Assessment in medical education: Evolving perspectives and contemporary trends

RITA SOOD, TEJINDER SINGH

ABSTRACT
Assessment is an important part of the educational process. It influences students’ learning. Traditionally, assessment has been used to measure the achievement of students for certification or selection. However, increasing attention is being paid to using formative assessment to improve learning. Miller’s pyramid provides a good conceptual model to assess clinical competence by providing tiered levels of assessment. All levels of the pyramid need to be included in the assessment process.

To be useful, assessment should be valid (measuring what it is supposed to measure), reliable (confidence in the results), acceptable (to various stakeholders), feasible and have a positive educational impact. The major attributes (validity and reliability) refer to judgements that we make from the assessment data and are not the inherent property of any test or tool. All validity is construct validity. There is often a trade-off between validity and reliability but an assessment which is low on one can still be useful by virtue of its being high on the other. No one tool is enough for assessing students and a combination of tools is preferred to get a composite picture of students’ attainment.

Assessment of knowledge is largely made by written assessment. The tests and questions should be contextual and look for application of knowledge rather than mere recall of facts. Assessment of clinical skills is done by traditional methods such as long and short cases and newer methods such as mini-clinical evaluation exercises, objective-structured clinical examinations, case-based discussions and portfolios. An adequate and representative sample of clinical tasks and direct observation of performance are the key to the validity and reliability of the assessment process. Faculty training is important for improving the quality of assessments.


INTRODUCTION
‘Examinations drive students’ learning’ describes one of the strongest relationships in education.1 Of all the different components of a medical education programme, the assessment strategies direct and influence the way students learn. A well-designed assessment sets clear expectations, establishes a reasonable workload and provides opportunities for students to self-monitor, rehearse, practise and receive feedback. Conversely, poorly designed assessments can mar the quality of learning.2 It is useful to view education as a set of components which interact with each other to produce the desired effect on student learning. These components are learning objectives, teaching methodology and assessment. All three components have a high degree of interdependence and one cannot sustain without the other.3

For academic staff, assessment is often the final consideration in their planning of the curriculum.4 Their primary concerns are designing learning outcomes and planning teaching and learning activities that will produce these outcomes. How assessment will be done is often considered once other decisions about the curriculum have been made. For most students, however, the curriculum is literally defined by the requirements of assessment. Students often work ‘backwards’ through the curriculum, focusing first and foremost on how they will be assessed and what they will be required to demonstrate.4 Therefore, recognizing the potent effects of assessment requirements on the study habits of students and capitalizing on the capacity of assessment for creating preferred patterns of study is a powerful means of re-conceptualizing and repositioning the use of assessment.

Assessment provides students with short-term goals, clarifies the tasks to be learned and provides feedback about learning.5 Since assessments tend to direct students’ learning efforts towards the intended learning outcomes (ILOs), they can be used as tools for increasing the transfer and retention of learning. Generally, learning outcomes at the level of understanding, interpretation and application are likely to be retained longer and have greater transfer value than outcomes at the level of recall.6

PURPOSES OF ASSESSMENT
Rowantree7 suggested six purposes of assessment: Selection of candidates for educational opportunity; maintenance of standards, particularly in relation to the final output from the system; motivation of students so that they are encouraged to learn; provision of feedback to students so that they know how they are performing; provision of feedback to teachers so that they know the strengths and weaknesses of their students’ learning (and their teaching); and preparation of the students for ‘real life’. While most teachers are well-versed with the summative or certifying purpose of assessment (assessment of learning), using assessment as an educational tool (assessment for learning) is a relatively recent phenomenon.8

EVOLUTION OF ASSESSMENT
There have been many distinct phases in the evolution of assessment over the past 50 years.6,9,10 Initially, educational testing focused on
measurement alone, with emphasis on objectivity and reproducibility of results.\textsuperscript{9,11} The focus then shifted to reliability, i.e. the degree of confidence that we can place on our results.\textsuperscript{12} The wheel of time seems to have gone full circle and now subjective assessments by experts are no longer viewed as being unreliable.\textsuperscript{12}

