Advances in Psychiatric Treatment (1999), vol. 5, pp. 405–414

Assessment and examination

Brian Jolly

A 19-year-old student, wandering the streets of a Northern city, is picked up at 3am by a taxi driver. The student requests delivery to a fictitious destination. The student recounts to the driver a series of events including abandonment by parents at the age of six in a forest in the West Country, and subsequent adoption by an elderly couple who live in Birmingham. The taxi driver takes the student to the central police station, where, after a brief interview, sectioning under the Mental Health Act 1983 takes place.

Over the next few days the student is interviewed for two hours by a senior registrar in psychiatry; interviewed for one hour by a consultant adolescent psychiatrist; observed over a continuous 72-hour period by a total of five psychiatric nurses who document her activities precisely; is given two series of extensive tests by a clinical psychologist; and is the subject of a two-hour case conference that involves all clinicians having contact with her. This includes reports from a general practitioner, interviews with parents about previous behaviour and a summary from an art therapist of the outcome of two sessions' work. At the end of this conference, the student is diagnosed as having a bipolar disorder and put on a treatment regime that includes medication. This treatment is highly successful – within three months the problem is controlled, and within six months the student is functioning well, and applies to repeat the year.

Four years later the same student takes her medical school finals. She prepares by spending long hours in the library and mugging up on banks of published multiple choice (MCQ) questions. She spends about four days on the wards but the competition for patient access is high and she can't face the hassle. The examination consists of three 3-hour MCQ papers, and two 30-minute 'long case' examinations of patients. The patient interaction or examination is not observed, but there is one 20-minute viva examination by two consultants for each patient. In each viva the student talks for approximately seven minutes. The examiners talk to the student for 11 minutes and the remaining two minutes is taken up by the examiners talking between themselves. The results of the two long cases are recorded and averaged, but the examiners do not meet and discuss the student. At the examiners' meeting, the issue of previous psychiatric disturbance is mentioned during a discussion totalling two minutes. As the student has made a seemingly good recovery, this information is deemed irrelevant. A pass in MBBS is awarded. Five weeks later she starts work on an acute cardio-thoracic surgical ward.

Each of these episodes entails a 'life or death' decision being taken by doctors about this student. In which decision would you have more confidence? Explain your reasoning.

A model answer

The remainder of this paper will explore and hopefully provide the background for what might be a good answer to the questions posed above. In very broad terms, I would hope that most people, all other things being equal, would have more confidence in the decision taken about the student's psychiatric disorder than that to license her as a medical practitioner. There are several reasons for this.

Generally, the samples of behaviour scrutinised in order to arrive at the two decisions are radically different in scope. The behaviour of the student was extensively observed in the psychiatric decision, but not at all in the educational one. The number of people involved in the first decision was also greater. Typically, they would have operated within certain

Brian Jolly joined the BMA's Centre for Health and Medical Education in 1972. In 1979 he transferred to Barts' and subequently became Head of the Academic Unit of Medical and Dental Education. His PhD, from the University of Maastricht, involved work on clinical teaching. He has been a consultant to organisations worldwide and is a founding member of the UK General Osteopathic Council and Chairman of its Education Committee. He has had a significant involvement in the GMC Performance Review project, and from 1996 to 1999 was Director of the Medical Education Unit, University of Leeds. Earlier this year he became one the founding Professors of a new Department of Medical Education at the University of Sheffield (Coleridge House, Sheffield S7 5AU).

standards of clinical practice and probably been forced to spend much of the time on salient issues. In the second scenario, the examiners spent more time talking than the candidate. Although we do not know the precise details of the patients examined by the student, they can have been exhibiting only a very small subset of the pathologies encountered in medical practice.

Neither scenario is particularly unique. Both descriptions are anchored in personal experience and reflected in research studies. Neither educational nor psychiatric decisions can be perfect. The purpose of this article is to help make educational decisions at least as robust as their clinical counterparts. After considering the purpose of assessment and then some definitions and constraints, this article will describe essential characteristics of assessment and ideas for the better application of assessment tools in clinical education.

