Available online at www.sciencedirect.com

ScienceDirect

Health Professions Education 5 (2019) 177–184

www.elsevier.com/locate/hpe

How 'Testing' Has Become 'Programmatic Assessment for Learning'

Lambert W.T. Schuwirth[a,b,*], Cees P.M. van der Vleuten[a,b]

[a]*Department of Education Development and Research, Maastricht University, Maastricht, The Netherlands*
[b]*Prideaux Centre for Research in Health Professions Education, Flinders University, Adelaide, Australia*

Received 22 June 2018; accepted 22 June 2018
Available online 27 June 2018

Abstract

Programmatic assessment for learning is a fundamentally different approach to assessment than the more traditional methods. Yet, it is a logical next step given the history of assessment. In this narrative and subjective review we describe our view on the historical developments in assessment and how they have logically led to the development of programmatic assessment for learning.

The early stages of assessment focussed on measurement of competence with an aim to develop the single best method for each aspect of competence. With the development of competencies the notion of integration and more meaningful assessment emerged but still reductionist issues remained. Programmatic assessment for learning currently seeks to assess students more holistically and meaningfully with rigorous attention to trustworthiness and credibility of the whole assessment process. As such, it may be a revolutionary development but it strongly builds on previous research and insights in the field.

Keywords: Programmatic assessment; Assessment for learning; History of assessment

Contents

*Corresponding author at: Prideaux Centre for Research in Health Professions Education, College of Medicine and Public Health, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia.
E-mail address: lambert.schuwirth@flinders.edu.au (L.W.T. Schuwirth).
Peer review under responsibility of AMEEMR: the Association for Medical Education in the Eastern Mediterranean Region

https://doi.org/10.1016/j.hpe.2018.06.005

1. Introduction

Let us warn you; this is not a systematic nor a thematic review, and perhaps you would not even call it a 'scientific' review. This paper describes the way we have seen and interpreted the developments in the field of assessment of medical competence over the past 50 years. It is our way of understanding where we are today with programmatic assessment for learning. Moreover, it is not even a paper that describes *every* part of the development of assessment of medical competence that has taken place, but merely our selection of what we have seen as the most important aspects. We therefore, do not seek to describe a unique historical truth to increase your knowledge, but to share with you our meaning making to further your understanding.

With respect to the history around assessment of medical competence we distinguish three phases of development and research: assessment as testing, the introduction of competencies and assessment as a programme for learning.

2. Assessment as testing

2.1. Conceptual framework

Clearly, the late 1960s and 1970s have been an important phase in the development of assessment of medical competence. Where before, unstructured oral examinations, long essays and bedside assessments took a significant place in many assessment programmes, a desire emerged to add more structure and standardisation to assessment. In order to achieve this, assessment developers mimicked the approaches from test psychology. Central at that time in test psychology was the notion of stable and generic personality traits. Traits such as 'extraversion', or 'intelligence' were seen as part of a person's personality and to be important factors in determining people's behaviour. In psychology many tests were developed to measure these personality traits and to compare individual test scores to population scores. Probably the most well-known personality test are the Minnesota Multiphasic Personality Inventory or MMPI for character traits and the Wechsler Adult Intelligence Scale or WAIS for intelligence. Given the success of this approach in psychology at that time, a similar approach was adopted to the assessment of medical competence with defining competence as a combination of stable and generic traits. The most popular set of traits was a distinction in 'knowledge', 'skills', 'problem-solving ability' and 'attitudes'. The underlying assumptions for this model were the same as in test psychology, namely that each of these could be assessed or measured individually and generically, and that each of these would be relatively stable aspects. For example, it was believed that one could measure 'problem-solving ability' independently of 'knowledge' and that somebody who was a good problem solver would be able to apply that skill in any given situation. In other words, it was assumed 'problem solving ability' would be a stable and reproducible characteristic of a student.

2.2. Implications for assessment development

This way of thinking had some important implications for the way assessment development was approached.

