2011; 33: 478–485 MEDICAL TEACHER

Programmatic assessment: From assessment of learning to assessment for learning

LAMBERT W. T. SCHUWIRTH & CEES P. M. VAN DER VLEUTEN
Maastricht University, The Netherlands

Abstract

In assessment a considerable shift in thinking has occurred from assessment of learning to assessment for learning. This has important implications for the conceptual framework from which to approach the issue of assessment, but also with respect to the research agenda. The main conceptual changes pertain to programmes of assessment. This has led to a broadened perspective on the types of construct assessment tries to capture, the way information from various sources is collected and collated, the role of human judgement and the variety of psychometric methods to determine the quality of the assessment. Research into the quality of assessment programmes, how assessment influences learning and teaching, new psychometric models and the role of human judgement is much needed.

Introduction

In the recent decades, a change in thinking about the role of assessment in education has occurred. This change is best characterised as a shift from assessment *of* learning to assessment *for* learning (Martinez & Lipson 1989). Behind this rather inconspicuous terminology hides nothing short of a revolution in the conceptual framework of assessment. In this article, we want to describe the implications of this change for our thinking and the practices of assessment, with a special focus on assessment in the context of medical education.

Most of us are most familiar with examinations that take place at the end of the instruction and are separated from the educational process. From the perspective of assessment of students, the almost exclusive purpose of such assessment is to determine whether the students have acquired sufficient knowledge, skills, etc. Assessment *for* learning, however, is an approach in which the assessment process is inextricably embedded within the educational process, which is maximally information-rich, and which serves to steer and foster the learning of each individual student to the maximum of his/her ability.

The idea of assessment *for* learning is not new; Martinez & Lipson (1989) already proposed it in 1989. Though their interpretation of assessment for learning is in its early developmental phase and does not surpass the notion of more dispersed test administrations and the use of more feedback, it is an early demonstration of a growing awareness that for assessment to be an integral and more relevant aspect of education, tests that merely try to classify and rank order students do not suffice anymore.

In the mean time, the theoretical perspective of assessment for learning programmes has evolved considerably. This is not illogical because originally the concept of assessment of learning had firm roots in the 20th century discourse of education and ability. Shepard (2009) describes the previously prevailing views on education as conceptually equivalent to a factory production process. Central in these views is a behaviouristic concept of learning, implying that becoming competent in a domain is the result of following a large number of small steps or modules, each of which has to be assessed at the end. Only after successful completion of a module can the student progress to the next. It follows then logically that assessment has to take a reductionist approach as well, viewing the total only as the sum of its constituent parts.

With the emergence of new – social constructivist – theories on learning and the notion of competencies as outcome indicators of the educational process the call for radical changes in the way we set up and use assessment is heard in the literature (Boud 1990; Brown 2004; van der Vleuten & Schuwirth 2005; Shute 2008). This was a highly needed antithetic movement against the traditional approaches.

In 2005, we advocated a more synthetic view on assessment incorporating both views and we suggested the notion of

Practice points

- In educational settings assessment for learning should take priority over assessment of learning.
- A programme of assessment should aim at building n:n relationships: each competency domain should be informed from various assessment sources and each assessment source should be used to inform about several competency domains.
- For programmatic assessment as part of assessment for learning, extensions to current psychometric approaches are needed.
- The role of human judgement in assessment should be re-appraised and studied.

Correspondence: L. W. T. Schuwirth, Department of educational development and research, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands. Tel: 31 43 3885731; fax: 31 43 3885779; email: l.schuwirth@maastrichtuniversity.nl

ISSN 0142–159X print/ISSN 1466–187X online/11/060478–8 © 2011 Informa UK Ltd.
DOI: 10.3109/0142159X.2011.565828

programmatic assessment (van der Vleuten & Schuwirth 2005). And although the literature on assessment for learning already acknowledged that a variety of instruments would be needed to obtain a more complete picture (Ram 1998; Prescott et al. 2001; Epstein & Hundert 2002; Davies et al. 2005; Carr 2006), the idea of programmatic assessment goes further. In programmatic assessment, modern approaches do not necessarily replace but rather supplement traditional ones (Prescott et al. 2002; van der Vleuten & Schuwirth 2005; Dannefer & Henson 2007; Fishleder et al. 2007). The central key is that the programme of assessment is set up to allow the whole picture of a student's competence to be obtained by a careful selection of assessment methods, formulation of rules and regulations and design of organisational systems. And although this notion has shown to be an appealing one for many educationalists, it still requires more concrete development.

It is the purpose of this article to describe further routes for the development of the idea of programmatic assessment in the context of assessment for learning.

Where are we now?

