Changing education, changing assessment, changing research?

Lambert W T Schuwirth & Cees P M van der Vleuten

BACKGROUND In medical education, assessment of medical competence and performance, important changes have taken place in the last 5 decades. These changes have affected the basic concepts in all 3 domains.

DEVELOPMENTS IN EDUCATION AND ASSESSMENT In education constructivism has provided a completely new view on how students learn best. In assessment the change from trait-orientated to competency- or role-orientated thinking has given rise to a whole range of new approaches. Certain methods of education, such as problem-based learning (PBL), and assessment, however, are often seen as almost synonymous with the underlying concepts, and one tends to forget that it is the concept that is important and that a particular method is but 1 way of using a concept. When doing this, one runs the risk of confusing means and ends, which may hamper or slow down new developments.

LESSONS FOR RESEARCH A similar problem seems to occur often in research of medical education. Here too, methods – or, rather, methodologies – are confused with research questions. This may lead to an overemphasis on research that fits well known methodologies (e.g. the randomised controlled trial) and neglect of what are sometimes even more important research questions because they do not fit well known methodologies.

CONCLUSION In this paper we advocate a return to the underlying concepts and a careful reflection of their use in various situations.

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

Correspondence: L W T Schuwirth, PO Box 616m 6200 MD Maastricht, The Netherlands. Tel: 00 31 43 388 1129; E-mail: l.schuwirth@educ.unimaas.nl.

KEYWORDS education; medical/*methods/standards; clinical competence/*standards; educational measurement/*standards; research design; reproducibility of results.

*Medical Education 2004; **38**: 805–812*
doi:10.1111/j.1365-2929.2004.01851.x

INTRODUCTION

Medical education, assessment of medical students, assessment of practising doctors and medical educational research have changed considerably during the last 5 decades. The plethora of new approaches, instruments and developments sometimes makes it difficult to see the wood for the trees. Still, there seem to be some generic elements in all these developments and some analogies which may be worthwhile exploring, especially with a view to discovering promising ways to advance and to prevent repetition of the mistakes of the past in the future. In this discussion paper we will first briefly describe some of the developments in medical education. We will then describe recent developments in assessment of competence and performance more extensively. Finally, we will end with some observations about current medical educational research. We will constantly try to highlight generic elements and analogies as we see them, and we will use them to underpin some indications and recommendations for future directions in development and research of assessment of medical competence and performance.

In 1996 van der Vleuten wrote a paper in which he suggested that the utility of assessment methods could be evaluated by looking at 5 elements: reliability, validity, educational impact, costs and acceptability of the method.[1] His main argument was that optimal utility is predicated on a carefully balanced compromise between these 5 criteria, and that it

Key learning points

Educational assessment methods are not synonymous with underlying concepts.

The underlying concepts, such as constructivism in education, role-orientated definition of medical competence and performance in assessment, are important; the methods are mere executions.

Ideally, the specific execution of a concept is chosen according to the specific demands of a situation, and not the other way round.

Applying this to research implies that the content of research questions is important, that the content and assumptions of certain methodologies should be considered, and that only after this should a specific study be set up.

would be ineffective to allow overenthusiasm to trap one into focussing on only 1 of them. This approach is important insofar as it illustrates a change in our thinking about assessment. Until that time assessment had been perceived mainly as a measurement problem, whereas currently it has come to be seen more as an issue of educational design.

In this paper, we will not explain in detail all the ins and outs of the 5 criteria, but it may be useful to first make some observations about 3 of them.

Reliability

Perhaps as a legacy of the age of the measurement view on assessment, reliability has always remained high on the agenda. But – strange though it may sound – a high Cronbach's alpha or a high g-coefficient is not a goal in itself. What is essential is the reproducibility of the decisions that are made on the basis of the results of tests. Any reliability *analysis* is a means and not an end. The results of the analysis inform teachers and the faculty that on the basis of the test results, certain decisions can and other decisions cannot be made with sufficient certainty. This means that the result of a reliability analysis should always be interpreted in relation to the actual data and in the light of the types of decisions one wants to make. The following analogy may help to clarify this. Imagine that a test is a yardstick. Then reliability would tell you whether that yardstick was able to determine the length of a beam with the precision of, say, inches or nanometres. The point is that lower reliability will necessitate more prudence in drawing conclusions from a test than higher reliability. Theoretically, one might envisage situations in which the results of an oral examination would be more reproducible than those of a multiple-choice examination. In real life, though, this will probably be a very rare occurrence. What is important is the relationship between the accuracy of the scoring and the level at which decisions are made. We should be interested not in the reliability of a test but in the reproducibility of the decisions made on the basis of the test.