The purpose of assessment

The purposes of assessment in a professional context are: to measure and/or make judgements about mastery of skills or knowledge; to measure improvement over time; to arrive at some definition of strengths and weaknesses; to rank people for selection or exclusion; and perhaps to motivate them. Many assessments are based on tests or examinations, or structured performance reviews (e.g. audit). The important principle to keep in mind is that the purpose of the assessment will determine its structure and interpretation. For example, for an assessment to reveal a student's weaknesses it needs to be designed to be comprehensive, informative, non-threatening, with low stakes for the student, and with little or no retribution involved. Hence, using an end-of-year hurdle or final examination, in which students would be at pains to conceal their faults, would not be very sensible in this capacity. If the assessment is too difficult, real weaknesses will not show up. Using an assessment designed to select high flyers will not be much use either.

The difference between assessment and examinations

Most assessment in medicine has traditionally been aimed at the measurement of competence. Students have grown up in a culture which is competitive and confrontational, both at the teaching and testing stages. Medical curricula have generally depended on examinations, in the sense of tests administered to a large group of students usually at the end of a course (e.g. finals), to make judgements about students' degree of competence. This has made 'assessment' synonymous with 'examinations' for many students and staff – an altogether unhelpful confusion. Furthermore, the emphasis on knowledge as an examinable feature of competence has resulted in overuse of banks of MCQs, often poorly constructed and administered. In fact, examinations are just one means of assessment. Others include audit, observation of real performance on the job, and peer review of one kind or another. This dependence on examinations has not helped the skills of self-assessment, now widely valued for lifelong learning, nor those of giving and receiving feedback. As a result, assessments have rarely been used to promote personal development or address deficiencies and they have been full of what Boud (1997) and others call 'final vocabulary' – "You have failed"; "That was no good"; "You can't do it"; or "You have passed". Such statements apportion value to an individual, but are no help in getting better. These trends have stemmed from an emphasis on certifying in order to establish standards and protect the public. In the aftermath of the 'Bristol case' (in which surgeons and a medical director were shown to have ignored signs warning that operations were not being appropriately used), there has understandably been even more pressure to assess competence and performance. Hence, assessment for license to practise and assessment for learning are in constant tension but, where possible, should be used symbiotically.

Characteristics of assessment

There are three major requirements of assessments used for licensing decisions. They need to be **reliable** (reproducible) and **valid** and have **appropriate effects** on the activity of those preparing to take them.

Reliability

In the student's psychiatric history above, it was important to reach a reliable or reproducible clinical diagnosis because effective management might well have depended on it. Some measures or elements of the history might have been gathered twice or more, and by different people. This probably partly accounted for the wealth of data collected. However, in educational measurement we sometimes inappropriately accept assessment data at face value without enquiring as to its reproducibility.

https://doi.org/10.1192/apt.5.6.405 Published online by Cambridge University Press

For any educational assessment to be useful in grading, selecting or accrediting candidates it should be reliable. An assessment or a test is 'reliable' if it measures something consistently – that is, it gives the same indications about candidates when used on different occasions, or with different markers, or in different forms. Reliability is a concept applicable to the same measure used in different circumstances, for example, a repeated skills test, or two replications of an audit. It does not apply to the relationship between different types of measures that could conceivably be measuring different capabilities, for example, an MCQ (measuring factual knowledge) and a clinical viva (measuring an indeterminate mixture of interpersonal, cognitive and clinical skills).

Efforts to ensure the reproducibility or consistency of data are common in clinical medicine, such as in blood sugar or blood pressure measurement. Reliability can also be represented numerically, for example, by the correlation between measurement on one occasion and a subsequent one (test–retest reliability). Because of the relative precision of biomedical measures compared to psychological ones, correlations between two consecutive estimates of, say, blood sugar will frequently be very high, for example, 0.9–0.96. For educational measurements, we would be happy if such correlations were at least of the order of 0.8.

While much is known about cyclical variation in biological indices like blood sugar, in education very little is known about how much an individual's day-to-day fluctuation in performance affects their assessments. Many students have left examinations only to discover that they *could* have answered the question better – either because they misinterpreted its meaning or because they just forgot some vital links. In educational assessment, students do not usually get a second chance to show how their performance might improve, unless they are near the fail point and are asked to resit. Psychological tests (such as personality measures or attitude tests) are frequently designed to cut through this diurnal variation. By contrast, tests of competence – especially of clinical competence – are often, by necessity, one-off and anchored in the 'here and now'. For this reason it is important to ascertain as much information about the reliabilities of such tests as is possible.

