Assessment Methods for Measuring Clinical Competence: Review on their psychometric properties

Gandes R Rahayu

ABSTRAK

Makalah ini mendiskusikan berbagai metode penilaian untuk digunakan dalam konteks klinis. Masing-masing metode mempunyai kelemahan dan kelebihan. Walaupun banyak metode assessment yang mungkin digunakan, namun demikian tidak ada single choice; pilihan terbaik yang diinginkan beberapa metode penilaian sangat diinginkan untuk mendapatkan fair judgement tentang kompetensi klinis mahasiswa. Penilaian metode penilaian seharusnya tidak hanya didasarkan pada psicometric properties dari masing-masing metode, namun juga didasarkan pada psychometric properties dari kesehatan metode penilaian yang diterapkan.

Kata kunci: assessment · clinical competency · psychometric properties · validity · reliability · feasibility

CLINICAL COMPETENCE

For many decades, the medical profession has been struggling with the question of competence.1,2,3,4 The motivation to do so has come from a variety of sources. Firstly, the influence of the profession itself. This comes in the form of the accreditation process, the licensing and certification process and specific training programs. Secondly, the influence of society. Modern society exerts an influence on the medical profession through formal and informal mechanisms, such as the general climate in which potential physicians grow up, including the level of public awareness about health and disease. Examples of formal mechanisms are specific laws and policies governing professional behaviour and the legal arrangements reflected in malpractice suits. Thirdly, the influence of research and development. The influence can be seen in the availability of sophisticated technology to practising physicians and in the rise of new fields of research, which relate to clinical competence. Fourthly, the globalisation era.

It is important to distinguish between “competence” and “performance”; the former meaning what a physician is capable of doing, and the latter meaning what a physician actually does in his day-to-day practice. There have been various techniques used in identifying competency e.g. reflective philosophical study, task analysis, descriptive studies, studies of diagnostic thinking, consumer opinion, belgian techniques, wise man approach, critical incident and epidemiological study. Many efforts have been done to define what competencies should be mastered by new medial graduates (see for example Bundere 12 outcomes, Brown University’s Nine Competencies, Indonesian KIPDI II). Although there are some variations, the basic competencies usually cover the following categories:

1. Clinical Skills

The ability to acquire clinical information and interpret the significance of the information obtained.
2. Basic Medical Knowledge and understanding
   The ability to use relevant knowledge about clinical conditions in order to provide effective and efficient care for patients.

3. Interpersonal attributes
   The expression of those aspects of a physician's personal and professional character that are observable in interaction with patients.

4. Problem-solving and clinical judgement
   Clinical reasoning skills is the cognitive process by which the information contained in a clinical case is synthesized, integrated with the physician's knowledge and experience, and used to diagnose and manage the patient's problem.

5. Technical Skills
   The ability to use special procedures and techniques in the investigation and management of patients.

Newble (1992) tried to define clinical competence as the mastery of a body of relevant knowledge and the acquisition of a range of relevant skills, which would include interpersonal, clinical and technical components. The relation of each component to clinical competence is shown in the following figure:

---

The following section is aimed to review psychometric properties, the strengths and weaknesses of the various established assessment methods if used to assess clinical competence.

1. Combination between Long Case or Short Case and Oral examination
   The oral or viva voce examination has for centuries been the predominant method, and sometimes the only method used to assess medical students, particularly in the clinical area. It is based on the belief that a clinician is the best clinical examiner.

   The use of oral examination to assess clinical competence is usually combined with a long case and/or short case. The purpose of the long case is for the student to obtain a full and comprehensive history from the patient and then to carry out a physical examination to elicit and interpret any physical signs that might be present. The main aim of the long case is that the student should be able to synthesise all of the information obtained into a problem list or differential diagnosis. The short case examination usually concentrates only on physical signs.