Many traditional examination programmes subdivided medical competence into four separate constructs: knowledge, skills, problem-solving skills and attitudes or professionalism. A good assessment programme in this view is composed of a combination of instruments for each of these constructs. The medical assessment literature in the 20th century is dominated either by papers presenting new instruments suggesting they measure one of the constructs better than previous methods or comparing different methods to prove the superiority of one of them. The original papers on triple jump exercises (Painvin et al. 1979), objective structured clinical examination (OSCE; Harden & Gleeson 1979), etc. are examples of the former; many papers comparing open-ended with multiple-choice questions are examples of the latter (e.g. Norman et al. 1996; Newble et al. 1979). This view, however, has important underlying assumptions which we will discuss here.

Each construct is treated as a stable and generic trait

Traits, here, are assumed to be both stable and generic. Much like, for example intelligence and extraversion. The intelligence of a person is assumed to be stable – at least in the short run – across measurements. Of course a person's actions and decisions may vary in their cleverness, but his/her intelligence is assumed to be stable. Logically, this variability in cleverness of the actions and decisions is almost invariably treated as error variance.

The traits are also assumed to be generic, one can be intelligent and introverted or intelligent and extraverted and vice versa. Similarly, there is no inherent relationship assumed between the four constructs; knowledge, skills, problem-solving skills and attitudes.

From this it follows naturally that reliability (or universe score representation) can best be determined by reproducibility of the test scores. So if a test of four items would be perfectly reliable, the score matrix of the results of students A, B and C would look like as shown in Table 1.

Table 1. Score matrix of a perfectly reliable test.

	1	2	3	4	Total
A	1	1	1	1	4
B	0.5	0.5	0.5	0.5	2
C	0	0	0	0	0

Table 2. Score matrix of the hypothesised parallel test.

	1′	2′	3′	4′	Total′
A	1	1	1	1	4
B	0.5	0.5	0.5	0.5	2
C	0	0	0	0	0

Table 3. Score matrices in real life.

	1	2	3	4	Total	1′	2′	3′	4′	Total′
A	0	0.5	0.5	0	1	1	0	0.5	1	2
B	1	0.5	0	1	2.5	1	0.5	0	0	2
C	1	1	0.5	1	3.5	1	0.5	0	0	1.5

A further assumption is that if these students were given another so-called parallel test (a test of equal difficulty on the same topics), the expected score matrix would be the same (Table 2).

Of course this is never the case; matrices look more like as shown in Table 3.

In this case, all the variance that does not fit the assumption of the stable trait is incorporated in the error variance.

Individual items or elements of a test are in principle meaningless

If performance on individual items can vary and this variability is seen as error, it is only logical that individual items in themselves can be treated as meaningless; their only value is the extent to which they contribute to the total score, and the total score is what can give meaning and validity to the assessment. In the case of a multiple-choice test, for example on internal medicine, most people would not have major objections to treating individual items as meaningless. The first item in Box 1 can be easily replaced by the second item, and if two students score 0.5 on the combination of both items pass and fail or even remediation decisions would not depend on whether they answered the first or the second item incorrectly.

It becomes more problematic if the two items are intuitively more meaningful, for example resuscitation and a communication station in an OSCE. Most people would question whether good communication skills can make up for poor resuscitation skills – unless perhaps the communication skills

Box 1. Two examples of items of an internal medicine examination.

Item 1: Ms. Smith is 72 years old. She has angina pectoris. Several times her blood pressure is taken and found to be 170/100 mmHg. Which antihypertensive drug is most indicated for her?
(a) captopril
(b) chlorthalidone
(c) metoprolol

Item 2: Mr. Johnson, 35 years old, consults his GP with complaints of chest pain. Without further information about Mr. Johnson the most likely origin of his chest pain is:
(a) the chest wall
(b) the lungs
(c) the myocardium
(d) the oesophagus

station is on breaking bad news. The traditional psychometrical approaches we use, however, lead us to treat individual stations as intrinsically meaningless.

Statistics are based on elimination of information

In view of education as a (factory) production process in which competence is acquired by carefully going through a sequence of discrete predefined steps – as discussed in the introduction – the assessment results have to indicate whether a step in the process was completed successfully or not. For this, procedures are designed to eliminate information as well as possible in order to arrive at the best possible dichotomous decision. Take the answers a student gives to a multiple-choice test. From the answers, it can be derived not only which correct answers were given but also which incorrect answers were given. But then the answers are compared to an answer key and converted to 1–0 scores. Now it is not known anymore what the incorrect answers were but only to which question an incorrect answer was given. Then the item scores are totalled. Now it is obscured to which items an incorrect and correct answer was given but only to how many items an incorrect or correct answer was given. This total score is then compared to a pass–fail score and now it is only known whether the number of correct answers was sufficient or not. The literature on scoring rubrics and standard setting methods is basically literature on how best to throw away assessment information (Cusimano 1996).

One single best instrument for each trait

The consequence of this is – as said before – that traditional examination programmes are built according to the one-best-instrument-for-each-trait model. The vast literature on whether open-ended questions are better than multiple-choice questions, the literature on proving that OSCEs are *the* best instrument for the assessment of skills are typical examples of these lines of thinking.