Validity

In the student's final examination we would hope that the examiners, in their 11 minutes of interrogating her, actually did attempt to investigate her clinical competence, or an aspect of it. This dimension of measurement is called 'validity' – the extent to which a test or assessment measures what it is supposed to measure and not something else. In order to be valid, of course, a measurement must be highly reliable.

The validity of educational measures is complex. Validity has four main components: content validity, construct validity, criterion validity and constraints on learning. Here we will concentrate just on content validity and on constraints on learning (also called consequential validity).

Content validity

Content validity is the extent to which an assessment systematically and representatively samples what it is supposed to be measuring. To the extent that the examiners' questions in the long case reflected important aspects of the curriculum, or those laid down in the examination regulations, we would say that the assessment was content-valid. A reasonable test of this content validity would be to chart the distribution of their questions by level and topic and compare it with the weighting of the different topics in the curriculum or in course documents. If we found gastroenterology or endocrinology to be vastly overemphasised in the examination, we might question its content validity, at least as a test of general medical knowledge. Notice that to establish content validity, we require a demonstrated relationship – for example in terms of distributions of items, hierarchy of skills, congruence of time allocation – between the test content and that of the knowledge/skills at issue. Merely alluding to the appearance of the test as being a reflection of the domain being tested is called 'face validity'. Face validity is not an adequate basis on which to interpret assessments of individuals.

Constraints on learning

When our student knew she was going to be examined as part of her finals, what did she do? How did she prepare? Was a large part of her activity devoted to reading books and journal articles, or was there a major effort to examine patients and interpret findings? How were her activities constrained by what the assessment appeared, to her, to be going to test? In fact, she was more worried about the questions from the examiners after the (non-observed) patient interaction so she prepared by hitting the library. Her rationale for this preparative strategy was that the patients can only have one or two problems each, but the examiners could ask her anything! The examiners, however, may well have assumed that preparation for such an examination was done largely on acute wards and in out-patients'.

This phenomenon is known as consequential validity (see Messick, 1995). It is about the impact

https://doi.org/10.1192/apt.5.6.405 Published online by Cambridge University Press

that the assessment has on learning activity or preparatory behaviour. Any tendency to prepare in a manner that is at odds with those capabilities that are supposed to be measured by the test will tend to invalidate the test. Usually, the effects are much more insidious than in our example. Depending upon how the assessment is interpreted by students, the effects can work to promote or inhibit that which an assessor intends. For example, Wakeford & Southgate (1992) found that changing a postgraduate examination format changed preparation for the examination in a desirable direction towards more critical reading of literature.

When planning assessments care should be taken to gauge the likely effect of the assessment on the student's activities: more book-based swotting of factual knowledge, more clinical activity, more time in the interview room, or more critical reading?

Creating an assessment blueprint

Assessing clinical competence is not a simple matter. Recent thinking has recognised the need to encompass all the aspects or dimensions of competence and to do this as far as possible for the whole range of clinical conditions or problems for which a doctor will be licensed (Dauphinee *et al*, 1994). This is because:

> "Research on assessing clinical competence has shown ... that ... fulfilling the requirements of a clinical task on one problem does not provide a basis for an accurate prediction of the ability to perform a similar task on a different problem" (Newble *et al*, 1994, p. 72).

In other words, the fact that a doctor can take a history for affective disorders does not mean that he or she can take one for psychosis or for alcohol problems. Tasks such as history-taking and patient examination involve traits or generic skills less than was originally thought, and they are more anchored to the clinical context in which they are expressed.

To assist in overcoming this difficulty all assessments should be constructed according to a predetermined blueprint based on the objectives (outcomes) specified for a particular course. *After* this blueprint has been defined, the methods (simulation, viva, essay or MCQ, etc.) employed to test each objective can be chosen.

