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

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

Correspondence: Cees P M van der Vleuten PhD, Department of Educational Development and Research, PO Box 616, 6200 MD Maastricht, The Netherlands. Tel: 00 31 43 388 1121; Fax: 00 31 43 388 4140; E-mail: C.vanderVleuten@educ.unimaas.nl

INTRODUCTION

Some years ago we proposed a conceptual model for defining the utility of an assessment method. The model derived utility by multiplying a number of criteria on which assessment instruments can be judged.[1] Besides such classical criteria as reliability and validity, the model included educational impact, the acceptability of the method to the stakeholders and the investment required in terms of resources. In the model the criteria were weighted according to the importance attached to each of them by a specific user in a specific situation and this defined the utility of the method. This means that the weights of the criteria depended on how the importance of each of the different criteria was perceived by those responsible for assessment in a certain assessment situation or assessment context.

Of course, this utility equation was merely intended as a conceptual model and by no means as an algorithm or new psychometric index. Neither were

Overview

What is already known on this subject

The utility of an assessment method depends on making a compromise between various quality parameters.

What this study adds

Any method – even those that are less structured and standardised – may have utility, depending on its use.

We need more methods that rely on qualitative information and thus require professional judgement.

Assessment is an educational design problem that needs a programmatic approach.

Suggestions for further research

Future research could report on evidence for the utility of integral assessment programmes.

all possible criteria included in the model, such as transparency, meaningfulness, cognitive complexity, directness and fairness.[2–4] Regardless of which criteria were included in the equation, the overriding message the model was intended to convey was that choosing an assessment method inevitably entails compromises and that the type of compromise varies for each specific assessment context. As an illustration, the weights attached to the criteria in a very high stakes assessment, for instance a certifying examination, will be very different from the distribution of weights among the criteria when the primary purpose of the assessment is to provide feedback to students in an in-training context. A second corollary of the 'formula' is that assessment is not merely a measurement problem, as the vast literature on reliability and validity seems to suggest, but that it is also very much an instructional design problem and includes educational, implementation and resources aspects. From this perspective, the utility model is useful, because it helps educators make considered choices in selecting, constructing and applying an assessment instrument.

In addition to its usefulness in deliberating on individual assessment methods, the model can also serve as an aid in the process of devising an overall assessment programme for a whole course. In this article, we will use the model for 2 purposes. Firstly, it will help us to summarise some developments in assessment that we regard as highly significant. Secondly, building on those views, we will argue that the model can serve as a guide to the design of integral assessment programmes. With respect to the first purpose, we will limit ourselves to the assessment characteristics of reliability, validity and educational impact. In discussing a more integral programmatic approach to assessment, we will attempt to achieve a conceptual shift so that instead of thinking about individual assessment methods, we think about assessment programmes.

EMPIRICAL AND THEORETICAL DEVELOPMENTS

For each of the first 3 criteria in the equation, we will describe some developments that we think are meaningful in the light of the future of assessment. We will not highlight, advocate or propose any individual (new) instrument, because we strongly believe that assessment instruments are not goals in themselves.[5] The degree to which the various quality criteria are attained is not an inherent, immutable characteristic of a particular instrument.[6,7] For example, a short, multiple-choice question examination will be unreliable for the assessment of a broad domain and an objective structured clinical examination (OSCE) will not be valid when it assesses trivial clinical activities in a postgraduate context. There is no such thing as *the* reliability, *the* validity, or any other absolute, immanent characteristic of any assessment instrument. We will try to shed more light on this issue in our deliberations below. The discussion will focus on some empirical outcomes and theoretical developments that we consider relevant for further progress in assessment.

Reliability

Reliability refers to the reproducibility of the scores obtained from an assessment. It is generally expressed as a coefficient ranging from 0 (no reliability) to 1 (perfect reliability). Often 0.80 is regarded as the minimal acceptable value, although it may be lower or higher depending on the examination's purpose (for instance, it will have to be higher for a licensing examination). Reliability can be

negatively affected by many sources of error or bias, and research has provided conclusive evidence that, if we want to increase reliability, we will have to ensure that our sampling takes account of all these unwanted sources of variance. A good understanding of the issues involved in sampling may offer us many more degrees of freedom in test development.