   When the long case and short case examination are combined with the oral examination, the student has a certain amount of time to interview and examine a patient, make the differential diagnosis and treatment plan, then report the findings and plan to an examiner before undergoing further interrogation. The student can expect not only to be questioned about the history and to demonstrate any physical signs, but also to discuss investigation, diagnosis, treatment, prognosis and social implications.
Appropriately implemented, the combination between the long case or short case and the oral examination can measure many attributes of clinical competence, such as history taking, physical examination, interpersonal attributes, clinical reasoning or problem solving and clinical judgment. This method offers some advantages. The most important advantage is the examiner can observe the student's clinical competence directly in taking a history and examining patient. The examiner can observe a proper approach to the patient, a correct attitude, and test for thoroughness and general proficiency, as well as courtesy, consideration and gentleness. Newble and Cannon (1987) added that it can be used to assess a number of complex skills which are more difficult to assess with strictly written tests; these include oral communication skills, the ability to think fast, and the ability to bring to bear other relevant information to the problem at hand. Oral examination provides an opportunity for the examinee to defend or explain his or her response. Furthermore, the opportunity exists for continued questioning on any given topic. If an examinee misinterprets a question, for example, he or she is not automatically scored as wrong; another question or rephrasing of the question may enable the examiner to demonstrate competence in the area being assessed. With face-to-face contact the examiners can explore both the breadth and depth of a student's understanding. Another value is its flexibility to allow direct feedback between examiner and students.  

Many criticisms, however, are directed at this method. It is often unstructured and the task and questions presented to the student and the discussion that accompanies it is usually in the hands of the examiner. It is also time-consuming for the individual and the examiners in terms of both preparation and administration, not practical if a large number of students are to be assessed, and is therefore costly. It contains significant elements of subjectivity, and is part unstructured and unstandardized. If the examiner does not observe the student while he/she is doing history taking and physical examination, the valuable advantages offered by direct observation will be lost. It also has been known for a long time that this method is often unreliable. The reason relates to the very small sample of course content, the variability of the tasks and questions, the variability of the marking, and the lack of agreement between examiners of the same candidate. Hubbard et al (1963) found that the lack of agreement between two independent ratings of a candidate in two different situations might be caused by the differences of the question or behaviour and skills observed. This interpretation is consistent with the finding that pairs of examiners observing the same piece of behaviour showed better inter-rater reliability in their ratings (0.75 to 0.89) than did individuals observing different sessions (0.25 to 0.45). When checklists and rating forms are used for the guidance of oral examiners, quite high inter-rater reliability coefficients have been reported, ranging from 0.79 to 0.92. Paupinee et al (1955) reported that the use of a sampling grid and the use of problems with defined questions and guidelines to answer could increase the level of reliability.

On the other hand, if materials are made overly uniform and questions and topics are defined for each candidate in the attempt to achieve consistency in rating, the flexibility, uniqueness and individualistic nature of the oral exam are lost. It is suggested that high reliability achieved by tight structure and standardisation of an oral exam may only allow the sampling of a very small part of a student's overall clinical competence. Thus the challenge is to standardise sufficiently to achieve acceptable reliability while not destroying the unique aspects of the oral. The experience of the American Board of Emergency Medicine suggests that it is possible to achieve acceptable reliability with relatively 'little structure, amounting to a standardisation of cases' and suggested key criteria of performance.

Although subject to many criticisms, particularly in terms of its unreliability, the oral examination continues to be used, primarily because of the strong conviction that oral examination measures some aspects of performance not measured by other methods of testing. Laube & Yang (1985) reported over 700 student/examiner situations that were recorded and categorized. The result shows that 20% of the entire examination process is to elicit general knowledge about the problem presented (disease, condition, medical procedures, etc.). Approximately 80% of the examination process
involves aspects of process-oriented thinking or the application of synthesis and analysis techniques. This included the collection and interpretation of data, making differential diagnoses and using clinical judgement while establishing clinical priorities.