The appropriate steps in constructing this blueprint are:

(a) Identify the clinical problems the trainee should (once qualified) be able to handle to a reasonable level of resolution. It is important, in a summative or qualifying assessment, to sample these problems from the stage to which the candidate is moving, not the one from which he/she has come – for example, to identify those problems with which a new consultant/specialist registrar is expected to deal relatively independently. Hospital or regional morbidity indices, departmental audits, practice profiles and prescription counts may all be helpful here.

(b) For each problem, define the clinical tasks in which the examinee is expected to be competent. These should match fairly closely the objectives of the course. If they do not (i.e. some are not represented), choose other problems or tasks, and next year adjust your course content appropriately. These tasks can be listed or broken down into sub-units of knowledge (that might be tested in written tests), skills (in clinical assessments) and attitudes.

The 'classical' or conventional tasks such as history-taking, examination, treatment and management options, and prevention strategies will all be important. However, others might be: the responsiveness of the trainee to the patient; the efficient use of resources; or complex communication skills related to giving bad news, patient comprehension or compliance.

Some tasks, such as physical examination, might be specified relatively infrequently (for psychiatry) in the blueprint. However, it would be important to list the many occasions on which such examinations might be important for particular clinical problems, for example, neurological examination for an alcoholic, or examination of a child in a non-accidental injury case.

This process will result in a blueprint comprised of a matrix or series of matrices such as that shown in Table 1. These matrices may be quite large depending on the range of problems, the nature of the tasks and the other dimensions that might be important. Dimensions of this blueprint might include the need to be representative on common psychiatric problems, patient age, severity, organ systems, therapeutics and management, etc. In particular, concentrate on those tasks that are most critical to the successful resolution of the clinical problem, or those that examinees will be expected to perform adequately in the clinical post for which you are helping them to qualify.

https://doi.org/10.1192/apt.5.6.405 Published online by Cambridge University Press

(c) Use this blueprint to guide in the selection of tasks and methods to be included in the assessment procedure. Such selection would be like the random or representative sampling process in an epidemiological or clinical study. The aim would be to reflect the population of clinical activities and problems to which the results of the examination could be generalisable. Other issues, like selecting particular tasks that may discriminate between competent and incompetent examinees, are very secondary considerations. If the test is constructed adequately, it will automatically discriminate between good, bad and satisfactory candidates.

In summary, the most important features of the assessment of clinical competencies are to ensure that:

(a) an adequate sample is taken of those activities and competencies;
(b) sampled activities are observed or recorded appropriately; and
(c) just as in patient assessment, these recorded data are used as the basis for informed judgement about the individual.

Choosing test methods

There are 3 principles to follow in doing this:

(a) The clinical task chosen should dictate the method by which it is to be tested – essay, MCQ, short answer, objective structured clinical examination (OSCE) or long case, etc.
(b) The best test method is that with the closest representation of reality (fidelity) appropriate to the clinical tasks being posed.
(c) This choice must of course be moderated by practical constraints (e.g. the number of examiners available, and testing time).

Applying the blueprint

In our scenario, a number of methods were used to evaluate the student's clinical competence. The student's educational achievements, monitored in her final examination, were intellectual and conversational skills concerning two clinical cases and her ability to configure her knowledge into a framework that would allow a series of judgements about the truth or falsehood of short descriptive statements (MCQs). This may not have been all that impressive when compared to those methods used to evaluate her clinical condition. A crucial contribution there was the use of detailed conversations with the student to elicit the way she thought and some data from art therapy that, although not very 'objective', actually reflected ideas used by the student.

If we applied the rules above to the construction of a valid and reliable educational assessment, what our student had to do might have looked very different and the outcome might have changed. Let us start with the most simple assessment – her knowledge.

Table 1. Example of a very simple blueprint for one dimension of a problem x task matrix. Each X represents a task that is likely to be important in the investigation and resolution of each problem.