The predominant condition affecting the reliability of assessment is domain- or content-specificity, because competence is highly dependent on context or content. This means that we will only be able to achieve reliable scores if we use a large sample across the content of the subject to be tested.[8] If the assessment involves other conditions with a potential effect on reliability – such as examiners and patients – careful sampling across those conditions is equally essential. With intelligent test designs, which sample efficiently across conditions (such as using different examiners for each station in an OSCE), reliable scores will generally be obtained within a reasonable testing time.

So far, this is nothing new. What is new, however, is the recent insight that reliability is not conditional on objectivity and standardisation. The fact that objectivity and reliability are often confused was addressed theoretically some time ago,[9] but the empirical evidence is becoming convincingly clear now and may point towards new directions in assessment. To illustrate our point, let us look at the OSCE. The OSCE was developed as an alternative to the then prevailing subjective and unreliable clinical assessment methods, such as vivas and clinical ratings. The main perceived advantage of the OSCE was objectivity and standardisation, which were regarded as the main underpinnings of its reliability. However, an abundance of study evidence has since shown that the reliability of an OSCE is contingent on careful sampling, particularly across clinical content, and an appropriate number of stations, which generally means that several hours of testing time are needed.[10] What actually occurred was that the brevity of clinical samples (leading to a larger sample overall than in previous methods) and the fact that students rotated through the stations (optimal sampling across patients and examiners) led to more adequate sampling, which in turn had a far greater impact on reliability than any amount of standardisation could have had. This finding is not unique to the OSCE. In recent years many studies have demonstrated that reliability can also be achieved with less standardised assessment situations and more subjective evaluations, provided the sampling is appropriate. Table 1 illustrates this by presenting reliability estimates for several instruments with differing degrees of

Table 1 Reliability estimates of different assessment instruments as a function of testing time

Instrument	Description	Reliability for different testing times			
		1 hour	2 hours	4 hours	8 hours
Multiple choice*[42]	Short stem and short menu of options	0.62	0.76	0.93	0.93
Patient management problem*[42]	Simulation of patient, full scenarios	0.36	0.53	0.69	0.82
Key feature case (write-in)*[43]	Short patient case vignette followed by write-in answer	0.32	0.49	0.66	0.79
Oral examination†[44]	Oral examination based on patient cases	0.50	0.69	0.82	0.90
Long case examination†[45]	Oral examination based on previously unobserved real patient	0.60	0.75	0.86	0.90
OSCE*[46]	Simulated realistic encounters in round robin format	0.54	0.69	0.82	0.90
Mini-clinical exercise (mini-CEX)‡[47]	Short follow-up oral examination based on previously observed real patient	0.73	0.84	0.92	0.96
Practice video assessment†[16]	Selected patient–doctor encounters from video recordings in actual practice	0.62	0.76	0.93	0.93
Incognito standardised patients‡[48]	Real consultations scored by undetected simulated patients 0.86	0.61	0.76	0.82	0.86

* One-facet all random design with items crossed with persons (pxi).
† Two-facet all random design with judges (examiners) nested within items within persons (j:i:p).
‡ One-facet all random design with items nested within persons (i:p).

standardisation. For comparative purposes, the reliability estimates are expressed as a function of the testing time needed.

The comparative data should not be interpreted too strictly as only a single study was included for each type of method and reliability estimations were based on different designs across studies. For our discussion it is irrelevant to know the exact magnitude of the reliability or which method can be hailed as the 'winner'. The important point is to illustrate that all methods require substantial sampling and that methods which are less structured or standardised, such as the oral examination, the long case examination, the mini-clinical evaluation exercise (mini-CEX) and the incognito standardised patient method, *can* be entirely or almost as reliable as other more structured and objective measures. In a recent review, a similar conclusion was presented for global clinical performance assessments.[11] They are not included in Table 1 as the unit of testing time is unavailable, but a sufficiently reliable global estimate of competence requires somewhere between 7 and 11 ratings, probably not requiring more than a few hours of testing time. All these reliability studies show that sampling remains the pivotal factor in achieving reliable scores with any instrument and that there is no direct connection between reliability and the level of structuring or standardisation.

