4 The psychometrics and edumetrics of performance assessment

Jean-Marie Degryse, Valéry Dory, Cees P. M. Van der Vleuten

4.1 Introduction: performance assessment – climbing Miller’s pyramid

For over more than twenty years now, Miller’s pyramid⁴⁹ has been used as a framework to define different levels of assessment of medical competence (see chapter 1). Historically, emphasis has been placed on assessment at the lower layers, directed at knowledge, application of knowledge and demonstration of skills. More recent developments are concentrated at the top: the “does” level. Performance assessment is predominantly assessment in the workplace. Workplace-based assessment⁵⁰ is likely to become an essential part of both licensure and re-certification procedures in family medicine. Figure 3 offers an overview of the different assessment methods that are commonly used at different levels of Miller’s pyramid.

Figure 3: Miller’s pyramid showing different levels of assessment and the associated assessment formats

From the psychometric perspective, workplace-based assessment offers new challenges. Some researchers point to the threats to reliability and validity from uncontrollable variables, such as patient mix, case difficulty and patient num-

bersⁱ. Others show that the utility of performance assessment results is compromised by lower inter-rater reliability and rater effects such as halo, leniency or range restriction⁵¹. As a consequence attempts to improve performance assessment typically focus on standardization and objectivity of measurement by adjusting rating scale formats and eliminating rater errors through rater training⁵².

Traditional psychometric approaches towards assessment tend to focus exclusively on quantitative properties of assessment outcomes. However in this chapter we will advocate that this rigorous psychometric approach might limit more meaningful approaches to performance assessment. Performance assessment is essentially a judgment and decision making process in which rating outcomes are influenced by interactions between individuals and the social context in which assessment occurs. Also, depending on the “stakes” of the performance assessment, the focus as well as the design of the procedure will be different. In a “low stake” context, emphasis will be put on the production of rich and meaningful (multi-source) feedback in order to support and enhance learning. While in a “high stake” context the focus will shift to the gathering and aggregation of information from as many measurements as possible in order to produce a *robust* final judgment and decision as an alternative to a single “final score”. Both designs require viewing of the individual doctor’s performance from different angles as visualized in the inverted Miller’s triangle, the Cambridge Model (see figure 2 in chapter 1).

Performance assessment in daily practice needs a theoretical framework to be scientifically sound. Therefore, the practical chapters 5–10 of EUPA are flanked by this theoretical chapter about the psychometrics and edumetrics of performance assessment. We will first provide an overview of the lessons that can be derived from the assessment literature and three decades of research. They have been synthesized in a paper by Van der Vleuten et al⁵³ and relate in particular to assessment of the three first layers of Miller’s pyramid. We will then confront those with the typical demands and constraints of workplace-based assessment and present some of the solutions and alternative approaches that have been proposed to assess medical competence at the top layer.

ⁱ We use Workplace-Based Assessment here as a term that relates to performance assessment in a “real life” professional environment (see chapter 11 – Glossary).

4.2 Lessons learned from three decades of research

4.2.1 *Competence is specific, not generic: Content specificity as a central issue in assessment of medical competence*

An issue that dominates (bedevils) every effort to develop a reliable assessment procedure at any level of Miller's pyramid is the problem of "content-specificity". In assessment of medical competence, more than in any other field, the scores on a test (or a case, or an item) that was designed to assess a specific content domain predict scores on another domain very poorly. This variability of performance of candidates across task appears to be one of the most consistent findings in all measurements of clinical competence⁵⁴. A direct consequence of this *content-specificity* is that assessment blueprints must ensure broad sampling across a huge variety of task in order to achieve sufficient reliability. Or in other words: tests containing a small sample of items (stations, observations, tasks, patient encounters) produce unstable or unreliable scores. Obviously this will also vary with the size of the domain being assessed, but even in smaller content domains the required sample size of test items/cases is usually high. In this respect some testing methods that allow broad sampling in less time might be considered as more reliable per unit of testing time than tests requiring more time per item. In order to produce adequate reliability coefficients within a high stake context, one should take into account that even efficient tests usually require several hours of testing time⁵⁵.

Research on a variety of assessment methods has consistently revealed that the most important determinant of score reliability is test length as is illustrated from the overview in table 1. Other determinants of variability that challenge the reliability of assessments, such as rater, patient or observer variability, are usually either less important or easier to control.

