

Clerkship assessment assessed

C.P.M. VAN DER VLEUTEN, A.J.J.A. SCHERPBIER, D.H.J.M. DOLMANS,
L.W.T. SCHUWIRTH, G.M. VERWIJNEN & H.A.P. WOLFHAGEN

University of Maastricht, The Netherlands

SUMMARY *This article reviews consistent research findings concerning the assessment of clinical competence during the clerkship phase of the undergraduate medical training programme on issues of reliability, validity, effect on training programme and learning behaviour, acceptability and costs. Subsequently, research findings on the clinical clerkship as a learning environment are discussed demonstrating that the clinical attachment provides a rather unstructured educational framework. Five fundamental questions (why, what, when, how, who) are addressed to generate general suggestions for improving assessment on the basis of the evidence on assessment and clinical training. Good assessment requires a thoughtful compromise between what is achievable and what is ideal. It is argued that educational effects are eminently important in this compromise, particularly in the unstructured clinical setting. Maximizing educational effects can be achieved in combination with improvements of other measurement qualities of the assessment. Two concrete examples are provided to illustrate the recommended assessment strategies.*

Introduction

A substantial part of learning medicine is to engage in real-life professional activities. Traditionally, apprenticeship learning has been relevant for the medical profession. Any undergraduate training programme will end with an extensive period of clinical rotations. By participating in the clinical care, by observing the clinician, and occasionally through formal and informal (bedside) teaching, students will build their clinical competence. Postgraduate programmes usually rely almost completely on the apprenticeship model. In a typical undergraduate programme students rotate through their clerkships in varying or fixed order depending on the medical school. At the end of each clerkship, students' performance is assessed. However, usually this assessment is not unproblematic (Streiner, 1995). Clerkship assessment typically relies on rather global evaluations from quite unstandardized testing situations and is often based on limited samples of (real-life) students' clinical behaviour. However, in the past few decades we have witnessed many developments in the field of assessment. Existing test methods have been investigated, and new methods have been proposed and investigated. By now there is a considerable body of literature on assessment in medical education. This article is intended as a contribution to the current debate on assessment of clinical competence during clerkships in the undergraduate medical curriculum. In it we offer some suggestions on how the available knowledge about assessment might be used to improve clerkship assessment.

Since assessment and education are inextricably linked, we also address education during the clerkships. The article

is divided into four parts. First, we give a brief overview summarizing the research consistencies found in studies on assessment of clinical competence. Subsequently, some studies on the didactic qualities of clinical clerkships are discussed. These results are the basis for general recommendations with respect to clerkship assessment. Finally, two examples illustrate how these suggestions might be translated into medical educational practice.

Outcomes of studies on assessment

Five characteristics can be used to evaluate achievement tests: reliability, validity, effect on training programme and learning behaviour, acceptability, and costs (Morgan & Irby, 1978, Van der Vleuten, 1996a). Each of these characteristics is addressed separately and general research consistencies are summarized. In selecting assessment methods to evaluate students' performance during clerkships all five characteristics should be considered.

Reliability

Reliability is concerned with the reproducibility of test results. Reproducibility can be impaired due to various sources of interference. A great many studies on reliability have been published. The general conclusion to emerge from these studies is that performance samples must be large to counteract interference. The main source of interference in all test formats is that competence measurement is situation-specific. When a patient contact is used to assess a student's competence in clinical problem solving, the resulting score appears to have little predictive value for the same student's score on a different patient problem. One sample of performance, whether it is with a real patient, a multiple-choice item, an essay or an OSCE station, consistently correlates poorly with another sample of performance. Competence varies considerably across content areas, independently of the test format (pencil-and-paper, oral, observation). We generally prefer achievement tests to make judgements about someone's competence irrespective of the specific test content used. For example, which patient is used in a long-case examination depends on chance: it could have been any other patient. The *content specificity problem* (Elstein *et al.*, 1978) requires any assessment to use a large sample of observations (items, patients, stations, essays, etc.) to arrive at judgements that are not content-dependent. As an inevitable consequence, assessment procedures are necessarily lengthy. The sample size required depends on the test

Correspondence: C.P.M. van der Vleuten, Department of Educational Development and Research, PO Box 161, 6200 MD Maastricht, The Netherlands. Tel: +31 43 3881111; fax: +31 43 3628799; email: vandervleuten@educ.unimaas.nl

format. Multiple-choice items cover a domain in a shorter time than does an observed structured clinical examination (OSCE). For an OSCE to attain an acceptable level of reliability, several hours of testing time are needed (Swanson *et al.*, 1987, Van der Vleuten *et al.*, 1994). Another source of variability concern the assessors (correctors, examiners, observers). For them the same applies as for content: as their subjective influence increases, so must the sample; that is, more assessors must participate in the assessment. The greater the number of different assessors, the better the test's reliability. Even with rather subjective assessments, the assessors' influence can be neutralized by an efficient assessment design, for instance by having different assessors for different parts of the test (Van der Vleuten *et al.*, 1991). Assessment guidelines, such as model answers or lists of criteria, will help to neutralize assessor effects to some extent, but a large sample of assessors is usually still required. Global or holistic assessments do not necessarily yield unreliable data. They turn out to be reliable when concrete behaviour is judged in a specific concrete clinical context, such as in an OSCE station, or when a student performs a physical examination in the clinic (Cunnington *et al.*, 1997; Regehr *et al.*, 1998). However, global ratings of general and broad aspects of behaviour or competence over a prolonged period of time, as is the case with most clerkship ratings, are generally unreliable (Streiner, 1995).

Validity

For a test instrument to be valid it must evaluate what it is intended to evaluate. This presupposes that we know what we want to measure. There is, however, no generally accepted definition of medical competence. The questions 'what is competence?', 'what constitutes medical expertise?', have led to quite a few fundamental cognitive psychological studies (Schmidt *et al.*, 1990; Regehr & Norman, 1996).