This insight has far-reaching consequences for the practice of assessment. Basically, the message is that no method is inherently unreliable and any method can be sufficiently reliable, provided sampling is appropriate across conditions of measurement. An important consequence of this shift in the perspective on reliability is that there is no need for us to banish from our assessment toolbox instruments that are rather more subjective or not perfectly standardised, provided that we use those instruments sensibly and expertly. Conversely, we should not be deluded into thinking that as long as we see to it that our assessment toolbox exclusively contains structured and standardised instruments, the reliability of our measurements will automatically be guaranteed.

Validity

Validity refers to whether an instrument actually does measure what it is purported to. Newer developments concerning assessment methods in relation to validity have typically been associated with the desire to attain a more direct assessment of clinical competence by increasing the authenticity of the measurement. This started in the 1960s with the assessment of 'clinical reasoning' by patient management problems and continued with the introduction of the OSCE in the 1970s. Authenticity was achieved by offering candidates simulated real world challenges, either on paper, in computerised forms or in a laboratory setting. Such assessment methods have passed through major developments and refinements of technique.[12] The assessment of higher cognitive abilities has progressed from the use of realistic simulations to short and focused vignettes which tap into key decisions and the application of knowledge, in which the response format (e.g. menu, write-in, open, matching) is of minor importance. The OSCE has similarly led to a wealth of research, from which an extensive assessment technology has emerged.[10] However, on top of the rapid progress in those areas, we see a number of interrelated developments, which may have a marked impact on the validity of our measurements in the future.

Firstly, we are likely to witness the continued progress of the authenticity movement towards assessment in the setting of day-to-day practice.[13] Whereas the success of the OSCE was basically predicated on moving assessment away from the workplace to a laboratory-controlled environment by providing authentic tasks in a standardised and objectified way, today, insights into the relationship between sampling and reliability appear to have put us in a position where we can move assessment back to the real world of the workplace as a result of the development of less standardised, but nevertheless reliable, methods of practice-based assessment. Methods are presently emerging that allow assessment of performance in practice by enabling adequate sampling across different contexts and assessors. Methods of performance assessment include the mini-CEX,[14] clinical work sampling,[15] video assessment[16] and the use of incognito simulated patients.[17] Such methods are also helpful in the final step of Miller's competency pyramid.[18] In this pyramid, assessment moves from the 'knows' stage via 'knows how' (paper and computer simulations) and 'shows how' (performance simulations such as the OSCE) to the final 'does' level of habitual performance in day-to-day practice.

A second development concerns the movement towards the integration of competencies.[19–21] Essentially, this movement follows insights from modern educational theory, which postulates that learning is facilitated when tasks are integrated.[22] Instructional programmes that are restricted to the 'stacking' of components or subskills of competencies are less effective in delivering competent professionals than methods in which different task components are

presented and practised in an integrated fashion, which creates conditions that are conducive to transfer. This 'whole-task' approach is reflected in the current competency movement. A competency is the ability to handle a complex professional task by integrating the relevant cognitive, psychomotor and affective skills. In educational practice we now see curricula being built around such competencies or outcomes.