A second practical consequence of this content specificity phenomenon is that when limited resources are available, the design of the assessment procedure should preferentially be adapted so that broad sampling of the domain that will be tested is guaranteed and a maximum number of cases/items is introduced. For instance when a limited number of observers are available as raters of performance in stations in an Objective Structured Clinical Examination (OSCE), it is preferable to spread those over double as many stations instead of appointing them in pairs to half the number of stations⁵⁶.

Testing Time in Hours	MCQ ¹	Case-Based Short Essay ²	PMP ¹	Oral Exam ³	Long Case ⁴	OSCE ⁵	Mini CEX ⁶	Practice Video Assessment ⁷	Incognito SPs ⁸
1	0.62	0.68	0.36	0.50	0.60	0.47	0.73	0.62	0.61
2	0.76	0.73	0.53	0.69	0.75	0.64	0.84	0.76	0.76
4	0.93	0.84	0.69	0.82	0.86	0.78	0.92	0.93	0.92
8	0.93	0.82	0.82	0.90	0.90	0.88	0.96	0.93	0.93

¹Norcini et al., 1985

²Stalenhoef-Halling et al., 1990

³Swanson, 1987

⁴Wass et al., 2001

⁵Petrusa, 2002

⁶Norcini et al., 1999

⁷Ram et al., 1999

⁸Gorter, 2002

Table 1: The reliability as a function of testing time (Van der Vleuten et al 2005⁹¹)

If observational test methods are used (as in level 3 or 4 of Miller's pyramid), other secondary factors might improve reliability. Standardization of the content of the assessment, as well as training of the observers, might contribute to a further improvement of the reliability.

A final issue that should be addressed here is that "reliability is relative" i. e. the reliability of a measure is intimately linked to the population to which one wants to apply the measure. There is literally no such thing as the reliability of a test. A reliability coefficient has meaning only when applied to a specific population. Reliability is not a fixed characteristic of an assessment method but will depend upon the "true variability" of the scores in a population.

In order to understand this statement, we should consider the way reliability is conceptualized in classical test theory. Reliability is defined as the relationship between the "true variance of the scores" to the observed variance of test scores, the latter being defined as the sum of the true variance and a variance linked to measurement error. In a formal way this can be expressed as:

$$reliability = \frac{\sigma_t^2}{\sigma_o^2} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}$$

Thus the reliability coefficient expresses the proportion of the total variance in the measurements ($\sigma_t^2 + \sigma_e^2$) that is due to "true differences between subjects" (σ_t^2). If σ_e^2 is zero, the reliability coefficient will be 1. Assuming that σ_e^2 remains constant, higher σ_t^2 will produce higher reliability coefficients.

4.2.2 Objectivity is not the same as reliability

Reliability does not co-vary with the objectivity of methods. The so called “subjective” tests can be reliable and objective tests can be unreliable, all depending on the sampling and the method⁵⁷. As an example we can refer to the historical development of the Objective Clinical Examination format. The OSCE was originally designed in order to overcome the problem of subjectivity that was present in classical clinical exams. The solution was sought in objectivity and standardization of the score forms and cases and in training of the observers. However the reliability of the OSCE turned out to be as dependent of the broad sampling of its content (i. e. the stations) as any other method⁵⁸. Table 2 shows the effect of test length (expressed as number of stations) on the reliability of a typical OSCE that is part of a national licensing examination of family physicians in Flanders.

Test length (minutes)	Number of stations	Reliability index	
		Norm referenced	Domain referenced
80	10	.60	.55
120	15	.69	.64
160	20	.75	.71
200	25	.79	.75
240	30	.82	.78
280	35	.84	.81
320	40	.86	.83

Table 2: Reliability indices of the OSCE as a function of testing time (Degryse 2003)

Another interesting finding from research on the OSCE is the strong correlation that was found between global rating scales and checklist ratings⁵⁹. Global ratings are associated with a slight decrease in inter rater reliability, but this was offset by a larger gain in inter station reliability.