Validity

Validity too must always be considered in the light of the goal of the test. The first question to ask about the validity of a test is: what is it valid for? A true/false test for factual knowledge can be much more valid than an objective structured clinical examination (OSCE) for clinical competence, depending on how the tests are set up. The main consideration must always be whether the test tests what it is purported to test, and it can always measure only 1 element of medical competence. That means that any claim that a certain test is valid as a measure of medical competence should be greeted with suspicion. Moreover, there is no such thing as *the* validity. Many different varieties of validity have been described in the literature. For our present purpose it may be helpful to distinguish 2 mainstream views in the thinking about validity. One view sees an educational test as a sort of psychological test and claims that it should be validated accordingly. This implies that validity is to be inferred from the way the scores behave in relation to theoretical notions. If the scores behave in accordance with expectations, the test is said to be valid. This view of validity is often referred to as indirect validity.[2] An example of this line of thinking is the determination of validity by checking whether experts perform better than novices.

The other view sees an educational test as a set of intrinsically meaningful and relevant assignments or items, which should be validated as such. This so-called direct validity implies that validity is to be inferred from the content of the test and not from the scores.[3] The implementation of this concept involves blueprinting, careful item review, quality control measures, etc. As usual, the truth probably lies somewhere in the middle. Establishing validity will always be a matter of gathering evidence from various sources rather than conducting a single analysis.[4,5]

Educational impact

The third element (and the last that we will discuss here) is the educational impact of assessment. We know that students are strongly influenced by assessment and we often chide them for it.[6] It would be better, however, to capitalise on this phenomenon. If we were to do so, we should first determine what type of study behaviour we want to foster. More often than not we appear to be in 2 minds about this. We say we want our students to have insight and understanding, but we are disappointed when they fail to reproduce some specialist factual detail. Only after we have decided how we want our students to learn, can we use assessment to steer their learning in the desired direction. For this purpose we can use the content and the format of the tests, but also programming and examination regulations. What influence is exerted by test content is often self-evident, but the impact of the test format may be less obvious. For instance, it would be too simple to say that open-ended questions induce more desirable study behaviour than do multiple-choice items, or that OSCEs lead to better learning compared with written tests. It all depends on what one wants to achieve, what the content of the test is, and how the tests are embedded in the assessment programme. Anyway, using a variety of methods in the examination programme is preferable to using a single type of instrument for all examinations.[7]

One of the threads running through our discussion so far is that elements of assessment methods should not be viewed as separate goals with an intrinsic, independent value. Rather, we should view them as interrelated aspects of a larger picture, whose roles are determined by what each particular assessment is intended to achieve. In addition, we have advocated that in assessment means should not be confused with goals, and methods should not be confused with concepts. We can illustrate this by pointing out that Cronbach's alpha is a measure of internal consistency and may be indicative of reproducibility of results and decisions, but it should not be taken to be synonymous with reproducibility. Ideally, one should calculate reliability, then go back to the original data and use the reliability to decide which decisions can be made and which decisions cannot be made with sufficient certainty. Too often, though, reliability is calculated and the decision about whether it is high enough or not is taken without any reference to the data. In such cases reliability is mistakenly used as a goal in itself and not as the tool that it really is.