Where do we want to go?

A search in the literature will most probably produce over 100 definitions of 'competency' (e.g. Albanese et al. 2008, Govaerts 2008). Apparently there is no completely agreed upon definition, but there is common ground, and the definitions converge on the notion of integration of knowledge, skills and attitudes/professionalism, the whole task performance. For the purpose of this article, we will define competencies as simple or more complex tasks a successful candidate must be able to handle, and during which s/he uses at the right time, the correct and relevant knowledge, skills, attitudes and meta-cognitions to manage the task successfully. Many official institutes issued their own set of competency domains or professional roles. The CanMeds (1996) contain the domains: Medical expert, Communicator, Collaborator, Manager, Health advocate, Scholar and Professional. The ACGME (2007) defined the domains: Patient care, Medical Knowledge, Practice-based Learning and improvements, Interpersonal and Communication skills, Professionalism and Systems-based practice. The first Dutch blueprint for medical education used four roles (Metz et al. 1994): Medical Expert, Scientist, Worker in the health care system and Person. We will use these four in the remainder of this article, not because we think they are better than the others, but simply because they are lean and easy to explain.[1]

The risk we as educators run now is that we would now be inclined to build an assessment programme in which one single best instrument is used for each of the domains. At our own institute, for example, the critically appraised topic (CAT) is almost exclusively used as an instrument to assess the role as a scientist, creating a one-instrument-to-one-competency domain programme. This way we would be making the same mistake as with the traditional assessment programmes, namely treating the domains as unidimensional, stable and generic entities. But then we would simply be replacing words ('traits' by 'competencies') instead of building a really new assessment programme. An important thing in innovative assessment programmes is that they are based on the notion of an n:n relationship. In other words, information of all assessment sources can be used to inform about all the competency domains, and all competency domains are informed by various information sources.

This may seem complex but in fact it is not. Especially for those who practice or have practiced medicine the analogy is simple. No clinician would convert the patient responses during history taking to numbers and average them and then add this average to the mean of the lab values to determine whether the patient is healthy or not, etc. Instead s/he takes the relevant information from history taking, physical examination, lab results, pathology reports, etc. to determine whether further diagnostics are needed, what therapy or management to start and whether the patient is healthy or not. This is exactly the n:n relationship we suggest to use in assessment programmes.

The traditional approach in most assessment programmes relies on adding the results on the communication skills station of an OSCE to the resuscitation skills, not because they can be combined rationally but simply because they have the same format (to use the analogy again: so do the sodium and potassium level). This is strange especially because a plethora of research has shown that it is not the format which determines what a test or an item assesses but the content

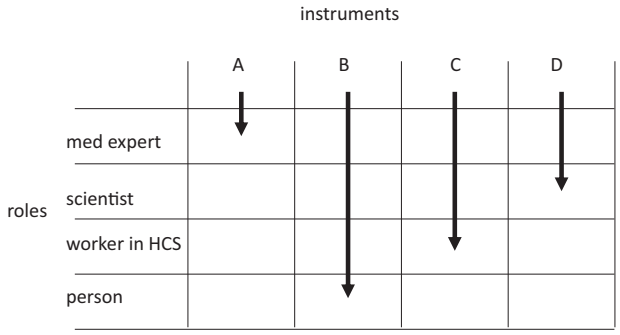

Figure 1. A typical 1:1 assessment competency-based programme.

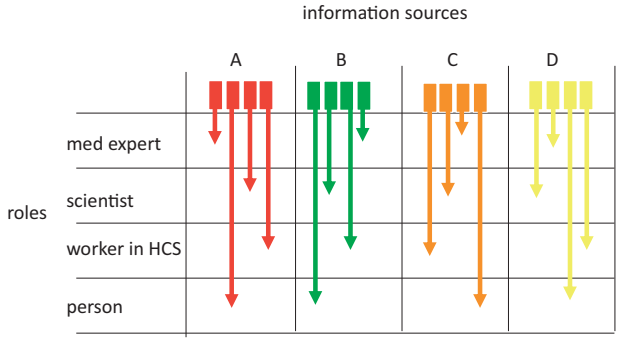

Figure 2. A typical n:n assessment competency-based programme.

(Ward 1982; Norman et al. 1985; Schuwirth et al. 1996). Theoretically, it is more logical to combine information that is similar in content and not because it is similar in format. Figures 1 and 2 demonstrate this difference.

In such an assessment programme, the constructs – the aspects we want to assess – do not have to be defined in stable and generic traits, some will have to be defined as variable and some as stable. Again, using the analogy with clinical work: some parameters are supposed to be so stable that one measurement suffices to determine them (sodium level, haemoglobin level) and some others are supposed to vary considerably (blood pressure, blood glucose level) that only repeated measurements or daily curves are informative enough.