Compared with the more analytical checklist scores, global holistic judgments tend to pick up on elements in candidates' performance which were more generalisable across stations. This is a clear and intriguing first indication that human expert judgment could add (perhaps even incrementally) a meaningful “signal” to measurement instead of “noise”.⁶⁰

Another issue that should be addressed here is that reliability is not necessarily related to agreement between judges/raters and in some cases it can even be inversely related to it! The scores produced by different raters in an observational test can vary consistently but still generate reliable mean scores. On the other hand, if all candidates on all occasions are rated above average, the agreement among raters is perfect but the reliability by definition is zero. What actually happens is an intriguing and somewhat paradoxical interaction between the problem of content specificity and assessor variability. The global nature of the judgment may be more subjective, but it dampens or attenuates the content specificity problem and therefore leads to improved reliability overall.

4.2.3 *Compromising and making informed choices*

Every assessment procedure is a by definition a compromise. The selection of a particular assessment method may involve factors that are not of a psychometric nature. Trade-offs between what is desirable and achievable or feasible are inevitable.

In order to clarify the compromises involved, Van der Vleuten et al.^{61,62} have presented some years ago a model to define the utility of an assessment method (figure 4). Reliability (R), Validity (V) and educational impact (E) should be part of the model for obvious reasonsⁱⁱ. However, in educational practice the choice of a particular assessment method will often be influenced by other considerations such as opinions, sentiments and traditions of teachers, students and institutions. That is why the model introduces two additional variables: the acceptability (A) and the cost (C) and logistic burden induced by the procedure. The utility of an assessment method is represented by an equation in which the relationship among variables is deliberately conceived as multiplicative. If one of the elements is zero, the utility will be zero.

In practice, one will always be required to compromise and assign different weights in different individual situations, depending on the context and purpose of the assessment. In situations where the assessment involves a high-stake examination with decisions having marked consequences on the future of ex-

ⁱⁱ It should be noted that from a psychometric perspective reliability and validity are not independent variables. Reliability is a prerequisite for validity. The reliability determines the maximum validity that can be achieved. The relation between reliability and validity can be expressed mathematically as follows: $r_{ig} = \sqrt{r_{ii}}$ where r_{ii} stands for reliability and r_{ig} for the validity of a test.

aminees, reliability will probably bear a heavier weight. While in the context of in-training assessment, where the final decision is based on many assessments, one would probably be prepared to compromise more on reliability in favour of the educational impact of the assessment⁶³. In other situations a higher weight should be attributed to the potential educational impact of a procedure even if one has to compromise on reliability. For instance in a CPD programme, a less reliable test procedure might be preferred due to the inherent signal it sends out which guides life-long learning of practitioners on a micro and macro level. In this context it is essential that the assessment procedure is at its most transparent in order to allow the communication of information and an optimal steering effect.

$$U = w_r R \times w_v V \times w_e E \times w_a A \times w_c C$$

U = Utility	R = Reliability	V = Validity
E = Educational impact	A = Acceptability	C = Cost
W = Weight		

Figure 4: The utility function of an assessment procedure

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

It has often been assumed that different assessment methods measure different aspects of medical competence. Summative assessment procedures and/or licensing exams have often purposefully been designed as a battery of tests with different formats. The underlying assumption has been for instance that Multiple Choice Question (MCQ) tests would focus on knowledge, OSCEs on clinical skills, and more sophisticated written tests or computer based tests on elements of clinical reasoning.

However, research has revealed that what is being measured is determined more by the stimulus format than by the response format. Provided the scores are reliable (which can be ensured by broad sampling), scores obtained on tests using different formats generally correlate with each other quite well. Cognitive activities follow the task that is posed in the stimulus format. A well designed

written knowledge test with rich contextual clinical vignettes can measure much more than factual knowledge, and conversely a poorly designed OSCE can target only rudimentary, decontextualized, and technical skills. Validity – what is being measured – is not so much determined by the response format as by the stimulus format⁶⁴.

A practical consequence is that test developers should worry more about designing appropriate stimulus formats than about appropriate response formats. In developing stimuli, *authenticity* is essential also, provided the stimulus is tuned to the appropriate level of complexity. The first OSCEs consisted of short stations assessing clinical skills in a fragmented manner (e.g. station 1 examination of the shoulder, station 2: abdominal examination, station 3: communication), which may be defensible at early stages of training. But, at a more advanced stage of training, integrated skills assessment is obviously a more appropriate stimulus format, since it provides a closer approximation of real clinical encounters.