In the past, definitions of competence were developed from the perspective that competence could be broken down into different attributes. Separate test methods were developed for each of these attributes. For example, it was assumed that problem solving and knowledge were distinct entities that could be tested separately. The attributes or skills were conceived to be quite generic. The skill of problem solving was considered to be general: the more a person possesses this skill, the more clinical situations he/she is able to handle.

It has been demonstrated, however, that such generic skills are virtually non-existent. The content specificity problem, described above, is a clear illustration. Knowledge and skills have been found to be highly dependent on specific content—or context. Similarly, separate attributes of competence have turned out to be hard to define. Even a relatively simple definition, such as the popular categorization of competences into knowledge, skills and attitude, proves problematic when attempts are made to develop mutually exclusive definitions. Competences appear to share many aspects. It has indeed been proven that seemingly different competences are not independent of each other. Assessments of different aspects of competence are found to be strongly correlated (if reliably assessed). Knowledge is an essential factor in all competences. The way knowledge is structured, its interrelationships, and the extent of 'automa-

tion' change with the building of medical expertise (Schmidt *et al.*, 1990).

There are nevertheless a few general insights into the validity of measurement instruments that can be derived from research results (Miller, 1976; Newble, 1976; Maatsch, 1978; Ward, 1982; McGuire, 1987; Elstein, 1993; Case, 1997). First, the test's content rather than the test method primarily determines what is being measured. Multiple-choice questions can also test problem-solving competence, while an oral examination can also test factual knowledge. What is being tested depends mainly on the tasks to which a candidate is exposed and less on the 'wrapping' of these tasks. This argument should not be reversed by assuming that format is of no consequence whatsoever. However, in practice we tend to overrate the significance of the format or method of assessment and some teachers are married to certain methods because of their, supposedly, inherent measurement potentials. The research has not yielded any evidence to support unique characteristics of certain measurement methods to assess certain competences. What is put *in* the method is more important than the method itself.

A second outcome is that, irrespective of what is being assessed, no single assessment format is ideal. To achieve a valid overall picture of clinical competence, a variety of different test methods is needed.

When clinical competence is split up into (numerous) attributes, and different methods are used for each attribute, an atomistic approach may be the result. At the end of a curriculum, such as with clerkship assessment in particular, we want to assess the integrated performance of clinical tasks. Splitting up tasks into different attributes and testing these separately will harm the authenticity of the assessment.

Effects on training programme and learning behaviour

The findings on how assessment affects students' learning behaviour are unequivocal: learning is primarily test-driven (Miller, 1976; Newble & Jaeger, 1983; Frederiksen, 1984; Messick, 1994). Tests can drive learning by their content, format and programming (Van der Vleuten, 1996a). Tests and examinations determine students' study success. Students tend to adopt learning behaviour that achieves the best results with maximum efficiency. Any strategy that appears to offer the best chance of success is likely to be adopted, whether it entails memorizing facts, knowing everything about one subject because it is the examiner's favourite topic, or neglecting all other educational activities to study for next week's test. For students, the real curriculum *is* the examination programme. If there is a mismatch between the objectives of the curriculum and the assessment programme, student learning will first and foremost be test driven.

The obvious conclusion should be that this relationship must be used instrumentally, that is tests should be used explicitly to steer student learning in the desired direction. However logical this may be, it nevertheless proves hard to implement, since effects on student learning are not entirely predictable. For instance, students' performances on an OSCE can be assessed using detailed checklists of criteria with appropriate operationalizations to provide adequate feedback to students. However, if students merely memorize

the criteria to maximize their score, this test format will be detrimental to their understanding (Van Luijk *et al.*, 1990). This example from educational practice demonstrates that even strategies that are considered educationally sound can have unexpected adverse effects. We should draw the conclusion that the strategic use of tests should be accompanied by permanent monitoring of how tests affect student learning. Students study for tests, but they usually learn little from them, because the educational value of tests is being neglected in educational practice. Educational programmes tend to emphasize the selective function of assessment methods, which tends to detract from their educational value. For instance, if a student is informed only whether he or she has passed or failed at the end of an educational unit (course, clerkship, block, etc.), the student will remain ignorant of his or her strengths and weaknesses and also of areas that require extra work. The feedback value of tests need not be limited to students, but should be extended to teachers. The results of tests indicate to what extent students are attaining the educational objectives. A multiple-choice item that yields many incorrect answers may be indicative of that particular subject being addressed inadequately during the course or not being described adequately in textbooks. Thus, test results can be used as a quality assurance tool (Downing & Haladyna, 1997; Verhoeven *et al.*, 1999). In day-to-day practice this use of tests proves to be very limited.

Acceptability

Education in general, and assessment in particular, is strongly determined by the expertise and experience of those involved. Teachers hold strong personal opinions on testing, fostered by their own educational history and experience. These opinions tend to be rather naive. Many assumed truths, such as what is being measured and the degree of certainty with which judgements can be made, may in reality prove to be less than true or even downright mistaken. Research outcomes that demonstrate the opposite of such naive intuitions are often not accepted (at least not immediately). As a result, discussions on assessment are dominated by tradition (and intuition) rather than research outcomes (Van der Vleuten, 1996b).

Independently of whether this is justifiable, opinions, feelings and tradition determine to what extent an assessment method is accepted by both teachers and students. Thus, acceptability is an important test characteristic on a par with other, more objective ones. It needs to be taken into account in designing and implementing assessment methods. Nevertheless, we do not believe that resistance to change is insurmountable. It is definitely possible to promote change by combating ignorance. Education, training, and staff development are prerequisites for change.

Costs

It is evident that good assessment is expensive. Constructing good tests is by no means easy and requires considerable time and energy, irrespective of the method used. Objections raised against these costs are often based on insufficient insight into the costs of educational development and the role of assessment in the educational programme. Quality assurance, with format-related, content-related and

statistical procedures as integral parts, is essential to the quality of assessment, but not very common in most medical schools (Downing & Haladyna, 1997). Such procedures are generally better left to central test committees than to individual departments. The costs of using central test committees may seem high compared with those for a decentralized test organization, but only because the latter are usually not explicitly calculated. The quality of assessment depends on investment in test development and quality assurance procedures.