The first and probably most important implication was the attempt to capture competence almost entirely by numerical outcomes; to view it as a psychometric measurement problem.[1] Logically then, the two main supporting factors for the quality of assessment were reliability and construct validity.[2] Typically, reliability was conceptually defined as test-retest reliability, i.e. would the outcomes on a given test be reproducible if the same group of students were given a similar but different – a so-called parallel - test? Construct validity related to the extent to which the test actually measured the psychological trait it was meant to measure,[2] after all personality traits cannot be observed directly and have to be inferred from the observations in assessment. These two approaches in themselves seem to be logical and coherent arguments; it is logical to assume that a test outcome should be reproducible and not depending

on the specific test occasion and plausible to require test developers to demonstrate that there test actually measures the trait.

However, even then there was some debate with respect to validity notably between Cronbach and Ebel. Ebel argued that tests in education differ substantially from psychological tests in that they have to consist of items or assignments which are intrinsically meaningful and therefore validity would have to be built into the test by careful blue printing, item construction and other quality control measures.[3] In other words, assessment measures more than invisible or latent traits, and observable behaviours could be valid in themselves. Cronbach on the other hand, argued that validity could only be established based on the way the test scores 'behaved', i.e. whether test scores would form patterns which could be expected based on the theoretical assumptions around the trait.[4] As a simple example, if a test is designed to measure clinical problem-solving expertise, one would expect the mean score of a group of experienced clinicians to be higher than the mean score of a group of final year medical students, regardless of the content of each specific item. But in its full extent construct validation meant that a series of theories about the nature of the trait had to be formulated and empirically tested (cf.[5]). These two adjacent views on validity have coexisted until the late 1990s and early 2000s when more integrative and nuanced validity theories were developed.[6,7]

As a flow-on effect of this thinking many of the developments in assessment focused on finding the single-best method for each trait, and it is fair to say that the literature of that period is filled with studies trying to demonstrate the superiority of one assessment method over another. Most well-known examples of such studies are the ones comparing open-ended with multiple-choice questions for the assessment of clinical reasoning and clinical decision-making[8,9] and of course the development and huge popularity of OSCE as *the* best instrument to measure skills.[10]

A further implication was that assessment had to be objective. As psychological personality inventories were paper and pencil, multiple-choice type tests that were seen to replace the unstructured and subjective human judgement of the psychologist, it was logical that assessment development would follow the same pathway. So, structuring and standardisation were seen as important ways to increase reliability of the assessment.

A final important assumption pertained to the purpose of assessment. As psychological tests were typically used to tell people apart, for example to distinguish between normal and abnormal personality traits, it was only logical that educational tests were designed to optimally tell competent students apart from incompetent students. Reflections of this thinking still exist in popular test psychometrics with parameters such as a Discrimination Index (DI), point biserial or item-total correlation (Rit). In an educational context, assessment developers, therefore, sought to develop assessment systems that would constantly select and allow the competent students to progress while withholding the incompetent students. These incompetent or not-yet-competent students would have to either resit the exam or retake part of the study. This was common practice under the assumption that this approach would automatically lead to graduating only highly competent students.

2.3. Limitations

Although this seemed a logical and coherent approach to assessment, dissatisfaction with this approach rose eventually. Probably the best way to explain this discontent is an analogy with healthcare; like most doctors are taught never to treat a patient on lab values only (so without history, physical examination, pathology reports, imaging, et cetera) educators increasingly felt that merely having numerical results from structured and standardised testing was not sufficient to determine a student's competence.

This was supported by what is perhaps the most important finding in the literature: traits could not be assessed or measured independently of each other. Especially the ability to solve problems was found to be highly dependent on having a well organised knowledge base.[11–13] Another important finding showed that performance did not generalise well across content and much better across assessment format.[9,14] For example, if similar content was asked it did not really matter whether open-ended questions on multiple-choice questions were used, as correlations between both test formats were extremely high (cf. for an overview[15]). This was even demonstrated for a written test on clinical skills and an actual OSCE.[16]

A final important realisation was that objectivity is not the same as reliability.[17] Reliability, or generalisability, is mainly a function of sampling and even the most subjective tests can become reliable when the sample is big enough. Actually, objectivity in assessment does not really seem to exist; assessment is always a combination of collecting information and judging that information, whether it be for the content of the test, the specific wordings of the items or the determination of the pass fail score, human judgement always plays a role.