4.2.5 *Assessment drives learning*

The impact of assessment, from selection to certification, on learning is significant⁶⁵. Before assessment even occurs, it influences learning by providing an external motivation and also by providing learners with cues which they will interpret in order to conduct and regulate their learning⁶⁶. Learners are generally strategic in their efforts. They will allocate time and select learning and regulation strategies according to their understanding of the task at hand. This can lead to negative effects, e.g. learners using surface strategies such as rote learning to prepare for MCQs, or indeed memorizing OSCE checklist items. This can lead to a trivialization of the learning endeavour and counteract the effects of other components of the curriculum^{67,68}.

During a test, learners are provided with an opportunity to practice a task which in itself can lead to improved performance. This type of impact is referred to as the testing effect and has been demonstrated for both tests of knowledge and tests of skills^{69,70}.

Assessment also has a significant impact by providing feedback to learners, what Norcini et al refer to as the catalytic effect of assessment⁷¹. Feedback has been shown to be conducive to learning providing a certain number of conditions are met: the feedback should be timely, specific, based on observation, pro-

vided by a credible person, and qualitative rather than numerical^{72,73,74}. Indeed feedback is not necessarily assimilated directly by learners. A study on general practitioners receiving multisource feedback found that assimilation depended on a process of reflection which balanced external feedback with evaluations based on self-assessment⁷⁵. This process was also influenced by the emotional reaction of learners to feedback. Negative feedback that was inconsistent with an individual's self-assessment was more likely to be discounted. The authors suggest that this reflective process could be facilitated by a coach: interpreting feedback and using these interpretations to steer learning should be viewed as a complex sense-making process influenced by social interactions and the learning climate.

In view of the significant impact of assessment on learning, several authors have called for a shift from assessment of learning to assessment for learning, in which assessment is viewed as a key component of the curriculum and is used strategically to foster desirable learning behaviours⁷⁶.

4.3 Assessing competence at the “does” level

Any assessment at the top level of Miller's pyramid in an authentic context, will rely to a large extent on “expert judgment”. The term expert should be interpreted broadly to include peers, supervisors, co-workers, teachers, and anyone knowledgeable about the work of the trainee/doctor, perhaps even the learner him or herself. The assessment consists of gathering these judgments in some quantitative or qualitative form. As with OSCEs the dominant response format is some form of observation structure (numerical rating scale, scoring rubrics, free text boxes) on which a judgment is based. Unlike the OSCE, however, the stimulus format is the authentic context, which is essentially unstandardized and relatively unstructured. In addition to scoring performance on rating scales, assessors are often invited to write narrative comments about the strengths and weaknesses of a learner's performance. Roughly sketched two types of assessment instruments have been used. The first involves judgment 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 from multiple sources over time⁷⁷ (table 3).

Direct performance measures	
<ul style="list-style-type: none"> • Individual encounter methods <i>Assessment is confined to a single concrete situation</i> 	
e. g. Mini-clinical-evaluation exercise (Mini-Cex)	
Direct observation of practice skills (DOPS)	
Professionalism mini-evaluation (P-Mex)	
Video-observation of clinical encounters	
<ul style="list-style-type: none"> • Long-term methods <i>Performance is assessed over a longer period of time</i> 	
e. g. Peer assessment	
Multisource or 360 ° feedback (MSF)	
Aggregation methods	
e. g. Logbook	<i>Continuous sampling of performance</i>
Portfolio	

Table 3: Overview of methods used to assess medical competence at the “does” level

4.3.1 A sample is required to achieve reliable inferences

All modern methods of assessment at the “does” level allow for frequent sampling across educational or clinical contexts and across assessors. The need to deal with content specificity (see above) means that sampling across a range of context remains invariably important⁷⁸. The subjectivity of expert judgments needs to be counterbalanced by additional sampling across assessors in order to produce “aggregated” information. Rather than targeting the production of a “reliable” final score, the final aim of any assessment at the “does level” will be to produce a robust judgment.