This period of innovation left us some important legacies.\textsuperscript{13} These are: the continuing sensitivity with regard to issues concerning what it is that a particular test really measures; recognition that competence is not a simple aggregate of homogeneous performances, but is multidimensional; willingness to admit that certain concepts such as generalized problem-solving ability may be convenient but misleading; and an ongoing concern with the impact of tests on student learning and the effects of different types of examinations on students’ priorities and methods of study.\textsuperscript{13}

Contemporary developments in assessment are based on more integrative concepts, in which the prominent features are not traits but roles or competences.\textsuperscript{14} The key element in this is the successful completion of a certain task or role, for which different aspects of medical competence have to come together and be integrated.\textsuperscript{15,16} Miller’s pyramid marked the beginning of this thinking.\textsuperscript{17} Miller introduced a conceptual framework in the form of a pyramid (Fig. 1) wherein various layers of the pyramid are defined not as traits but as verbs or actions, which are observable and can be judged and thus used for assessment. These are—'knows’ (factual knowledge), ‘knows how’ (analysis, application and interpretation of knowledge), ‘shows how’ (actual application and practical demonstration in a simulated situation) and ‘does’ (perform in real situations).

Clinical competence is defined as the ability to assume a combination of well-defined roles.\textsuperscript{18} These roles are: provider of direct patient care, worker in the healthcare system, scientist, educator and person.\textsuperscript{18} In contemporary assessment programmes, various instruments are used to obtain information about a student’s competence in each of those roles. The ‘one instrument for one trait’ approach has now become a ‘multi-instrument for multiple roles’ approach.\textsuperscript{19} The growing interest in quality improvement, bolstered by increasing demands for public accountability, has shifted the focus to assessment of work.\textsuperscript{20} The assessment of actual performance in practice is essential to quality management. This is what is referred to as ‘performance’ or ‘work-based assessment’.\textsuperscript{20}

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{Fig1.png}
\caption{Miller’s pyramid\textsuperscript{17}}
\end{figure}

\textbf{EFFECTIVE STUDENT ASSESSMENT}

The following guidelines provide a general framework for effective student assessment:\textsuperscript{3}

1. Effective assessment requires a clear conception of all intended learning outcomes.
2. It requires a variety of assessment procedures to be used.
3. The instructional relevance of the procedures is considered.
4. It requires an adequate sample of student performance.
5. It requires that the procedures be fair to everyone.
6. The criteria for judging successful performance need to be specified.
7. It requires feedback to students that emphasizes strengths of performance and weaknesses to be corrected.
8. It must be supported by a comprehensive grading and reporting system.

\textbf{ATTRIBUTES OF A GOOD ASSESSMENT}

Van der Vleuten suggested that the utility of assessment methods could be conceptualized by looking at the five attributes of the method—reliability, validity, educational impact, feasibility and acceptability.\textsuperscript{14} His argument was to carefully balance the compromise between these five criteria, rather than focus on any one of them. The utility is expressed as a product of all the five attributes. This conceptual model helps to decide on the appropriate tool. A tool may be low on reliability but can be useful if it is high on another attribute, say, educational impact.

\textbf{Validity}

Validity is concerned with the interpretation and use of assessment results. It refers to the evidence presented to support or refute the meaning or interpretation assigned to the results of assessment.\textsuperscript{21} The traditional view that there were several different types of validity has been replaced by the view that validity is a unitary concept; it is a matter of degree and is inferred from various forms of evidence.\textsuperscript{21} It needs to be emphasized here that validity refers to interpretation of results rather than being an inherent quality of the tool.

Reliability is now considered a part of validity, although on its own reliability may not be sufficient to make an assessment valid. Messick states that all validity is in fact construct validity.\textsuperscript{22} It follows that construct irrelevance (e.g. asking a theory question in an objective-structured clinical examination [OSCE]) or construct under-representation (e.g. not observing the student during history taking) are the biggest threats to validity.

