Revisiting 'Assessing professional competence: from methods to programmes'

Cees P M van der Vleuten

Editor's note: As part of our 50th volume celebrations, Medical Education is looking back at its most impactful articles, as defined by citation count. The most cited articles from each 5-year interval were identified and the original authors of one of them (or other knowledgeable scholars if the original authors could not be found) were asked to comment on the state of the field at the time of the publication, the impact of the article, and what we have learned since then. The article illustrated in Figure 1 was one of the most cited articles in our journal in the 2002-2006 period. To see the other top-cited articles from Volumes 1–50 please view the interactive PDF by visiting www.mededuc.com.

The 2005 paper to which the title of this piece refers represented a landmark in my thinking around assessment.[1] It started, however, with an earlier paper published in 1996, which has also been cited frequently.[2] The earlier paper identified five distinct quality characteristics of any assessment method: reliability; validity; educational impact; acceptability, and costs.[2] Although these five criteria have been used frequently in research on assessment, this was not the main message of the paper.[2] Many other quality characteristics are possible and are also mentioned in the literature.[3,4] The central tenet of the paper[2] was that any single assessment method can never be perfect on all criteria and in reality assessment always involves a compromise. Good assessment involves a mindful choice about where and how to compromise. The nature of the compromise will depend on the purpose of the assessment and the assessment context. For example, the compromises made in relation to a certification examination will differ from those required in an in-training assessment.

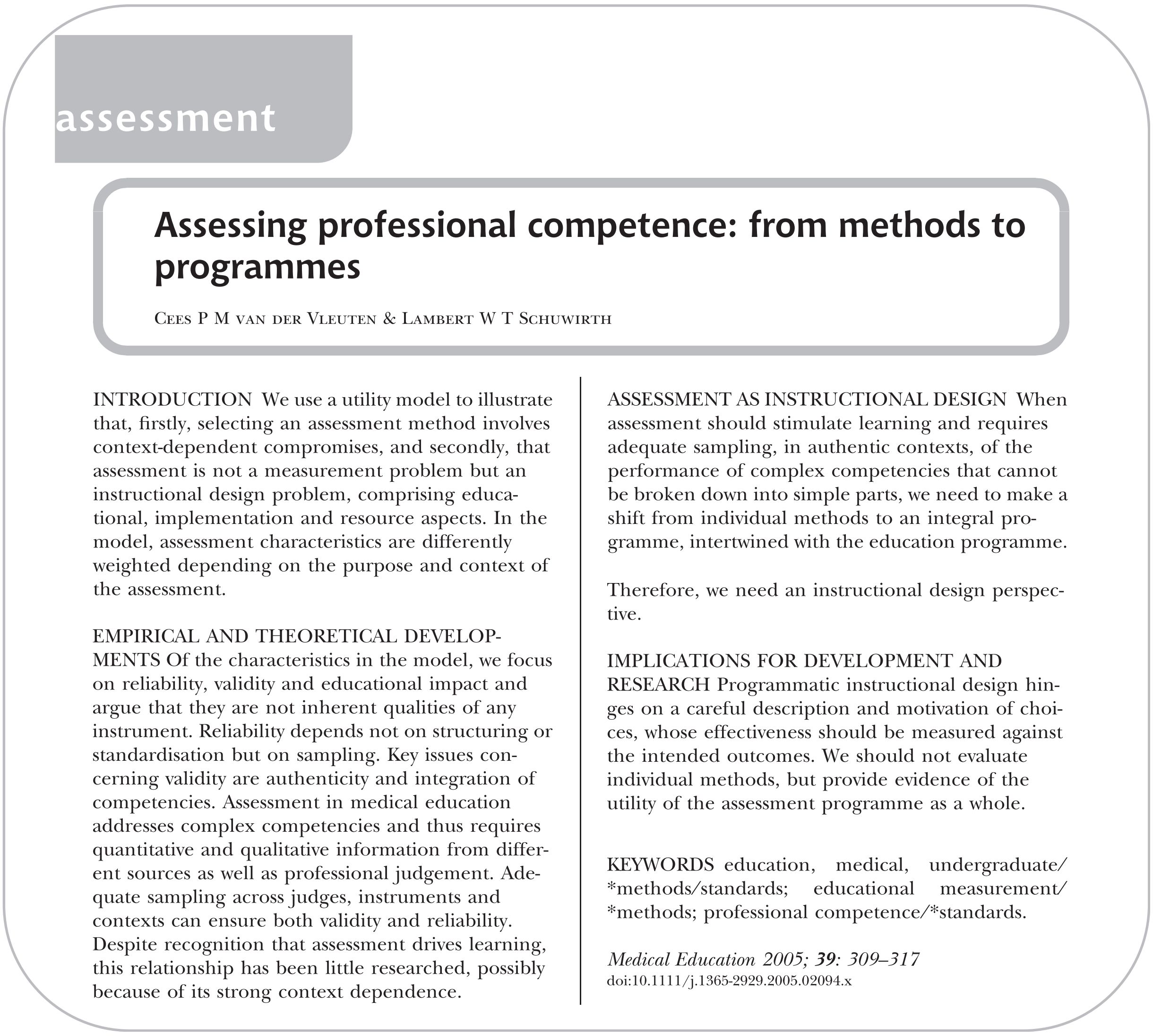
> *Any single assessment method can never be perfect on all criteria and in reality assessment always involves a compromise.*

