even be qualified as 'established' assessment technology. Assessment at the top ('does') level is predominantly assessment in the workplace. This article first discusses general lessons from the research of assessment at the bottom three layers and then concentrates on the top layer. The lessons are summarised in the 'Practice points'.

The first three layers: 'Knows', 'Knows how' and 'Shows how'

Competence is specific, not generic

This is one of the best-documented empirical findings in the assessment literature.[4] In medical education, it was first described in the research on so-called patient management problems (PMPs).[5,6] PMPs are elaborate, written patient simulations, and candidates' pathways and choices in resolving a problem are scored and taken as indications of competence in clinical reasoning. A quite disconcerting and counterintuitive finding was that candidates' performance on one case was a poor predictor of performance on any other given case, even within the same domain. This phenomenon was later demonstrated in basically all assessment methods, regardless of what was being measured. It was termed the 'content specificity' problem of (clinical) competence. A wealth of research on the Objective Structured Clinical Examination (OSCE) exposed content specificity as the dominant source of unreliability. All other sources of error (i.e., assessors, patients) either had limited effects or could be controlled.[7] The phenomenon of content specificity is not unique to medical expertise; it is also found elsewhere, often under the name of task variability.[8] How surprising and counterintuitive this finding was (and sometimes still is) is easier to understand when it is realised that much of the thinking about

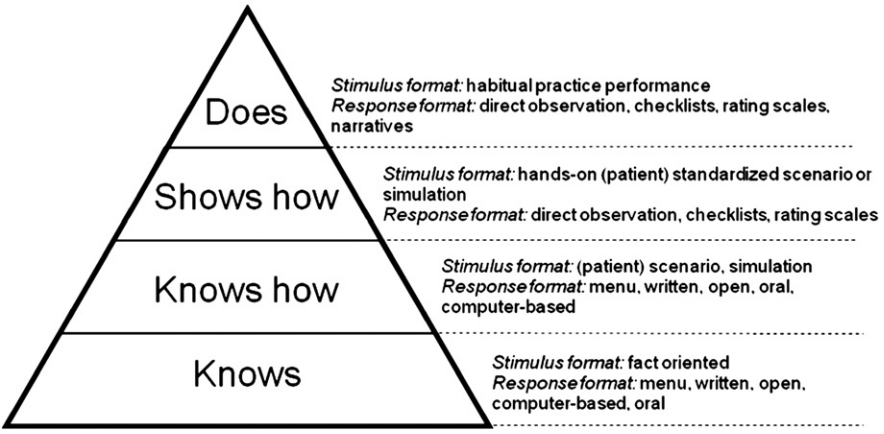

Fig. 1. Miller's pyramid and types of assessment used for assessing the layers.

competencies and skills was based on notions from research on personality traits. Personality traits are unobservable, 'inferred', stable traits, distinct from other traits and characterised by monotonous linear growth. A typical example of a trait is intelligence. It cannot be observed directly, so it has to be inferred from behaviour; it is independent of other personality traits, etc. The trait approach was a logical extension of psychometric theory, which had its origins in personality research. However, empirical research in education contradicted the tenets of the personality trait approach, revealing that the expected stability across content/tasks/items was very low at best. Moreover, when sufficient cases or subjects were sampled to overcome the content specificity problem, scores tended to correlate across different methods of assessment, thereby shattering the notion of the independence of measurements (it is later seen that this led to another insight). Content specificity resonated with findings from cognitive psychology, where much earlier transfer was identified as a fundamental problem in learning.[9] This sparked a great deal of research in cognitive psychology, providing insights on how learners reason through problems, how eminently important knowledge is therein, how information is chunked, automated and personalised as a result of personal experience and how people become experts through deliberate and sustained practice.[10,11] Viewed from the perspective of cognitive psychology, the phenomenon of content specificity thus becomes understandable as a quite logical natural phenomenon.

The consequences of content specificity for assessment are far-reaching and dramatic. It would be naïve to rely on small samples across content. Large samples are required to make reliable and generalisable inferences about a candidate's competence. In other words, short tests can never be generalisable. Depending on the efficiency of the methods used to sample across content (a multiple choice test samples more efficiently than a 'long case' oral examination such as used in the British tradition), estimations show that at least 3–4 h of testing time are required to obtain minimally reliable scores.[2] In short, one measure is no measure, and single-point assessments are not to be trusted. The wisest strategy is to combine information across content, across time and across different assessment sources.

Objectivity does not equal reliability

This insight is closely related to the previous one, and it is central to our thinking around assessment. Another discovery emerged from increasing numbers of publications on the reliability of assessment methods: reliability does not co-vary with the objectivity of methods; so-called subjective tests can be reliable and objective tests can be unreliable, all depending on the sampling within the method.[12] It became clear that content specificity was not the only reason to sample widely across content. When another factor, such as the subjective judgements of assessors, influenced measurement, it was usually found that sampling across that factor also improved the reliability of the scores. To illustrate this, even the notoriously subjective, old-fashioned, oral examination can be made reliable by wide sampling across content and examiners.[13,14] The concept of the OSCE arose to combat the subjectivity of the then-existing clinical assessment procedures. The solution was sought in objectivity and in standardisation, hence the 'O' and 'S' in the acronym. However, as research accumulated, the OSCE turned out to be as (un)reliable as any method, all depending on the sampling within the OSCE.[15] Apparently, reliability depended less on objectivity and standardisation than on sampling of stations and assessors. Further research around the OSCE revealed yet another piece of the puzzle: a strong correlation between global rating scales and checklist ratings.[16,17] Admittedly, global ratings were associated with a slight decrease in inter-rater reliability, but this was offset by a larger gain in inter-station reliability. Apparently, compared with the more analytical checklist scores, global, holistic judgements tended to pick up on elements in candidates' performance, which were more generalisable across stations. In addition, global rating scales proved to be more valid: they were better able to discriminate between levels of expertise.[18,19] This was a clear and intriguing first indication that human expert judgement could add (perhaps even incrementally) meaningful 'signal' to measurements instead of only 'noise'.

The notion that objectivity is not synonymous with reliability has far-reaching practical consequences. Most importantly, it justifies reliance on (expert) human judgement. Obviously, this is primarily relevant to those assessment situations where we cannot do without it, but, later in this

article, we will argue that reliance on human judgement is at least as relevant to many of the modern competency definitions that are being developed around the world. It is reassuring to know that, provided our sampling is adequate, we have no reason to ban subjective and holistic judgements from our assessment repertoire. In our view, this justifies the return of assessment to the clinical environment, which it had abandoned when the OSCE was introduced. Only this time, the move is scientifically underpinned by assessment theory.

