Medical education quartet

Assessment of clinical competence

Val Wass, Cees Van der Vleuten, John Shatzer, Roger Jones

Tests of clinical competence, which allow decisions to be made about medical qualification and fitness to practise, must be designed with respect to key issues including blueprinting, validity, reliability, and standard setting, as well as clarity about their formative or summative function. Multiple choice questions, essays, and oral examinations could be used to test factual recall and applied knowledge, but more sophisticated methods are needed to assess clincial performance, including directly observed long and short cases, objective structured clinical examinations, and the use of standardised patients. The goal of assessment in medical education remains the development of reliable measurements of student performance which, as well as having predictive value for subsequent clinical competence, also have a formative, educational role.

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

Key issues that underpin any test

Key issues	Description
Summative/formative	Be clear on the purpose of the test.
Blueprinting	Plan the test against the learning objectives of the course or competencies essential to the speciality.
Validity	Select appropriate test formats for the competencies to be tested. This action invariably results in a composite examination.
Reliability	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.
Standard setting	Define endpoint of assessment. Set the appropriate standard—eg, minimum competence—in advance.

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 FRCGP, 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.[5] 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.[6]

The consistency of candidate performance across the cases (intercase reliability) is perhaps the most important issue in clinical competence testing. Doctors do not perform consistently from task to task.[7] 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 the multicase format with the development of the objective structured clinical examination (OSCE), consisting of a series of tasks and encounters (stations). Many stations and sufficient testing time are essential to achieve adequate intercase reliability for the test. Whatever the test format, length is critical to the reliability of any clinical competence test. Figure 1 shows reported reliabilities for various tests each lasting for 4 h.[6,8–11]

Validity, on the other hand, focuses on whether a test actually succeeds in testing the competencies that it is designed to test. No valid assessment methods that measure all facets of clinical competence have been designed. The pyramid of competence (figure 2), introduced by Miller,[12] is a simple conceptual model, which outlines the issues involved when analysing validity.

The pyramid conceptualises the essential facets of clinical competence. The base represents the knowledge components of competence: knows (basic facts) followed

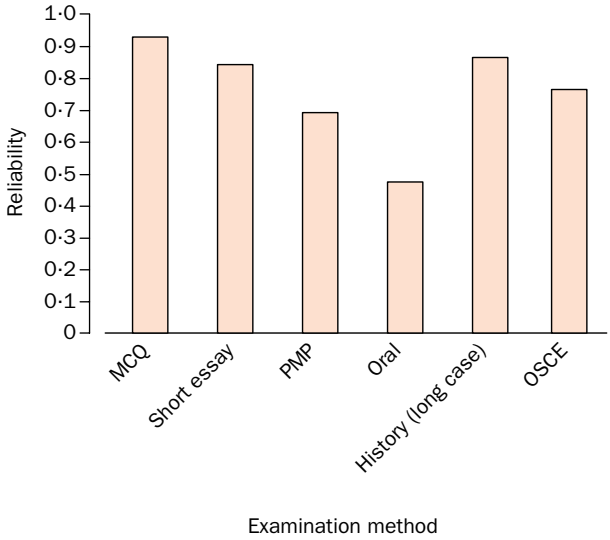

Figure 1: **Reported reliability when 4 h testing times are used for different test formats**
MCQ=multiple-choice examination; PMP=patient management problem; OSCE=objective structured clinical examination.

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 work place. 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

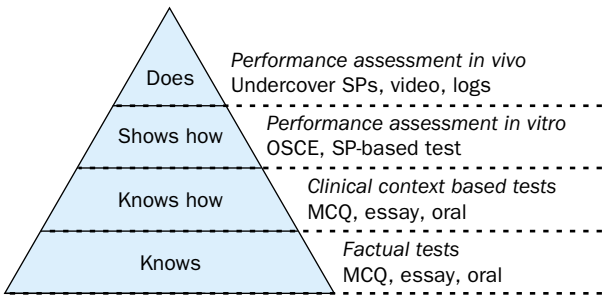

Figure 2: **Miller's pyramid of competence**
SP=simulated patients; OSCE=objective structured clinical examination; HCQ=multiple-choice questions.

deliver. Multiple-choice formats are the most widely used and are popular. Although time consuming to set, these tests have high reliability, because of the large number of items that can be easily tested and marked. However, criticism of the validity of multiple-choice questions has stimulated much research into alternative options, such as true or false, single best answer, extended matching,[13] and long and short menus of options.[14] Many people argue that only trivial knowledge can be tested in this way. By giving options, candidates are cued to respond and the active generation of knowledge is avoided.