In the 2005 paper,[1] we argued that any single assessment has such severe limitations that any single measurement is really no measurement. For example, the paper contained a table showing reliability data for classic and modern assessment methods across all layers of Miller's pyramid as a function of testing time. Any assessment, old or new, objective or subjective, standardised or unstandardised, requires at least 3–4 hours of testing time to achieve minimal reliability. Even with a reliability criterion of 0.80, we should realise that 20% of the pass/fail decisions we make may be false positives and negatives (depending on the distribution of scores in relation to the pass/fail cut-off score). This naturally leads to the question of when to optimise what. If we can't have it all in a single measure, can we then have it all across measures? Shouldn't we stop trying

Medical Education 2016: 50: 885–888, doi: 10.1111/medu.12632

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

Correspondence: Cees P M van der Vleuten, Department of Educational Development and Research, Maastricht University,
PO Box 616, Maastricht 6200 MD, the Netherlands.
Tel: 00 31 43 388 6725;
E-mail: c.vandervleuten@maastrichtuniversity.nl



assessment

Assessing professional competence: from methods to programmes

Cees P M van der Vleuten & Lambert W T Schuwirth

INTRODUCTION We use a utility model to illustrate that, firstly, selecting an assessment method involves context-dependent compromises, and secondly, that assessment is not a measurement problem but an instructional design problem, comprising educational, implementation and resource aspects. In the model, assessment characteristics are differently weighted depending on the purpose and context of the assessment.

EMPIRICAL AND THEORETICAL DEVELOPMENTS Of the characteristics in the model, we focus on reliability, validity and educational impact and argue that they are not inherent qualities of any instrument. Reliability depends not on structuring or standardisation but on sampling. Key issues concerning validity are authenticity and integration of competencies. Assessment in medical education addresses complex competencies and thus requires quantitative and qualitative information from different sources as well as professional judgement. Adequate sampling across judges, instruments and contexts can ensure both validity and reliability. Despite recognition that assessment drives learning, this relationship has been little researched, possibly because of its strong context dependence.

ASSESSMENT AS INSTRUCTIONAL DESIGN When assessment should stimulate learning and requires adequate sampling, in authentic contexts, of the performance of complex competencies that cannot be broken down into simple parts, we need to make a shift from individual methods to an integral programme, intertwined with the education programme.

Therefore, we need an instructional design perspective.

IMPLICATIONS FOR DEVELOPMENT AND RESEARCH Programmatic instructional design hinges on a careful description and motivation of choices, whose effectiveness should be measured against the intended outcomes. We should not evaluate individual methods, but provide evidence of the utility of the assessment programme as a whole.

KEYWORDS education, medical, undergraduate/*methods/standards; educational measurement/*methods; professional competence/*standards.