What is being measured is determined more by the format of the stimulus than by the format of the response

Any assessment method is characterised by its stimulus and response formats.[20] The stimulus is the task presented to the candidate, and the response determines how the answer is captured. A stimulus format may be a written task eliciting a fact, a written patient scenario prompting a diagnostic choice or a standardised scenario portrayed by a simulated patient (SP), who is interviewed and diagnosed in an OSCE. Responses can be captured by short multiple-choice questions (MCQ) or long menu answers, a write-in, an essay, an oral situation, direct observation reported in a checklist, etc. Although different response formats can be used with one method, assessment methods are typically characterised by their response formats (i.e., MCQs, essays, orals, and so on). What empirical research revealed, surprisingly, was that validity – what is being measured – was not so much determined by the response format as by the stimulus format.[20] This was first demonstrated in the clinical reasoning literature in repeated reports of strong correlations between the results of complex paper-based patient scenarios and those of multiple-choice questions.[21,22] Like case specificity, this finding seemed highly counterintuitive at first sight. In fact, among test developers, it remains a widely accepted notion that essays tap into understanding and multiple-choice questions into factual knowledge. Although there are certain trade-offs (as we pointed out in relation to checklists and rating scales), there is no denying that it is the stimulus format and not the response format that dictates what is being measured. Studies in cognitive psychology, for example, have shown that the thought processes elicited by the case format differ from those triggered by a factual recall stimulus.[23,24] Moreover, there is evidence that written assessment formats predict OSCE performance to a large extent.[25]

The insight that the stimulus format is paramount in determining validity has first of all a practical implication: we should worry much more about designing appropriate stimulus formats than about appropriate response formats. An additional, related, insight concerns the stimulus format: authenticity is essential, provided the stimulus is pitched at the appropriate level of complexity. Extremely elaborate and costly PMPs, for example, did not add much compared with relatively simple short patient scenarios eliciting a key feature of a problem. Thus, short scenarios turned out to be not only relatively easy to develop, but they were quite efficient as well (good for wide sampling). It is no coincidence that written certifying examinations in the US and Canada have completely moved from measuring 'Knows' to measuring 'Knows how', using short scenario-based stimulus formats.[26,27] Pitching formats at the appropriate level of authenticity is relevant for OSCEs too. The classic OSCE consists of short stations assessing clinical skills in fragmentation (e.g., station 1: abdominal examination, station 2: communication). This is very different from the reality of clinical practice, which the OSCE was designed to approximate in the first place. Although fragmented skills assessment may be defensible at early stages of training (although one might question that too), at more advanced stages of training, integrated skills assessment is obviously a more appropriate stimulus format, since it provides a closer approximation of the real clinical encounter. The importance of pitching the stimulus at a suitable level of complexity is supported by cognitive load theory,[28] which posits that, when a learning task is too complex, short-term memory quickly becomes overloaded and learning is hampered as a result. This probably applies equally to assessment tasks. Authenticity therefore needs to be carefully dosed and fitted to the purpose of the assessment. However, the core lesson is that across all assessment methods, it is not the response format but the stimulus format on which we should focus.

A second implication of the significance of the stimulus format is more theoretical, although it has practical implications as well. When we aggregate information across assessments, we should use meaningful entities, probably largely determined by or related to the content of the stimulus format. This signifies a departure from the single-method-to-trait match (i.e., written tests measure knowledge,

PMPs measure clinical reasoning and OSCEs measure clinical skills), which is in line with the trait approach and still characteristic of many assessment practices: it is easy to aggregate within one method. This tenet becomes questionable if we accept that the stimulus is the crucial element. Is the method (the response format) really the most meaningful guide for aggregation? For example, does it make sense to add the score on a history-taking station to the score on the next station on resuscitation? Clearly, these stations measure very different skills. Why does similarity of method warrant aggregation? We see no legitimacy. Perhaps the current competency movement can provide a more meaningful framework. Nonetheless, in our view, the prominence of the stimulus implies that we should aggregate information across sources of information that are meaningfully similar and make sense. It also implies that similar information is not by definition information derived from identical assessment methods. We will address the practical pay-off of this insight when we discuss assessment programmes.

Validity can be 'built-in'

The general notion here is that assessment is not easy to develop and is only as good as the time and energy put into it. Good assessment crucially depends on quality assurance measures around both test development and test administration. Quality appraisal of tests during the developmental stage is imperative. Peer review is an essential ingredient of efforts to improve the quality of test materials significantly.[29] Unfortunately, it is not uncommon for test materials in medical schools to go unreviewed both before and after test administration. Not surprisingly, the quality of test materials within schools is often poor.[30] The same holds for test administration. For example, it is important to train SPs and assessors for an OSCE, because it makes a difference in terms of preventing noise in the measurement. Ebel, one of the early theorists on educational achievement testing, highlighted the difference between assessment in education and trait measurement in psychology. He argued that, while the latter is concerned with unobservable latent variables, assessments in education have direct meaning, can be discussed and evaluated, and directly optimised.[31] Ebel also argued that validity can be a 'built-in' feature of an assessment method. We take the view that all assessment at the three bottom layers of Miller's pyramid can be controlled and optimised: materials can be scrutinised, stakeholders prepared, administration procedures standardised, psychometric procedures put in place, etc. The extent to which this is actually done will ultimately determine the validity of the inferences supported by the assessment. Later, we will discuss how built-in validity is different at the top end of the pyramid.

The logical practical implication is to invest as much time and effort in test construction and administration processes as resources will allow. Another implication is that we should consider about sharing resources. Good assessment material is costly, so why not share it across schools and institutions? Not sharing is probably one of the biggest wastes of capital in education. Within our own context, five medical schools in the Netherlands have joined forces to develop and concurrently administer a comprehensive written test (Progress Test).[32] Laudable international initiatives to share test material across institutions are the IDEAL Consortium (http://www.hkwebmed.org/idealweb/homeindex.html, accessed 4 November 2009) and the UK UMAP initiative (http://www.umap.org.uk/accessed 4 November 2009).

Assessment drives learning

By now, it has almost become a cliché in assessment that assessment drives learning. The idea that assessment affects learning, for better or for worse, is also termed 'consequential validity'.[33] It has been criticised by some who argue that it negates intrinsic motivation.[34] Without any doubt, learners are also intrinsically motivated and not all learning is geared to assessment, but at the same time, academic success is defined by summative assessment, and learners will try to optimise their chances of success, much as researchers allow impact factors to drive their publication behaviour. If certain preparation strategies (reproductive learning, for instance) are expected to maximise assessment success, one cannot blame learners for engaging in these strategies. Nevertheless, the relationship remains poorly understood (what happens, to whom and why?) and we will revisit this issue in our suggestions for further research. For the time being, we note that many issues around assessment (format, regulations, scheduling, etc.) can have a profound impact on learners.

The immediate implication is that we should monitor assessment and evaluate its effect on learners. Assessment has been known to achieve the opposite effect to that intended. For example, when we introduced OSCEs within our school, students immediately started memorising checklists, and their performance in the OSCE was trivialised.[35] This reinforces the point we made about quality control, and extends it beyond test administration. A second, potential consequence is that we might use assessment strategically to achieve desired effects. If assessment drives learning in a certain (known) way, we might actually use this to promote positive learning effects.

No single method can do it all

No single method can be the magic bullet for assessment. Single-point assessments have limitations and any form of assessment will be confined to one level of Miller's pyramid. This realisation has inspired us to advocate 'Programmes of Assessment'.[2,36] Each single assessment is a biopsy, and a series of biopsies will provide a more complete, more accurate picture.

Thinking in terms of programmes of assessment has far-reaching consequences, particularly in relation to the governance of assessment programmes. We see an analogy here with a curriculum and how it is governed. A modern curriculum is planned, prepared, implemented, co-ordinated, evaluated and improved. We believe the same processes should be in place for an assessment programme. Such a programme needs to be planned and purposefully arranged to stimulate students to reflect at one point, to write at another, to present on certain occasions, to demonstrate behavioural performance at other arranged points, etc. Committees should be appointed to oversee test development, support should be arranged for test administration, evaluations should be carried out, and necessary improvements should be implemented. In a programme of assessment, any method can have utility, depending on its fitness for purpose. In our earlier reviews, we argued in favour of mindful utility compromises, allowing, for example, inclusion of a less reliable assessment method to make use of its beneficial effect on learning.[1] We propose that decisions about learners should never be based on a few assessment sources but rely on many. Information is preferably aggregated across the programme, and, as we argued earlier, across meaningful entities. This hinges on the presence of an overarching structure to organise the assessment programme.