3. The introduction of competencies

3.1. Conceptual framework

In the 1990s conceptual thinking about the nature of competence and the way it should be assessed changed dramatically. A good illustration of this change of thinking is illustrated by a position paper by Boud published in 1990 in which he argues that the academic values we, as educational institutes, consider essential, are actually not promoted by the way we assess competence,[18] or that they are even discouraged. He argued that values such as independence, thoughtfulness and critical thinking and the ways academics handle their own contributions to knowledge are at odds with the then current practice of testing/assessment. This is not to say that the notion that assessment drives learning was unknown until then; in his review Frederiksen already described the different ways of how testing can be used to drive student learning.[19] However, the influence of summative testing on student learning was strongly rooted in a behaviourist -stick and carrot - framework and had very little to do with the students' own meaning making – the constructivist drivers of assessment.[20,21]

In order to be able to use assessment more meaningfully, the definition of competence had to become more meaningful and more directly practice relevant. This led to movement away from defining competence as a set of independently measurable traits to a combination of competency domains or competencies. The most well-known definitions are the CanMeds[22] and the ACGME competency frameworks.[23] The precise definitions in the literature about what competency is, differ from source to source but they all tend to converge on the notion that a competency is the ability to successfully manage a professional problematic situation using just-in-time knowledge and its application, in conjunction with appropriate skills, attitudes and metacognitive abilities.[24,25]

3.2. Implications for assessment development

This change of thinking had a huge impact on the approaches to assessment design. Where formerly the focus had been on trying to find *the* optimal test for each trait, now the realisation grew that every assessment method has its unique strength and weaknesses, or as an analogy to healthcare its indications, side-effects and contraindications. Van der Vleuten published an opinion paper mid-1990s on this. He suggested evaluating the utility of each assessment method by judging the contributions of various aspects of the method, and proposed using 'reliability', 'validity', 'educational impact', 'cost effectiveness' and 'acceptability'.[25] He argued that no single assessment method could be perfect in all five aspects and therefore a trade-off would have to be made specific to the purpose of the assessment context. National licensing bodies with a remit of selection would therefore have to make different choices than medical schools with a remit of education. Especially in the latter context, the educational impact became a factor of more importance. So, understanding how assessment would drive learning in a much broader view then the mere behaviourist 'stick and carrot' became important.[20,21,26]

With a rapid expansion of the assessment 'toolkit' it became necessary to have a fresh look at how the credibility and trustworthiness of assessment results could be supported other than merely through reproducibility and construct validity. In the testing era, psychometric theories were sufficient to argue for the trustworthiness of assessment outcomes, but with the advent of more qualitative or narrative assessment methods and portfolios this was no longer the case. A good illustration for this change of thinking is a paper by Delandshere and Petrosky in which they make explicit and critically question all the assumptions that have to be met in order for an examiner to reduce the rich information from an examination to a score.[27] This is translation from what an examiner observes and their conclusion is extremely important. The reason for this is that in modern validity theory the inference from observation to 'score' is the starting point for any validity argumentation and without being extremely explicit and thoughtful about the assumptions that go into this process validity cannot exist.[7,28] Also, the assumptions underlying traditional reliability as the sole support for generalisability of assessment results were questioned in an opinion paper by Schuwirth and Van der Vleuten. More specifically they questioned the need to use the concept of a relatively stable trait, and the ensuing requirement of test retest reproducibility.[29] They argued that not all aspects of competence needed to be stable, or in fact that in a successful educational context change of competence is a much more likely phenomenon than stability. They suggested to the medical education community that alternative methods to establish generalisability of test results – actually the second inference from Kane's validity theory – should be developed, for example saturation of information or 'seeing the whole picture'. This thinking culminated further in the work on portfolios in which of course huge panels of judges per

portfolio were neither feasible nor desirable. Driessen at al. then suggested to incorporate lessons learned from qualitative research methodology in the models to support the trustworthiness of assessment results.[30]

3.3. Limitations

Despite these important steps there were still questions as to how to use a competency framework to design a comprehensive assessment programme. The notion of the stable and independent traits and the need for reductionist approaches still underpinned the thinking around competencies. Discussions for example, as to whether separate competencies should be measured separately and to what extent compensation between competencies would be allowable were and are still firmly rooted in the notion of separate traits. From a modern viewpoint it seems defensible to claim that a student can only graduate if they are able to demonstrate sufficient achievement in all necessary competency domains, but this 'conjunctive' perspective does not sit well with the psychometric notion of reliability which argues that compensatory models generally produce better reliabilities than conjunctive ones.

