Assessment drives learning. Many people argue that this statement is incorrect and that the curriculum is the key in any clinical course. In reality, students feel overloaded by work and respond by studying only for the parts of the course that are assessed. To promote learning, assessment should be educational and formative—students should learn from tests and receive feedback on which to build their knowledge and skills. Pragmatically, assessment is the most appropriate engine on which to harness the curriculum.

Additionally, with an increasing focus on the performance of doctors and on public demand for assurance that doctors are competent, assessment also needs to have a summative function. Tests of clinical competence, which allow a decision to be made about whether a doctor is fit to practise or not, are in demand. This demand raises a challenge for all involved in medical education. Tests that have both a formative and summative function are hard to design. Yet, if assessment focuses only on certification and exclusion, the all-important influence on the learning process will be lost. The panel shows the key measurement issues that should be addressed when designing assessments of clinical competencies.1

Blueprinting

If students focus on learning only what is assessed, assessment in medical education must validate the objectives set by the curriculum. Test content should be carefully planned against learning objectives—a process known as blueprinting.2 For undergraduate curricula, for which the definition of core content is now becoming a requirement,3 this process could be easier than for postgraduate examinations, where curriculum content remains more broadly defined. However, conceptual frameworks against which to plan assessments are essential and can be defined even for generalist college tests.4

Assessment programmes must also match the competencies being learnt and the teaching formats being used. Many medical curricula define objectives in terms of knowledge, skills, and attitudes. These cannot be properly assessed by a single test format. All tests should be checked to ensure that they are appropriate for the objective being tested. A multiple-choice examination, for example, could be a more valid test of knowledge than of communication skills, which might be best assessed with an interactive test. However, because of the complexity of clinical competence, many different tests should probably be used.

Standard setting

Inferences about students’ performance in tests are essential to any assessment of competence. When assessment is used for summative purposes, the score at which a student will pass or fail has also to be defined. Norm referencing, comparing one student with others, is frequently used in examination procedures if a specified number of candidates are required to pass—ie, in some college membership examinations. Performance is described relative to the positions of other candidates. As such, variation in the difficulty of the test is compensated for by the use of standard setting.

Key issues that underpin any test

<table>
<thead>
<tr>
<th>Key issues</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summative/formative</td>
<td>Be clear on the purpose of the test.</td>
</tr>
<tr>
<td>Blueprinting</td>
<td>Plan the test against the learning objectives of the course or competencies essential to the specialty.</td>
</tr>
<tr>
<td>Validity</td>
<td>Select appropriate test formats for the competencies to be tested. This action invariably results in a composite examination.</td>
</tr>
<tr>
<td>Reliability</td>
<td>Sample adequately. Clinical competencies are inconsistent across different tasks. Test length is crucial if high-stakes decisions are required. Use as many examiners as possible.</td>
</tr>
<tr>
<td>Standard setting</td>
<td>Define endpoint of assessment. Set the appropriate standard—eg, minimum competence—in advance.</td>
</tr>
</tbody>
</table>

Lancet 2001; 357: 945–49

Department of General Practice and Primary Care, Guy’s, King’s and St Thomas’ School of Medicine, Weston Education Centre, London SE5 9RJ, UK (V Wass FRCP, Prof R Jones DM); Department of Educational Development and Research, University of Maastricht, Maastricht, Netherlands (Prof C Van der Vleuten PhD); and Johns Hopkins University School of Medicine, Baltimore, MD, USA (J Shatzer PhD)

Correspondence to: Dr Val Wass
(e-mail: valerie.wass@kcl.ac.uk)
for. However, differences in the abilities of student cohorts sitting the test are not accounted for. Therefore, if a group is above average in ability, those who might have passed in a poorer cohort of students will fail. Norm referencing is clearly unacceptable for clinical competency licensing tests, which aim to ensure that candidates are safe to practise. A clear standard needs to be defined, below which a doctor would not be judged fit to practise. Such standards are set by criterion referencing. In this case, the minimum standard acceptable is decided before the test. However, although differences in candidate ability are accounted for, variation in test difficulty becomes a key issue; standards should be set for each test, item by item. Various time-consuming but essential methods have been developed to do this, such as the techniques of Angoff and Ebel. The choice of method will depend on available resources and on the consequences of misclassifying examinees as having passed or failed.

**Validity versus reliability**

Just as summative and formative elements of assessment need careful attention when planning clinical competence testing, so do the issues of reliability and validity.

Reliability is a measure of the reproducibility or consistency of a test, and is affected by many factors such as examiner judgments, cases used, candidate nervousness, and test conditions. Two aspects of reliability have been well researched: inter-rater and inter-case (candidate) reliability. Inter-rater reliability measures the consistency of rating of performance by different examiners. The use of multiple examiners across different cases improves inter-rater reliability. In an oral examination, the average judgment of ten examiners, each assessing the candidate on one question, produces a more reliable test than that of one examiner asking ten questions.

The consistency of candidate performance across the cases (inter-case reliability) is perhaps the most important issue in clinical competence testing. Doctors do not perform consistently from task to task. Broad sampling across cases is essential to assess clinical competence reliably. This observation might not be surprising given the differences in individual experiences encountered during training and practice, but it challenges the traditional approach to clinical competence testing, whereby the competence of the candidates was assessed on a single case. Tests of clinical skills have moved into a series of tasks and encounters (stations). Objective structured clinical examination (OSCE), the multicase format with the development of the pyramid of competence introduced by Miller, is a simple conceptual model, designed to test. No valid assessment methods that actually succeed in testing the competencies that it is supposed to assess individuals and carry out necessary procedures. However, the ultimate goal for a valid assessment of clinical competence is to test what the doctor actually does in the workplace. Over the past four decades, research in this area has focused on developing valid ways of assessing the summit of the pyramid—ie, a doctor’s actual performance.