\textbf{Reliability}

As for other experimental data, all assessment data must be reproducible in order to be interpreted meaningfully. Would the same results be obtained if a different sample of the same type of task were used, or if the assessment were at a different time? If a performance assessment is being rated, would different assessors rate the performance in the same manner? Reliability refers to the consistency or reproducibility of assessment results over time or instances.\textsuperscript{22} Like validity, reliability is a characteristic of the result or outcome of the assessment and not the measuring instrument itself.

The factors that lower the reliability of test scores (written assessments) include too few test items, too easy or too difficult items, a narrow range of scores, lack of objectivity in scoring and suboptimal testing conditions.\textsuperscript{1} In performance assessments, the greatest threat to reliable measurement is case specificity.
Increasing the tasks used to assess a student is therefore a very useful way of improving both validity and reliability.

The reliability must be high (0.80–0.89) in high-stakes examinations (e.g. certifying or selection examinations). For assessments with lower consequences (e.g. formative assessments), a lower reliability (0.70–0.79 or lower) may be acceptable. 23

**Educational impact**

It is known that students are strongly influenced by assessment. 24 Assessment should help to guide students towards better learning. If any assessment method promotes unhealthy learning or encourages shortcuts in learning, then it may be low on this attribute of validity. We are all aware of the distorted learning as a result of postgraduate entrance examinations in India. 25,26

**Practicability or feasibility**

Feasibility is a relative term. Theoretically, every type of test should be feasible. However, the time and effort involved in developing, administering, scoring, interpreting and reporting a test needs to justify its use. There will always be constraints on the resources available to conduct assessments. Expertise and creativity are required to develop the best compromise between the ideal and the practical. Newble stressed that most medical schools will have to be prepared to spend more time and resources on their assessments if they are to achieve minimally acceptable standards of validity and reliability. 26 This applies particularly to the assessment of clinical skills, where much longer or more frequent observations of student performance are required than for other assessments.

**ASSESSING CLINICAL COMPETENCE**

A proper definition of clinical competence and its components is important to serve as a criterion for validating medical educational programmes and to assure a minimum level of competence at the end of medical school and during residency.

Newble defined clinical competence in terms of what a student or doctor should be able to do at an expected level of achievement (e.g. at the beginning of an internship) and clinical performance as what a student or doctor actually does in real clinical practice. 27 Epstein and Hundert proposed a definition of professional competence as the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values and reflection in daily practice for the benefit of the individual and community being served. 28 Competence builds on a foundation of basic clinical skills, scientific knowledge and moral development.

The American Board of Internal Medicine distinguished between the four different dimensions of clinical competence, where problem solving was the core aspect. 29 These included abilities (i.e. knowledge, technical skills and interpersonal skills), problem-solving skills (i.e. data gathering and diagnosis), the nature of the medical illness (problems encountered by physicians) and social and psychological aspects of the patients’ problems. In a later report, more elements were added—communication skills, professionalism (e.g. ethical practice, understanding diversity, responsible attitude), and system-based practice (i.e. understanding of the healthcare system to improve and optimize healthcare). 30

In the USA, the assessment of residents, and increasingly of students as well, is largely based on a model that was developed by the Accreditation Council for Graduate Medical Education (ACGME). 31 This model uses six areas of competence and some means of assessing them (ACGME Outcome Project 2000). These are—medical knowledge, patient care (including clinical reasoning), professionalism, interpersonal and communication skills, practice-based learning and improvement (including information management) and systems-based practice (including health economics and teamwork). 31

In India, the Medical Council of India provides broad guidelines and national and institutional goals for the medical graduate training programme (Regulations on Graduate Medical Education 1997). 32 This document delineates the objectives of training in various disciplines and skills to be learnt during the course, including internship. It also gives broad guidelines for teaching–learning methods and an assessment plan. The document is in the process of revision. There is a call for making assessments more reliable and skill-based. 33

With this background, we now discuss some of the methods used for assessment in medical education.