Armed with the lessons and insights on assessment, which we have discussed so far, we are now ready to tackle the top end of Miller's pyramid. Pivotal in this move are the determination to strive towards assessment in authentic situations and the broad sampling perspective to counterbalance the unstandardised and subjective nature of judgements in this type of assessment.

Assessing 'Does'

Any assessment method at the 'does' level is characterised one way or another by reliance on information from knowledgeable people to judge performance. Obviously, this includes the assessee too. For now, we will park self-assessment to return to it later. Essentially, all assessment in natural settings relies on knowledgeable others or on 'expert' judgements. Sometimes reliance is indirect, as when assessment primarily relies on artefacts (e.g., prescription records, chart review, procedures done), but, ultimately, artefacts will have to be judged by one or more suitable assessors. The term 'expert' should be interpreted broadly to include peers, superiors, co-workers, teachers, supervisors, and anyone knowledgeable about the work or educational performance of the assessee. The assessment consists of gathering these judgements in some quantitative or qualitative form. As with OSCEs, the dominant response format is some form of observation structure (rating scale, free text boxes) on which a judgement is based. Unlike the OSCE, however, the stimulus format is the authentic context, which is essentially unstandardised and relatively unstructured. The response format is usually more or less generic and is not tailored to a specific assessment context. Predominantly, judgements take the form of global ratings of multiple competencies, often followed by oral feedback and discussion. In addition to scoring performance on rating scales, assessors are often invited to write narrative comments about the strengths and weaknesses of a student's performance.

The authentic context can be 'school based'. An example is the assessment of professional behaviour in tutorial groups in a problem-based learning environment. The authentic context can also be 'work-based', that is, medical practice at all levels of training (undergraduate, postgraduate and continuous professional development).[g] Particularly in the work-based arena, we have witnessed a recent explosion of assessment technologies. At the same time, we see a proliferation of competencies that are to be assessed. Increasingly, integral competency structures are proposed for modern assessment programmes, including the well-known general competencies from the US Accreditation Council of Graduate Medical Education[37]and the Canadian 'CanMEDS' competencies.[38] What they have in common is their emphasis on competencies that are not unique to the medical domain but have equal relevance to other professional domains. An example is the CanMEDS competency 'Collaborator' or 'Communicator', which has wide applicability. Although these competencies are generic to some extent, we immediately acknowledge that, for assessment purposes, they are just as context-specific as any other skill or competency. It is interesting that these frameworks should heavily emphasise more generic competencies, and they probably do so for all the right reasons. Typically, when things turn bad in clinicians' performance, it is these competencies that are at stake. Research shows that success in the labour market is more strongly determined by generic skills than by specific domain-specific skills.[39] Recent research in the medical domain shows that issues around problematic professional performance in clinical practice are associated with detectable flaws in professional behaviour during undergraduate medical training.[40–42] Therefore, it is imperative that generic skills are assessed. Unfortunately, these competencies are as difficult to define as their assessment is indispensable. An illustration in point is professionalism, a competency that has given rise to a plethora of definitions.[43] Detailed definitions and operationalisations can be incorporated in a checklist, but the spectre of trivialisation looms large.[44] At the same time, all of us have an intuitive notion of what these competencies entail, particularly if we see them manifested in concrete behaviour. We would argue that, to evaluate domain-independent competencies, we have no choice but to rely on assessment at the top of the pyramid, using some form of expert judgement. It follows that expert judgement is the key to effective assessment at the 'does' level.

Clinical professionals in a (postgraduate) teaching role traditionally gauge the professional maturity of trainees by their ability to bear clinical responsibility and to safely perform clinical tasks without direct supervision. It has been advocated that a summative assessment programme at the 'does' level should result in statements of awarded responsibility (STARs).[45] These STARs, representing competence to practise safely and independently, would be near the top end of the Miller pyramid, but below its highest level: a physician's track record in clinical practice. This is where the ultimate goal of competence, good patient care, comes into play.

All modern methods of assessment at the 'does' level allow for or apply frequent sampling across educational or clinical contexts and across assessors. The need to deal with content specificity means that sampling across a range of contexts remains invariantly important. At the same time, the subjectivity of expert judgements needs to be counterbalanced by additional sampling across experts/assessors. The aggregate information must theoretically suffice to overcome the subjectivity of individual assessments. At this point, we will bypass instruments that do not allow for wide sampling.

First, we will discuss the organisation of assessment procedures at the 'does' level and then derive some general notions based on the current state of affairs in the literature. Assessment at the top of the pyramid is still very much a work in progress. Systematic reviews of these assessment practices invariably lead to the conclusion that hard scientific evidence is scarce and further research needed.[46,47] Nevertheless, we believe that some generalisations are possible.

We will make a distinction between two types of assessment instruments. The first involve judgement of performance based directly on observation or on the assessor's exposure to the learner's performance. The second consists of aggregation instruments that compile information obtained from multiple sources over time. These two types will be discussed separately.

[g] We note a different use of work-based assessment in North-America and Europe. In North-America, this term is associated with work after completion of training. In Europe, it refers to all (learning) contexts that take place in a workplace. This may include undergraduate clinical rotations and postgraduate residency training programmes. We use the term here in the latter sense.

Direct performance measures

Within direct performance measures, we make another distinction between two classes of assessment methods, characterised by the length of the period over which the assessment takes place. In 'Individual Encounter' methods, performance assessment is confined to a single concrete situation, such as one (part of a) patient encounter. Instruments that are found here include the Mini-Clinical Evaluation Exercise (Mini-CEX[48]), Direct Observation of Practical Skills (DOPS[49]), the Professionalism Mini-evaluation (P-Mex[50]) and video observation of clinical encounters.[51] In a concrete, time-bound, usually short (hence the 'mini' epithet), authentic encounter, performance is appraised by an assessor using a generic rating form often reflecting multiple competencies, such as the competency frameworks discussed earlier. Observation is generally followed by discussion or feedback between assessor and assessee. For individual trainees, this assessment procedure is repeated across a number of encounters and assessors.

The second class of methods we propose are longer-term methods, in which performance is assessed over a longer period of time, ranging from several weeks to months or even years. Instead of judging individual encounters, assessors here rely on their exposure to the learner's work for an extended period of time. Examples of these methods include peer assessment[52] and multisource feedback.[53] Multisource, or 360°, feedback (MSF) is an extension of peer feedback. It often includes a self-assessment and assessments from a range of others who are in a position to give a relevant judgement of one or more aspects of the candidate's performance. These may include peers, supervisors, other health-care workers, patients, etc. The evaluation format usually involves a questionnaire with rating scales, which, again, evaluate multiple competencies. In many cases, additional narrative information is provided as well. Concrete procedures around MSF may vary. In some implementations, the learner selects the assessors; in others, the learner has no say in this. Sometimes the assessors remain anonymous and sometimes their identity is disclosed to the learner. Sometimes the feedback from MSF is mediated, that is, by a discussion with a supervisor or facilitator. This class of performance-appraisal methods can also be seen to comprise classic in-training evaluations by a supervisor, programme director or teacher. Unlike all other performance-appraisal methods, in-training evaluation is based on a single assessor. This does not mean that it is less useful, it only means that it should be treated as such. Naturally, it can be part of a larger assessment programme (remember any method can have utility depending on its function within a programme). It should also be noted that, with sufficient sampling across assessors, there is no reason why these global performance evaluations cannot be reliable.[54]

Aggregation methods

The second class of methods comprises aggregation methods, sampling performance across a longer period of time or even continuously. Two much-used instruments are the logbook and the portfolio. Portfolios have become particularly popular as an aggregation instrument. Just like 'OSCE', the term portfolio is an umbrella term that covers many manifestations, purposes of use and procedures surrounding it. Van Tartwijk and Driessen classify portfolios in terms of the functions they can serve: monitoring and planning, coaching and reflection, and assessment.[55] In fact, one might classify a logbook as a particular kind of portfolio with an exclusive focus on monitoring and planning. Portfolios can be used for a short time span and for a very limited set of competencies, even for a single competency. They can play a minor part in a larger assessment programme or they can be the main method to aggregate and evaluate all assessments at the 'does' level.[56] Alternatively, they can be the single method of assessment across the entire curriculum.[57] Obviously, it is hard to generalise across all these manifestations to provide general conclusions around validity and reliability. However, recent reviews have made such attempts, resulting in clear recommendations.[55,58–60] We will partly use these to infer our general notions. For specific details on portfolios, we refer to the reviews. For our thinking here, it is important to be aware that portfolios tend to work best if functions are combined,[55] in other words, when the portfolio is used for planning, coaching 'and' assessment. Portfolios also tend to work best if they perform a very central function (rather than peripheral) in guiding learning, in coaching and in monitoring longitudinal competency development.