DEVELOPMENTS IN EDUCATION

The first major change in medical education took place in the 1960s. Educational approaches were developed in which students played bigger and more active roles in acquiring medical competence. At the same time a large body of research in cognitive psychology demonstrated that the acquisition of knowledge was indeed an active process, and that learning in a relevant context enabled more efficient acquisition and better retention of knowledge.[8] These are only 2 elements of the so-called constructivist theory of learning. One of the most famous implementations of this theory is problem-based learning (PBL), which has since become extremely popular.[9,10] Unfortunately, the implementation of the theory is all too often mistaken for the theory itself, in that PBL is seen as synonymous with constructivism. Sometimes even 1 specific type of PBL, used in 1 particular medical school, is simply 'transplanted' to another school in another country, with no consideration being given to whether that specific implementation and the receiving institution are a good match for one another. Once again, this exemplifies our inclination to think in methods rather than in underlying concepts. The underlying concepts of constructivism, such as active learning, learning in context, and learning in collaboration are essential, but it is equally essential that the methods we use to put those concepts into practice are tailored to a specific setting and circumstances. Ideally, we should first make a careful analysis of the content of the curriculum and its goals and only after we have done so will we be able to make well balanced, specific choices for certain implementations. This approach would guarantee that each implementation is tailored to the specific goals of the individual parts of the curriculum.

To summarise, the method of implementing PBL is not the same as its underlying concepts, and it is the concepts that count. For a strong curriculum a variety of educational methods will probably be the most effective.

DEVELOPMENTS IN ASSESSMENT OF MEDICAL COMPETENCE

For a long time the majority of developments in assessment have lagged behind developments in teaching and learning. Not until the 1990s did developments such as authentic and integrated assessment become more popular, and more

attention was paid to student involvement in assessment.[11] These developments too have their basis in what we currently know about knowledge acquisition and knowledge application.

From trait-based to role-based theories

Today, it has become clearer that the traditional model of medical expertise as a combination of separate and generic constructs is no longer tenable. The most popular model used 'knowledge', 'skills', 'problem-solving skills' and 'attitudes'. However, these features turned out to be not generic and not as independent as was assumed initially. Repeatedly, it was found that there was more variation within an instrument (from 1 case to another, from 1 OSCE station to another) than there was between different instruments.[12] The correlation between a written test of skills (knowledge) and an OSCE station on the same skill may be higher than that between 2 OSCE stations on different skills.[13] It was also shown that it is not the method but the content that is important in determining what competence is being measured.[14,15] For example, a multiple-choice question linked to a case description which asks for a decision may provide a better assessment of clinical reasoning than a context-free, open-ended question that does not ask for a decision.[16] Perhaps concluding that there are no generic traits is overstating the case. However, it has by now become quite obvious that thinking in such generic, latent attributes has not been very successful in the development of assessment, not least because trait-orientated thinking has often culminated in a '1 instrument for 1 trait' approach. Such an approach meant that for each trait a separate instrument had to be developed, which had to be one that was superior to all other instruments. It is superfluous to say that such an instrument has never been found.

Modern developments have abandoned this type of thinking. They are founded on more integrative concepts, in which the prominent features are not traits but roles or competences. The key element in this line of thinking is that for the successful completion of a certain task or role, different aspects of medical competence have to come together, and have to be integrated. Miller's pyramid marks the beginning of this thinking.[17] The various layers are defined not as traits but as verbs or actions ('knows', 'knows how', 'shows how' and 'does') which are observable and can be judged and thus used for assessment. Current approaches go slightly further in that they define medical competence as the ability to assume a combination of well defined roles. A popular subdivision defines these roles as: provider of direct patient care, worker in the health care system, scientist, educator and a person.[18] In modern assessment programmes various instruments are used to obtain evidence about a student's competence in each of those roles. So the '1 instrument for 1 trait' approach has now become a 'multi-instrument for multiple roles' approach.

Authenticity in relation to validity

Another factor in current assessment approaches is high authenticity. Authenticity is often confused with validity, but they are different. The purpose of high authenticity has its roots in the notion of context specificity or encoding specificity of knowledge acquisition and application. This refers to the finding that people are better at reproducing and applying knowledge and skills if the context in which they have to do so resembles the context in which the knowledge and skills were first learned.[19] In this sense, the concept of authentic assessment is inseparably linked to constructivist learning theory.

We cannot stress enough that validity and authenticity are not interchangeable. There are many examples of more authentic methods (such as patient management problems) being less valid for their purpose (e.g. problem solving) than less authentic methods (such as extended matching questions or key feature problems).[20]

Integration and authenticity of assessment are not goals in themselves but ways to ensure optimal congruence between assessment on the one hand and educational goals and the demands of future practice on the other.