The second dilemma emerged between the holistic and integrative nature of competencies per se and the need to make them more tangible and therefore reductionist for assessment. It was clear from the beginning that assessing at the level of 6 or 8 large competency domains would not work but subdividing them into hundreds of little sub-sub-competencies would also defeat the purpose. This of course, led to the question asking how much reductionism is enough. A useful contribution to this whole debate was made by the introduction of entrustable professional activities.[31] This redefinition of aspects of competencies had the huge advantage of being in a jargon that the average medical teacher or clinician is more familiar with than the psychological/educational jargon of the previous years. For example, in workplace-based assessment a simple change of jargon into a more familiar one was shown to lead to better psychometric qualities.[32] Still however, even with entrustable professional activities the question remained as to how many is enough to capture competence as a whole.

4. Assessment as a programme

4.1. Conceptual framework

In the mid-2000s the realisation emerged that comprehensive and holistic assessment of a complex phenomenon such as competence can only be done at the level of a complete assessment programme.[33,34] This was a fundamental change. Up until now competence was deconstructed into separate assessable chunks and the million dollar question was how to reconstitute these chunks most meaningfully back into 'competence' again. Programmatic assessment does not seek to deconstruct but to keep competence integrated and using the competencies as lenses on competence rather than as separate traits. The underlying question in programmatic assessment was therefore how to build a meaningful holistic narrative or conclusion about students' competence rather than merely a set of individual chunks. As an illustration, formerly assessment focused on the pixels and then had to reconstruct the picture, programmatic assessment keeps the whole picture and homes in on those pixels that require further attention. So, where previously the focus had been on the development of individual assessment instruments, now the value of individual instruments is perceived as extent to which they contribute as data points to the whole programme.

This is a fundamentally different view on assessment as it allows, or better requires, assessment information to be combined *across* assessment methods[35] rather than solely within methods. Previously for example, student performance on an OSCE station on neurological examination would be combined – and compensated – with the other stations, e.g. a station on shoulder examination. The implicit thought behind this was still firmly rooted in the idea that a given method would measure a certain trait, and that the individual items in an assessment could be treated as intrinsically meaningless as long as they contributed to the validity of the total score.[2,4] As we explained above, this is a construct validity thinking stemming from the 1960s. In programmatic assessment information about a student's competence is combined, triangulated and evaluated across assessment instruments in an attempt to draw content-based meaningful conclusions.[36] In programmatic assessment, the contention would be that poor performance on a shoulder examination would have to be triangulated with the performance on, for example, the shoulder anatomy part of a larger multiple-choice test and perhaps feedback on shoulder examination in the context of a mini CEX.

This may seem counterintuitive at first, because assessment practice has been different for such a long time, but is actually well supported by the literature which repeatedly demonstrates that assessment results generalise much better across assessment formats with similar content than across content when the format is similar.[15,16] So, combining assessment results across

methods to achieve meaningful conclusions is actually more in line with the research outcomes. In addition, such a collation of information aligns more intuitively with the way clinicians collate information about their patients. For a clinician compensating the history finding 'no abdominal pain' with 'considerable neck ache' into 'average pain' simply would not make sense[37] but combining 'neck pain' with 'fever' and 'high White Blood Cell Count in the cerebrospinal fluid' into possible meningitis does make sense. Finally, this approach aligns better with modern views on validity as well in which validity is not merely numerically and experimentally derived but seen as a series of convincing narratives – which may contain numbers and experimental outcomes – building the argument for validity.[7,38]

Not only does programmatic assessment strive to allow the assessor to draw more meaningful conclusions, it also would enable the provision of more meaningful feedback to the learner. But, in order to be able to draw meaningful conclusions from various assessment instruments, and thus various qualities of information, *expert* judgement is needed.[39] The realisation that an assessment programme requires examiners with assessment literacy[40] has had considerable impact on our thinking about ensuring quality of these narrative conclusions. Still a need for standard psychometrics for the 'testy' part of the assessment programme remained but it had to be complemented with qualitative methodology approaches and organisational procedures and check and balances, such as clear procedures, transparency and audit trails. Programmatic assessment added the realisation of the need for expert judgement in the assessment process.[39,41]