So what general notions can we infer from the work published so far regarding direct performance measures?

A feasible sample is required to achieve reliable inferences

Recent reviews of direct observations in individual encounters summarise a number of studies, some based on large samples,[49] which examine how many observations are needed for adequate reliability.[47,61] Similar findings have been published for peer evaluations and multisource feedback instruments where assessment ranges across a longer period of time.[53,62–67] Despite variation between studies, we conclude that reliable inferences can be made with very feasible samples. The magical number seems to be somewhere between 8 and 10, irrespective of the type of instrument and of what is being measured (except when patient ratings are used; then many more are needed). This is a very clear confirmation that reliability is a matter of sampling, not of standardisation or structuring of assessment. Compared with other methods, the reliabilities actually appear to be somewhat better than those of standardised assessments.[2] One may speculate that this could be an indication that global performance appraisals pick up more generalisable competencies. Further research will be needed to answer this question, but it is an interesting thought that global expert judgement might bring more unique information to assessment, information that is not, or to a lesser extent, captured by more analytical methods.

Bias is an inherent characteristic of expert judgement

Adequate reliability does not preclude bias in global judgements. Indeed, global judgements are prone to bias, probably much more so than more structured, analytical methods.[68] With direct observation methods, inflation of scores has been noted.[69,70] In multisource feedback, selection of assessors or the background of assessors can introduce worrisome biases.[71] Another potentially important source of bias is the assessment context. Assessors' propensity to use only (the positive) part of the scale is heavily influenced by their desire not to compromise the relationship with the learner or to avoid more work (and trouble) consequent to negative evaluations. We need more research on biases in global judgements and why and how they operate, but, for the time being, we must be aware of their presence and take appropriate precautions wherever possible. For example, in those situations where the learner and the assessor have a relationship (very instrumental in any good learning,[72]) we would suggest that measures be taken to protect the assessor. In direct observation methods, such measures could entail removing the summative aspect of the assessment from the individual encounter. The assessor's task is not to judge if the learner is a good doctor, but to judge what happens in a specific encounter, to feed this back in a way that helps the learner to improve performance and, finally, to document this in an appropriate way for later meaningful review by the learner and by others. This is not to imply that the information cannot be used summatively somewhere somehow, later in the process, but the point is to remove the pass/fail decision from the individual encounter. A high-stakes decision should be based on multiple sources of assessment within or across methods, and robustness lies in the aggregation of all that rich information. Wherever possible, we would encourage relieving the assessor of potentially compromising, multiple roles. In making high-stake decisions based on aggregated information, protection could be provided by installing procedures that surpass the 'power' of the individual assessor. We will revisit this issue later.

Another important bias stems from self-assessment. The literature is crystal clear: we are very poor self-assessors,[73–77] equally likely to underestimate as to overestimate ourselves.[78] From a sampling perspective, this is not surprising. Self-assessment is inherently confined to a single assessment. In fact, the validity of a single self-assessment may not be so bad when it is compared with other single assessments. Nevertheless, sample size in self-assessment cannot be increased. The implication is that self-assessment can never stand on its own and should always be triangulated with other information. A continuous process of combining self-evaluations with information from others – such as in multisource feedback or in the reflection part of a portfolio – will hopefully pay off in the long run, and stimulate lifelong learning skills. However, even in continuous professional development, it is suggested that self-assessment should always be complemented by other assessments, an approach sometimes referred to as 'directed self-assessment'.[79]

Validity resides more in the users of the instruments than in the instruments that are used

We feel particularly strongly about this issue, because it is central and unique to assessment at the 'does' level and has profound practical implications. It complements our view on the earlier 'built-in validity' issue. In the lower layers of Miller's pyramid, we can control much around test development and test administration. We can 'sharpen' the instrument as much as we can, but at the 'does' level, assessment can only be as good as the job done by the assessors using the instrument. For example, the utility of an assessment will depend not so much on the operationalisation of the rating scale used in the direct observation, but much more on the way the assessor and the learner deal with the information that emerges from the encounter. Conscientiousness is essential to the process of assessment and determines its value. Increased control of the noisy real world by standardising, structuring and objectifying is not the answer. On the contrary, it will only harm and trivialise the assessment. To improve we must 'sharpen' the people rather than the instruments. Therefore, the quality of the implementation will be the key to success.[80] Published research so far seems to indicate we can do a much better job here: assessors are only rarely trained for their task and if they are, training is a brief and one-off event.[47] Receiving and giving feedback requires skills that need to be trained, honed and kept up-to-date. From personal experience with assessor training, we know that the skills required are very similar to the skills for the doctor–patient encounter. Nevertheless, like communication skills, they are not part of every teacher's make-up: they can and must be fostered.

Formative and summative functions are typically combined

In the preceding section, we already noted that in assessment at the 'does' level, the summative functions are typically linked with the formative functions. Indeed, we would argue that without formative value the summative function would be ineffective, leading to trivialisation of the assessment. As soon as the learner sees no learning value in an assessment, it becomes trivial. If the purpose is narrowed to doing eight summative Mini-CEXs, learners will start to play the game and make their own strategic choices regarding moments of observation and selection of assessors.[81] If the assessors join in the game, they will provide judgement without adequate information and return to their routines. If the main objective of the reflections in the portfolio is to please the assessment committee, the portfolio will lose all significance to the learner. We have seen similar things happen with logbooks.[82] We argue that whenever assessment becomes a goal in itself, it is trivialised and will ultimately be abandoned. Assessment has utility insofar as it succeeds in driving learning, is integrated in a routine, and ultimately comes to be regarded as indispensable to the learning practice. For assessment to be effective, certain conditions need to be met. We know that feedback is often ignored and fails to reach the intended recipient,[83] positive feedback has more impact than negative feedback,[84] (not implying that negative feedback has no value) feedback directed at the individual should be avoided and task-oriented feedback is to be preferred.[85] We know the rules of feedback[86] and we know that a positive learning climate is essential.[87] The literature suggests that successful feedback is conditional on social interaction,[58] such as coaching, mentoring, discussing portfolios and mediation around multisource feedback,[88] and this principle may even extend to all assessment at the 'does' level. It stipulates that assessment should be fully integrated in the learning process, firmly embedded within the training programme and serves a direct function in driving learning and personal development. For that matter, the principle that assessment drives learning is strongly reinforced by evidence around assessment, but we would argue that at the top of the pyramid it is the *sine qua non* of effective assessment.