So, individual elements of the assessment can be meaningful in themselves. The low score on the item 'history taking' in a mini-CEX is meaningful in itself and can lead to remedial actions. For this, it does not have to be added to the rest of the items of the same mini-CEX (Norcini et al. 1995). On the other hand, individual items or elements can acquire meaning in a combination with elements of other tests. A failed abdominal examination station in an OSCE will have different repercussions for the student if s/he has also performed poorly on test items on abdominal anatomy than for the student whose patient communication is poor. Let us – again – use the clinical analogy: a haemoglobin level of 7.5 mmol/L is a positive sign in a female patient two days after a delivery with considerable bleeding after iron therapy, but it is an ominous sign in a 55-year-old male who visits you with rectal blood loss.

In this light, it is important to highlight the relationship between objectivity/subjectivity and reliability/unreliability (van der Vleuten et al. 1991). There is a widespread misconception that only 'objective' tests can be reliable and that 'subjective' tests are unreliable. Unfortunately, this kind of thinking is not very helpful in improving the quality of the assessment. A single-item multiple choice test on internal medicine would be a so-called objective test, but it can hardly be a reliable test as one item is simply too small a sample. On the other hand, a collection of expert opinions on a certain performance (e.g. musical artistry) can be highly reliable, as long as there are multiple experts and multiple pieces of music played, and perhaps observations at various occasions. This distinction between improving the sampling qualities of an assessment and attempts to make it objective is important. There are many cases in which assessment designers in their pursuit of objectivity have unnecessarily trivialised the assessment, for example by designing scoring rubrics for portfolios (Koretz 1998). In programmes of assessment, subjective elements should not be trivialised but should be assessed by optimising the sampling procedure (Schuwirth et al. 2002; Driessen et al. 2005).

Of course this does not make the whole assessment process easier, quite the contrary. If we cannot break down the results into little pieces which arithmetically add up to a total score, human judgement is needed to collect and collate information, especially if – in a programme of assessment – information from various types of assessment needs to be combined. When human judgement is central in the assessment process, it may be clear that the quality and expertise of the person who is making the judgement is decisive for the quality of the assessment. Where in multiple choice test the quality of the assessment is built into the paper – it really does not matter who hands out the test forms and a computer can do the calculations – in assessments involving human observation and evaluation the quality of the user is central (and the form basically serves only to support the users).

To ensure the quality of the assessment then, the exclusive focus on construct validity and reproducibility do not suffice anymore. Concepts such as fairness, trustworthiness and dependability also need to be included (Driessen et al. 2005). Such concepts of course can only be established on the programme level and mainly through organisational procedures, such as second opinions, independent observations, careful note taking, interdisciplinary consultations, etc.

This way, the assessment programme can be tailored specifically to the individual needs of each student. First, this enables the teacher/mentor to advise that specific assessment information be collected for an individual student to ensure a complete picture of his/her competence. For a student who has had seven excellent independent mini-CEX judgements on all criteria, further collection of data is probably not useful, whereas in the case of seven highly variable judgements more information may be necessary. This could be called the 'diagnostic' decisions in assessment for learning. Also, as described above, a tailored advice for remediation can be given for each student, which could be called the 'therapeutic' decision. Finally, a prognostic decision – is the student on the

right track to sufficient competency – can be made about each student.

In summary, assessment for learning is an information-rich approach in which a programme of assessment is used to collect and combine information from various sources to inform about the strengths and weaknesses of each individual student, with the purpose to optimise their learning. So, the central goal is not whether John is better than Jill (or better than a cut-off score, which essentially is a specific 'Jill'), but to determine whether John is maximally better today than he was yesterday or whether Jill will be maximally better tomorrow than she is today, and how to achieve this.

Implications for research

It is clear that such a change in thinking about assessment must have implications for the research agenda. Of course, the most obvious item on the agenda would be research into feedback. This, fortunately, is a topic that is already being researched extensively (Shute 2008). The outcomes of the many studies have provided us with valuable insights regarding the value of feedback, how best to provide feedback, how to pace and schedule feedback sessions, etc. We want to discuss here research paths that, at least in medical education, are less well trotted.

What constitutes high-quality assessment programmes?