*Medical Education 2005; **39**: 309–317*
doi:10.1111/j.1365-2929.2005.02094.x

Figure 1 Title page from ‘Assessing professional competence: from methods to programmes’[1]

to optimise everything in a single measure and instead optimise the collection of methods? This line of thinking explains the second part of the title: ‘from methods to programmes’. If we think about optimising programmes and each method is embedded in such a programme, we may actually arrive at different compromises for each individual method. For example, if we combine assessment information across methods in an in-training assessment programme, we might compromise more on the reliability of individual methods and less on educational impact. I am aware of accreditation practices which involve inspection of the reliabilities of individual measures: if these are not high enough, the schools in question are in trouble. I shiver when I hear about such an absolute use of a single psychometric measure. Here, one measure (the reliability coefficient) is no measure. Rather, we should make a mindful choice of a combination of methods in which compromises are justified in light of the educational context and the purpose of the whole programme.

If we can’t have it all in one single measure, can we then have it all across measures?

The notion of optimising assessment programmes has resonated in the assessment field, as evidenced by the many citations of this paper.[1] However, the paper[1] also denoted an agenda for further research and development. All assessment research during the period prior to its publication focused on individual methods, and publications typically addressed reliability and validity issues. The 2005 paper[1] not only represented a call for a shift in thinking, but also a call for a shift in our research and development agenda.

Taking up this agenda, I worked with Liesbeth Baartman on a set of criteria for assessment in competency-based education programmes.[4] This led to a self-assessment instrument for evaluating the quality of assessment programmes.[5] Later this work was complemented by that of Joost Dijkstra when we developed a set of education-neutral design guidelines for assessment programmes.[6] These guidelines are appropriate for any assessment context that has at least two or more assessment elements. They are also appropriate for a certification context. They are truly guidelines, not prescriptions. As an example, a very basic guideline is: 'Decisions (and the consequences) should be proportional to the quality of the information on which they are based.'[6] A more specific guideline is: 'A rationale should be provided for the standard-setting procedures.'[6] Such guidelines serve as prompts for thinking, designing and evaluating an assessment programme. In the future, we hope to condense these guidelines further so that they can be used as a framework for more formal evaluations or accreditations of assessment programmes.

In the meantime, assessment research in medical education has continued to be very productive. In an analysis of papers published in six high-impact journals in the field over a 22-year period, assessment represented the topic most frequently addressed.[7] I noticed consistencies in that research, such as when a qualitative researcher encounters triangulation and even some saturation. Together with a group of colleagues, I published these consistencies in 2010, designating them as 'principles of assessment' and perhaps as 'building blocks' for the further development of theory in the assessment of professional competence.[8] The principles were divided into two classes for, respectively, standardised (the first three layers of Miller's pyramid) and unstandardised (the top of the pyramid) assessment. An example of the first is 'Validity can be built in', which points to the need for quality assurance around item and test development. An example of the latter is 'Validity resides more in the users of the instruments than in the instruments that are used', which points to the need to carefully prepare users of the instruments (e.g. assessors and learners) for their roles in the assessment. The principles should more or less universally apply to all assessment situations.

These principles led me to think towards a more theoretical framework around programmes of assessment, as well as through my experiences in actual education practice. I was directly involved in setting up new assessment programmes in my own school and I also served as a consultant to many institutions. Some of these practices were very inspiring.[9] My experiences in educational practice shaped my view on assessment and learning. 'Educational consequences' were deliberately included in the 2005 paper,[1] and the principle of 'assessment drives learning' was very prominent in the 2010 paper.[8] In the assessment literature, the notion of assessment *for* learning emerged.[10] Assessment of learning often leads to negative effects on learning and the educational system: learners strive for grades; learners ignore feedback; learners engage in tick-box exercises, and learners beat the system by playing the game. Many of our assessment practices are rather reductionist and trivialise learning as a result. The most conventional approach to assessment is to have a course and an end-of course summative assessment. In the event of a fail, we take a mindless decision: repeat the test. We don't look at what the problem is, we simply say: show us again whether you can surpass a minimum (!) standard. If another failure ensues, we again take a mindless decision: repeat the course. If a learner passes, the pass is eternal and thus to that course and the topics covered in it the learner is considered to be 'immune for life'. There is very little information in such an assessment system about the learner. This may be appropriate for a mastery learning conception of learning, which basically reflects a behaviourist view of learning, but it does not accord with modern views on learning. Modern education is more constructivist in nature or is based on socio-cultural learning theories. Learners construct knowledge, and apply, experience and practise knowledge in action. Feedback, metacognition, reflection, self-monitoring and self-directing are important concepts for lifelong learning. Modern curricula address complex behavioural outcomes that are learned in longitudinal and vertically integrated curricular lines. The development of learners in time is essential. However, given the driving effect of assessment, these modern programmes will fail if they continue to adhere to an outdated assessment model.