However, in assessment we tend to persist in our inclination to break down the competency that we wish to assess into smaller units, which we then assess separately in the conviction that mastery of the parts will automatically lead to competent performance of the integrated whole. Reductionism in assessment has also emerged from oversimplified skills-by-method thinking,[1] in which the fundamental idea was that for each skill a single (and only a single) instrument could be developed and used. We continue to think in this way despite the fact that experience has taught us the errors of our simplistic thinking. For example, in the original OSCE, short, isolated skills were assessed within a short time span. Previous validity research has sounded clear warnings of the drawbacks of such an approach. For example, the classic patient management problem, which consisted of breaking down the problem-solving process into isolated steps, has been found to be a not very sensitive method for detecting differences in expertise.[23] Another example can be derived from OSCE research that has shown that more global ratings provide a more faithful reflection of expertise than detailed checklists.[24] Atomisation may lead to trivialisation and may threaten validity and, therefore, should be avoided. Recent research that shows the validity of global and holistic judgement thus helps us to avoid trivialisation. The competency movement is a plea for an integrated approach to competence, which respects the (holistic or tacit) nature of expertise. Coles argues that the learning and assessing of professional judgement is the essence of what medical competence is about.[25] This means that, rather than being a quality that augments with each rising level of Miller's pyramid, authenticity is present at all levels of the pyramid and in all good assessment methods. A good illustration of this is the way test items of certifying examinations in the USA are currently being written (http://www.nbme.org). Compared with a few decades ago, today's items are contextual, vignette-based or problem-oriented and require reasoning skills rather than straightforward recall of facts. This contextualisation is considered an important quality or validity indicator.[26] The validity of any method of assessment could be improved substantially if assessment designers would respect the characteristic of authenticity. We can also reverse the authenticity argument: when authenticity is not a matter of simply climbing the pyramid but something that should be realised at all levels of the pyramid, we can also say that similar authentic information may come from various sources within the pyramid. It is, therefore, wise to use these multiple sources of information from various methods to construct an overall judgement by triangulating information across these sources, a fact that supports the argument that we need multiple methods in order to make a good job of assessment.

A final trend is also related to the competency movement. The importance of general professional competencies – which are not unique to the medical profession – is acknowledged. These competencies include the ability to work in a team, metacognitive skills, professional behaviour, the ability to reflect and to carry out self-appraisal, etc. Although neither the concepts themselves nor the search for ways to assess them are new, there is currently a marked tendency to place more and more emphasis on such general competencies in education and, therefore, in assessment. New methods are gaining popularity, such as self-assessment,[27] peer assessment,[28] multi-source feedback or 360-degree feedback[29] and portfolios.[30] We see the growing prominence of general competencies as a significant development, because it will require a different assessment orientation with potential implications for other areas of assessment. Information gathering for the assessment of such general competencies will increasingly be based on qualitative, descriptive and narrative information rather than on, or in addition to, quantitative, numerical data. Such qualitative information cannot be judged against a simple, pre-set standard. That is why some form of professional evaluation will be indispensable to ensure its appropriate use for assessment purposes. This is a challenge to which assessment developers will have to rise in the near future. In parallel to what we have said about the dangers of reductionism, the implications of the use of qualitative information point to a similar respect for holistic professional judgement on the part of the assessor. As we move further towards the assessment of complex competencies, we will have to rely more on other, and probably more qualitative, sources of information than we have been accustomed to and we will come to rely more on professional judgement as a basis for decision making about the quality and the implications of that information. The challenge will be to make this decision making as rigorous as possible without trivialising the content for

'objectivity' reasons. There is much to be done in this regard.[31]

Impact on learning

The impact of assessment on learning has also been termed 'consequential validity',[4] which is incorporated in the formal definition of validity by the American Educational Research Association.[32] We prefer to use it as a separate criterion, simply because of its importance in any balanced utility appraisal. This brings us to 2 somewhat paradoxical observations.

The first observation is that the notion of the impact of assessment on learning is gaining more and more general acceptance. Many publications have acknowledged the powerful relationship between assessment and learning. Recognition of the concept that assessment is the driving force behind learning is increasingly regarded as one of the principles of good practice in assessment.[33] Unfortunately, this does not mean that changes are easy to achieve in practice or that changes in assessment will no longer be the last item on the agenda of curriculum renewal.[34]

The second observation is that there is a paucity of publications that shed light on the relationship between assessment and learning.[35] From our daily experience in educational practice we are familiar with some of the crucial issues in this respect: how to achieve congruence between educational objectives and assessment; how to provide and increase feedback from assessment; how to sustain formative feedback; how to combine and balance formative and summative assessment; how much assessment is enough; how to spread assessment over time, etc. Unfortunately, published information that can further our thinking and progress in this area is hard to come by.

An explanation of this scarcity may be that it is almost impossible to study the impact of assessment on learning without knowing about the context of the assessment. For example, a recent paper showed that students' performance on an OSCE station had a much stronger relationship with the students' momentary context (the rotation they were in) than with their past experience with the subject.[36] The concept that a characteristic of an assessment method is not inherent in the method but depends on how and in what context assessment takes place is even more applicable in the case of its impact on learning than for any of the other characteristics in the utility equation. Similar methods may lead to widely differing educational effects, depending on their use and place in the overall assessment programme. This means that we are badly in need of more studies to address the issues mentioned above, research that will inevitably require more specification of the assessment context.