Involving students in their assessment

Congruence between assessment and education is also an important reason for student involvement in assessment, such as peer and self-assessment. If education is student-centred and requires students to take an active role, it would be strange if assessment were to remain the exclusive domain of the teachers. Moreover, many curricula expect their graduates to continue to assess their own learning needs after graduation, as a prerequisite for lifelong learning. The least these curricula should do is to provide their students with the opportunity to gain experience with self-assessment during their training. Finally, there is a validity argument in favour of self- and peer assessment, in that students are often better able to assess their peers on certain aspects than are the

teachers. For instance, unlike the teachers, the students have first hand experience of how well the students in their group collaborated on a certain project.

Lessons and warnings

All these new theories and concepts have led to the development of new instruments, such as portfolio assessment. The expectations of this instrument are high, but we feel that some words of warning may be in order. As so often before, we are at risk of confusing means and ends. We should remind ourselves that integration, authenticity and student involvement are not goals in themselves, but ways to achieve better congruence between assessment and education (and, we hope, practice demands). Another familiar risk is that we will once again start to think in terms of methods (the portfolio as the holy grail) rather than in terms of underlying concepts. When this happens, the danger looms that we will be attempting to prove that the portfolio is the single superior instrument that will bring the definitive solution to all our assessment problems. We should steer well clear of this trap. The really important issues concern the content of the portfolio, how we ensure that portfolio assessment is fair and honest (or reproducible and valid), and how the portfolio is embedded in the existing assessment programme. Simply using a portfolio without carefully addressing these issues will not lead to better assessment.

DEVELOPMENTS IN PERFORMANCE ASSESSMENT

Apart from developments in assessment of medical competence there is increasing interest in assessment of performance. The former can be defined as what doctors are able to demonstrate in staged and obtrusive conditions, the latter as what doctors do on a day-to-day basis in unobtrusive situations.[21]

From trait-based to role-based theories

In setting up assessment of performance we run the risk of making similar mistakes to those seen in competence assessment. For instance, goals might be set in the form of traits, such as communication skills, collaboration skills, scholarship, etc.[22] It is encouraging that there are many examples of developments where great care is taken to avoid such mistakes, such as in CANMEDS 2000[23] and in the UK General Medical Council's *Good Medical Practice*.[24]

A trait-orientated approach to performance assessment would also expose us to the risk of succumbing to the temptations of '1 instrument for 1 trait' assessment with attempts to develop *the* single instrument that will outperform all other instruments for performance measurement. Already video assessment, assessment by incognito simulated patients and by interviews have on occasion been hailed as such.

This is slightly puzzling, as we have known for some time now from research into assessment of competence that such traits are not separate and generic entities and that such a trait-orientated approach has been tried and found unsuccessful in the past. In order to avoid the pitfall of trait orientation, we would advocate that, like competence, performance should be defined in terms of roles. For the assessment of performance, information or evidence should be collected for each role by means of a variety of instruments. The roles can be similar to those used in competence assessment.

Objectivity and subjectivity versus reliability and unreliability

Assessment programmes that focus on these roles can never be solely based on the use of so-called objective or highly structured instruments. Not only will a variety of instruments be necessary to see the whole picture, but some of the instruments will yield more subjective data than others.[22] There is a widespread misconception about the relationship between subjectivity and reliability. It is often believed that subjective measures are inevitably unreliable, and that objective measures are reliable by definition. This is not true.[25] Objective measures can be unreliable and subjective measures can yield reproducible results. Suppose that we compose 10 pieces of music ourselves and that we select 10 pieces of music by Mozart. We then submit them to a panel of 10 experts who will judge the musical artistry of all 20 pieces. Finally, we ask the experts to award a prize for the best composer. In all likelihood the panel will give the prize to Mozart (probably unanimously). The outcomes of this test would be unaffected by a different selection from both our own and Mozart's music, or by a different panel of experts. The decision is generalisable to all sources of variance (and, nevertheless, highly subjective). Essential in this matter is that careful sampling has taken place, in the form of samples of our music, of Mozart's music and of panellists. In such a situation the underlying concept of reproducibility is applied and it leads to reproducible outcomes.