Assessment as compromise

It will be evident from the above that it is impossible for a test to fully meet all demands related to all five characteristics. It is even more evident that perfection on all characteristics is impossible to achieve. Improvement of the quality of one characteristic usually goes at the expense of the quality of another characteristic. For instance, resources are generally limited, which means that test-related costs reduce the budget available for other activities. In making decisions on assessment, one needs to weigh the importance of the five characteristics described above against each other. No single characteristic can be neglected. What compromise is reached depends on the context in which assessment takes place and will therefore vary from situation to situation. Depending on the particular situation the weights assigned to each characteristic will vary. For instance, if the test result has far-reaching consequences, such as that of a licensing examination, reliability will weigh heavily. It will be clear that reaching a compromise is no predetermined process. It is literally a matter of balancing the impact and defensibility of specific choices given a specific educational and assessment context.

Research findings on clerkships

The clinical clerkships in undergraduate medical training are modelled after the old apprenticeship model. Working and following the master's example, the apprentice experiences a learning situation that cannot be offered by any other educational format. Despite participation in real work, the emphasis in the clerkships is placed firmly on learning. This changes after graduation. In postgraduate programmes the emphasis shifts to work. Despite general agreement on the importance of clerkship education, little appears to be known about how students learn during clerkships (Jolly, 1994). The straightforward master-apprentice relationship is actually hard to identify in day-to-day clinical practice. Students find themselves confronted with a wide range of medical professionals: house officers, registrars, consultants, and paramedical workers, fellow students, and patients, in the context of a strict formal and informal hierarchy inherent in a complex health care organization, with expectations that more often than not remain implicit. In addition, all these aspects tend to vary from place to place and from clerkship to clerkship. Little is known about how students actually learn, and which aspects make an effective contribution to students' competence. From an educational perspective, clerkship training is for the most part a black box. However, the clerkship as an educational environment is gradually receiving more attention. Some important findings have emerged from studies on clerkships:

- There are far fewer patient contacts than is generally assumed (Schamroth & Haines, 1992; Dolmans *et al.*, 1999).
- Considerable time is spent on activities with little educational value (Cook *et al.*, 1992; Schamroth & Haines, 1992).
- Individual students differ substantially with regard to the activities they perform (Friedman *et al.*, 1978; McKerwgow *et al.*, 1991; Cook *et al.*, 1992; Tortolani *et al.*, 1997).
- Students have educational contacts mainly with house officers and registrars and fellow students and to a lesser degree with senior staff (Remmen *et al.*, 1998).
- The nature and quality of supervision shows great variability (Martens *et al.*, 1999).
- Students are only rarely observed during patient contacts (Ende, 1983; Kernan & O'Gonnor, 1997; Szenas, 1997; Dolmans *et al.*, 1999).
- The average level of expectancy on reaching the learning objectives at the end of a clerkship differs only marginally from the expectations at the beginning of the clerkship (Scherpbier, 1997; Tortolani *et al.*, 1997). Probably the entrance expectations are too high. Moreover, there is little consensus on the clerkship objectives (Scherpbier, 1997).
- A wide gap is experienced between the previous theoretical training and clerkship teaching (Boshuizen, 1996; Prince *et al.*, 1999). In making the transition from theoretical education to the clerkships students appear to experience a huge chasm between these two curricular phases. The shock of practice is a phrase that is being used. The changed context results in students forgetting (or being unable to apply) what they already know, being unable to understand clinicians' language, and needing considerable time before they can start to learn anything at all. Knowledge often appears to be constructed inadequately: a patient presents with complaints and symptoms, whereas most textbooks focus on diagnoses (Prince *et al.*, 1999). The change is also characterized by a (in some cases difficult) socialization process. And finally, all these aspects show large variations across individual students.

Although it is unclear how students learn during clerkships, it should be said that there is no evidence that the above-mentioned findings constitute a barrier to effective learning. However, it seems safe to assume that clinical clerkship is a relatively unstructured educational format. In light of the above findings and the importance of clerkships to medical education, improvements can and must be made.

Clerkship assessment

Bearing in mind the above evidence about assessment and medical training in clerkships, we propose a number of recommendations with respect to clerkship assessment. We will structure these by addressing five fundamental questions to be posed concerning assessment (Harden, 1979).

Why?

"Why" is asked to determine the function of assessment. Do we assess in order to take decisions over student promotion (the selective function), to provide feedback to the

student (the formative function) or to monitor the quality of the educational programme (the accountability function)?

Assessment in clinical clerkships typically concerns the final stage of undergraduate medical training. Thus, selection of students should play no or only a very minor role in the assessment. Its proper place is at an earlier stage in the curriculum. The emphasis in clerkship assessment must be on the formative role of assessment (Bloom *et al.*, 1971; Miller, 1976; Leeder *et al.*, 1979; Ende, 1983; Rolfe & McPherson, 1995).

By saying so, we explicitly argue against the most common use of assessment. It is often proclaimed that the clinical situation is the ideal situation to decide whether a student is fit to be a doctor. Irrespective of the high demands that such a decision places on the quality of the assessment tool, we consider it ethically unacceptable to postpone such a far-reaching decision until the end of the medical curriculum. Those in favour of selection of students at this stage argue that students have no or only limited clinical experience before this stage, and that attitudinal aspects in particular can be assessed only in clinical practice. Apart from the complexity of measuring attitude, we believe that this too should have been done earlier in the curriculum. However, the problem of doing so is of another nature. Most undergraduate medical curricula are divided into two relatively distinct parts: the theoretical phase (the first years) and the clinical phase (the last years). This resembles a so-called 'H-model'. Although there are exceptions in some innovative schools and in some countries, the H-model is rather strictly used in many undergraduate medical training programmes. It is indeed not surprising that students experience their entrance into the second phase as a dramatic change rather than a gradual transition. Better integration of theory and practice would ease the transition. However, this would entail real patient contacts much earlier in the curriculum. In such a curriculum, more resembling a 'Z-model', clinical contacts would be offered from the very start of the undergraduate programme, gradually increasing as theoretical education decreases over the course of the curriculum. A curricular change that would introduce clinical contacts at a much earlier stage would offer excellent opportunities for earlier selection decisions.