ASSESSMENT AS INSTRUCTIONAL DESIGN

The paragraph on reliability indicated that there are no inherently inferior assessment methods and that reliability depends on sampling. However, it is also fair to admit that most of the assessment methods commonly used within regular medical training programmes are used unreliably. The section on validity showed that we cannot expect a single method to be able to cover all aspects of competencies of the layers of Miller's pyramid, but that we need a blend of methods, some of which will be different in nature, which may mean less numerical with less standardised test taking conditions. Professional judgement is important, both for the assessment tasks that we design as well as for the assessor who appraises task performance. As for the impact of assessment on learning, we have made it clear that any method of assessment can have any sort of influence on learning (positive or negative), depending on how it is used and in what context. It is our view that the preceding discussion constitutes a strong plea for a shift of focus regarding assessment, that is, a shift away from individual assessment methods for separate parts of competencies towards assessment as a component that is inextricably woven together with all the other aspects of a training programme. From this point of view, the instructional design perspective, the conceptual utility model should be applied at the level of the integral assessment programme. Assessment then changes from a psychometric problem to be solved for a single assessment method to an educational design problem that encompasses the entire curriculum. Keeping in mind what is acceptable in a given context (i.e. level of expertise of staff, past experience in assessment, student and staff beliefs) and the available resources, the challenge then becomes how to design an assessment programme that fulfils all the assessment criteria. This approach offers considerably more degrees of freedom in the use of a variety of methods. One could try and cover the entire competency pyramid, deliberately incorporating 'hard' measures in some instances on reliability grounds and 'softer' ones in other instances in order to deliberately steer

learning in a certain direction. One can diversify the selection of methods, using some that elicit verbal expression and others that call for writing skills. Instead of designing course-related assessments only, one could think of longitudinal, course-independent measures targeted at individual students' growth or personal development. The issue then is not whether one uses 'old-fashioned' or 'modern' methods of assessment, but much more why and how we should select this or that method from our toolbox in a given situation. Trustworthy and credible answers will ultimately determine the utility of the assessment.

A programmatic, instructional design approach to assessment surpasses the autonomy of the individual course developer or teacher. It requires central planning and co-ordination and needs a well written master plan. Essentially, this notion follows that of modern curriculum design. No curriculum renewal will be successful without careful orchestration and planning.[37] The same holds for an assessment programme. Another likeness to curriculum design is the need for periodic re-evaluation and re-design. The effect of assessment on learning can be quite unpredictable and may change over time. For example, the way regulations are set or changed may result in dramatic strategic effects for learners. This means that ongoing evaluation and adjustment of the assessment programme will be imperative.

We know that many psychometric issues are involved in collating assessment information and combining scores from different sources. We cannot say that the use of multiple measures will automatically increase reliability and validity. With every decision that is made within an assessment programme, reliability is at stake and decision errors are made and may accumulate. When we combine information from totally different sources, we may seem to be adding apples to oranges in a way that will inevitably complicate the evaluation of the validity. Yet making pass or fail decisions is something that – again – should be evaluated at the level of the programme. We think that this too will require professional judgement. We should move away from the 1-competence–1-method approach to assessment.[5] A good assessment programme will incorporate several competency elements and multiple sources of information to evaluate those competencies on multiple occasions using credible standards. The information obtained will have to be aggregated into a final (promotion) decision. When all sources point in the same direction, the information is consistent and the decision is relatively straightforward. With conflicting information, decision making is more problematic and a defensible judgement will require additional information, by obtaining more information, by adding more decision makers, by a conditional promotion decision or by postponing the decision. Such a decision-making procedure bears far greater resemblance to a qualitative approach that continues to accumulate information until saturation is reached and a decision becomes trustworthy and defensible.[31] A good assessment programme should have a feedback mechanism in place to ensure that any final decision does not come as a total surprise to the learner. If the latter is the case, it is indicative of a failure of the feedback mechanism somewhere along the way, which should be remedied.