In 1996, van der Vleuten (1996) published a paper in which he advocated to evaluate the quality of individual assessment instruments as a trade-off between various criteria. In this paper, he suggested to include reliability, validity, educational impact, cost-effectiveness and acceptability as such criteria. Although these criteria have been shown to be useful for individual instruments their applicability to an assessment programme as a whole is limited. Dijkstra et al. (2009) have endeavoured on a research line that will provide us with more insight into what constitutes quality of assessment programmes. He takes here another angle than earlier work by Baartman (2008). Her work is focussed on the evaluation of the quality competency assessment programmes as a post hoc measurement, whereas Dijkstra tries to establish design guidelines or criteria for building or adapting a programme of assessment. A first study (Dijkstra et al. 2009) was done with two focus groups of international assessment experts in which their opinions and knowledge were sought on about good practices and on new ideas concerning programmes of assessment. After an extensive iterative analysis and member checking, a model emerged which incorporated of course the most obvious design criteria, namely those concerning the programme in action (collecting assessment information, combining assessment information from various instruments, valuing the resultant information to come to decisions, and taking action upon these decisions). In addition, however, a series of different layers were defined: criteria concerning the supporting aspects for a programme, criteria for documentation and dissemination of information about the assessment programme, measures for continuous improvement of the programme and the quality of all procedures supporting the accountability of the programme. A second study started using the same data but now extended with a series of individual, structured interviews with the international experts has led to the definition of over 65 specific design criteria or guidelines for developing or improving programmes of assessment (Dijkstra et al. submitted). Further research to validate the model and the resulting criteria is underway.

How does assessment influence learning?

This of course is a central question if programmes of assessment are to be established in the context of assessment for learning. Amazingly, the amount of research actually studying this educational impact is scarce, especially in light of the strength of the shared opinion that assessment does impact on learning and teaching. Cilliers et al. (2010) have conducted a series of individual in-depth interviews with students and teachers at the University of Stellenbosch in South Africa. Using a grounded theory approach, a model emerged which relates the qualities of the assessment programme to the outcomes of learning. Three main elements were identified: sources of impact, mechanisms by which these sources impact on student learning and the consequences of the impact.

Mechanisms of impact constituted the ways students appraised the impact of the assessment programme, their own learning response, their own perceptions of agency and contextual factors. As sources, main factors were task demands, imminence of assessment, the design of the assessment system and the cues (Cilliers et al. 2010). As consequences, the main groups were defined by cognitive and meta-cognitive regulation activities (Cilliers et al. 2011). Future work in this research line will incorporate further validation/generalisation of the model to different contexts and populations. It will hopefully provide enough insight to help us design assessment programmes for learning in a way that we can actually predict better how programmatic design decisions will impact on the educational process.

Extension of psychometric models

In 2006, we published a plea for the extensions of the psychometric models used to determine the quality of assessment instruments (Schuwirth & van der Vleuten 2006). In this paper, we highlighted some of the major concessions that would have to be made in some assessment situations, and we advocated that new methods would be developed that cater better to more observation-based instruments, such as mini-CEX, 360° feedback and portfolios. It turns out that there have already been developments in this area in the 1960s and 1970s with respect to criterion-referenced tests (Berk 1980; Rickets 2009). Since then, psychometric theory and resulting insights have changed dramatically. Validity for example has evolved from a uniform (instead of a unifying) construct validity theory to a much more eclectic and at the same time rigorous theory, thanks to the important work by Messick (1994) and Kane (2006). Especially Kane highlights the need for an argument based set of inferences from observations eventually to conclusions about the target domain. This approach is eclectic

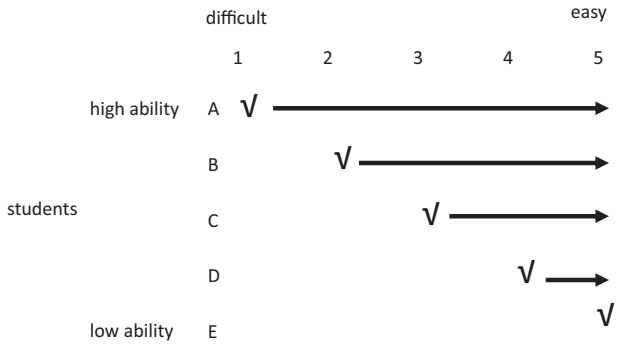

Figure 3. A typical example of a Guttman scale.

because it requires the researcher and/or assessors to make explicit assumptions about the nature of the target domain, and it does not automatically subsume a stable trait. It is rigorous because it does not allow the researcher to rely on standard tricks or select the most convenient methods of validation, but instead requires him/her to build a complete set of arguments supported by research testing the most critical assumptions with the highest priority. As such, it does not differ very much from the original requirements Cronbach & Meehl (1955) put forward, but it is a theoretical notion that is considerably more versatile.

One of the important inferences to make is the one from observed score to universe score (as a sort of reliability). For this, a conceptualisation of the universe score is needed and this conceptualisation can take different forms. Most of our thinking is more or less based on the notion of the Guttman scale. Figure 3 shows an example of such a Guttman scale.