Emphasizing the formative value of assessment during clerkships becomes even more relevant against the background of the lack of educational structure of clerkships as shown above. Feedback from tests might offer students better guidance. The problem that students experience in applying theoretical knowledge demonstrates the crucial importance of feedback. Students explicitly ask for more direct observation and feedback regarding their activities. This is fully compatible with a more formative approach to clerkship assessment.

The clerkship as a teaching programme also benefits from more emphasis on the educational or formative impact of clerkship assessment. If we formulate clear clerkship objectives and assess students in relation to these expectations, shortcomings in the curriculum can be identified and anticipated and/or remedied (Wijnen, 1981; Miller, 1976). In this way assessment stimulates improvement of the quality of clerkship from a didactic perspective.

The emphasis on the formative value of clerkship assessment also allows us greater latitude for reaching a compromise between the demands of the other assessment characteristics. In a more selective approach one cannot compromise too much on the psychometric qualities of the assessment, such as reliability and validity. It can actually easily be demonstrated that the psychometric quality of most of our within-school assessments tends to be substandard (Van der Vleuten *et al.*, 1994).

What?

In discussing validity, we stated that medical competence is not easy to define. We propose a pragmatic approach involving a definition of competence not so much in terms of student abilities, but in terms of *medical content* relevant to the objectives of the clerkship (Newble *et al.*, 1994). Rather than speaking about a student's clinical reasoning ability, we would prefer to consider the student's ability to manage independently patients with a range of complaints related to a multidisciplinary series of diseases/diagnosis. This can be achieved by determining more systematically which medical problems a student should encounter during clerkship rotations and at what level a student should be able to deal with them. Consensus on the objectives is not easy but by no means impossible to achieve (Martens *et al.*, 1997). In this approach, the debate would be one on medical content conducted by medical professionals rather than on inevitably fuzzy definitions of the abilities underlying clinical competence. The literature offers a number of examples, for instance from family medicine, neurology and nephrology (Mancall *et al.*, 1987; Martens *et al.*, 1997). The benefit to be gained from outlining the content of clerkships lies in the provision of guidelines to both students and teachers, which would enhance the educational structure of clerkships.

In the debate on clerkship assessments considerable emphasis is often placed on the evaluation of attitude. Attitude is a complicated concept. Psychologists define attitude as the propensity for a certain behaviour. It has an inside (a latent person variable) and an outside (the resulting behaviour) (Batenburg, 1997). The inside is difficult to measure and consequently untestable. The resulting behaviour can be observed and therefore assessed. Communication aspects or interpersonal behaviour are relatively easy to assess using the many instruments that are available. More global, holistic assessments may be useful, however, but only when the assessment is anchored in concrete (clinical) situations. The same rule applies again: minimize 'psychologizing' and maximize stating concrete and clinically meaningful actions. 'The attitude is not sufficiently patient oriented' is less meaningful than 'the student fails to explain to the patient what physical examination he or she is going to perform'. Caution should always be exercised in being too normative in an approach. Differences of opinion concerning the aspects of an 'appropriate' attitude are numerous, even when these aspects are translated into concrete medical behaviour. In addition, the 'masters' attitude', i.e. that of the clinicians, is not always entirely exemplary, and we should beware of forcing students to adopt 'our' attitudes through the assessment. Particularly in the area of attitude assessment, it is even more important that formative assessment should pave the way for a compromise on the quality of the other characteristics of

assessment. To our knowledge, there is no existing instrument or format to test attitude that meets rigorous psychometric requirements.

A clarifying and useful perspective on clinical competence has been proposed by Miller (Miller, 1990). He described a pyramid with various layers, as shown in Figure 1. The first layer at the base of the competence pyramid consists of knowledge or facts (knows). One level higher we find the ability to apply knowledge in concrete situations (knows how). This means that knowledge can be activated and used to tackle problems, follow arguments, etc., but only at the cognitive level. The next level up represents the ability to use this knowledge to perform concrete actions (shows how). At the top of the pyramid we find performance in day-to-day practice, which is the final objective of competence (does).

In the literature a distinction is also made between competence, i.e. what a person can do, and performance, i.e. what a person actually does (Rethans *et al.*, 1990). Thus, competence refers to the first three layers of the pyramid, and performance to the top. The relevance of Miller's pyramid to the discussion on clerkship assessment is that the pyramid illustrates that clerkship assessment should be geared to the top rather than to the lower layers of the pyramid, seeing that the focus of clerkship learning is the application of knowledge and skills in clinical practice. The assessment should reflect this. Most of our assessment instruments are geared towards the lower layers, even in clerkship assessment, or, if directed at higher layers, their quality is poor.

When?

The usual approach to assessment programming is planning a test at the end of a course or programme. From a summative perspective—have students achieved the objectives—this approach appears to be logical. From a formative perspective, however, the logic is less obvious. Students are hardly able to remedy their deficiencies at the end of a course and feedback at this stage is of little use (particularly if the results are limited to a mere pass or fail). Programming tests midway through a clerkship or at the start, or two tests, one at the beginning and one at the end of a clerkship (longitudinal testing), would be more logical. Generally, information on a student's performance during a

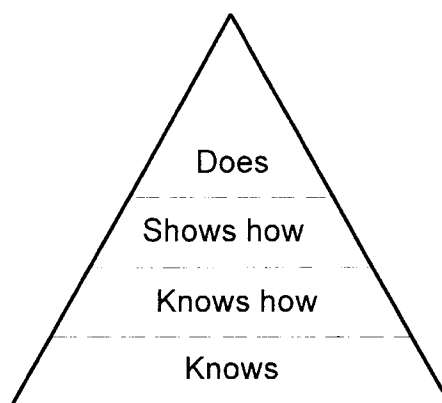


Figure 1. Miller's pyramid of competence.

specific clerkship is only available within the department or clerkship concerned. It is not customary for information to be passed on to the departments of subsequent clerkship rotations. Again, from a feedback perspective, this does not make sense. In summary, if clerkship assessment is regarded as formative testing, more extensive programming should be used instead of the conventional tests at the end of every clerkship and information should be carried through longitudinally. More frequent assessment (i.e. longitudinal assessment) will also better reveal defects in the clerkships, thereby enabling adjustment of the programme (Miller, 1976; Wijnen, 1981; Ende, 1983).