4.3.2 Bias in expert judgment

Global judgments are prone to bias, probably much more than structured analytical methods. Two examples: with direct observations, inflation of scores has often been noted⁷⁹, and in multisource feedback, careless selection of assessors can induce important bias⁸⁰. The context in which the assessment takes place is another important potential source of bias in particular when the assessor(s) have an educational relationship with the trainee or junior doctor. They might be tempted to inflate scores to avoid negative impacts on learners, the relationship or themselves (e. g. having to justify failing a learner)⁸¹. One solution to reduce this type of bias could be to remove the summative aspect of the assessment from the individual encounter. The tutor/assessor should not assess whether the learner is a good doctor, but concentrate on what happens in a specific encoun-

ter in order to produce meaningful feedback. Summative decisions (pass/fail) should be based on multiple sources of assessment within and across methods and should preferentially involve external assessors. As mentioned earlier: the robustness of the final decision will rely on the meaningful aggregation of different sources of information.

There is a growing interest in understanding rater cognition. Initial studies suggest that in the same way as doctors use illness scripts and examples of actual patients to diagnose diseases, raters use mental models or scripts of learner performance and perhaps exemplars of prior learners to gauge learner performance^{82,83,84,85}. This may explain the somewhat idiosyncratic nature of judgments based on personal experience with learners. Training may help communities of raters develop shared mental models and improve the consistency of ratings across raters.

4.3.3 The role of self-assessment

Trainees as well as practicing doctors are very poor self assessors⁸⁶. Broadly speaking, self-assessment is a judgment one makes about oneself. Different authors have put forth different categories of self-assessment. Eva et al.⁸⁷ have proposed to distinguish broad self-assessment (e.g. do I have a good sense of humour? Am I good enough in managing congestive heart failure?) from self-monitoring which is a more specific moment-to-moment evaluation of how one is doing during an activity. They suggest that broad self-assessment is mainly relevant to continuing professional development which often relies on doctors to assess their learning needs and select appropriate activities, whereas self-monitoring is mainly relevant to autonomous practice (e.g. knowing when to look something up or refer a patient to someone else). Dory et al.⁸⁸ have proposed a four-category classification which distinguishes further between each of the two levels proposed by Eva et al., with the most general level referring to sweeping evaluations such as self-concept and self-esteem, and the most specific referring solely to metacognitive monitoring, i.e. monitoring of one's mental processes. Self-assessment, particularly at the most general levels, has been found to be inaccurate. This is partly due to issues related to aggregating information from memories of several relevant episodes, but also to judgment biases such as the "above-average effect", i.e. most people believe that they are above average in a wide variety of domains, which of course cannot be the case⁸⁹.

The practical implication is that self-assessment can never stand on its own and should always be triangulated with other information. Self-assessment should not on the other hand be discarded altogether. As a previously cited study has indicated, self-assessment influences reflection on feedback and coaches should pay deliberate attention to self-assessment when facilitating this process of reflection⁹⁰.

4.3.4 Formative and summative functions are typically combined

Within a classical psychometric framework a clear distinction is made between formative and summative assessment procedures since both the design and the psychometric requirements of the test depend on the purpose of the assessment.

Within a more “edumetric” framework that underpins assessment at the “does” level, an integration of formative and summative functions is advocated. Without formative value, the summative function would be ineffective, leading to trivialization or alienation of the assessment and introduce the risk of a negative effect on the learning process. Also, “no single method can do it all”. Every assessment method and every single point assessment has its own limitations.

Van der Vleuten & Schuwirth⁹¹ therefore strongly advocate a “programmatic approach” of assessment in which assessment and learning are salient. Assessment in general and performance assessment in particular should be embedded in the curriculum (for trainees) or in daily practice (for licensed doctors). Such programmes of assessment cannot be improvised and should be planned, prepared, implemented, evaluated and improved. An assessment programme can be conceptualized as a purposeful collection of assessment moments, i.e. assessment data points. From the sum of those judgments a global picture will emerge that is more than the sum of the individual measurements. The programme should be aligned to the curriculum objectives (for trainees) and/or the job description (for practicing doctors). It should both foster learning and allow sound decisions to be reached. Within such programmes a meaningful aggregation across the available multiple assessment sources is advocated, that is *triangulated* with complementary judgments by external assessors ultimately leading to a defensible and robust summative pass/fail decision.