Other test formats have been used to assess factual knowledge. Essay and oral examinations are still popular in the UK and other European countries, despite being excluded for more than 20 years from assessments in North America on grounds of unreliability. Many argue that the ability to recall and synthesise information can be judged best in a face-to-face encounter. However, essays are notoriously difficult to mark consistently[15] and orals are unreliable because of issues already discussed; lack of standardisation of questions, insufficient judges, and lack of sufficient testing time.[6] Unfortunately, the reconciliation of validity arguments with reliability issues is almost impossible.

These difficulties have led to an increase in focus on testing "knows how"—ie, the assessment of knowledge as applied to problem solving or clinical reasoning in specific contexts. All the test formats mentioned above can be adapted to test knowledge across different contexts. Clinical scenarios can be applied to multiple-choice formats such as single best answer or extended matching, and to orals. However, it is difficiult to set questions that test application rather than straight delivery of factual knowledge.[16] Problem solving seems to be closely linked to knowledge and also to be content-specific—ie, a candidate's problem-solving ability is inconsistent across different tasks. As in all areas of clinical competence testing, intercase reliability becomes an issue.[16] This dilemma is most easily solved in written formats in which a large number of questions can be covered relatively quickly. However, the use of orals and essays is hampered by the low generalisability of problem solving skills. The generalisability coefficient is an estimate of the reproducibility of the measurement on a scale from 0–1·0, and 0·8 is seen as the minimum requirement for reliable measurement. Swanson[6] calculated that to achieve an acceptable degree of generalisability (0·8) would take 12–16 case histories in oral examinations. Similar difficulties apply to orals done by trained examiners.[17]

The "knows how" of ethical situations and attitudinal issues can also be explored through orals and essays. The examination for membership of the Royal College of General Practitioners in the UK uses orals to assess decision-making in a structured question-setting format; an area of focus is professional and personal development. A recent publication draws attention to the difficulties of ensuring a fair assessment in this arena—for example, if English is not the candidate's first language.[18]

In the past, UK medical schools have used short orals or vivas for candidates with borderline results on written tests, to decide on whether they should pass or not. To base this crucial pass or fail decision of a borderline written test on an oral, one of the most unreliable assessment tools, is not a good idea and this practice is gradually being abandoned. The focus should be on making the written test as reliable as possible to give confidence in the pass or fail decision.

Difficulties in setting up "knows how" tests involve combining the application of knowledge with the large range of problems required to achieve good intercase reliability, and distinguishing between responses cued by straight factual knowledge ("knows") from thought processes stimulated by the problem ("knows how"). Formats have been developed such as the "key feature" test developed in Canada.[19] These require short uncued answers to clinical scenarios, but limit the assessment to key issues to enable a large number of scenarios to be covered. Similarly, computer simulations can replace written or verbal scenarios and, hopefully, with the development of multimedia, can be used to raise the standard of clinical testing.[20] In the past, simulations have been complicated. Dynamic and complex situations have been created that require enormous resources rarely available in medical schools. A focus on short simulations to produce the required breadth for tests, which stimulate rather than cue responses, remains a challenge for those developing them.