IMPLICATIONS FOR DEVELOPMENT AND RESEARCH

In a programmatic, instructional design approach to assessment, 'simple' psychometric evaluation will not suffice. We should probably start with more and proper descriptions of such assessment programmes. To our knowledge, there are only a few good examples of these,[19,21,38,39] and we definitely need many more. These programme specifications should motivate the choices that are made and provide sufficient contextual information. From these descriptions, commonalities could then be inferred. What constitutes a good programme, what factors contribute to it and what are the pitfalls? How is information combined and how are decisions made? Further empirical research could investigate whether the intended programme does or does not work in practice. Are the intended learning effects actually being achieved? How do the stakeholders perceive the programme? This kind of research will be less psychometrically oriented (although there are some good examples[40,41]) and will probably bear more resemblance to curriculum research.

It is our opinion that the assessment literature is overly oriented towards the individual assessment method and too preoccupied with exclusively psychometric issues. We advocate the perspective that any method can have utility, depending on its usage and the programmatic context. There are no inherently bad or good assessment methods. They are all relative. What really matters is that the assessment programme should be an integrated part of the curriculum and this should be the main focus of our attention and efforts. The crucial question concerns the utility of the assessment programme as a whole.

Contributors: both authors contributed to the views expressed in this article. CvdV wrote the paper, assisted by suggestions from LS.
Acknowledgements: none.
Funding: none.
Conflicts of interest: none.
Ethical approval: not applicable.

REFERENCES

1 van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ* 1996;**1** (1):41–67.
2 Gielen S, Dochy F, Dierick S. Evaluating the consequential validity of new modes of assessment: the influence of assessment on learning, including pre-, post- and true assessment effects. In: Segers M, Dochy F, Cascallar E, eds. *Optimising New Modes of Assessment: in Search of Qualities and Standards*. Dordrecht: Kluwer Academic Publishers 2003.
3 Linn RL, Baker E, Dunbar SB. Complex, performance-based assessment: expectations and validation criteria. *Educational Res* 1991;**16**:1–21.
4 Messick S. The interplay of evidence and consequences in the validation of performance assessments. *Educational Res* 1995;**23**:13–23.
5 Schuwirth LWT, van der Vleuten CPM. Changing education, changing assessment, changing research. *Med Educ* 2004;**38**:805–12.
6 Downing SM. The metric of medical education. Validity: on the meaningful interpretation of assessment data. *Med Educ* 2003;**37** (9):830–7.
7 Hodges B. Validity and the OSCE. *Med Teach* 2003;**25** (3):250–4.
8 Swanson DB, Norman GR, Linn RL. Performance-based assessment: lessons from the health professions. *Educational Res* 1995;**24** (5):5–11.
9 van der Vleuten CPM, Norman GR, de Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ* 1991;**25**:110–8.
10 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;673–709.
11 Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med* 2003;**15** (4):270–92.
12 Schuwirth LWT, van der Vleuten CPM. The use of clinical simulations in assessment. *Med Educ* 2003;**37** (1):65–71.
13 Rethans JJ, Norcini JJ, Baron-Maldonado M *et al*. The relationship between competence and performance: implications for assessing practice performance. *Med Educ* 2002;**36** (10):901–9.
14 Norcini JJ, Blank LL, Arnold GK, Kimbal HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med* 1995;**123**:795–9.
15 Turnbull J, van Barneveld C. Assessment of clinical performance: In-training evaluation. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education*. Dordrecht: Kluwer Academic Publishers 2002;793–810.
16 Ram P, Grol R, Rethans JJ, Schouten B, van der Vleuten CPM, Kester A. 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):447–54.
17 Gorter S, Rethans JJ, Scherpbier A *et al*. How to introduce incognito patients into outpatient clinics of specialists in rheumatology. *Med Teach* 2001;**23** (2):138–44.
18 Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;**65** (9):63–7.
19 Epstein RM, Hundert EM. Defining and assessing professional competence. *JAMA* 2002;**287** (2):226–35.
20 Harden RM. Developments in outcome-based education. *Med Teach* 2002;**24** (2):117–20.
21 Smith SR, Dollase RH, Boss JA. Assessing students' performance in a competency-based curriculum. *Acad Med* 2003;**78**:97–107.
22 van Merriënboer JJG. *Training Complex Cognitive Skills*. Englewood Cliffs, New Jersey: Educational Technology Publications 1997.
23 Swanson DB, Norcini JJ, Grosso LJ. Assessment of clinical competence: written and computer-based simulations. *Assess Eval Higher Educ* 1987;**12** (3):220–46.
24 Regehr G, MacRae H, Reznick R, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med* 1998;**73** (9):993–7.
25 Coles C. Developing professional judgement. *J Contin Educ Health Prof* 2002;**22** (1):3–10.
26 Jozefowicz RF, Koeppen BM, Case SM, Galbraith R, Swanson DB, Glew RH. The quality of in-house medical school examinations. *Acad Med* 2002;**77** (2):156–61.
27 Dochy F, Segers M, Sluijsmans D. The use of self-, peer and co-assessment in higher education: a review. *Studies Higher Educ* 1999;**24** (3):331–50.
28 Norcini JJ. The metric of medical education: peer assessment of competence. *Med Educ* 2003;**37** (6):539–43.
29 Lockyer J. Multisource feedback in the assessment of physician competencies. *J Contin Educ Health Prof* 2003;**23**:2–10.
30 Snadden D. Portfolios – attempting to measure the unmeasurable? *Med Educ* 1999;**33**:478–9.
31 Driessen EW, van der Vleuten CPM, Schuwirth L, van Tartwijk J, Vermunt JD. The use of qualitative research criteria for portfolio assessment as an alternative for reliability evaluation: a case study. *Med Educ* 2005;**38**:214–20.