Problem	Alcohol abuse	Obsessive disorder	Bedwetting	Etc. …	Psychosis (es)
Task					
History	X	X	X		X
Physical examination					
Investigations	X		X		
Management					
Psychotherapy		X			
Etc. …	X		X		
Etc. …	X		X		X
Patient education	X		X		

https://doi.org/10.1192/apt.5.6.405 Published online by Cambridge University Press

Knowledge

Frequently, assessments of knowledge include true/false or 1 from 5 type MCQ items, or sometimes short answer questions. It should be noted that traditional true/false items have some disadvantages. They are difficult to write well; they have variable amounts of cueing; they require writers to produce an examination in which 50% of the content is wrong; conventionally they need 'negative' marking; and they lead to behaviour on the part of candidates that may sometimes be bizarre (e.g. learning 'knowledge' in MCQ format). However, they are still one of the most efficient test methods (for a given unit of testing time a wide range of content can be covered) and are machine-scorable. (For more information on how to write MCQs well and on some novel formats, there is a downloadable booklet available on the USA National Board of Medical Examiners' (NBME) website at http://www.nbme.org/new.version/item.htm (see p. 11)). Nevertheless, it is not clear exactly what traditional true/false MCQs measure (do we construe life in terms of true/false decisions?). It has been shown that some candidates guess more frequently than others. 'Guessers' generally do better. Women tend not to guess as much as men. Hence, these are factors unrelated to the test that determine outcome (Newble *et al*, 1994; Wood, 1993). In many instances, items are just reiterated from a bank that someone constructed four to five years ago. Frequently, they test knowledge in isolation. Yet cognitive research (Regehr & Norman, 1996; Dolmans *et al*, 1998) has shown contextual issues to be paramount in the retrieval and use of knowledge, especially in experts. Accordingly, test constructors are now attempting to reflect this in item design.

As a simple example, Box 1 delineates problem-/task-related basic science taken from the NBME. The authors state:

> "[Box 1] illustrates alternate approaches to posing tasks for examinees... The top item illustrates the type of test material that has given written assessment a bad name (the item is shown in a short-answer response format, but it could easily be converted to multiple choice – and it would still be a poor item: the examinee is simply required to recall an isolated piece of information)" (Case & Swanson, 1998).

By contrast, items B and C are more clinically oriented and more useful but, unfortunately, require a little more effort to mark. However, new item formats that are machine-scorable known as 'extended matching items' (EMIs) are also being developed. An example of such an item is reproduced in Box 2.

Case & Swanson (1993) give the following advice on constructing EMIs.

(a) Include items that require examinees to make clinical decisions rather than recall isolated facts.
(b) Questions need not require examinees to provide essay-style justifications.
(c) Focus items on key features of clinical decision-making situations.
(d) Use clinical vignettes rich in undigested patient findings.
(e) Specify the number of responses required for each vignette.
(f) Multiple choice forms (extended matching) are better than free response.
(g) In option lists, provide all the relevant options that examinees might make, although many more than five are ideal, and the number of options can vary from one set of items to the next.

Further information about writing EMIs and other types of items is available on the NBME web site (see earlier).

Clinical skills

Our student encountered what has become, worldwide, the standard way of assessing clinical skills –

Box 1. Alternate examinee tasks – recalling isolated facts *versus* making decisions in context *versus* justifying decisions

A Which nerve enervates the triceps muscle?

B A 20-year-old man is stabbed in the arm with a knife. There is anaesthesia of the dorsum of the forearm and the dorsum of the hand between the thumb and forefinger. The extensors of his wrist are paralysed, and he cannot extend his thumb at the metacarpophalangeal or interphalangeal joints. Which nerve has been damaged?

C A 20-year-old man is stabbed in the arm with a knife. There is anaesthesia of the dorsum of the forearm and the dorsum of the hand between the thumb and forefinger. The extensors of his wrist are paralysed, and he cannot extend his thumb at the metacarpophalangeal or interphalangeal joints. Which nerve has been damaged? Justify your answer.

https://doi.org/10.1192/apt.5.6.405 Published online by Cambridge University Press

the long case; a form of viva examination. Unfortunately, unstructured oral examinations, as practised in any discipline let alone in most UK medical schools, are unreliable and of doubtful validity (Wiersma & Jurs, 1990, pp. 196 & 216; Van der Vleuten *et al*, 1994, pp. 112–114). It is not clear what capacities are being measured in an oral, it is too short to be reproducible, it is clouded by examiner effects (each student gets different examiners, different patients and different and unstructured questions), and candidates frequently find it unacceptable for psychological reasons.