**METHODS OF ASSESSMENT**

All methods of assessment have strengths as well as limitations. 34 Some commonly used methods of assessment in medical education along with their strengths and weaknesses are described below. The role of an assessment planner is to use different methods such that the benefits can be maximized.

**Written assessments**

Written examination questions are typically classified according to whether they are the selection type, e.g. multiple choice questions (MCQ) or the supply type, e.g. short answers or traditional essays. They can also be ‘context-rich’ or ‘context-poor’. 35 Questions with rich descriptions of the clinical context invite more complex cognitive processes that are characteristic of clinical practice. On the other hand, context-poor questions can test factual knowledge but not its transferability to real clinical problems. 36

**MCQs**

The multiple choice test is a flexible assessment format that can be used to measure knowledge, abilities, values, thinking skills, etc. Such a test usually consists of a number of items that pose a question to which students must select an answer from among a number of choices. Items can also be statements to which students must find the best completion.

Various formats of MCQs have been in use. The formats that ask the students to choose the best answer from a list of possible answers are most commonly used (‘single or one best option’ type). The other formats are ‘true or false’, ‘multiple true or false’, ‘matching’ and the ‘extended matching’ type questions.

MCQs can provide a large number of examination items that encompass many content areas, so a broad domain can be covered. They can be administered in a relatively short period, can be graded by a computer and have a high reliability per hour of testing time. MCQs that are rich in context (problem-based) are difficult to write but can become very effective tests for assessing higher levels of knowledge. The reliability of an MCQ test is a consequence of a wider sampling of content. 35,37 Guidelines regarding the content, formatting, style and writing the stem and distractors for construction of MCQs need to be followed. 38

‘True or false’ questions. The main advantage of such questions is that they can be written easily and cover a wide range of content. However, there are two major disadvantages. They are difficult to construct flawlessly—the statements have to be absolutely true or false. When a student answers a ‘false’ question correctly, we can only conclude that the student knew the statement was false, not
that he or she knew the correct fact. Also, they are weak in discriminating between high and low performers. Their place in assessment has been questioned.38

‘Multiple true or false’ questions. In this format, a question with more than one correct answer can be asked. These take slightly longer to answer than the previously discussed formats. The scoring procedure too can be complicated.

Newer formats of MCQs. Certain newer formats of MCQs allow for a better assessment of diagnostic reasoning. The ‘key feature’ question focuses on critical decisions in particular clinical cases.46 These questions are best used for testing the application of knowledge and problem solving in ‘high stakes’ examinations. The ‘extended matching items’ consist of several questions, all with the same long list of possible answers. The elements of extended matching questions are a list of options, a ‘lead-in’ question, and some case descriptions or vignettes.41 An option may be correct for more than one vignette, and some options may not apply to any of the vignettes. This format is in use by the United States Medical Licensure Examination (USMLE) for many years.

‘Short answer’ questions

A short answer question (SAQ) is similar to a well-stated MCQ without the alternatives. Here, the assessees are required to provide rather than select the answer. These open-ended questions are more flexible in that they can test issues that require creativity and spontaneity. However, these have a lower reliability.37 Because answering open-ended SAQs is much more time-consuming than answering MCQs, they are less suitable for broad sampling. A good open-ended question should include a detailed answer key for the person marking the paper.3

Essay questions

Essays are ideal for assessing how well students can summarize, hypothesize, find relations and apply known procedures to new situations. These can provide insights into more complex cognitive processes, ability to process information, different aspects of the ability to write and more contextualized answers.35

When constructing essay questions, it is essential to define the criteria on which the answers will be judged. Essay-type questions are time-consuming to grade and need more work to establish inter-rater reliability. As answering them is time-consuming, a long testing time is required to include a variety of domains. This may often result in a smaller sample, thereby limiting reliability. When clear grading guidelines are in place, structured essays can be psychometrically robust.