Qualitative, narrative information carries a lot of weight

If feedback is central to assessment and if social interaction mediates effective feedback, numerical and quantitative information has obvious limitations, while narrative, qualitative information has benefits. This is also reported in empirical studies: narrative, descriptive and linguistic information is often much richer and more appreciated by learners.[89,81] Inescapably, narrative and qualitative information is something the assessment field will have to get used to. The assessment literature is strongly associated with quantification, scoring, averaging, etc., what Hodges calls the 'psychometric

discourse'.[90] It is quite clear that a rating of 2 out of 5 on counselling skills in a patient encounter should raise some concern with the learner, but a mere numerical rating fails to disclose what the learner actually did and what she should do to improve. To provide richness to the assessment to a greater extent, we have an excellent tool: language. We would argue that effective formative assessment is predicated on qualitatively rich information. We should encourage instrument developers to ensure that all their instruments have built-in facilities to elicit qualitative information (e.g., space for narrative comments) and we should stimulate assessors to routinely provide and document such information. This argument has even more relevance if we wish to assess difficult to define, domain-independent competencies, such as professionalism. These competencies, in particular, have much to gain from enriched narrative information.

Summative decisions can be rigorous with non-psychometric qualitative research procedures

Looking beyond the psychometric discourse is also imperative if we wish to strengthen decisions based on information that is aggregated across assessment sources. Within the conventional psychometric discourse, we typically quantify: we calculate and average scores and grades, and determine the reliability and validity of decisions. However, as soon as information of different kinds is aggregated across all kinds of sources, psychometric evaluation is bound to fall short.[91] We argue that aggregation in a programme of assessment (either at the 'does' level or across the full pyramid) depends on expert judgement. There are few situations in which purely quantitative strategies suffice, requiring no further judgement strategies. As soon as one source of information is qualitative, quantitative strategies will be found wanting. In trying to force quantification, similar to any individual method, we inevitably incur the risk of trivialisation.

In our efforts to proceed beyond the psychometric discourse, we find inspiration in methodologies from qualitative research. As in quantitative research, rigour is built into qualitative research, but the terminology and procedures are different.[92,93] Rigour depends on 'trustworthiness' strategies replacing conventional notions of internal validity by credibility, external validity by transferability, reliability by dependability and objectivity by conformability. For each of these notions, methodological strategies are proposed that bring rigour to the research: prolonged engagement, triangulation, peer examination, member checking, structural coherence, time sampling, stepwise replication, audit and thick description. With some creativity, we can apply these strategies to assessment to achieve rigour of decision making. In Table 1, we list some examples of assessment strategies that mirror these trustworthiness strategies and criteria.

An example may serve to further explain our ideas about qualitative rigour. An illustration from assessment practice is given by Driessen et al.[94] To achieve rigour in the judgement of a learner's portfolio in a work-based setting, it is wise to have a committee judge the portfolio (structural coherence and peer examination). The committee receives input from a mentor who is familiar with the learner and his or her portfolio (prolonged engagement). Depending on how much one wants to protect the learner–mentor relationship this input may be limited, for example, to a declaration of the mentor that the portfolio provides authentic evidence of the learner's progress. The committee uses predefined criteria to make their judgement more transparent, for example, in the form of rubrics describing decision categories (audit). The committee deliberates and justifies its decisions in a written motivation (audit). If the decision is difficult to make, the committee deliberates more and justifies more and perhaps even invites additional committee members or consults relevant parties (triangulation). In preparing the portfolio for submission, the learner is aware of the criteria and will have had feedback on earlier drafts of the portfolio with some form of social interaction (i.e., with peers or a mentor) so that the committee's judgement will only rarely come as a complete surprise to the learner (and mentor) (member checking). Both learner and mentor are trained for their tasks; committee members are (re)trained (periodically) and use benchmark portfolios to calibrate their decision making (prolonged engagement and member checking). Committee decisions are documented (audit), and appeal procedures for learners are in place (audit). The more procedures and measures, the more trustworthy the resulting decision will be. To some extent, this resonates with the validity discussion around standard setting procedures in assessment, where, in the absence of a gold standard, arbitrariness is always part of any standard and the resulting decisions. A standard is more or

Table 1
Illustrations of potential assessment strategies related to qualitative research methodologies for building rigour in assessment decisions.

Strategies to establish trustworthiness	Criteria	Potential Assessment Strategy
Credibility	Prolonged engagement	Training of assessors. The persons who know the student the best (a coach, peers) provide information for the assessment. Incorporate in the procedure intermittent feedback cycles.
	Triangulation	Many assessors should be involved and different credible groups should be included. Use multiple sources of assessment within or across methods. Organize a sequential judgement procedure in which conflicting information necessitates the gathering of more information.
	Peer examination (sometimes called Peer debriefing)	Organize discussion between assessors (before and intermediate) for benchmarking and discussion of the process and the results. Separate multiple roles of the assessors by removing the summative assessment decisions from the coaching role.
	Member checking	Incorporate the learner's point of view in the assessment procedure. Incorporate in the procedure intermittent feedback cycles.
	Structural coherence	Organize assessment committee to discuss inconsistencies in the assessment data.
Transferability	Time sampling	Sample broadly over different contexts and patients.
	Thick description (or Dense description)	Incorporate in the assessment instruments possibilities to give qualitative, narrative information. Give narrative information a lot of weight in the assessment procedure.
Dependability	Stepwise replication	Sample broadly over different assessors.
Dependability/ Confirmability	Audit	Document the different steps in the assessment process (a formal assessment plan approved by an examination board, overviews of the results per phase). Organize quality assessment procedures with external auditor. Give learners the possibility to appeal to the assessment decision.

less credible, depending on due diligence of the procedures.[95] Credibility and defensibility are the operative terms in qualitative assessment. Attention is deflected from the question whether the decision procedure is psychometrically sound to the question whether the decision is credible or defensible to anyone who may challenge it, ultimately even to a judge in court.

We would advocate that these strategies should be increasingly incorporated in our assessment practice for any programme of assessment (whether at the 'does' level or elsewhere). We think it is possible to use more diversified assessment information – richer materials, softer information, subjective information, expert information and artefacts – and still reach decisions that are credible and defensible. We see good examples in the literature.[57,94,96] At the same time, such strategies can prevent trivialisation and enhance the formative function of the assessment system to maximise learning.

Discussion

The general notions around assessment, which we have considered may be turned into principles of assessment. We think that both research and educational practice have created a solid base for these principles. We acknowledge that they also strongly testify to our interpretation of research and practice, but we hope they will be scrutinised in further debate and preferably in further research. This brings us to suggestions for the latter.

Based on the arguments we have presented and in line with others,[97] we advance the use of expert judgement for assessment purposes as an indispensable source of information both within methods of assessment and in the aggregation of information in a programme of assessment. To some extent, this should be comforting, since expert judgement is our daily business in clinical practice. Nevertheless, we must also realise that (clinical) expert judgement is fallible and open to criticism. There is a wealth of research in many diverse professional areas showing that experts make poorer judgements than actuarial or statistical models and that even replacing the expert by herself/himself in a model can lead

to more accurate judgement.[98,99] This research strongly advocates the 'scaffolding' of judgement with probabilistic and empirical information. This resonates with clinical decision making and the development and use of guidelines.[100] Naturally, in assessment, as in clinical practice, guidelines must be interpreted and tailored to individual learners. We need to reconcile and benefit from various research traditions such as psychology of judgement and decision making,[98] cognition and medical expertise development[10] and naturalistic decision making.[101] We should engage in research that studies cognition and expertise development in performance appraisals, and we need to understand the biases, differences in expertise, conditions affecting judgement and the facilities or interventions that can support judgement.