4.2. Implications for assessment development

This conceptual change enabled a stronger support for the design of assessment programmes *for* learning.[42] 'Assessment for learning' is not to be confused with formative assessment as it is distinctly different. In formative assessment the focus is on feedback to the student without any stakes. So there is a reliance on the student being able and willing to translate the feedback into successful learning activities. If the student is not able or willing to use the feedback there are no consequences and it its effectiveness is therefore quite unpredictable.[43] In assessment for learning on the other hand, the onus is put on the student to collect, collate and analyse information about their own competence and translate this into doable and effective learning goals *and* make them happen. So, in programmatic assessment for learning all assessments, be it formal tests, workplace-based assessments or any informal feedback, are *in*formative and have to be meaningfully and convincingly analysed, triangulated and described by the student with help of a coach or mentor and translated into learning goals.[36] The quality of the learning goals, the plans to enact them and the demonstration of achieving them is the summative aspect of the assessment programme. For this, the student has to learn how to analyse the information and extract effective learning goals, as this does not come naturally to all students. Typically, assessment for learning programmes have coaches or mentors who guide students through this process using a dossier or portfolio in which to collect and collate all assessment information[44].

4.3. Where are we now?

While assessment in the early phases of testing was purely seen as a measurement problem, and even in the era of competencies the quality of assessment tools was seen as intrinsic to the tool, in programmatic assessment the value of the assessment instrument is viewed as the interaction between the instrument and the user. Perhaps a mundane example to illustrate this would be a normal toolkit. In the testing era, thinking about assessment was focused on demonstrating that a hammer is a better tool than a screwdriver. In the competencies phase, the purposes and downsides of hammers and screwdrivers were studied to illustrate their utility. In the programmatic assessment era, finally, the value of assessment is seen as the combination of the quality of the hammer (the affordance of the tool) and expertise of the carpenter (the effectivities of the user).[34] So, assessment has changed from being a purely measurement problem to becoming an educational design problem to becoming a staff expertise development problem.

It is logical that current research focuses on the 'validity' of the user and their way of interacting with the assessment instrument rather than purely the validity of the instrument. Current research focuses for example on the development of examiner expertise analogous to research into clinical diagnostic expertise.[45–47] Also, studies seek to gain a better understanding of the narratives that students and their examiners use in their judgements,[48–50] and the way the organisational culture plays a role in this.[51] Finally, research seeks to understand how different stakeholders add valuable perspectives[52] and how summative and formative functions of assessment can and should be combined in different organisational cultures.[53,54]

A final and more philosophical change in thinking that is currently happening is a move from a purely logical positivist epistemology – competence is an existing phenomenon which is clearly definable and measurable – to including a constructivist view as well – competence is a phenomenon that exists through observation in the here and now and which changes over time.[34] In the light of this change of views it is even more logical that the focus of research has moved away from a purely measurement perspective to a human judgement and narratives perspective.

5. Epilogue

Although there are some examples of existing programmatic assessment for learning curricula,[44,55] evidence as to whether it leads to the types of graduates still has to be collected. However, for us the development of programmatic assessment for learning and emerging research are logical developments firmly rooted in the changes of thinking and developments of the past decades.

Disclosures

The authors declare no conflict of interest. Ethical approval for this narrative review paper was not necessary.