How?

This question addresses the assessment format or method. It is questionable whether the most commonly used formats really do assess what we want to assess during clerkships. We would like to test the 'does' of Miller's pyramid and it is open to question as to whether this is the case in most current training programmes. Oral and paper-and-pencil tests are suitable for assessing the first two layers of the pyramid. The popular patient-based examination, i.e. in a long or short case, is intended to test the 'shows how', but in many instances there is no direct observation of the patient examination, which limits this test method to the 'knows how'. With regard to the top of the pyramid, the assessment is usually restricted to a general evaluation of a student's functioning by a clerkship supervisor, but this is hardly specific (has little feedback value) and is mostly based on second- or third-hand information rather than on direct observation of the student (which is rare) (Ende, 1983). Assessment is limited to a general impression of the student, which we know to be unreliable and uninformative. It is not surprising that in practice these evaluations yield little differentiation.

The popular objective structured clinical examination (OSCE) uses direct observation of students in simulated clinical situations under standardized conditions (Harden & Gleeson, 1979). The assessment, however, remains limited to the 'shows how'. In conclusion, there is a wide gap between the educational objectives and the assessment of clerkship training. Considering the systematic relationship between learning and assessment, we have every reason for concern in this respect.

What we know from studies on assessment and clerkships suggests various ways in which we could improve the situation. From the perspective of Miller's pyramid, we might conclude that different assessment formats are needed, covering the different layers of the pyramid. There have been studies of formats for evaluating the top of the pyramid in particular, such as longitudinal videotaping of patient contacts and hidden simulated patients, but as yet these appear difficult to implement because of logistic and financial constraints (Rethans *et al.*, 1991; Ram, 1999). This leads to the inevitability of more direct observation of students during clinical work to assess clerkship performance (Ende, 1983). This approach would have the additional advantage of meeting students' complaints that they experience direct observation only rarely.

A second prerequisite is a large sample of student performances. The traditional long-case examination involving one single patient contact fails to meet this criterion

and is thus unreliable, even if all examiner effects were to be eliminated. As evaluations in a direct observation situation are more subjective, the sample will have to be even larger. The unreliability of using (inevitably) rather unstandardized direct observation on the ward or in the outpatient clinic as a basis for assessment will have to be compensated for by a larger sample of observations. In this way the errors per observation will be spread across the total assessment. A large sample will also enhance the formative value (more feedback) of tests in addition to offering increased reliability.

Whichever assessment format is used, some form of quality assurance is essential. This might be done by monitoring (the construction of) test material, with regard to both content and statistical aspects, and by regular evaluation of the test's educational effects, thus enabling adjustment if needed.

Who?

Three different considerations need to be taken into account in answering the question of who should be involved in clerkship assessment.

In the first place all partners involved in clerkship training should be involved in assessment. Nurses are likely to be better judges of infusion skills during clinical contacts than consultants. Other people can contribute to assessment besides the formal clinical teacher.

Second, it is advisable to separate the roles of teacher and assessor as much as possible in assessments of a predominantly summative nature. Combining the two roles causes problems to both student and teacher. The student's relationship with the teacher will change if the latter is also judging him or her at the end of a clerkship. Conversely, it is difficult for the teacher to be an objective judge when he or she has been directly involved in 'producing' the outcome.

Thirdly, there is a need for more collaboration in assessment. Two aspects are relevant. First and foremost, a systematic and planned approach to clerkship assessment requires better coordination and central planning—even more so when multidisciplinary and longitudinal aspects are incorporated into the assessment process. This will inevitably have consequences for the organization. Furthermore, collaboration makes sense from a cost perspective. Each discipline in each medical school designing its own different tests for its own educational programme means a substantial loss of capital. Developing adequate test procedures is costly and there are no valid arguments against collaboration and coproduction. Between-school exchange or collaboration would achieve substantive cost reductions, besides which, all parties involved would learn from the experience.

Illustrations of clerkship assessment

So far we have given rather general directions for clerkship assessment on the basis of the five basic questions. We will now present two concrete illustrations of how these recommendations could be implemented. The examples illustrate only two possible ways out of a myriad of possibilities to implement the recommendations.

A surgical examination

The surgical clerkship covers eight weeks with new students entering the clerkship every four weeks. The clerkship can be done in the academic hospital or in an affiliated hospital. The general objective of the clerkship is to prepare students for independent patient contacts on the ward or in the outpatient clinic. This didactic mini-concept of the surgical clerkship should be reflected in the assessment procedure. The specific clerkship objectives are specified in a list of patient problems and skills included in a logbook issued to each student at the start of the clerkship. The logbook contains information about the problems students must encounter during the clerkship and also indicates the level of mastery.

Figure 2 gives an overview of the assessment programme. An entrance test is given in the first week. A multiple-choice test is administered to test relevant knowledge on anatomy and physiology as well as knowledge taught in the short introductory surgical course earlier in the preclinical curriculum. The test aims to refresh knowledge. Students are advised to study the literature, depending on their scores on the various domains. Test items being reusable, there is no need to construct new test items (the questions are not kept secret) and the entrance test is relatively cheap. It does not contribute to the final assessment.