It should also be a matter of extreme concern to examiners that the most unreliable methods, namely unstructured vivas, are the ones frequently chosen in the UK to discriminate for honours or pass/fail decisions.

There is another drawback to the long case. Note that in Table 1 it can quickly be seen that not every clinical task is either possible or relevant in every case. This immediately throws into doubt the assumption that the long case assessment is capable of demonstrating a student's capacity to 'put it all together'. There will be many cases in which it is just plain impossible or rather silly to engage in some aspects of the clinical process.

Most postgraduate examinations have adopted orals structured around examinees' practice activities (Feletti *et al*, 1994, pp. 156–157), usually collected in a logbook. These seem generally to be a more dependable method of assessment than the unstructured oral. There is no reason why undergraduates could not also keep a log of their last 10 patients for their examinations.

Alternatively, an attempt can be made to structure the long case. This might be more appropriate in psychiatry where, traditionally, patients are interviewed over a lengthy period, and hence the psychiatrist's skill might not be apparent from a brief sample of behaviour.

(a) The case chosen should fit comfortably into the examination blueprint.
(b) A standard scoring schedule for the tasks involved should be drawn up and agreed before the examination. It helps if these schedules include descriptions of candidate activity across a range of abilities from outstanding to bad. External examiners should be made fully aware of these schedules. Examiners should mark the case according to this schedule, independently completing all sections. An example of such a scoring schedule is given in Table 2. Further examples are given in Jolly & Grant (1998).
(c) During the candidate–patient interaction, examiners should observe only. Intervention should be undertaken only if the patient's

Box 2. Extended matching item on the theme of fatigue

A	Acute leukaemia	H	Hereditary spherocytosis
B	Anaemia of chronic disease	I	Hypothyroidism
C	Congestive heart failure	J	Iron deficiency
D	Depression	K	Lyme disease
E	Epstein–Barr virus	L	Microangiopathic haemolytic anaemia
F	Folate deficiency	M	Miliary tuberculosis
G	Vitamin B_{12} deficiency	N	Glucose 6-phosphate dehydrogenase deficiency

For each patient with fatigue, select the most likely diagnosis. Each option can be used once, more than once or not at all

Questions

1. A 19-year-old woman has had fatigue, fever and sore throat for the past week. She has a temperature of 38.3°C, cervical lymphadenopathy and splenomegaly. Initial laboratory studies show a leucocyte count of 5×10^9/l. Serum aspartate aminotransferase activity is 200 U/l. Serum bilirubin concentration and serum alkaline phosphatase activity are within normal limits.

2. A 15-year-old female has a two-week history of fatigue with back pain. She has widespread bruising, pallor, and tenderness over the vertebrae and both femurs. Complete blood count shows a haemoglobin concentration of 7.0 g/dl, a leucocyte count of 2×10^9/l and a platelet count of 15×10^9/l.

https://doi.org/10.1192/apt.5.6.405 Published online by Cambridge University Press

well-being is at risk. At the end of the candidate's activity, examiners should confer about what the candidate did, focusing on apparent strengths and weaknesses. The candidate's actual performance in relation to the patient should form the focus of this discussion, not other incidental material or past experiences with the candidate.

(d) Irrespective of what experienced examiners may think, the examination of content (knowledge alone) in the clinical long case is an activity that can introduce error. The long case is designed to measure clinical competence, not knowledge *per se*. If the investigation of content areas is appropriate or essential for a case, all candidates seeing any one case should be asked a standard set of questions about the case.

(e) Long cases should preferably be used with other larger samples of behaviour (see below), and the scores generated combined with these.

(f) A pool of patients of equivalent difficulty, or standardised patients, should be used. It is important to keep in mind the notion of fairness when choosing different patients.

Another approach at undergraduate level has been the use of the OSCE.

Objective structured clinical examination

The OSCE is not actually a method of examination in the way that MCQs and essays are. It is better viewed as a framework for sampling essential clinical tasks. Each clinical task in the OSCE is called a station – in the sense of a stop on a route – and the OSCE is more suitably referred to as a multiple-station examination. In the UK, the Royal

Table 2. Example of a scoring schedule for long case

The case and task for the candidate

This 56-year-old factory worker has been referred by his doctor for persistent haemoptasis over the last three weeks. He is a heavy smoker (approximately 40 per day, and has been smoking for the past 42 years). Take a full history and perform a clinical examination of all relevant systems. Be prepared to discuss with the examiners common and rarer differential diagnoses, and prepare in your own mind how you would communicate with the patient the likely diagnosis and probable management and prognosis for this.