Assessment of "shows how"

Traditional long and short cases

Although abandoned for many years in North America, the use of unstandardised real patients in long and short cases to assess clinical competence remains a feature of both undergraduate and postgraduate assessment in the UK. Such examinations are increasingly challenged on the grounds of authenticity and unreliability. Long cases are often unobserved, the assessment relies on the candidate's presentation, representing an assessment of "knows how" rather than "shows how". Generally, only one long case and three or four short cases are used. The low generalisability of clinical skills that are content-specific is clearly of concern when applying this test format. Yet little psychometric research on long cases has been published. Initial data (figure 1) suggest that test length is again the key to improving reliability in this form of testing and that ten observed, history-taking long cases, rather than one, could achieve the test reliability required of a high-stakes examination.[10]

Attempts have been made to improve the long-case format. A more structured presentation of an unobserved long case, the Objective Structured Long Examination Record (OSLER), has been developed, which includes some direct observation of the candidate interacting with the patient.[21] Observed long cases are also used in some institutions.[22,23] Decreasing the length of time available to assess a case and allowing more cases to be assessed within a given testing time could also be an option.[24] However, this requires balancing the station length that provides the most reliable results with a length similar to that in clinical practice. Unless the format of long and short cases is improved by direct observation, and the test length is extended to include many more cases, the unreliability of this traditional format does not justify its use. An unreliable test cannot be valid.

Objective structured clincial examination

As a potential solution to the difficulties of adequate sampling and standardisation of cases, the OSCE[25] has gained increasing importance on both sides of the Atlantic. Candidates rotate through a series of stations based on clinical skills applied in a range of contexts. Wide sampling of cases and structured assessment improve reliability, but this examintion format is expensive and labour-intensive. Validity can be lost at the expense of reliability, since complex skills, requiring an integrated professional judgment, become fragmented by the relatively short station length (generally 5–10 min).

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. Collyer[26] 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.[27] 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.[28,29] Scoring against a checklist of items is not as objective as originally supposed.[30] 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 McLaughlin[27] 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-findings questionnaire minimises 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 liscensing 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.[33] 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 comprehensivesness 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 role[34] is a gold standard yet to be achieved.

References

1 Neufeld VR, Norman GR. Assessing clinical competence, vol 7. New York: Springer, 1985.

2 Dauphinee D. Determining the content of certification examinations. In: Newble D, Jolly B, Wakeford R. The certification and recertification of doctors: issues in the assessment of clinical competence. Cambridge: Cambridge University Press, 1994: 92–104.

3 The General Medical Council Education Committee. Tomorrow's doctors: recommendations on undergraduate medical education. London: General Medical Council; 1993.

4 Hays RB, van der Vleuten C, Fabb WE, Spike NA. Longitudinal reliability of the Royal Australian College of General Practitioners certification examination. *Med Educ* 1995; **29:** 317–21.

5 Cusimano MD. Standard setting in medical education. *Acad Med* 1996; **71** (suppl 10): S112–20.

6 Swanson DB. A measurement framework for performance based tests. In: Hart IR, Harden RM, eds. Further developments in assessing clinical competence. Montreal: Can-Heal, 1987: 13–45.

7 Swanson DB, Norman GR, Linn RL. Performance-based assessment: lessons learnt from the health professions. *Educ Res* 1995; **24:** 5–11.

8 Norcini JJ, Swanson DB, Grosso LJ, Shea JA, Webster GD. Reliability, validity and efficiency of multiple choice questions and patient management problem items formats in the assessment of physician competence. *Med Educ* 1985; **19:** 238–47.

9 Stalenhoef-Halling BF, van der Vleuten CPM, Jaspers TAM, Fiolet JFBM. The feasibility, acceptability and reliability of open-ended questions in a problem based learning curriculum. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwiestra RP, eds. Teaching and assessing clinical competence. Groningen: Boekwerk, 1990: 1020–31.

10 Wass V, Jones R, van der Vleuten CPM. Standardised or real patients to test clinical competence? The long case revisited. *Med Educ* (in press).

11 Newble DI, Swanson DB. Psychometric characteristics of the objective structured clinical examination. *Med Educ* 1996; **22:** 325–34.

12 Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990; **65:** 563–67.

13 Case SM, Swanson DB. Extended matching items: a practical alternative to free response questions. *Teach Learn Med* 1993; **5:** 107–15.