Halfway through the clerkship students take a mock OSCE. Students tend to find an OSCE quite stressful. To prepare for the OSCE, the students participate in the OSCE that is the final surgical assessment of the preceding group of clerks. The OSCE is organized in the academic hospital. It consists of four 15-minute stations: three with integral (simulated) patient contacts and one using models to test skills such as suturing or infusion. Clinical staff of both the academic hospital and the affiliated hospitals use checklists to rate students' performances. To stimulate feedback, the last few minutes of each station are set aside for feedback by the examiner. After the OSCE one of the surgeons conducts a debriefing, which also covers students' general experiences in the surgical clerkship. The OSCE requires considerable staff effort and resources.

In the clerkship's final week students take a computerized test (Schuwirth *et al.*, 1996). Students can log onto one of the available computers whenever they want to take the test. The test comprises a series of short realistic presentations of surgical patients, some with sound and video. The presentation of the case is followed by one or more questions, in different formats (depending on the problem), addressing the essential steps to be taken in managing the problem. For each student an individual test is generated by

the computer from a large database of cases using a fixed blueprint to balance for test content. The test score is announced immediately after the test and the student receives detailed feedback for all content-related dimensions. If the student fails, the test can be repeated after one day. The case database is used by all surgical departments in the country. The development of the database was funded jointly by all departments and involved a considerable investment. The bank is constantly being updated. Investment in equipment was moderate and test administration costs are nil (in addition, the case bank has given the affiliated hospitals access to the university computer network, including the library, email and the Internet). Clinical staff receive periodic feedback on their teaching performance in the form of a web-based summary of the test results on the different surgical domains.

During the clerkship so-called logbook assessments are given. Students must be observed during at least six patient contacts followed by an evaluation by the clinician. It is the students' own responsibility to make sure that they are observed. They are allowed to ask clinical staff to observe them. A form has been developed to rate the contacts on a five-point scale in respect of quality of history taking, differential diagnosis, physical examination, management and communication with the patient. There is ample space for any additional narrative comments from both teacher and student.

The final assessment of the clerkship is determined by a combination of tests. One of the clerkship supervisors gives recommendations for improvement in a written report. This report is included in the student's portfolio, which the students take with them to the following clerkships.

A psychiatric examination

The second example concerns the clerkship in psychiatry. This clerkship takes four weeks, two in an ambulatory care setting and two in a psychiatric hospital. The general objective of the clerkship is to offer students a general orientation on mental health care. The objective is kept limited on purpose; students are not expected to have independent patient contacts. The specific objective is to get acquainted with the main psychiatric theories concerning psychopathology and the consequences of these theories for the therapeutic management of clients.

Figure 3 shows an overview of the test procedure. Students are given a homework assignment at the start of the clerkship. During the four-week clerkship students are required to do a write-up of four patients. This report should contain: the results of the psychiatric history; arguments

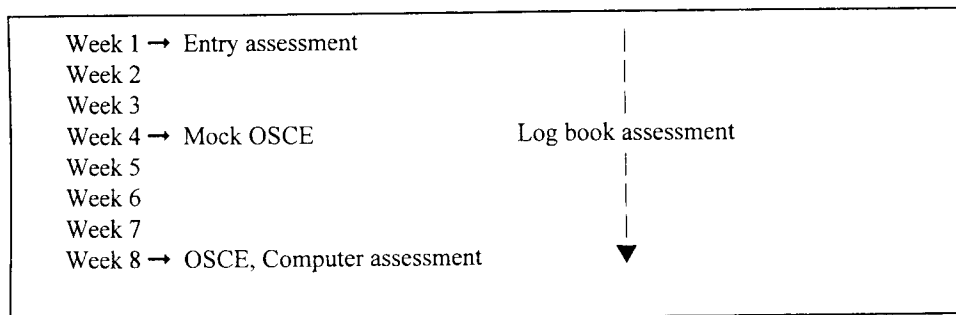


Figure 2. An illustration of an assessment programme during a surgical clerkship.

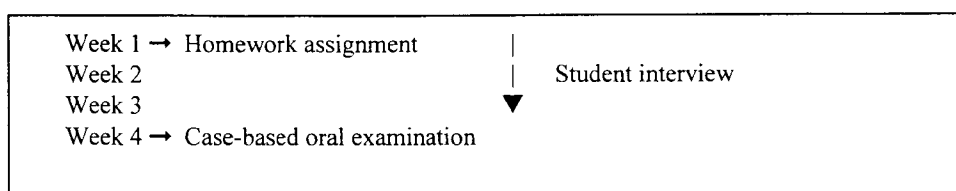


Figure 3. An illustration of an assessment programme during a psychiatry clerkship.

supporting the choice of theoretical model that is most appropriate for the complaints; an explanation of the psychiatric picture and the proposed management strategy. The four reports should deal with several different theoretical models.

Students have an interview with their personal supervisor/mentor at the end of the first, second and third week. Subjects for discussion cover the student's functioning, including emotional experiences, progress of the homework assignment and the patients encountered by the student.

At the end of the clerkship two (of the four) reports are randomly selected at the last minute and provide the subject of an oral examination. Each of two different pairs of examiners, for which the student's supervisor is not eligible, asks questions about one case report. Each of the four examiners rates the student's performance independently. If the average judgement across cases and examiners is borderline, or if the marks given by the different examiners are too far apart, consensus must be reached on the student's mark. All individual examiners' marks are stored in a file. Periodically, an average is calculated for each of the examiners to gauge interexaminer reliability.

Conclusion

This article presents a summary of research results on assessment and clerkships. These findings were the basis for general recommendations for clerkship assessment. The recommendations were illustrated by two examples of how final clerkship assessment might be organized. These examples underscore the importance of considering the educational effects of assessment as well as its effects on students, teachers and the entire curriculum. These effects should be monitored continuously to ensure that students are given the opportunity and are being encouraged to achieve the educational objectives the clerkship is intended to provide. In developing our examples we have tried to strike a balance between the attainable and the desirable. Our prime guideline in this has been that assessment should be an important and powerful educational tool.