In such a scale, the underlying assumption is that some items are inherently easier than others. For example, my eldest daughter can ride her bike without training wheels. Therefore, it is safe to assume that she is also able to ride a bike with training wheels, and following from this that she can ride a tricycle, and that she is able sit straight, etc. If a student is able to perform an abdominal examination successfully in a 65-year-old highly obese patient with severe abdominal complaints and s/he is still able to come to correct conclusions, s/he is most likely also to be able to perform a successful abdominal examination in a normal healthy young adult. Sets of items that behave well according to such a Guttman scale are very well suited for certain assessment approaches, such as computer adaptive testing. The logical consequence is that any variance not in accordance with the assumed scale is measurement error. So if at one day my daughter is observed being able to ride without training wheels and subsequently fail to ride the bike with trainers, the only logical assumption is this to be measurement error. Universe representation is only sufficient if the various observations agree enough about the level of ability of the student in relation to the extent to which the data allow us to distinguish between the levels of abilities of the students. It subsumes a homogeneous universe from which the sample is drawn.

One can of course wonder whether all aspects of assessment in a programme of assessment are best modelled this way. We know that the population of the Netherlands is roughly 16.5 million. This does not automatically imply that we know what the population of the US or the UK is, or even Belgium for that matter. If we know that surfactant is produced by type II pneumocytes, this does not automatically imply that we know which cells produce calcitonin, or even where the type II pneumocytes are located. In such cases measures of universe representation need to describe the probability of a new observation providing new information about the representation of the universe, but it does not prescribe homogeneity of the universe. The main focus now is not on determining the position of a student on the ability scale, but to determine the proportion of relevant knowledge, skills, etc. the student has. We have given a more detailed description of both models and the implications for assessment in a specific paper (Schuwirth et al. submitted).

How to scaffold human judgement?

It is inevitable that in a programme of assessment human judgement is involved. This will probably not only happen at the level of individual observation and assessment but also in combining information from various sources. Traditionally in examination systems, information from qualitative sources is quantified – for example, the evaluation of professional behaviour 'good' is translated into an '8' – whereas in clinical practice, quantitative information is traditionally qualified – a sodium level of 132 mmol/L is translated into a 'low-normal sodium level'. In assessment *for* learning programmes, in which feedback and information-rich procedures are required, information needs to be combined in a qualitative way. This involves inevitably human judgement. Unfortunately though, human judgement is often considered fallible, especially when compared to actuarial methods (Dawes et al. 1989). This is of course logical, because in such comparisons conscious bottom-up processing (starting with evaluating all the individual data to arrive at a conclusion) is required, which is intended to lead to hard data conclusions and which can be numerically modelled. It is obvious that this is exactly what humans with their limited short-term memory processing capacities are not good at (Van Merrienboer & Sweller 2010). In such cases they necessarily have to revert to processing only a limited part of the whole data set. Typical heuristics humans apply in such cases are overvaluing first or last impressions (primacy and recency effects), combining data to form one single entity (halo effects), etc. (cf. Plous 1993 for an interview). In such cases human processing can only be seen as severely biased.

However, we are also capable of processing enormous amounts of information. Estimates, especially those including information from the visual system, are in the range of between 10 and 20 million bits per second. The research into naturalistic decision making focuses on human decisions in areas where the outcomes are not hard or numerical, but judgemental (Klein 2008), where too precise modelling of the data often leads to more inaccurate prediction than more superficial modelling (Marewski et al. 2009). In other words, why is human judgement with such an overload of information to process and vague outcomes still so good? It is clear that for such judgements processes more top-down processing

activities are needed. Still, however, methods for reduction of cognitive load are required. In this view, an incomplete representation of the information is not necessarily a bad representation, provided the essential important elements are in the representation. This bears a striking resemblance with scripts in the theory on expertise. (Schmidt & Boshuizen 1993) The implication of this – somewhat theoretical – expose is that we should not try to eliminate these types of biases (there are others like framing-related, cognitive dissonance-type and strategic behaviours minimising the concordance between private and public judgements that need to be addressed and counteracted in the programme of assessment), but we should train our judges to produce better representations instead of more 'objective' or more structured. In this case judgement tasks in assessment are diagnostic expertise tasks, and this has huge implications for teacher training programmes. A simple briefing, workshop or training cannot produce enough expertise for the job (like a single training in clinical reasoning does not work), but training on the job, constant feedback, supervision would be needed.

Govaerts et al. (2007) have paved the way in this direction. In their paper on the role of human judgement in assessment they have described the outcomes of preliminary studies in this field in the business literature and the medical education literature. Subsequently, they have studied the thought processes of experience GP supervisors and novice supervisors in a think aloud study (Govaerts et al. submitted). Members of both groups were individually shown DVD-recordings of a student performing a consultation with a patient. The medical content of both cases – a simple atopic dermatitis and a classical chest pain case – was no problem for all the participants, but the judgement of the student performance was. In the first case, the performance was clearly substandard, and in the second case it was marginally satisfactory/unsatisfactory (complex case). She found that in complex case experts needed more time than novices, but in the simple case they were faster. Experts make more inferences/interpretation whilst observing the performance, while novice provides more literal descriptions of the process, experts use more contextual cues and considerations and make more evaluations. This is all highly in concordance with the findings about diagnostic expertise (Schmidt & Boshuizen 1993; Eva 2004). Surprisingly, she found no difference in self-monitoring activities, which is different from findings in the expertise literature. She concludes that raters are not interchangeable measurement instruments, that richer processing and use of contextual cues leads to richer and more holistic feedback and judgements, that experts have better performance scripts which enable them to superior top-down processing. She suggests that training of raters should therefore incorporate maximally the characteristics of deliberate practice (Ericsson & Charness 1994).