Planning of assessment: assessment as an educational design problem

If we apply these notions to performance assessment we have to consider sources of variance, such as methods, judges, domains, tasks, patients, contexts, timeframes, authenticity levels, educational consequences, etc. When so many elements have to be taken into account, it is wise to first devise a plan of action, preferably a document setting out the purposes of the assessment, the goals, which methods will be used and why, how sampling will take place, which quality control mechanisms are in place and how the results of various measurements will be combined.[22] Such a programme would recognise that assessment is an issue of instructional or educational design rather than a mere measurement problem. It would also recognise that there is no 1 single instrument that can provide the whole picture, but that a combination of instruments is needed. In addition, it acknowledges that good psychometric characteristics do not come from overstructuring and overemphasising 1 instrument, but from a carefully composed combination of methods. Finally, such a programme would ensure the best possible connection between assessment, education and practice.

DEVELOPMENTS IN RESEARCH IN ASSESSMENT

It appears as if thinking in assessment research has not yet caught up with the thinking about assessment methods. We often see that more traditional methodologies, which were perfect for more traditional assessment methods, are now being used for more modern assessment methods. One might wonder, for example, whether we should really investigate the reproducibility of portfolio assessment using, for instance, a standard ($P \times J$) generalisability analysis. In addition, most research is mainly aimed at construct validity and reliability from a quantitative angle. This is not wrong in itself. What is unfortunate, however, is that far less research is theory building, focuses on direct validity, on educational impact, or uses a more qualitative angle. Especially lacking is research into assessment programmes. Studies (including some research into portfolios) are too often still aimed at single instruments, trying to prove a particular instrument to be a holy grail. Sometimes methodologies of questionable suitability for the purpose are used, such as a randomised controlled trial for a comparison between curricula.[26] In our opinion, it is important not to think in terms of methods but in terms of concepts or research questions when setting up assessment research. This means that in designing a study one should not start from the methodology ('Let's do a randomised controlled trial'), but from the research questions ('Does active learning lead to better retention of knowledge than passive learning?').

Thinking in terms of underlying concepts has – in our opinion – some implications for medical educational research. A first implication is that higher statistical power is not by definition better. For some research questions a series of small replications may be preferable, especially if these replications can be performed in different centres with comparable contexts or even with different contexts. In the medical education community there is a sufficiently open attitude towards collaboration – both on national and international levels – to make such studies possible. These studies would also ensure that the results are useful in a framework broader than the individual context of the authors. Too often studies are performed comparing a psychometric parameter of local instrument X to that of local instrument Y. Such a study is often only locally relevant. A second implication concerns the use and application of measurement instruments. The choice and construction of the instruments must be based on the specific research question. Too often, though, the development of the instrument is the tail that wags the dog, and not enough care of its development and validation is taken. Simply transplanting an instrument that has proven validity within another context to one's own study is not advisable. Issues such as local usefulness, validity, production and quality control have to be addressed. Therefore, not only should samples of the instrument be provided, but also the question why this particular instrument is the best method for this particular study in this particular context must be answered. A final implication is that we should not be afraid to develop our own methodologies, if they are defensible, effective and fit the concepts of our research questions.

In summary, we want to advocate the following approaches for future research agenda:

- thinking in terms of research questions rather than in terms of methodologies;
- embedding the research topics in the existing literature and then deciding exactly what research question to formulate;

- conducting studies to explore the value or utility of assessment programmes rather than the value or utility of individual instruments, and
- accepting that a single study is never going to provide the definitive answer to a big question, and that for this a collation of results from various – perhaps smaller – studies will be needed.

CONCLUSION

In all developments in education, assessment and research there is a tendency to sometimes lose track of the general concepts or the bigger picture. Successful methods are often promoted to the status of holy grails that will resolve all our problems. Such an approach, however, is not very helpful in the advance of education, assessment and research. Therefore, we want to summarise our observations in some caveats. A method is only a realisation of an underlying concept and it is the concept that is important. The method is not synonymous with the concept; there are often various ways to implement a concept. Rather than adopting a method that has been successful in a certain situation, one should adopt its underlying concepts and translate them to fit the unique demands of the local situation.

In addition, a single method will never solve the whole problem, not in education, nor assessment, nor research. On the contrary, strength will come from a carefully balanced combination of methods involving educational programmes, assessment programmes and research programmes.