Given the driving effect of assessment, these modern programmes will fail if they keep adhering to an outdated assessment model

In my thinking, learning started to drive assessment. Again with very influential colleagues, I published a model or theoretical framework for assessment programmes in 2012.[11] Based on the assessment principles and on a wish to bring into and use more learner-centred information in the assessment system, we presented a model which optimises both the learning and the decision-making function of the system as a whole. Any assessment is seen as but one data point. Pass/fail decisions are decoupled from individual data points. Each data point is maximally informative to the learner and is information-rich. Decisions are taken on the basis of many data points. Learners are required to self-analyse and are mentored as they do so. The number of data points required is proportional to the importance of the decision. High-stake decisions are taken on many data points and a lot of rich information. This information should tell a story about the learner. For most learners this story will make progression decisions easy, but for some the story is less clear and a decision requires either much deliberation or the collection of further data points. The model may serve as a basis for development and research, such as in a design-based research approach. The model may shape practices and the practices may shape the model further.

> *Assessment information should tell a story about the learner*

Although this model for programmatic assessment is well received in educational practice, it is difficult to implement. It requires a cultural change in our thinking around assessment. Teachers or supervisors are given different roles in assessment that may not lie within their existing repertoire. Such a change requires a shift from a positivist view of assessment to a more constructivist–interpretivist approach to assessment.[12] Cultural changes in education are not made overnight. Problem-based learning, which required a similar cultural change, has taken many years to develop. Innovations move slowly, and so will programmatic assessment.

Overall, the 2005 paper 'Assessing professional competence: from methods to programmes'[1] represented an important step in a chain of thinking around assessment. It is my firm belief that medical education is quite unique in the area of assessment. We have covered substantial ground. It is exciting to be part of this community and to watch as the field evolves. I wonder what the next element in the chain will be.

REFERENCES

1 van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ* 2005;**39** (3):309–17.

2 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.

3 Messick S. The interplay of evidence and consequences in the validation of performance assessments. *Educ Res* 1995;**23**:13–23.

4 Baartman LKJ, Bastiaens TJ, Kirschner PA, van der Vleuten CPM. The wheel of competency assessment. Presenting quality criteria for competency assessment programmes. *Stud Educ Eval* 2006;**32** (2):153–70.

5 Baartman LKJ, Prins FJ, Kirschner PA, van der Vleuten CPM. Determining the quality of assessment programmes: a self-evaluation procedure. *Stud Educ Eval* 2007;**33** (3):258–81.

6 Dijkstra J, Galbraith R, Hodges BD, McAvoy PA, McCrorie P, Southgate LJ, van der Vleuten CP, Wass V, Schuwirth LW. Expert validation of fit-for-purpose guidelines for designing programmes of assessment. *BMC Med Educ* 2012;**12**:20.

7 Rotgans JI. The themes, institutions, and people of medical education research 1988–2010: content analysis of abstracts from six journals. *Adv Health Sci Educ Theory Pract* 2012;**17** (4):515–27.

8 van der Vleuten CP, Schuwirth LW, Scheele F, Driessen EW, Hodges B. The assessment of professional competence: building blocks for theory development. *Best Pract Res Clin Obstet Gynaecol* 2010;**24** (6):703–19.

9 Dannefer EF, Henson LC. The portfolio approach to competency-based assessment at the Cleveland Clinic Lerner College of Medicine. *Acad Med* 2007;**82** (5):493–502.

10 Black P, McCormick R, James M, Pedder D. Learning how to learn and assessment for learning: a theoretical inquiry. *Res Papers Educ* 2006;**21** (2):119–32.

11 van der Vleuten CP, Schuwirth LW, Driessen EW, Dijkstra J, Tigelaar D, Baartman LK, van Tartwijk K. A model for programmatic assessment fit for purpose. *Med Teach* 2012;**34** (3):205–14.

12 Govaerts M, van der Vleuten CPM. Validity in work-based assessment: expanding our horizons. *Med Educ* 2013;**47** (12):1164–74.