Conclusion

This summary of development and research is by far incomplete. The field is so rich and there are so many developments and research activities that it would be impossible for us to describe them in full detail in a single paper, even if we assumed that we were completely informed about all of them. The purpose of this article was to raise awareness of the changes in the thinking of assessment, to show what huge implications this has for the practice of assessment and for the research agenda. For this we have tried to bring together several lines of development and research, necessarily omitting aspects the reader may find important or highlighting elements other readers may find irrelevant. Nevertheless, we hope to have been able to provide an overview that is helpful enough to stimulate new ideas for development and research and to foster future research endeavours.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Notes on contributors

Lambert W.T. Schuwirth, MD, PhD, is a Professor for Innovative Assessment at the Department of Educational Development and Research at the University of Maastricht in the Netherlands.

Cees P.M. van der Vleuten, PhD, is a Professor for Medical Education and Chair of the Department of Educational Development and Research at the University of Maastricht in the Netherlands.

Note

1. The new revised blueprint has adopted the seven domains of the CanMeds.

References

ACGME 2007. [Published 2008 September 28]. Available from: http://www.acgme.org/outcome/comp/compCPRL.asp.Chicago

Albanese MA, Mejicano G, Mullan P, Kokotailo P, Gruppen L. 2008. Defining characteristics of educational competencies. Med Educ 42:248–255.

Baartman LKJ. 2008. Assessing the assessment. Heerlen, The Netherlands: ETEC, Open University.

Berk RA. 1980. A consumers' guide to criterion-referenced test reliability. J Educ Meas 17:323–349.

Boud D. 1990. Assessment and the promotion of academic values. Stud High Educ 15:101–111.

Brown S. 2004. Assessment for learning. Learn Teach Higher Ed 1:81–89.

CanMeds. 1996. Available from http://rcpsc.medical.org/publications/index.php#canmeds. Ottawa.

Carr S. 2006. The foundation programme assessment tools: An opportunity to enhance feedback to trainees? Postgrad Med J 82:576–579.

Cilliers FJ, Schuwirth LWT, Adendorff HJ, Herman N, Van Der Vleuten CPM. 2010. The mechanisms of impact of summative assessment on medical students' learning. Adv Health Sci Educ Theory Pract 15(5):695–715.

Cilliers FJ, Schuwirth LWT, Herman N, Adendorff HJ, van der Vleuten CPM. A model of the sources, consequences and mechanism of impact of summative assessment on how students learn. Adv Health Sci Educ. DOI: 10.1007/s10459-011-9292-5.

Cronbach LJ, Meehl PE. 1955. Construct validity in psychological tests. Psychol Bull 52:281–302.

Cusimano MD. 1996. Standard setting in medical education. Acad Med 71:S112–S120.

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

Davies H, Archer J, Heard S, Southgate L. 2005. Assessment tools for foundation programmes – a practical guide. BMJ Career Focus 330:195–196.

Dawes RM, Faust D, Meehl PE. 1989. Clinical versus actuarial judgment. Science 243:1668–1674.

Dijkstra J, Galbraith R, Hodges B, Mcavoy P, McCrorie P, Southgate L, van der Vleuten CPM, Wass V, Schuwirth LWT. Development and validation of guidelines for designing programmes of assessment: A modified Delphi-study (submitted).

Dijkstra J, Van Der Vleuten CPM, Schuwirth LWT. 2009. A new framework for designing programmes of assessment. Adv Health Sci Educ Theory Pract 15(3):379–393.

Driessen E, Van Der Vleuten CPM, Schuwirth LWT, Van Tartwijk J, Vermunt J. 2005. The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: A case study. Med Educ 39:214–220.

Epstein RM, Hundert EM. 2002. Defining and assessing professional competence. JAMA 287:226–235.

Ericsson KA, Charness N. 1994. Expert performance. Am Psychol 49:725–747.

Eva KW. 2004. What every teacher needs to know about clinical reasoning. Med Educ 39:98–106.

Fishleder AJ, Henson LC, Hull AL. 2007. Cleveland Clinic Lerner College of Medicine: An innovative approach to medical education and the training of physician investigators. Acad Med 82:390–396.

Govaerts MJB. 2008. Educational competencies or education for professional competence? Med Educ 42:234–236.