Examiner rating schedules (examples only)

The doctor–patient consultation

Tasks	Poor	Excellent
1. Introduction and establishment of rapport	Does not introduce self, does not verify patients name and/or number and is brusque No rapport established Appears disinterested	Introduces self by name, uses appropriate patient name(s) Is friendly and/or caring Establishes excellent rapport and patient confidence
2. History of presenting complaint	Collects inadequate data or misses vital information such as date of onset, duration, quantity of blood, colour of sputum, etc.	Thoroughly charts progress and development of symptoms, acquiring all essential data and critical dates.
3. Social history	Ignores, inadequately follows up, or is judgmental about smoking/alcohol consumption. Does not inquire about job, hours, stress, sexual/emotional circumstances, children etc.	Obtains in non-threatening manner all relevant social history details including, cigarette consumption, alcohol history, family circumstances, etc.

Other tasks as necessary, e.g. examination of CVS/respiratory system

Reproduced, with permission, from Jolly & Grant (1998)

https://doi.org/10.1192/apt.5.6.405 Published online by Cambridge University Press

College of General Practitioners, the Royal College of Anaesthetists and the Royal College of Surgeons have all implemented, or are experimenting with, multiple-station examinations. Each station has a pre-defined task, a scoring schedule and, usually, controlled patient characteristics, either through the use of simulated patients or 'primed' real patients. Research has shown that to be reliable, OSCEs should be of at least two hours' duration and contain 20–25 stations. However, the unreliability of clinical assessment methods comes from the unpredictability of performance by candidates across different clinical cases and *not* from the frameworks themselves. Nevertheless, it is obvious under these circumstances that some frameworks will be more reliable than others. For example, the OSCE (many cases and examiners) has been shown to be much more reliable than the unitary long case (one case and two examiners) or traditional oral examinations (two examiners). The reason is simple: the OSCE framework samples more aspects of behaviour and more cases at the same time – hence, error due to sampling is low and also, as a by-product, content validity is high. Research has also shown that comparatively little is gained in OSCEs by having two examiners per station, probably because the cases are standardised and because examiners are trained and provided with appropriate scoring schedules. It is better to have more stations and one examiner per station, than half the number of stations with two observers.

In summary, the idea behind the OSCE is that it should sample, in a standardised manner, a broad range of clinical skills, most of which are observed by examiners in an 'objective' manner. It should not be used for testing those blueprint components more easily tested by other methods – for example, identification of pathological specimens or lab data interpretation – but should focus on clinical activity.

An OSCE has been used in assessment of psychiatric skills (Hodges *et al*, 1998). There is, however, a compelling argument that where detailed psychiatric history-taking is concerned, attempting an OSCE, in its traditional 5–10 minute station format, may be difficult. For example, candidates might need to spend three days taking histories! Also, an OSCE format might have poor consequential validity – students might be lulled into condensing the psychiatric interview into shorter chunks in order to prepare for the examination. But the important factor is to sample as broadly as possible, over many cases. Hence, this may be a good rationale for using more job-related performance assessment, or it may be that in training, assessments must figure more prominently to accomplish appropriate sampling (Jolly *et al*, 1994).

Assessing the student – resumé

Reflecting on the assessment of our student, we can see that her final examinations left a great deal to be desired, both in the sampling of her knowledge and skills, and in the standardisation of the process. Probably the most noticeable difference between the assessment of her psychological state and her clinical competence was the amount of data that was gathered in the former, the time spent doing it and the depth of discussion involved. The finals examiners may also have been assessing her attitudes, but we have no insight into that aspect of the examination, or the other bits of information that might have been used to make the final decision. There might be an argument for suggesting that if this student were to be assessed appropriately, she would suffer from stress related to *over-assessment*. Such stress usually occurs when students have multiple competing tasks to accomplish, when they do not get adequate feedback on their progress and when the stakes of the assessments are very high.