References

1. Swanson DB. A measurement framework for performance-based tests. In: Hart I, Harden R, editors. *Further Developments in Assessing Clinical Competence*. Montreal: Can-Heal publications; 1987. p. 13–45.
2. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955;52(4):281–302.
3. Ebel RL. The practical validation of tests of ability. *Educ Meas: Issues Pract* 1983;2(2):7–10.
4. Cronbach LJ. What price simplicity?. *Educ Meas: Issues Pract* 1983;2(2):11–12.
5. Benson J. Developing a strong program of construct validation: a test anxiety sample. *Educ Meas: Issues Pract* 1998;17(2):10–17.
6. Messick S. The interplay of evidence and consequences in the validation of performance assessments. *Educ Res* 1994;23(2): 13–23.
7. Kane MT. Validation. In: Brennan RL, editor. *Educational Measurement*. Westport: ACE/Praeger; 2006. p. 17–64.
8. Newble DI, Baxter A, Elsmlie RG. A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Med Educ* 1979;13:263–268.
9. Norman G, Swanson D, Case S. Conceptual and methodology issues in studies comparing assessment formats, issues in comparing item formats. *Teach Learn Med* 1996;8(4):208–216.
10. Van der Vleuten CPM, Swanson D. Assessment of clinical skills with standardized patients: state of the art. *Teach Learn Med* 1990;2(2):58–76.
11. Chi MTH, Glaser R, Rees E. Expertise in problem solving. In: Sternberg RJ, editor. *Advances in the psychology of human intelligence*. Hillsdale NJ: Lawrence Erlbaum Associates; 1982. p. 7–76.
12. Norman G, Tugwell P, Feightner, J, et al. Knowledge and clinical problem-solving. *Med Educ* 1985;19:344–356.
13. Norman GR, Smith EKM, Powles, AC, et al. Factors underlying performance on written tests of knowledge. *Med Educ* 1987;21: 297–304.
14. Ward WC. A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Appl Psychol Meas* 1982;6(1): 1–11.
15. Schuwirth LWT, Van der Vleuten CPM, Donkers HHLM. A closer look at cueing effects in multiple-choice questions. *Med Educ* 1996;30:44–49.
16. Van der Vleuten CPM, Van Luyk SJ, Beckers HJM. A written test as an alternative to performance testing. *Med Educ* 1988;22: 97–107.
17. Norman GR, Van der Vleuten CP, De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Med Educ* 1991;25(2):119–126.
18. Boud D. Assessment and the promotion of academic values. *Stud High Educ* 1990;15(1):101–111.
19. Frederiksen N. The real test bias: influences of testing on teaching and learning. *Am Psychol* 1984;39(3):193–202.
20. Cilliers FJ, Schuwirth LWT, Adendorff, HJ, et al. The mechanisms of impact of summative assessment on medical students' learning. *Adv health Sci Educ* 2010;15:695–715 http://dx.doi.org/10.1007/s10459-010-9232-9.
21. Cilliers FJ, Schuwirth LWT, Herman, N, et al. A model of the pre-assessment learning effects of summative assessment in medical education. *Adv health Sci Educ* 2012;17:39–53 http://dx.doi.org/10.1007/s10459-011-9292-5.
22. Canmeds. ⟨http://www.royalcollege.ca/portal/page/portal/rc/canmeds⟩ Ottawa2005. Accessed 26 July April 2017 .
23. ACGME. ⟨http://www.acgme.org/What-We-Do/Accreditation/Milestones/Overview⟩ Chicago2007; April 2017.
24. Albanese MA, Mejicano G, Mullan, P, et al. Defining characteristics of educational competencies. *Med Educ* 2008;42 (3):248–255.
25. Govaerts M. Educational competencies or education for professional competence?. *Medical Education* 2008;42(3):234–236.
26. Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ* 1996;1(1):41–67.
27. Gielen S, Dochy F, Dierick S. Evaluating the consequential validity of new modes of assessment: the influences of assessment on learning, inlcuding pre-, post- and true assessment effects. In: Segers M, Dochy F, Cascallar E, editors. *Optimising New Modes of Assessment: in Search of Qualities and Standards*. Dordrecht: Kluwer Academic Publishers; 2003. p. 37–54.
28. Delandshere G, Petrosky AR. Assessment of complex performances: limitations of key measurement assumptions. *Educ Res* 1998;27(2):14–24.
29. Kane M. Current concerns in validity theory. *J Educ Meas* 2001;38(4):319–342.
30. Schuwirth LWT, Van der Vleuten CPM. A plea for new psychometrical models in educational assessment. *Med Educ* 2006;40(4):296–300.