We recommended meaningful aggregation of information across assessment sources within programmes of assessment. Across these sources, information about the learner is triangulated until a firm, defensible decision can be inferred. However, when is 'enough' enough?[102] When do we stop gathering additional information? Qualitative research would say, 'when saturation is reached'. To some extent, this is the counterpart of reliability or generalisability in psychometric research. However, across multiple sources of information, particularly if the information is different in nature (i.e., quantitative and qualitative), psychometric theory falls short.[91] From a philosophical perspective, matters get even worse if we challenge the assumption, underlying psychometric theory, of the existence of a 'true score'. If we have to rely on expert judgement, we rely on judgements that are idiosyncratically constructed realities unique to individual judges. Multiple judges therefore have multiple constructed realities, which may not or only partly coincide. Does this make them less useful? We think not. It may actually be highly relevant and beneficial to individual learners to be exposed to different perspectives. We therefore prefer triangulation and saturation of information as concepts to guide aggregate decision making. When the probability of finding new information is low, saturation is achieved and this justifies discontinuation of the search for further evidence. Nevertheless, can this process be further formalised? Can we think of certain probabilistic rules to guide this decision making? Bayes' theorem seems an attractive model, at least in theory, because it interprets the value of new information in the light of prior information. However, attempts to apply it to assessment decisions are non-existent, at least to our knowledge. Thinking about formalised models of decision making might guide our efforts to make decisions at the aggregated programmatic level more robust, and such models are therefore interesting to explore further in research.

A third area for research is how assessment drives learning. This relationship is poorly understood and we badly need more empirical and theoretical input. Laboratory studies have convincingly shown that assessment enhances retention and performance,[103] but studies of summatively oriented assessment programmes, on the other hand, have revealed quite a few imposing, surface-oriented, negative effects on learning.[104–107] The effect of learning is mediated by the learner's perceptions of the assessment programme,[108] and these perceptions and the resulting learning strategies can be very resistant to change.[109] Perceptions of learners and teachers may actually be quite opposite and conflicting.[107] In all, traditional summative programmes of assessment appear to have quite a negative effect on learning. The question then is how to change? From reviews on feedback studies, we learn that grades provide poor feedback and hardly influence learners.[85] Some data even suggests that grades impair learning.[110] Solutions need to be sought in integral programmatic systems of intensive formative feedback[57] with careful implementation strategies to ensure that learning behaviour is fundamentally influenced through the formative assessment.[80] How to balance formative and summative assessment, how to implement effective formative strategies to change the perceptions and behaviour of learners and assessment developers – these are important questions that need to be addressed. What is clear is that assessment is the key to (not) achieving deeper learning strategies. What we need to learn is how to use this key.

Finally, we need more research and development to inform the design of assessment programmes. Virtually all literature on assessment is concerned with individual assessment methods. However, how do we construct and organise assessment programmes? Pioneering work has been done to define quality characteristics of assessment programmes,[36] which have been operationalised in a useful self-assessment instrument.[111] Next on the agenda is the design of guidelines. Recently, we proposed a model for developing such guidelines,[112] and a next step would be to actually formulate and reach a consensus on proposed guidelines. We believe this work is important to shape our thinking on how to

advance assessment; we need to bridge the gap between test construction and curriculum construction, also in the complicated world of postgraduate training with its specific demands.

To summarise, ultimately, all the principles of assessment are interrelated and interacting. We need to work on assessment programmes that foster learning effectively, that use a mixture of methods and procedures informed by evidence of their utility, and that promote societal accountability by providing rich and robust information to support learner quality, safe and independent practice and ultimately the proof of good patient care. The principles we discussed need further elaboration, discussion and research. To us, they form the building blocks for theory development on assessment. Ultimately, such a theory can guide us in realising the best possible assessment programmes for the future of medical education.

Practice points

Assessing 'knows', 'knows how', 'shows how'

- Competence is specific, not generic.
- Objectivity is not the same as reliability.
- What is being measured is determined more by the stimulus format than by the response format.
- All methods of assessment can have 'built-in' validity.
- Assessment drives learning.
- No single method can do it all.

Assessing 'Does'

- A feasible sample is required to achieve reliable inferences.
- Bias is an inherent characteristic of expert judgement.
- Validity lies in the users of the instruments, more than in the instruments.
- Formative and summative assessment are typically combined.
- Qualitative, narrative information carries a lot of weight.
- Summative decisions can be rigorous by using non/psychometric qualitative procedures.

Research agenda

- How do assessors reason and arrive at judgements?
- Can we model information seeking and decision making in formalised models?
- How does assessment drive learning?

References

1. Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract* 1996; **1**(1)**:** 41–67.
2. Van der Vleuten CPM & Schuwirth LWT. Assessment of professional competence: from methods to programmes. *Med Educ* 2005; **39:** 309–317.
3. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990; **65**(9)**:** S63–S67.
4. Eva KW. On the generality of specificity. *Med Educ* 2003 Jul; **37**(7)**:** 587–588.
5. Elstein AS, Shulmann LS & Sprafka SA. *Medical problem-solving: an analysis of clinical reasoning.* Cambridge, MA: Harvard University Press, 1978.
6. Elstein AS, Shulman LS & Sprafka SA. Medical problem solving: a ten year retrospective. *Eval Health Prof* 1990; **13:** 5–36.
7. Van der Vleuten CPM & Swanson D. Assessment of clinical skills with standardised patients: state of the art. *Teach Learn Med* 1990; **2**(2)**:** 58–76.