CONTRIBUTORS

LWTS wrote the first draft of this paper. CPMvdV commented. The final version is the result of agreement between both authors.

ACKNOWLEDGEMENT

The content of this paper is a summary of a keynote address presented at the 2003 Association for the Study of Medical Education Conference on Medical Education, Edinburgh.

FUNDING

None.

ETHICAL APPROVAL

Ethical approval was not required for this study.

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 Cronbach LJ. What price simplicity? *Educational Measurement: Issues Pract* 1983;**2** (2):11–2.
3 Ebel RL. The practical validation of tests of ability. *Educational Measurement: Issues Pract* 1983;**2** (2):7–10.
4 Flanagan JC. A rational rationale. *Educational Measurement: Issues Pract* 1983;**2** (2):12.
5 Gardner EF. Intrinsic rational validity: necessary but not sufficient. *Educational Measurement: Issues Pract* 1983;**2** (2):13.
6 Frederiksen N. The real test bias: influences of testing on teaching and learning. *Am Psychol* 1984;**39** (3):193–202.
7 Van der Vleuten CPM, Scherpbier AJJA, Dolmans DHJM, Schuwirth LWT, Verwijnen GM, Wolfhagen HAP. Clerkship assessment assessed. *Med Teacher* 2000;**22** (6):592–600.
8 Schmidt HG. Foundations of problem-based learning: some explanatory notes. *Med Educ* 1993;**27** (5):422–32.
9 Albanese MA, Mitchell S. Problem-based learning: a review of literature on its outcomes and implementation issues. *Acad Med* 1993;**68** (1):52–81.
10 Vernon DT, Blake RL. Does problem-based learning work? A meta-analysis of evaluative research. *Acad Med* 1993;**68** (7):550–63.
11 Boud D. Assessment and the promotion of academic values. *Studies Higher Educ* 1990;**15** (1):101–11.
12 Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet* 2001;**357**:945–9.
13 Van der Vleuten CPM, Van Luyk SJ, Beckers HJM. A written test as an alternative to performance testing. *Med Educ* 1988;**22**:97–107.
14 Ward WC. A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Appl Psychol Measurement* 1982;**6** (1):1–11.
15 Norman GR, Smith EKM, Powles AC, Rooney PJ, Henry NL, Dodd PE. Factors underlying performance on written tests of knowledge. *Med Educ* 1987;**21**:297–304.
16 Schuwirth LWT, Verheggen MM, Van der Vleuten CPM, Boshuizen HPA, Dinant GJ. Validation of short case-based testing using a cognitive psychological methodology. *Med Educ* 2000;**35**:348–56.
17 Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;**65** (9):63–7.
18 Hays R, Davies H, Beard J *et al.* Selecting performance assessment methods. *Med Educ* 2002;**36**:910–7.
19 Regehr G, Norman GR. Issues in cognitive psychology: implications for professional education. *Acad Med* 1996;**71** (9):988–1001.

20 Schuwirth LWT. *An Approach to the Assessment of Medical Problem Solving: Computerised Case-Based Testing*. Maastricht: University of Maastricht 1998.
21 Rethans J, Norcini J, Báron-Maldonado M *et al*. The relationship between competence and performance: implications for assessing practice performance. *Med Educ* 2002;**36**:901–909.
22 Schuwirth LWT, Southgate L, Page GG *et al*. When enough is enough: a conceptual basis for fair and defensible practice performance assessment. *Med Educ* 2002;**36**:925–30.
23 CANMEDS. Skills for the new millenium CANMEDS 2000 project. 2000. http://www.rcpsc.medical.org/publications/index.php. [Accessed October 2003.]
24 General Medical Council. Good *Medical Practice*. 2001. http://www.gmc-uk.org/med_ed/default.htm. [Accessed October 2003.]
25 Van der Vleuten CPM, Norman GR, De Graaf E. Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ* 1991;**25**:110–8.
26 Norman G. RCT = results confounded and trivial: the perils of grand educational experiments. *Med Educ* 2003;**37**:582–4.

Received 27 November 2003; editorial comments to authors 23 December 2003; accepted for publication 26 January 2004