31. Driessen E, 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.
32. Ten Cate O. Entrustability of professional activities and competency-based training. *Med Educ* 2005;39:1176–1177 http://dx.doi.org/10.1111/j.1365-2929.2005.02341.x.
33. Weller JM, Misur M, Nicolson, S, et al. Can I leave the theatre? A key to more reliable workplace-based assessment. *Br J Anaesth* 2014;112(6):1083–1091.
34. Van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ* 2005;39 (3):309–317.
35. Durning SJ, Artino A, Pangaro, L, et al. Redefining context in the clinical encounter: implications for research and training in medical education. *Acad Med* 2010;85(5):894–901.
36. Van der Vleuten CPM, Schuwirth LWT, Driessen, EW, et al. 12 Tips for programmatic assessment. *Med Teach* 2015;37(7):641–646.
37. Van der Vleuten CPM, Schuwirth LWT, Driessen, EW, et al. A model for programmatic assessment fit for purpose. *Med Teach* 2012;34:205–214 http://dx.doi.org/10.3109/0142159X.2012.652239.
38. Schuwirth LWT, Van der Vleuten CPM, Durning SJ. What programmatic assessment for learning in medical education can learn from healthcare. *Perspect Med Educ* 2017: 1–5 http://dx.doi.org/10.1007/s40037-017-0345-1.
39. Schuwirth LWT, Van, der Vleuten CPM. Programmatic assessment and Kane's validity perspective. *Med Educ* 2012;46(1): 38–48 http://dx.doi.org/10.1111/j.1365-2923.2011.04098.x.
40. Govaerts MJB, Van der Vleuten CPM, Schuwirth, LWT, et al. Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Adv Health Sci Educ* 2007;12(2):239–260.
41. Popham WJ. Assessment literacy for teachers: faddish or fundamental?. *Theory Pract* 2009;48:4–11.
42. Schuwirth LWT, Van, der Vleuten CPM. Assessing competence: extending the approaches to reliability. In: Hodges BD, Lingard L, editors. *The Question of Competence*. Ithaka New York US: Cornell University Press; 2012.
43. Schuwirth LWT, Van, der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach* 2011;33(6):478–485.
44. Shute V. Focus on formative feedback. *Rev Educ Res* 2008;78(n): 153–189.
45. Schuwirth LWT, Ward H, Heeneman S. Assessment for Learning. In: Joy Higgs, Julie Baldry-Currens, Gail, Jensen, editors. *Realising ExemplaryPractice-based Education*. Rotterdam: Sense Publishers; 2013. p. 143–150.
46. Govaerts MJB, Schuwirth LWT, Van der Vleuten, CPM, et al. Workplace-based assessment: effects of rater expertise. *Adv Health Sci Educ* 2011;16(2):151–165.
47. Govaerts MJB, Van de Wiel MWJ, Schuwirth, LWT, et al. Workplace-based assessment: raters' performance theories and constructs. *Adv Health Sci Educ* 2012:1–22 http://dx.doi.org/10.1007/s10459-012-9376-x.
48. Berendonk C, Stalmeijer RE, Schuwirth LWT. Assessors' perspectives on assessment: 'i think you call it expertise'. *Adv Health Sci Educ* 2012 http://dx.doi.org/10.1007/s10459-012-9392-x. [published Online First: 31 July 2012].
49. Ginsburg S, Regehr G, Lingard, L, et al. Reading between the lines: faculty interpretations narrative evaluation comments. *Med Educ* 2015;49:296–306.
50. Cook DA, Kuper A, Hatala, R, et al. When assessment data are words: validity evidence for qualitative educational assessments. *Acad Med* 2016;91(10):1359–1369.
51. Ginsburg S, Vleuten CP, Eva, KW, et al. Cracking the code: residents' interpretations of written assessment comment. *Med Educ* 2017;51:401–410 http://dx.doi.org/10.1111/medu.13158.
52. Watling C, Driessen E, Van der Vleuten, CPM, et al. Beyond individualism: professional culture and its influence on feedback. *Med Educ* 2013;47(6):585–594.
53. Gingerich A. *Questioning the Rater Idiosyncrasy Explanation for Error Variance by Searching for Mulitple Signals within the Noise*. Maastricht University; 2015.
54. Harrison CJ, Könings KD, Dannefer, EF, et al. Factors influencing students' receptivity to formative feedback emerging from different assessment cultures. *Perspect Med Educ* 2016;5:276–284.
55. Harrison CJ, Könings KD, Schuwirth, L, et al. Barriers to the uptake and use of feedback in the context of summative assessment. *Adv Health Sci Educ* 2015;20(1):229–245 http://dx.doi.org/10.1007/s10459-014-9524-6.
56. Dannefer E, Henson L. The portfolio approach to competency-based assessment at the Cleveland Clinic Lerner College of Medicine. *Acad Med* 2007;82(5):493–502.