Newer formats

The other methods of written assessment that have been used in the past include modified essay questions (MEQs) and patient management problems (PMPs) for testing problem-solving and decision-making.46 A MEQ presents a typical problem faced in daily practice. The use of both MEQs and PMPs encourage more complex thinking skills among students and help develop problem-solving skills. They are especially suitable for emergency or high-risk problems where a student cannot be allowed to handle real problems.

Assessments by supervising clinicians

Observation by supervising clinicians and their impression of students over a specific period remains the most common tool used to evaluate performance.

Long case examination

This has been in use for a long time, is highly valued by many clinicians and closely resembles some of the tasks doctors undertake in their everyday practice.44 The format of the long case can vary; traditionally a student is given an uninterrupted and unobserved time, usually 30–45 minutes, to interview and examine a real patient. The student then presents his/her findings and plan to one or more examiners who then conduct an unstructured oral examination with the student about the patient’s problem and related topics. Given the logistics of long case examinations, medical schools traditionally assess students on a single case. This limits the generalizability of the results.44 Also, in long cases, differences among cases appear to contribute more to variation in scores than differences among examiners.44 The traditional long case examination has been almost abandoned in most of the western world.44 However, it is still widely used in India in almost all clinical assessments of undergraduates and postgraduates.46,47

The reliability of the long case can be improved by increasing the number of cases seen by each student.39 Using the generalizability theory, Wass et al. predicted that examining a student on 10 cases would achieve a reliability factor of 0.8.19 Another problem in the long case is the unstandardized nature of the patients. Standardizing patients enables accurate blueprinting of the test. However, content specificity appears to be more important than standardization.39

Observation is another strategy to improve the long case examination.49,49 Rather than relying on the presentation alone, observation while the candidate is eliciting the history and carrying out the physical examination would be a more valid assessment of the candidate’s competence.

Gleeson developed a more structured presentation of an unobserved long case, the objective-structured long examination record (OSLER), which includes structuring of the long case and direct observation of the candidate interacting with the patient for a small component, e.g. explaining a procedure.39 Gleeson suggested that reliability would also improve if the number of judgements within the case are increased. OSLER seems to be a powerful tool for providing feedback and therefore has great potential to increase clinical competence.51,52 A study by Gleeson indicated that OSLER had a profound effect on student learning in all areas of competence.53 Wass and van der Vleuten made an argument for the retention of and further research into the long case clinical examination.43 Its utility for high-stake examinations, however, remains questionable.43

Objective-structured clinical examination

As a potential solution to the difficulties of adequate sampling and standardization of cases, the OSCE has gained importance worldwide. Candidates rotate through a series of timed stations. All candidates move from station to station in sequence on the same schedule. OSCEs have used standardized patients as the primary assessment tool. However, other assessment items such as data interpretation exercises using clinical cases and clinical scenarios with mannequins have also been used to assess technical skills. The observing faculty or tutor or patient uses either a checklist of specific behaviours or a global rating form to evaluate the student’s performance.34,35 OSCEs have been used in most British and American medical schools, many residency programmes and by the licensure boards of Canada and USA for many years. This format provides a standardized means to assess a variety of clinical skills. These include physical examination and history-taking skills, communication skills with patients and
family members, breadth and depth of knowledge, ability to summarize and document findings, ability to make a differential diagnosis or plan treatment and clinical judgement based upon patients’ notes. A separate performance score is derived for each task performed at a station and scores are combined across stations or tasks to determine a pass/fail score.

Like MCQs, the reliability of the OSCE stems from its wider content sampling rather than from its objectivity or standardization. A minimum of 10 stations, which the student usually visits over the course of 3–4 hours, is necessary to achieve a reliability of 0.85–0.90.28 There is increasing evidence that global ratings, especially by physicians, are as reliable as checklists.54,55 However, extensive training of judges is required to ensure consistency.