Govaerts MJB, Schuwirth LWT, Van Der Vleuten CPM, Muijtjens AMM. 2011. Workplace-based assessment: Effects of rater expertise. Adv Health Sci Educ 16:151–165.

Govaerts MJB, Van Der Vleuten CPM, Schuwirth LWT, Muijtjens AMM. 2007. Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. Adv Health Sci educ 12:239–260.

Harden RM, Gleeson FA. 1979. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 13:41–54.

Kane MT. 2006. Validation. In: Brennan RL, editor. Educational measurement. Westport: ACE/Praeger. pp 17–64.

Klein G. 2008. Naturalistic decision making. Hum Factors 50:456–460.

Koretz D. 1998. Large-scale portfolio assessments in the US: Evidence pertaining to the quality of measurement. Assessment Educ 5:309–334.

Marewski JN, Gaissmaier W, Gigerenzer G. 2009. Good judgements do not require complex cognition. Cogn Process 11:103–121.

Martinez ME, Lipson JI. 1989. Assessment for learning. Educ Leader 47:73–75.

Messick S. 1994. The interplay of evidence and consequences in the validation of performance assessments. Educ Res 23:13–23.

Metz JCM, Pels Rijcken Van Erp Taalman Kip EH, Van De Brand Valkenburg BWM. 1994. Raamplan 1994, eindtermen van de artsenopleiding [Blueprint 1994, endgoals for the education of doctors], Nijmegen, The Netherlands: Univesitair Publicatiebureau.

Newble DI, Baxter A, Elsmlie RG. 1979. A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. Med Educ 13:263–268.

Norcini J, Blank LL, Arnold GK, Kimball HR. 1995. The Mini-CEX (Clinical Evaluation Exercise); A preliminary investigation. Ann Intern Med 123:795–799.

Norman G, Swanson D, Case S. 1996. Conceptual and methodology issues in studies comparing assessment formats, issues in comparing item formats. Teach Learn Med 8:208–216.

Norman G, Tugwell P, Feightner J, Muzzin L, Jacoby L. 1985. Knowledge and clinical problem-solving. Med Educ 19:344–356.

Painvin C, Neufeld V, Norman G, Walker I, Whelan G. 1979. The "triple jump" exercise: A structured measure of problem solving and self-directed learning. Proceedings of the Eighteenth Annual Conference on Research in Medical Education, November 6–7, Washington, DC, 18:73–77.

Plous S. 1993. The psychology of judgment and decision making. New Jersey: McGraw-Hill Inc.

Prescott L, McKinlay P, Rennie J. 2001. The development of an assessment system for dental vocational training and general professional training: A Scottish approach. Br Dent J 190:41–44.

Prescott L, Norcini J, Mckinlay P, Rennie J. 2002. Facing the challenges of competency-based assessment of postgraduate dental training: Longitudinal evaluation of performance (LEP). Med Educ 36:92–97.

Ram P. 1998. Comprehensive assessment of general practitioners. General practice. Maastricht: University of Maastricht.

Rickets C. 2009. A plea for the proper use of criterion-referenced test in medical assessment. Med Educ 53:1141–1146.

Schmidt HG, Boshuizen HP. 1993. On acquiring expertise in medicine. Special issue: European educational psychology. Educ Psychol Rev 5:205–221.

Schuwirth LWT, Muijtjens AMM, Ricketts C. The matrix unloaded: Various models to determine reliability or universe representation (submitted).

Schuwirth LWT, Southgate L, Page GG, Paget NS, Lescop JMJ, Lew SR, Wade WB, Baron-Maldonado M. 2002. When enough is enough: A conceptual basis for fair and defensible practice performance assessment. Med Educ 36:925–930.

Schuwirth LWT, Van Der Vleuten CPM. 2006. A plea for new psychometrical models in educational assessment. Med Educ 40:296–300.

Schuwirth LWT, Vleuten CPMVD, Donkers HHLM. 1996. A closer look at cueing effects in multiple-choice questions. Med Educ 30:44–49.

Shepard L. 2009. The role of assessment in a learning culture. Educ Res 29:4–14.

Shute VJ. 2008. Focus on formative feedback. Rev Educ Res 78:153–189.

Van Der Vleuten CPM. 1996. The assessment of professional competence: Developments, research and practical implications. Adv Health Sci Educ 1:41–67.

Van Der Vleuten CPM, Norman GR, De Graaf E. 1991. Pitfalls in the pursuit of objectivity: Issues of reliability. Med Educ 25:110–118.

van der Vleuten CPM, Schuwirth LWT. 2005. Assessing professional competence: From methods to programmes. Med Educ 39:309–317.

Van Merrienboer JJ, Sweller J. 2010. Cognitive load theory in health professional education: Design principles and strategies. Med Educ 44:85–93.

Ward WC. 1982. A comparison of free-response and multiple-choice forms of verbal aptitude tests. Appl Psych Meas 6:1–11.