Notes on contributors

C.P.M. VAN DER VLEUTEN is Professor and Chair Department of Educational Development and Research at the University of Maastricht.

A.J.J.A. SCHERPBIER is Professor, Director of the Educational Institute Faculty of Medicine at the University of Maastricht.

D.H.J.M. DOLMANS is Assistant Professor, Department of Educational Development and Research at the University of Maastricht.

L.W.T. SCHUWIRTH is Assistant Professor, Department of Educational Development and Research at the University of Maastricht.

G.M. VERWIJNEN is Director of the Skillslab, Faculty of Medicine at the University of Maastricht.

H.A.P. WOLFHAGEN is Associate Professor, Department of Educational Development and Research at the University of Maastricht.

References

- BATENBURG, V. (1997). *Medical students' attitude.*, University of Utrecht, Utrecht.
- BLOOM, B. S., HASTINGS, J. T., & MADAUS, G. E. (Eds.). (1971). *Handbook on formative and summative evaluation of student learning.* New York: McGraw-Hill.
- BOSHUIZEN, H. P. A. (1996). *The shock of practice: effects on clinical reasoning.* Paper presented at the AERA, New York.
- CASE, S. M. (1997). Assessment truths that we hold as self-evident and their implications. In A. J. J. A. SCHERPBIER & C. P. M. VAN DER VLEUTEN & J. J. RETHANS & A. F. W. VAN DER STEEG (Eds.), *Advances in medical education.* Dordrecht: Kluwer Academic Publishers.
- COOK, R. L., NOECKER, R. J., & SUITS, G. W. (1992). Time allocation of students in basic clinical clerkships in a traditional curriculum. *Academic Medicine*, 67(4), 279-281.
- CUNNINGTON, J. P. W., NEVILLE, A. J., & NORMAN, G. R. (1997). The risk of thoroughness: reliability and validity of global ratings in checklists in an OSCE. *Advance in Health Sciences*, 1, 227-233.
- DOLMANS, D., SCHMIDT, A., VAN DER BEEK, J., BEINTEMA, M., & GERVER, W. (1999). Does a student log provide a means to better structure clinical education? *Medical Education*, 33, 89-94.
- DOWNING, S. M., & HALADYNA, T. M. (1997). Test Item Development: Validity Evidence From Quality Assurance Procedures. *Applied Measurement in Education*, 10(1), 61-82.
- ELSTEIN, A. E. (1993). Beyond Multiple Choice Questions and Essays: The Need for a New Way to Assess Clinical Competence. *Academic Medicine*, 68(4), 244-248.
- ELSTEIN, A. S., SHULMANN, L. S., & SPRAFKA, S. A. (1978). *Medical Problem-solving: An Analysis of Clinical Reasoning.* Cambridge, MA: Harvard University Press.
- ENDE, J. (1983). Feedback in clinical medicine. *JAMA*, 250(6), 777-781.
- FREDERIKSEN, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, 39(3), 193-202.
- FRIEDMAN, C. P., MURPHY, G. C., SMITH, A. C., MATTERN, W. D. (1994). Exploratory study of an examination format for problem-based learning. *Teaching and Learning in Medicine*, 6, 194-198.
- HARDEN, R. M. (1979). How to assess students: an overview. *Medical Teacher*, 1(2), 65-69.
- HARDEN, R. M., & GLEESON, F. A. (1979). Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education*, 13(1), 41-54.
- JOLLY, B. C. (1994). *Bedside manners: Teaching and learning in the hospital setting.*, University of Maastricht, Maastricht.
- KERNAN, W. N., & O'CONNOR, P. G. (1997). Site accommodations and preceptor behaviors valued by 3rd-year students in ambulatory internal medicine clerkships. *Teaching and Learning in Medicine*, 9(2), 96-102.
- LEEDER, S. R., FELETTI, G. I., & ENGEL, C. E. (1979). Assessment/help or hurdle. *Programmed Learning and Educational Technology*, 16(308-315).
- MAATSCH, J. L. (1978). *Towards a testable theory of physician competence: an experimental analysis of a criterion-referenced specialty certification test library.* Paper presented at the 17th Annual Conference on Research in Medical Education, Washington.