OSCEs are very useful for measuring specific clinical skills and abilities. The performance of a task is actually observed, thus improving the validity of interpretation. Immediate feedback is possible, thus helping improvement in learning. OCSEs can also provide feedback to teachers and help in correcting teaching-learning errors.77 However, they also have weaknesses. The examination format is labour-intensive and expensive.59 We found OSCEs to be unsustainable on a regular basis due to the lack of resources and faculty time.68 It can also be a challenge when real patients are used as they can get difficult when the same questions are asked repeatedly, and standardization is lost.68 Complex skills, requiring an integrated professional judgement, become fragmented by the relatively short station length (generally 5–10 minutes), thereby leading to loss of validity at the cost of reliability.61 Aspects of competence such as the ability to perform procedures and manage life-threatening clinical situations or abilities in continuity of care cannot be tested using OSCEs.62

Contrary to popular notion, the high reliability of OSCEs do not depend on their objectivity or structure. It is the wider sampling of content and skills which help in achieving high reliability. In fact, the reliability of a one-hour OSCE and a one hour-long case is almost the same.63 An important implication of this is to use OSCE for wider sampling rather than keeping only a few stations, and depend on its objectivity to give a high reliability.77

**Multisource or 360-degree assessments**

These evaluations consist of measurement tools completed by several people in a person’s sphere of influence. Evaluators completing rating forms in a 360-degree evaluation are usually peers, other members of the clinical team and patients, who can provide insights into the trainee’s work habits, capacity for teamwork, and interpersonal sensitivity.64,65 A multisource feedback is most effective when it includes narrative comments as well as statistical data, when the sources are recognized as credible, when the feedback is framed constructively and when the entire process is accompanied by good monitoring and follow-up.66 Peer assessments depend on trust and require scrupulous attention to confidentiality, failing which they can be undermining, destructive and divisive.

The strengths of multisource assessments are that ratings generally encompass habitual behaviours rated by credible sources, and assessments often correlate with future academic and clinical performance. However, in most clinical settings, conducting 360-degree assessments with large numbers of evaluators, compiling and reporting confidentially to students/residents is challenging, though an electronic system may make it feasible.

**Portfolio**

The portfolio is a framework containing evidence of achievement of learning outcomes over time. Portfolios have been used to support learning and for assessment purposes, both formative and summative.69 Reflecting upon what has been learnt forms an important part of constructing a portfolio.69

A portfolio typically contains written documents but can include video or audio recordings, photographs and multimedia, and can be maintained in an electronic format. In a graduate medical education, a portfolio might include a log of clinical procedures performed, a summary of research literature reviewed when selecting a treatment option, ethical dilemmas faced and how they were handled, etc. Portfolio assessment is intimately linked to self-directed learning and is most useful for evaluating mastering of competences that are difficult to evaluate otherwise, such as practice-based improvement, use of scientific evidence in patient care, professional behaviour and patient advocacy. Cole asserts that portfolios can differ in purpose, components and processes.70

Portfolios are excellent tools for assisting formative assessment and professional development. They have the potential to assess clinical performance over a period of time, constituting one form of authentic assessment.80 However, assessment through portfolios is labour-intensive and requires staff development. This is also cumbersome for comparative assessments as they are essentially non-standardized.

**Chart-stimulated recall oral examination (CSR)/case-based discussions (CbDs)**

In a chart-stimulated recall oral examination (CSR) examination, patients of the examinee are assessed in a standardized oral examination. A trained and experienced physician examiner questions the examinee about the care provided, probing for reasons behind the work-up, diagnosis, interpretation of clinical findings and treatment plans. The examiner rates the examinee using a well-established protocol and scoring procedure.73 CSRs are also called case-based discussions (CbDs) in some settings. A trainee should be assessed in this manner at least six times in a year. These are different from the traditional case presentations as the discussion centres on what the trainee has already done.74 Thus, they test different skills compared to traditional case presentations.