14 Case SM, Swanson DB. Constructing written test questions for the basic and clincial sciences. Philadelphia: National Board of Examiners, 1996.

15 Frijns PHAM, van der Vleuten CPM, Verwijnen GM, Van Leeuwen YD. The effect of structure in scoring methods on the reproducibility of tests using open ended questions. In: Bender W, Hiemstra RJ, Scherbier AJJA, Zwierstra RP, eds. Teaching and assessing clinical competence. Gromingen: Boekwerk, 1990: 466–71.

16 Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ* 1996; **1:** 41–67.

17 Wakeford R, Southgate L, Wass V. Improving oral examinations: selecting, training and monitoring examiners for the MRCGP. *BMJ* 1995; **311:** 931–35.

18 Roberts C, Sarangi S, Southgate L, Wakeford R, Wass V. Oral examinations, equal opportunities and ethnicity: fairness issues in the MRCGP. *BMJ* 2000; **320:** 370–74.

19 Page G, Bordage G, Allen T. Developing key-feature problems and examinations to assess clinical decision-making skills. *Acad Med* 1995; **70:** 194–201.

20 Schuwirth LWT, van der Vleuten CPM, de Kock CA, Peperkamp AGW, Donkers HHLM. Computerised case-based testing: a modern method to assess clinical decision making. *Med Teach* 1996; **18:** 294–99.

21 Gleeson F. The effect of immediate feedback on clinical skills using the OSLER. In: Rothman AI, Cohen R, eds. Proceedings of the sixth Ottawa conference of medical education 1994. Toronto: University of Toronto Bookstore Custom Publishing, 1994: 412–15.

22 Newble DI. The observed long case in clinical assessment. *Med Educ* 1994; **25:** 369–73.
23 Price J, Byrne GJA. The direct clinical examination: an alternative method for the assessment of clinical psychiatric skills in undergraduate medical students. *Med Educ* 1994; **28:** 120–25.
24 Shatzer JH, Wardrop JL, Williams RC, Hatch TF. The generalisability of performance of different station length standardised patient cases. *Teach Learn Med* 1994; **6:** 54–58.
25 Harden RM, Gleeson FA. ASME medical educational booklet no 8: assessment of medical competence using an objective structured clinical examination (OSCE). *J Med Educ* 1979; **13:** 41–54.
26 Colliver JA, Willis MS, Robbs RS, Cohen DS, Swartz MH. Assessment of empathy in a standardized-patient examination. *Teach Learn Med* 1998; **10:** 8–11.
27 Williams RG, McLaughlin MA, Eulenberg B, Hurm M, Nendaz MR. The patient findings questionnaire: one solution to an important standardized patient examination problem. *Acad Med* 1999; **74:** 1118–24.
28 van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardised paients: state of the art. *Teach Learn Med* 1990; **2:** 58–76.
29 Colliver JA, Verhulst SJ, Williams RG, Norcini JJ. Reliability of performance on standardised patient cases: a comparison of consistency measures based on generalizability theory. *Teach Learn Med* 1989; **1:** 31–37.
30 Reznick RK, Regehr G, Yee G, Rothman A, Blackmore D, Dauphinee D. Process-rating forms versus task-specific checklists in an OSCE for medical licensure. *Acad Med* 1998; **73:** S97–99.
31 Regehr G, MacRae H, Reznick R, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med* 1998; **73:** 993–97.
32 Schwatz MH, Colliver JA, Bardes CL, Charon R, Fried ED, Moroff S. Global ratings of videotaped performance versus global ratings of actions recorded on checklists: a criterion for performance assessment with standardized patients. *Acad Med* 1999; **74:** 1028–32.
33 Ram P, Grol R, Rethans JJ, Schouten B, van der Vleuten CPM, Kester A. Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of validity, reliability and feasibility. *Med Educ* 1999; **33:** 447–54.
34 van Der Vleuten C. Validity of final examinations in undergraduate medical training. *BMJ* 2000; **321:** 1217–19.