**Assessment of “knows” and “knows how”**

The assessment of medical undergraduates has tended to focus on the pyramid base: “knows”—ie, the straight factual recall of knowledge, and “knows how”—ie, the application of knowledge to problem-solving and decision-making. This method might be appropriate in early stages of the medical curriculum, but as skill-teaching is more vertically integrated, careful planning of assessment formats becomes crucial. Various test formats of factual recall are available, which are easy to devise and by knows how (applied knowledge). These can be more easily assessed with basic written tests of clinical knowledge such as multiple-choice questions. Clearly, this test format cannot assess the more important facet of competency required by a qualifying doctor—ie, the shows how. This factor is a behavioural rather than a cognitive function and involves hands on, not in the head, demonstration. A senior student about to start work with patients must be able to show an ability to assess individuals and carry out necessary procedures. However, the ultimate goal for a valid assessment of clinical competence is to test what the doctor actually does in the workplace. Over the past four decades, research in this area has focused on developing valid ways of assessing the summit of the pyramid—ie, a doctor’s actual performance.

**Figure 1:** Reported reliability when 4 h testing times are used for different test formats

<table>
<thead>
<tr>
<th>Examination method</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCQ</td>
<td>0.8</td>
</tr>
<tr>
<td>Short essay</td>
<td>0.7</td>
</tr>
<tr>
<td>PMP</td>
<td>0.6</td>
</tr>
<tr>
<td>Oral</td>
<td>0.5</td>
</tr>
<tr>
<td>History (long case)</td>
<td>0.4</td>
</tr>
<tr>
<td>OSCE</td>
<td>0.3</td>
</tr>
</tbody>
</table>

**Figure 2:** Miller’s pyramid of competence

- **Does**
  - Performance assessment in vivo
    - Undercover SPs, video, logs
  - Performance assessment in vitro
    - OSCE, SP-based test

- **Shows how**
  - Clinical context based tests
    - MCQ, essay, oral

- **Knows how**
  - Factual tests
    - MCQ, essay, oral

**Knows**
- Basic facts

**Figure 3:** Table of examination methods

<table>
<thead>
<tr>
<th>Examination method</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCQ</td>
<td>0.8</td>
</tr>
<tr>
<td>Short essay</td>
<td>0.7</td>
</tr>
<tr>
<td>PMP</td>
<td>0.6</td>
</tr>
<tr>
<td>Oral</td>
<td>0.5</td>
</tr>
<tr>
<td>History (long case)</td>
<td>0.4</td>
</tr>
<tr>
<td>OSCE</td>
<td>0.3</td>
</tr>
</tbody>
</table>
For personal use only. Reproduce with permission from The Lancet Publishing Group.
Assessment of communication skills and attitudes can also be attempted in the OSCE. Interestingly, these are proving to be case-specific and have low generalisability across clinical contexts. Collyer20 has shown that to assess empathy reliably, as many as 37 different scenarios could be needed. However, patient perception questionnaires that include aspects of physician communication are used in many standardised patient assessments and are quite reliable.21 Who should be the judge: the patient or the examiner?

Whether assessment is done by the patient or the examiner does not seem to matter.20,26 Scoring against a checklist of items is not as objective as originally supposed.20 There is increasing evidence that global ratings, especially by physicians, are as reliable as checklists.31,32 However, extensive training of judges is required to ensure consistency. Neither global nor checklist ratings offer a true “gold standard” of judging ability. Although checklists cannot capture all aspects of the physician-patient interaction, global ratings could be subject to other rater biases, and the choice to use them should depend on the resources available during the assessment. A third alternative uses information collected at “post-encounter” stations, now used in North America, where students spend 5–7 min recording their findings from the simulated patient encounter. Williams and McLaughlin27 explored the use of a patient-findings questionnaire, comparing it with the checklist performance record of the standardised patient. Both instruments gave similar data-acquisition scores and pass or fail decisions at both the item and test level. The authors argue that the patient(examinee) scoring against a checklist of items is not as objective as originally thought.20 The estimations in some of the shortcomings of checklists and relies solely on the data-collection abilities of the examinee to rate ability.

Standardised patients

The use of standardised patients versus real patients remains an area of interest. Simulations are the norm in North America. Extensive training to ensure reproducibility and consistency of scenarios is carried out. Given the high reliabilities required of the North American licensing tests, the high costs of training can be justified but, perhaps, at the cost of validity. Performance on an OSCE might not be the same as performance in real life.21 Clearly this question is the focus of testing at the very tip of the pyramid—eg, “performance”.

Assessment of “does”

The real challenge lies in the assessment of a student’s actual performance on the wards or in the consulting room. Increasing attention is being placed on this type of assessment in postgraduate training, because revalidation of a clinician’s fitness to practise and the identification of badly performing doctors are areas of public concern. Any attempt at assessment of performance has to balance the issues of validity and reliability, and there has been little research into possible approaches in undergraduate medical schools. Some UK schools use in-course assessment and student portfolios to assess student performance in the final year, rather than a summative final examination. Whether such methods are robust enough to cover the issue of content-specificity and have the necessary comprehensiveness of the assessments discussed above remains to be seen.

Further research into the format and reliability of composite medical examinations and the use of portfolio assessment is important. Current examination formats tend to focus too heavily on knowledge-based competencies. Assessment at the apex of Miller’s pyramid, the “does”, is the international challenge of the century for all involved in clinical competence-testing. The development of reliable measurements of student performance with predictive validity of subsequent clinical competencies and a simultaneous educational role28 is a gold standard yet to be achieved.

References