**Checklist evaluation**

Checklists consist of essential or desired specific behaviour, activities or steps that make up a more complex competency. Typical response options on these form a check or ‘yes’ to indicate that the behaviour occurred or options to indicate the completeness or correctness of the action. The forms provide information about behaviour, but for the purpose of making a judgement about the adequacy of the overall performance, standards need to be set that indicate different levels of performance.39 Checklists are useful for evaluating any competency or its component that can be broken down into specific behaviours or actions. However, their use is limited by the fact that checklists assume a fixed sequence of actions which is not often the case in reality. There may be different valid sequences to perform the same task effectively. To ensure the validity of content and scoring rules, checklist development requires consensus by several experts with agreement on essential behaviour, sequencing and criteria for evaluating performance.75

**Procedure, operative or case logs**

Procedure, operative or case logs document each patient encountered by the student for the medical conditions seen and procedures or surgical operations performed. Patient case logs
involves recording of some number of consecutive cases in a designated time frame. Case and/or procedure logs are useful for determining the scope of patient-care experience.

**Patient surveys**

Patient surveys to assess satisfaction with the hospital or clinic visits typically include questions about the physician’s care. The questions pertain to general aspects of physicians’ care such as time spent with the patient, overall quality of care, competence of the physician, courtesy, interest and empathy shown. Specific aspects of care would include listening skills, explanation about the problem, its diagnostic tests, treatment planned and side-effects of drugs. A patient survey asks the patients to rate their satisfaction using qualitative rating categories or agreements with statements describing the scenario. Each rating is given a value and the final satisfaction score is calculated by averaging scores across responses to generate a single overall score.

Patient surveys can be used for formative assessments and performance improvement. The American Board of Internal Medicine reports that 20–40 patient responses were needed to obtain a reliability of 0.70–0.82 on individual resident ratings.

**Record review**

A review of patient’s records can provide evidence about clinical decision-making, follow through in patient management and appropriate use of clinical facilities and resources. However, the documented care may often not be directly attributable to a single resident but to the whole clinical team. A sample of 8–10 patient records is sufficient for a reliable assessment of care for a diagnosis or procedure.

**Workplace-based assessment (WPBA)**

Assessment of a student’s actual performance in the wards or in the consulting rooms poses a real challenge for teachers. Increasing attention is being placed on this type of assessment (highest level of Miller’s pyramid) because of its possible high consequential and predictive validity. Attempts at performance assessment have to balance issues of validity and reliability. Several methods of assessment have been developed to address these issues and many of these are variations on the traditional bedside oral examination. Though these methods have been developed focusing on practising doctors, they also apply to medical students, particularly to postgraduates. Some of the WPBA methods are described below.

**Mini-clinical evaluation exercise (mCEX)**

In the mCEX, a faculty member observes a trainee–patient encounter in any healthcare setting. The encounters are intended to be short (about 15 minutes) and focused. The trainee is expected to conduct a focused history-taking and/or physical examination during this time and then provide the assessor with a diagnosis and treatment plan. The performance is scored using a structured form and thereafter educational feedback is provided. Trainees are expected to undertake at least six such encounters during the year, with a different assessor for each encounter representing a different clinical problem, appropriately sampled from a list of patient problems. A number of studies have been reported using mCEX and it has been found to be a useful tool for formative assessment of residents in medicine. Its use in India has also been reported.

**Direct observation of procedural skills (DOPS)**

DOPS is a variation on the mCEX in which the assessor observes while the trainee is doing a procedure (e.g. venepuncture, giving an injection, etc.), rates the performance and then provides feedback. Like mCEX, trainees need to undertake six observed encounters during the year, each with a different assessor, and the procedures need to be sampled from an approved list. Research specific to DOPS is limited but in general four to eight encounters are sufficient for reasonable confidence in the final results.

**Mini-peer assessment tool (mPAT)**

In this method, eight assessors are nominated by the trainee from among his/her supervisors or peers, including nurses and other health professionals. The assessors are required to fill out a questionnaire concerning the technical and interpersonal skills of the trainee. These data are then shared with the trainee and the educational supervisor so that there is an agreement about the strengths and weaknesses of the trainee and a plan for improvement is developed. The use of peer assessment in higher education has been supported through research. Ramsey et al. showed that a reasonable reliability was obtained with eight to twelve peers and about five questions.