4.3.5 Qualitative, narrative information carries a lot of weight

The assessment literature has been dominated over the last two decades by the psychometric paradigm and has focused on the development of reliable and valid (normative) measurement instruments that produce scores based on quantifiable concept and quantitative methods. Within the complementary edumetric framework the main focus of assessment is shifted to producing meaningful feedback for learners. Successful feedback is conditional on social interaction. Feedback that consists of nothing more than quantitative information or a score is not very meaningful. Narrative and qualitative information should complement and enrich the feedback. Effective formative assessment is predicated on qualitatively rich information. As a practical consequence rating forms should provide the possibility for assessors to provide complementary narrative feedback.

4.3.6 Summative decisions can be rigorous by using non-psychometric qualitative research procedures

Within the proposed edumetric framework rigor can be defined in a similar way as in qualitative research. The concept of “trustworthiness” has been put forth⁹², and the conventional notion of internal validity is replaced by credibility, external validity by transferability, reliability by dependability and objectivity by conformability.

Table 4 presents some examples of assessment strategies that mirror these trustworthiness strategies and criteria as they have been proposed by Van der Vleuten et al⁹³.

4.4 Miscellaneous issues

4.4.1 Noise and signal in performance assessment

A fundamental issue, far beyond any psychometric reasoning, is the question of the “nature” of medical competence and performance in family medicine. Is the GP who provides good quality care for the 90 percent of daily “trivial” problems performing better than the GP who demonstrates particular acumen in dealing with 10 per cent of particularly challenging conditions/clinical problems s/he will be confronted with? This philosophical question is at the heart of the matter in the debate on how performance should be measured, judged and evaluated.

Strategy to establish trustworthiness	Criteria	Potential Assessment Strategy (sample)
Credibility	Prolonged engagement	Training of examiners
	Triangulation	Tailored volume of expert judgment based on certainty of information
	Peer examination	Benchmarking examiners
	Member checking	Incorporate learner view
	Structural coherence	Scrutiny of committee inconsistencies
Transferability	Time sampling	Judgment based on broad sample of data points
	Thick description	Justify decisions
Dependability	Stepwise replication	Use multiple assessors who have credibility
Confirmability	Audit	Give learners the possibility to appeal to the assessment decision

Table 4: Potential strategies related to qualitative research methodologies for building rigor in assessment decision

We advocate that within a performance assessment programme both the core competencies as well as the more specific/advanced competencies should be looked at, reflecting the complexity of real practice.

4.4.2 What is “unacceptable” performance?

From a psychometric point of view medical competence has often been conceptualized as a “continuous variable” and measured by more or less sophisticated instruments that produce a reproducible and valid score – an approach that can be challenged and debated. Is unacceptable or extremely poor performance a “normative” problem that has to be identified by a “cut-off” score (i.e. a minimum standard)? Or should “poor-unacceptable performance” of a licensed experienced doctor be assessed using a more defensible qualitative approach, i.e. conceptualised as a categorical or nominal variable? Such an approach is based on the assumption that “unacceptable performance” is an extra-ordinary performance that cannot be captured within the logic of a “continuum”. Therefore the term “poor performance” might have to be substituted by another term: for instance “aberrant” performance. We think the instruments aimed at identifying (or screening for) “aberrant” performance in a “rigorous” way, should be different from traditional instruments that have been put forth so far in the literature

and that such judgment should rely on holistic judgments by external assessors, using multi-source feedback, chart review, critical incident analysis and a personal interview.

4.5 Conclusion

From the psychometric perspective workplace-based assessment offers new challenges. Traditional psychometric approaches tend to focus exclusively on quantitative properties of assessment outcomes. Within the proposed complementary edumetric framework, inspiration is found in methodologies from qualitative research and the focus shifts to the gathering of rich contextual information and the role of assessment as a production tool of meaningful feedback to the learner. Broad sampling of meaningful and rich contextual information across educational or clinical contexts and across assessors remains essential in order to produce aggregated information from which – depending on the stakes of the performance assessment – a “robust” final judgment and decision can be made.

Finally, performance assessment should be embedded in the curriculum (for trainees) or in daily practice (for licensed doctors). Such programmes of assessment cannot be improvised and should be planned, prepared, implemented, evaluated and improved. The EURACT Performance Agenda (EUPA) may help to shape and sharpen this process.