8. Shavelson RJ, Baxter GP & Gao X. Sampling variability of performance assessments. *Journal of Educational Measurement* 1993; **30:** 215–232.
9. Norman G. Teaching basic science to optimise transfer. *Med Teach* 2009 Sep; **31**(9)**:** 807–811.
10. Regehr G & Norman GR. Issues in cognitive psychology: implications for professional education. *Acad Med* 1996; **71**(9)**:** 988–1001.
11. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Acad Med* 2004; **79**(Suppl. 10)**:** S70–S81.
12. Van der Vleuten CPM, Norman GR & De Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ* 1991; **25:** 110–118.
13. Swanson DB. A measurement framework for performance-based tests. In Hart I & Harden R (eds.). *Further developments in assessing clinical competence*. Montreal: Can-Heal Publications, 1987, pp. 13–45.
14. Wass V, Wakeford R, Neighbour R et al. Achieving acceptable reliability in oral examinations: an analysis of the royal college of general practitioners membership examination's oral component. *Med Educ* 2003 Feb; **37**(2)**:** 126–131.
15. Petrusa ER. Clinical performance assessments. In Norman GR, Van der Vleuten CPM & Newble DI (eds.). *International handbook for research in medical education*. Dordrecht: Kluwer Academic Publishers, 2002, pp. 673–709.
16. Rothman AI, Blackmore D, Dauphinee WD et al. The use of global ratings in OSCE station scores. *Adv Health Sci Educ Theory Pract* 1997; **1:** 215–219.
17. Regehr G, MacRae H, Reznick R et al. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med* 1998; **73**(9)**:** 993–997.
18. Hodges B, Regehr G, McNaughton N et al. OSCE checklists do not capture increasing levels of expertise. *Acad Med* 1999 Oct; **74**(10)**:** 1129–1134.
19. Norman G. Editorial-Checklists vs. ratings, the illusion of objectivity, the demise of skills and the debasement of evidence. *Adv Health Sci Educ Theory Pract* 2005; **10**(1)**:** 1–3.
20. Schuwirth LW & Van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? *Med Educ* 2004 Sep; **38**(9)**:** 974–979.
21. Ward WC. A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Appl Psychol Meas* 1982; **6**(1)**:** 1–11.
22. Swanson DB, Norcini JJ & Grosso LJ. Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education* 1987; **12**(3)**:** 220–246.
23. Schuwirth LW, Verheggen MM, Van der Vleuten CP et al. Do short cases elicit different thinking processes than factual knowledge questions do? *Med Educ* 2001; **35**(4)**:** 348–356.
24. Skakun EN, Maguire TO & Cook DA. Strategy choices in multiple-choice items. *Acad Med* 1994; **69**(10 Suppl)**:** S7–S9.
25. Van der Vleuten CPM, Van Luijk SJ & Beckers HJM. A written test as an alternative to performance testing. *Med Educ* 1989; **23:** 97–107.
26. Case SM & Swanson DB. *Constructing written test questions for the basic and clinical sciences*. Philadelphia: National Board of Medical Examiners, 2002.
27. Page G & Bordage G. The medical council of Canada's key features project: a more valid written examination of clinical decision-making skills. *Acad Med* 1995; **70**(2)**:** 104–110.
28. Sweller J. Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction* 1994; **4:** 295–312.
29. Verhoeven BH, Verwijnen GM, Scherpbier AJJA et al. Quality assurance in test construction: the approach of a multi-diciplinary central test committee. *Educ Health* 1999; **12**(1)**:** 49–60.
30. Jozefowicz RF, Koeppen BM, Case SM et al. The quality of in-house medical school examinations. *Acad Med* 2002; **77**(2)**:** 156–161.
31. Ebel RL. The practical validation of tests of ability. *Educational Measurement: Issues and Practice* 1983; **2**(2)**:** 7–10.
32. Van der Vleuten CP, Schuwirth LW, Muijtjens AM et al. Cross institutional collaboration in assessment: a case on progress testing. *Med Teach* 2004 Dec; **26**(8)**:** 719–725.
33. Messick S. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 1995; **23:** 13–23.
34. McLachlan JC. The relationship between assessment and learning. *Med Educ* 2006 Aug; **40**(8)**:** 716–717.
35. Van Luijk SJ, Van der Vleuten CPM & Van Schelven RM. The relation between content and psychometric characteristics in performance-based testing. In Bender W, Hiemstra RJ, Scherpbier AJJA et al (eds.). *Teaching and assessing clinical competence*. Groningen: Boekwerk Publications, 1990. pp 202–207.
36. Baartman LKJ, Bastiaens TJ, Kirschner PA et al. The wheel of competency assessment. Presenting quality criteria for competency assessment programmes. *Studies in Educational Evaluations* 2006; **32**(2)**:** 153–170.
37. (ACGME) ACoGME. Common program requirements: general competencies [updated 2009; cited]; Available from, http://www.acgme.org/outcome/comp/GeneralCompetenciesStandards21307.pdf; 2009.
38. Frank JR & Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. *Med Teach* 2007 Sep; **29**(7)**:** 642–647.
39. Meng C. Discipline-specific or academic? acquisition, role and value of higher education competencies. Maastricht: PhD Dissertation, Maastricht University, 2006.
40. Papadakis MA, Hodgson CS, Teherani A et al. Unprofessional behavior in medical school is associated with subsequent disciplinary action by a state medical board. *Acad Med* 2004 Mar; **79**(3)**:** 244–249.
41. Papadakis MA, Teherani A, Banach MA et al. Disciplinary action by medical boards and prior behavior in medical school. *N Engl J Med* 2005 Dec 22; **353**(25)**:** 2673–2682.
42. Tamblyn R, Abrahamowicz M, Dauphinee D et al. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *JAMA* 2007 Sep 5; **298**(9)**:** 993–1001.
43. Ginsburg S, Regehr G, Hatala R et al. Context, conflict, and resolution: a new conceptual framework for evaluating professionalism. *Acad Med* 2000 Oct; **75**(10 Suppl)**:** S6–S11.
44. Norman GR, Van der Vleuten CPM & De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Med Educ* 1991; **25:** 119–126.

45. Ten Cate O & Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad Med* 2007 Jun; **82**(6)**:** 542–547.
46. Lurie SJ, Mooney CJ & Lyness JM. Measurement of the general competencies of the accreditation council for graduate medical education: a systematic review. *Acad Med* 2009 Mar; **84**(3)**:** 301–309.
47. Kogan JR, Holmboe ES & Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA* 2009 Sep 23; **302**(12)**:** 1316–1326.
48. Norcini JJ, Blank LL, Arnold GK et al. The mini-CEX (Clinical Evaluation Exercise): a preliminary investigation. *Ann Intern Med* 1995; **123:** 795–799.
49. Wilkinson JR, Crossley JG, Wragg A et al. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Med Educ* 2008 Apr; **42**(4)**:** 364–373.
50. Cruess R, McIlroy JH, Cruess S et al. The professionalism mini-evaluation exercise: a preliminary investigation. *Acad Med* 2006 Oct; **81**(10 Suppl)**:** S74–S78.
51. Ram P, Grol R, Rethans JJ et al. Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of validity, reliability and feasibility. *Med Educ* 1999; **33**(6)**:** 447–454.
52. Norcini JJ. The metric of medical education: peer assessment of competence. *Med Educ* 2003; **37**(6)**:** 539–543.
53. Lockyer J. Multisource feedback in the assessment of physician competencies. *J Cont Educ Health Prof* 2003; **23:** 2–10.
54. Williams RG, Klamen DA & McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med* 2003; **15**(4)**:** 270–292.
55. Van Tartwijk J & Driessen EW. Portfolios for assessment and learning: AMEE Guide no. 45. *Med Teach* 2009 Sep; **31**(9)**:** 790–801.
56. Scheele F, Teunissen P, Van Luijk S et al. Introducing competency-based postgraduate medical education in the Netherlands. *Med Teach* 2008; **30**(3)**:** 248–253.
57. Dannefer EF & Henson LC. The portfolio approach to competency-based assessment at the Cleveland Clinic Lerner College of Medicine. *Acad Med* 2007 May; **82**(5)**:** 493–502.
58. Driessen E, Van Tartwijk J, Van der Vleuten C et al. Portfolios in medical education: why do they meet with mixed success? a systematic review. *Med Educ* 2007 Dec; **41**(12)**:** 1224–1233.
59. Butler PA. *Review of the literature on portfolios and electronic portfolios*. Palmerston North, New Zealand: Massey University College of Education, 2006. Contract No.: Document Number.
60. Tochel C, Haig A, Hesketh A et al. The effectiveness of portfolios for post-graduate assessment and education: BEME Guide No 12. *Med Teach* 2009 Apr; **31**(4)**:** 299–318.
61. Pelgrim EAM, Kramer AWM, Mokkink H et al. In training assessment, using direct observation of patient encounters, a systematic literature review. Under editorial review.
62. Falchikov N & Goldfinch J. Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks. *Rev Educ Res* 2000; **70**(3)**:** 287–322.
63. Atwater LE, Waldman DA & Brett JF. Understanding multi-source feedback. *Hum Resour Manage* 2002; **41**(2)**:** 193–208.
64. Archer JC, Norcini J & Davies HA. Use of SPRAT for peer review of paediatricians in training. *BMJ* 2005 May 28; **330**(7502)**:** 1251–1253.
65. Wood L, Hassell A, Whitehouse A et al. A literature review of multi-source feedback systems within and without health services, leading to 10 tips for their successful design. *Med Teach* 2006 Nov; **28**(7)**:** e185–e191.
66. Whitehouse A, Hassell A, Bullock A et al. 360-degree assessment (multisource feedback) of UK trainee doctors: field testing of team assessment of behaviours (TAB). *Med Teach* 2007 Mar; **29**(2-3)**:** 171–176.
67. Davies H, Archer J, Bateman A et al. Specialty-specific multi-source feedback: assuring validity, informing training. *Med Educ* 2008 Oct; **42**(10)**:** 1014–1020.
68. Plous S. *The psychology of judgment and decision making*. New Jersey: McGraw-Hill Inc., 1993.
69. Govaerts MJ, Van der Vleuten CP, Schuwirth LW et al. Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Adv Health Sci Educ Theory Pract* 2007; **12:** 239–260.
70. Dudek NL, Marks MB & Regehr G. Failure to fail: the perspectives of clinical supervisors. *Acad Med* 2005 Oct; **80**(10 Suppl)**:** S84–S87.
71. Bullock AD, Hassell A, Markham WA et al. How ratings vary by staff group in multi-source feedback assessment of junior doctors. *Med Educ* 2009 Jun; **43**(6)**:** 516–520.
72. Boor K, Teunissen PW, Scherpbier AJ et al. Residents' perceptions of the ideal clinical teacher-a qualitative study. *Eur J Obstet Gynecol Reprod Biol* 2008 Oct; **140**(2)**:** 152–157.
73. Gordon MJ. A review of the validity and accuracy of self-assessments in health professions training. *Acad Med* 1991; **66** (12)**:** 762–769.
74. Davis DA, Mazmanian PE, Fordis M et al. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA* 2006 Sep 6; **296**(9)**:** 1094–1102.
75. Eva KW & Regehr G. Self-assessment in the health professions: a reformulation and research agenda. *Acad Med* 2005 Oct; **80**(10 Suppl)**:** S46–S54.
76. Eva KW & Regehr G. Knowing when to look it up: a new conception of self-assessment ability. *Acad Med* 2007 Oct; **82**(10 Suppl)**:** S81–S84.
77. Colthart I, Bagnall G, Evans A et al. The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice: BEME Guide no. 10. *Med Teach* 2008; **30**(2)**:** 124–145.
78. Evans AW, Leeson RM & Petrie A. Reliability of peer and self-assessment scores compared with trainers' scores following third molar surgery. *Med Educ* 2007 Sep; **41**(9)**:** 866–872.
79. Sargeant J, Mann K, Van der Vleuten C et al. "Directed" self-assessment: practice and feedback within a social context. *J Contin Educ Health Prof* 2008 Winter; **28**(1)**:** 47–54.
80. Segers M, Gijbels D & Thurlings M. The relationship between students' perceptions of portfolio assessment practice and their approaches to learning. *Educ Stud* 2008; **34:** 35–44.
81. Sargeant J, Eva KW, Lockyer J et al. The Processes and Dimensions of Informed Self-Assessment: A Conceptual Model. Acad Med in press, doi:10.1097/ACM.0b013e3181d85a4e.