The success of WPBA depends on its ability to assess competencies which are not amenable to testing by conventional methods and making use of its formative function. With our present understanding, WPBA is not recommended for summative or high-stakes examinations due to issues of equivalence (i.e. how to compare performance of two students on two different cases). Students showing unsatisfactory performance on one tool should be offered other tools for diagnosis and remediation. Assessor training would go a long way in maintaining the quality of WPBA.

**CHOOSING THE RIGHT TOOL**

Assessments do not happen in a vacuum. All assessments are done with a context and a purpose. A number of problems related to assessment because an inappropriate tool is chosen. Many times, people tend to use certain tools because they are ‘in’ or are being used by others. The validity or reliability of assessment is not an inherent property of the tools. Thus, not all knowledge assessment may happen with essay questions and not all skills may be assessed by mPAT. A good framework to decide on the suitable tools is to ask who is going to use the results, when and for what purpose. If the purpose is selection, then we have to choose more reliability, but if the purpose is formative, then tools with higher educational impact should be chosen. Classifying the proposed use of assessment in relation to the utility criteria (viz. validity, reliability, acceptability, educational impact and feasibility) can help in making a useful choice. However, it is important to realize that not all tools are applicable in all contexts. It is also useful to keep in mind that in practice, there is always some trade-off between validity and reliability. Placing too much emphasis on reliability can take its toll on validity by reducing the authenticity of assessment. In general, it is better to build validity in the assessment, even though sometimes it may mean a little compromise with reliability.

To summarize, assessment requires careful planning. It must be strategically designed for its educational effects. The assessment tasks should be in close alignment with the curricular objectives and the instructional methods. All assessment methods have limitations and no one method can assess all types of knowledge and skills. A good assessment uses a mix of methods depending upon the context. A good assessment must demonstrate validity and reliability. It should be feasible and must be acceptable to all...
stakeholders, including students, faculty and licensing bodies. To effectively use some of the newer methods of assessment, there is a need for training examiners. Assessments that are bad can promote unhealthy approaches to learning.

REFERENCES


11 Sood R, Paul VK, Mittal S, Adkoli BV, Sahni P, Kharbanda OP, et al. Effectively use some of the newer methods of assessment, there is


15 Norcini JJ. Current perspectives in assessment: The assessment of performance at


42 Norman G. The long case versus objective structured clinical examinations: The long case is a bit better, if time is equal. BMJ 2002;324:748–9.


SGPGI-NIH Workshops on ‘Scientific paper writing’, ‘Biostats’ and ‘Observational Studies’

As part of an Indo-US Collaboration, the National Institutes of Health, USA has sponsored a series of workshops since 2006 on various aspects of clinical research (with an emphasis on clinical trials), including biostatistics, study design and randomization issues, data management, research ethics, and regulatory aspects. As a continuation of this series, three workshops are planned at Sanjay Gandhi Post Graduate Institute of Medical Sciences (SGPGI), Lucknow during 2013 as follows:

(i) Workshop on ‘Scientific Paper Writing’ on April 19–21, 2013,

(ii) Workshop on ‘Basic Biostatistics’ on July 25–28, 2013, and

(iii) Workshop on ‘Observational Studies’ on September 29 to October 2, 2013.

All the workshops are aimed at active biomedical researchers who hold faculty positions. Investigators involved in clinical research, who are in a position to lead clinical research studies, should find this workshop useful.

Only a limited number of applicants will be accepted for each workshop. Applicants should download an application format (asks for summary of their experience and expertise in clinical research in a structured format) from https://sites.google.com/site/sgpgininhcourses/ and email it as an email attachment to sgpgi.courses@gmail.com. The last dates for applications for the three courses are March 3, May 31 and July 31, respectively. A selection committee will notify the successful applicants of acceptance about 4 weeks before each course.

A limited number of scholarships for travel and lodging are available for qualified applicants whose institution cannot cover their expenses.