- MANCALL, E. L., MURRAT, T. J., SWICK, H. M., MILLER, J. Q., SMITH, D. B., & WEISS, M. (1987). A modal clinical neuroscience curriculum. *Neurology*, 37, 1697-1699.
- MARTENS, F. M. J. G., VAN DER VLEUTEN, C. P. M., GROL, R. P. T. M., OP 't ROOT, J. M. H., CREBOLDER, H. F. J. M., & RETHANS, J. J. (1997). Educational objectives and requirements of an undergraduate clerkship in general practice: The outcome of a consensus procedure. *Family Practice*, 14, 153-159.
- MARTENS, F. M. J. G., VAN DER VLEUTEN, C. P. M., RETHANS, J. J., GROL, R. P. T. M., CREBOLDER, H. F. J. M., & OP 't ROOT, J. M. H. (1999). Time spent by undergraduate general practice teachers on educational interactions. *Under editorial review*.
- MCGUIRE, C. (1987). Written methods for assessing clinical competence. In I. R. HART & R. M. HARDEN (Eds.), *Further Developments in Assessing Clinical Competence* (pp. 46-58). Montreal: Can-heal Publications.
- MCKERWGOW, T., EGAN, A. G., & HEATH, C. J. (1991). Student contact with patient in hospital: frequency duration and effects. *Medical Teacher*, 13(1), 39-47.
- MESSICK, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(3), 13-23.
- MILLER, G. E. (1976). Continuous assessment. *Medical Education*, 10, 611-621.
- MILLER, G. E. (1990). The Assessment of Clinical Skills/Competence/Performance. *Academic Medicine*, 65(9), S63-67.
- MORGAN, K. H., & IRBY, D. H. (1978). *Evaluating clinical competence*. Saint Louis: CV Mosby Company.
- NEWBLE, D., DAWSON, B., DAUPHINEE, D., PAGE, G., MACDONALD, M., SWANSON, D., MULHOLLAND, H., THOMSON, A., & VAN DER VLEUTEN, C. P. M. (1994). Guidelines for assessing clinical competence. *Teaching and Learning in Medicine*, 6, 213-220.
- NEWBLE, D. I. (1976). The evaluation of clinical competence. *The Medical Journal of Australia*, 2, 180-183.
- NEWBLE, D. I., & JAEGER, K. (1983). The effect of assessments and examinations on the learning of medical students. *Medical Education*, 17, 165-171.
- PRINCE, C. J. A. H., VAN DE WIEL, M. W. J., SCHERPBIER, A. J. J. A., VAN DER VLEUTEN, C. P. M., & BOSCHUIZEN, H. P. A. (1999). A qualitative analysis of the transition from theory to practice in medical education. *Academic Medicine*, *In press*.
- RAM, P., GROL, R., RETHANS, J. J., SCHOUTEN, B., VAN DER VLEUTEN, C. P. M., & KESTER, A. (1999). Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of validity, reliability and feasibility. *Medical Education*, 33(6), 447-454.
- REGEHR, G., MACRAE, H., REZNICK, R., & SZALAY, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine*, 73(9), 993-997.
- REGEHR, G., & NORMAN, G. R. (1996). Issues in Cognitive Psychology: Implications for Professional Education. *Academic Medicine*, 71(9), 988-1001.
- REMMEN, R., DENEKENS, J., SCHERPBIER, A. J. J. A., VAN DER VLEUTEN, C. P. M., HERMANN, I., VAN PUUMBROECK, H., & BOSSAERT, L. (1998). Evaluation of clinical skills training during clerkships using student focus groups. *Medical Teacher*, 20, 428-431.
- RETHANS, J. J., STURMANS, F., DROP, M. J., & VAN DER VLEUTEN, C. P. M. (1991). Assessment of performance in actual practice of general practitioners. *British Journal of General Practice*, 41, 97-99.
- RETHANS, J. J., VAN LEEUWEN, Y., DROP, M., VAN DER VLEUTEN, C. P. M., & STURMANS, F. (1990). Competence and performance: two different constructs in the assessment of quality of medical care. *Family Practice*, 7, 168-174.
- ROLFE, I., & MCPHERSON, J. (1995). Formative assessment: how am I doing? *The Lancet*, 345, 837-839.
- SCHAMROTH, A. J., & HAINES, A. P. (1992). Student assessment of clinical experience in general surgery. *Medical Teacher*, 14(4), 355-262.
- SCHERPBIER, A. J. J. A. (1997). *Kwaliteit van vaardigheid gemeten (The quality of clinical skills assessed)*. University of Maastricht, Maastricht.
- SCHMIDT, H., NORMAN, G., & BOSCHUIZEN, H. (1990). A cognitive perspective on medical expertise: theory and implications. *Academic Medicine*, 65(10), 611-622.
- SCHUWIRTH, L. W. T., VAN DER VLEUTEN, C. P. M., DE KOCK, C. A., PEPPERKAMP, A. G. W., & DONKERS, H. H. L. M. (1996). Computerized case-based testing: a modern method to assess clinical decision making. *Medical Teacher*, 18(4), 295-300.
- STREINER, C. (1995). Clinical ratings-ward rating. In S. SHANNON & G. NORMAN (Eds.), *Evaluation methods: a resource handbook* (pp. 29-32). Hamilton: Program for Educational Development McMaster University.
- SWANSON, D. B., NORCINI, J. J., & GROSSO, L. J. (1987). Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education*, 12(3), 220-246.
- SZENAS, P. (1997). The role of faculty observation in assessing students' clinical skills. *Contemporary Issues in Medical Education*, 1(1), 1-2.
- TORTOLANI, A. J., LEITMAN, I. M., & RUSICCI, D. A. (1997). Student perceptions of skills acquisition during the surgical clerkship. *Teaching and Learning in Medicine*, 9(3), 186-191.
- VAN DER VLEUTEN, C. P. M. (1996a). The assessment of Professional Competence: Developments, Research and Practical Implications. *Advances in Health Science Education*, 1(1), 41-67.
- VAN DER VLEUTEN, C. P. M. (1996b). *Beyond intuition*. Maastricht: Maastricht University Press.
- VAN DER VLEUTEN, C. P. M., NEWBLE, D. I., CASE, S. M., HOLSGROVE, G., MCCANN, B., MCGRAE, C., & SAUNDERS, N. (1994). Methods of Assessment in Certification. In D. I. NEWBLE & B. JOLLY & R. WAKEFORD (Eds.), *The Certification and Recertification of Doctors, Issues in the Assessment of Clinical Competence* (pp. 105-125). Cambridge: Cambridge University Press.
- VAN DER VLEUTEN, C. P. M., NORMAN, G. R., & DE GRAAFF, E. (1991). Pitfalls in the pursuit of objectivity: Issues of reliability. *Medical Education*, 25, 110-118.
- VAN LUIJK, S. J., VAN DER VLEUTEN, C. P. M., & SCHELVEN, R. M. (1990). The relation between content and psychometric characteristics in performance-based testing. In W. BENDER & R. J. HIEMSTRA & A. J. J. A. SCHERPBIER & R. P. ZWIERSTRA (Eds.), *Teaching and Assessing Clinical Competence*. (pp. 202-207). Groningen: Boekwerk Publications.
- VERHOEVEN, B. H., VERWIJNEN, G. M., SCHERPBIER, A. J. J. A., SCHUWIRTH, L. W. T., & VAN DER VLEUTEN, C. P. M. (1999). Quality assurance in test construction: the approach of a multidisciplinary central test committee. *Education for Health*, 12(1), 49-60.
- WARD, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement*, 6(1), 1-11.
- WIJNEN, W. H. F. W. (1981). Formative use of assessment. In J. C. M. METZ & J. MOLL & H. J. WALTON (Eds.), *Examinations in medical education*. (pp. 116-131). Utrecht: Bunge.