82. Dolmans D, Schmidt A, Van der Beek J et al. Does a student log provide a means to better structure clinical education? *Med Educ* 1999; **33:** 89–94.
83. Kluger AN & DeNisi A. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychol Bull* 1996; **119:** 254–284.
84. Hattie J & Timperley H. The power of feedback. *Rev Educ Res* 2007; **77:** 81–112.
85. Shute VJ. Focus on formative feedback. *Rev Educ Res* 2008; **78:** 153–189.
86. Pendleton D, Scofield T, Tate P et al. *The consultation: an approach to learning and teaching*. Oxford: Oxford University Press, 1984.
87. Boor K, Scheele F, Van der Vleuten CP et al. How undergraduate clinical learning climates differ: a multi-method case study. *Med Educ* 2008 Oct; **42**(10)**:** 1029–1036.
88. Sargeant J, Mann K, Sinclair D et al. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Adv Health Sci Educ Theory Pract* 2008 Aug; **13**(3)**:** 275–288.
89. Govaerts MJ, Van der Vleuten CP, Schuwirth LW et al. The use of observational diaries in in-training evaluation: student perceptions. *Adv Health Sci Educ Theory Pract* 2005 Aug; **10**(3)**:** 171–188.
90. Hodges B. Medical education and the maintenance of incompetence. *Med Teach* 2006 Dec; **28**(8)**:** 690–696.
91. Schuwirth LWT & Van der Vleuten CPM. A plea for new psychometric models in educational assessment. *Med Educ* 2006; **40:** 296–300.
92. Guba EG & Lincoln YS. *Naturalistic inquiry*. Thousand Oaks: Sage, 1985.
93. Schwandt TA. Judging interpretations: but is it rigorous? trustworthiness and authenticity in naturalistic evaluation. *New Directions for Evaluation* 2007; **114:** 11–25.
94. Driessen EW, Van der Vleuten CPM, Schuwirth LWT et al. The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. *Med Educ* 2005; **39**(2)**:** 214–220.
95. Norcini JJ & Shea JA. The credibility and comparability of standards. *Applied Measurement in Education* 1997; **10**(1)**:** 39–59.
96. Papadakis MA, Osborn EH, Cooke M et al. A strategy for the detection and evaluation of unprofessional behavior in medical students. University of California, San Francisco School of Medicine Clinical Clerkships Operation committee. *Acad Med* 1999 Sep; **74**(9)**:** 980–990.
97. Coles C. Developing professional judgment. *J Cont Educ Health Prof* 2002; **22**(1)**:** 3–10.
98. Hastie R & Dawes RM. *Rational choice in an uncertain world*. Thousand Oaks: Sage Publications, 2001.
99. Karelia N & Hogarth RM. Determinants of linear judgment: a meta-analysis of lens model studies. *Psychol Bull* 2008; **134** (3)**:** 404–426.
100. Feder G, Eccles M, Grol R et al. Clinical guidelines: using clinical guidelines. *BMJ* 1999 Mar 13; **318**(7185)**:** 728–730.
101. Klein G. Naturalistic decision making. *Hum Factors* 2008; **50:** 456–460.
102. Schuwirth LW, Southgate L, Page GG et al. When enough is enough: a conceptual basis for fair and defensible practice performance assessment. *Med Educ* 2002 Oct; **36**(10)**:** 925–930.
103. Larsen DP, Butler AC & Roediger 3rd HL. Test-enhanced learning in medical education. *Med Educ* 2008 Oct; **42**(10)**:** 959–966.
104. Harlen W & Crick RD. Testing and motivation for learning. *Assessment in Education* 2001; **10:** 169–207.
105. Harlen W. Teachers' summative practices and assessment for learning – tensions and synergies. *The Curriculum Journal* 2005; **16:** 207–223.
106. Cillier F. Assessment impacts on learning, you say? please explain how. The impact of summative assessment on how medical students learn. In preparation.
107. Al Kadri HMF, Al-Moamary MS & Van der Vleuten C. Students' and teachers' perceptions of a clinical assessment program: a qualitative study in a PBL curriculum. Under editorial review.
108. Segers M, Nijhuis J & Gijselaers W. Redesigning a learning and assessment environment: the influence on students' perceptions of assessment demands and their learning strategies. *Studies in Educational Evaluation* 2006; **32:** 233–242.
109. Gijbels D, Segers M & Struyf E. Constructivist learning environments and the (im)possibility to change students' perceptions of assessment demands and approaches to learning. *Instructional Science* 2008; **36:** 431–443.
110. Nyquist JB. *The benefits of reconstruing feedback as a larger system of formative assessment: a meta-analysis*. Nashville, Te: Vanderbilt University, 2003.
111. Baartman LKJ, Prins FJ, Kirschner PA et al. Determining the quality of assessment programs: a self-evaluation procedure. *Studies in Educational Evaluation* 2007; **33:** 258–281.
112. Dijkstra J, Van der Vleuten CP & Schuwirth LW. A new framework for designing programmes of assessment. *Adv Health Sci Educ Theory Pract* 2009 Oct 10.