Recently, at the other end of the scale from our student, the General Medical Council has developed an assessment procedure for poorly performing doctors needing remedial training, or in danger of being removed from the register, that includes:

- an extended personal portfolio of career development;
- interviews with 13 colleagues or associates;
- a review of between 50 and 150 of the doctor's medical records;
- a case-based structured viva;
- observation of actual practice;
- a structured interview; and
- clinical skills and knowledge testing .

Perhaps if assessment were so thorough at an earlier stage, such procedures would not be in such demand. For those readers wanting more information on the assessment methods mentioned in this article and some that were not, but that might still prove useful, Jolly & Grant (1998) is a useful source.

References

Boud, D. (1997) Assessment and learning: contradictory or complementary. In *Assessment for Learning in Higher Education*, pp. 35–48. London: Kogan Page.

Case, S. & Swanson, D. (1993) *Item Writing Workshop (Course Manual)*. Philadelphia, PA: National Board of Medical Examiners.

https://doi.org/10.1192/apt.5.6.405 Published online by Cambridge University Press

—— & —— (1998) *Constructing Written Test Questions for the Basic and Clinical Sciences*. Philadelphia, PA: National Board of Medical Examiners.

Dauphinee, D., Fabb, W., Jolly, B., *et al* (1994) Determining the content of certifying examinations. In *The Certification and Recertification of Doctors: Issues in the Assessment of Clinical Competence* (eds D. I. Newble, B. C. Jolly & R. E. Wakeford), pp. 92–104. Cambridge: Cambridge University Press.

Dolmans, D. H., Wolfhagen, I. H. & Van Der Vleuten, C. P. (1998) Motivational and cognitive processes influencing tutorial groups. *Academic Medicine*, **73** (suppl.), S22–S24.

Feletti, G., Cameron, D., Dawson-Saunders, B., *et al* (1994) In-training assessment. In *The Certification and Recertification of Doctors: Issues in the Assessment of Clinical Competence* (eds D. I. Newble, B. C. Jolly & R. E. Wakeford), pp. 151–166. Cambridge: Cambridge University Press.

Hodges, B., Regehr, G., Hanson, M., *et al* (1998) Validation of an objective structured clinical examination in psychiatry. *Academic Medicine*, **73**, 910–912.

Jolly, B. C. & Grant, J. (1998) *The Good Assessment Guide*. London: Joint Centre for Education in Medicine.

——, Newble, D. I. & Wakeford, R. E. (1994) Requirements for action and research in certification and recertification. In *The Certification and Recertification of Doctors: Issues in the Assessment of Clinical Competence* (eds D. I. Newble, B. C. Jolly & R. E. Wakeford), p. 241. Cambridge: Cambridge University Press.

Messick, S. (1995) Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, **14**, 5–8.

Newble, D. I., Dauphinee, D., Dawson-Saunders, B., *et al* (1994) Guidelines for the development of effective and efficient procedures for the assessment of clinical competence. In *The Certification and Recertification of Doctors: Issues in the Assessment of Clinical Competence* (eds D. I. Newble, B. C. Jolly & R. E. Wakeford), pp. 69–91. Cambridge: Cambridge University Press.

Regehr, G. & Norman, G. R. (1996) Issues in cognitive psychology: implications for professional education. *Academic Medicine*, **71**, 988–1001.

Van Der Vleuten, C. Newble, D. I., Case, S., *et al* (1994) Methods of assessment in certification. In *The Certification and Recertification of Doctors: Issues in the Assessment of Clinical Competence* (eds D. I. Newble, B. C. Jolly & R. E. Wakeford), pp. 105–125 Cambridge: Cambridge University Press.

Wakeford, R. E. & Southgate, L. (1992) Postgraduate medical education: modifying trainees study approaches by changing the examination. *Teaching & Learning in Medicine*, **4**, 210–213.

Wiersma, W. & Jurs, S. G. (1990) *Educational Measurement and Testing* (2nd edn). Boston, MA: Allyn and Bacon.

Wood, R. (1993) *Assessment and Testing: A Survey of Research*. Cambridge: Cambridge University Press.

https://doi.org/10.1192/apt.5.6.405 Published online by Cambridge University Press