32 Anonymous. *Standards for Educational and Psychological Testing*. Washington DC: American Educational Research Association 1999.

33 Crossley J, Humphris G, Jolly B. Assessing health professionals. *Med Educ* 2002;**36**:800–4.

34 Mennin SP, Kalishman S. Student assessment. *Acad Med* 1998;**73** (9):S46–54.

35 Gibbs G, Simpson C. *How Assessment Influences Student Learning – a Conceptual Overview:* Heerlen, the Netherlands: Open University. Centre for Higher Education Practice 2002.

36 Chibnall JT. The influence of testing context and clinical rotation order on student OSCE performance. *Acad Med* 2004;**79** (6):597–601.

37 Davis WK, White CB. Managing the curriculum and managing change. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education*. Dordrecht: Kluwer Academic Publishers 2002;917–44.

38 Epstein RM. *Comprehensive Assessment Manual*. Rochester, New York: University of Rochester School of Medicine and Dentistry 2001.

39 Accreditation Council for Graduate Medical Education. Outcome project. http://www.acgme.org/Outcome/. Last accessed 25 January 2005.

40 Hays RB, Fabb WE, van der Vleuten CPM. Reliability of the fellowship examination of the Royal Australian College of General Practitioners. *Teach Learn Med* 1995;**7**:43–50.

41 Wass V, McGibbon D, van der Vleuten C. Composite undergraduate clinical examinations: how should the components be combined to maximise reliability? *Med Educ* 2001;**35**:326–30.

42 Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Med Educ* 1985;**19**:238–47.

43 Hatala R, Norman GR. Adapting the key features examination for a clinical clerkship *Med Educ* 2002;**36**:160–5.

44 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;13–45.

45 Wass V, Jones R, van der Vleuten C. Standardised or real patients to test clinical competence? The long case revisited. *Med Educ* 2001;**35**:321–5.

46 van der Vleuten CPM, van Luijk SJ, Swanson DB. *Reliability (Generalisability) of the Maastricht Skills Test*. Proceedings of the 27th Annual Conference on Research in Medical Education (RIME). Chicago: American Association for Medical Colleges 1988.

47 Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. *Ann Intern Med* 2003;**138** (6):476–481.

48 Gorter S, Rethans JJ, van der Heijde D *et al.* Reproducibility of clinical performance assessment in practice using incognito standardised patients. *Med Educ* 2002;**36** (9):827–32.

Received 1 March 2004; editorial comments to authors 4 May 2004; accepted for publication 19 July 2004