A combination between long case or short case and oral examinations is appealing because of its high face validity, its flexibility and the possibility that it measures aspects of clinical competence that are perhaps not tapped in written assessment. But there are a number of serious problems in using this combination method. First, it is usually impossible to examine all candidates at the same hospital, on the same patients, and with the same examiners, even though this type of standardisation has been demonstrated as being crucial for establishing the reliability of the technique. Second, the oral assessment period usually is short and stressful, thus not entirely mirroring real life. Third, it is difficult to marshal, or ever find, sufficient numbers of patients in the large specialities for clinical orals. Fourth, the actual cost of this method should be considered.

Finally, if the combination between the long case or short case with the oral examination is to be retained, Newble (1992) suggested some considerations that may be possible to improve its quality:

1. Use it only to test those aspects of competence that cannot easily measured by more objective methods, such as an OSCE.
2. Define more clearly the attributes one aims to assess. This should allow the adoption of a more structured approach to the marking.
3. Actively observe the student perform a complete history taking and examination.
4. Increase the number of cases. One case is certainly untenable and the more the better.
5. Attempt to reduce the variability in the case presented to the students. Van der Vleuten and Swanson (1990a) reported that in some places simulated patients are being used to assist in this process.

2. Multiple Choice Question (MCQ)

The MCQ examination came to the attention of the medical community in the early 1950s. In assessing clinical competence the MCQs have low face validity. There is little resemblance between the writing of MCQ examinations and the work of a clinician. Actually its role in medical education is limited to testing factual knowledge. The role of MCQs is illustrated as:

"MCQs glorify the facts, many of which are detailed, and some not necessary for medical student but only for an expert in the field. Only minimal thinking is needed. Problem solving is virtually absent."

A large MCQ examination can sample a broad spectrum of knowledge, and appropriate sampling methods can ensure that it is representative of a designated knowledge base. In this sense, MCQs have considerable content validity. It appears, however, that MCQs are not usually representative of the full range of intellectual activities, which it is assumed medical students and physicians should demonstrate. When looking at any list of the components of clinical competence, it is acknowledged that MCQs can be used with confidence for only knowledge-related components.

Several forms of reliability have been devised to analyse the MCQ assessment. Given the nature of the MCQs, in particular the characteristic that the test setters agree in advance about the correct answer option of a question, it is of no great surprise that it is reasonably reliable.

Studies of the concurrent validity of MCQs have been described at various points along the continuum of medical education, undergraduate training, certification at the end of postgraduate training and continuing medical education. Swanson and Case (1992) found a correlation of 0.83 between MCQs and essays. In one study (Norman et al, 1987), MCQ and Modified Essay Question versions of a test were constructed. The true correlation across test format was equal to one. Rothman and Kerneky (1986) compared the correlation between MCQs and practical examination. The correlation was only 0.46. Pinetti et al (1992) also found a poor correlation between the MCQ examination of clinical clerkship students in medicine, obstetric and gynaecology, paediatric and surgery department and faculty perception of their ability. Respectively the correlation coefficients were 0.42, 0.08, 0.05 and 0.13.
O’Gorman (1992) supported this finding. She found a very low correlation between the MCQ and the OSCE used to assess the course in Psychiatry in fourth-year students. The coefficient was 0.12.

Studies to examine correlation between grades in undergraduate medical education as measured by MCQs and physician performance as measured by a variety of performance assessment methods show low correlation. It means that the MCQs do not measure the clinical skills measured by performance assessment methods.

In terms of feasibility, it is relatively easy to prepare, administer, and score the MCQs. The preparation requires a degree of expertise, but various books, instructional manuals, and workshops are available for this purpose.

Finally, the following conclusion may be drawn in terms of the use of MCQs in assessing clinical competence:

1. The face validity of MCQs are vary low. There is little apparent resemblance between clinical performance and test taking performance.

2. In terms of clinical competence, MCQs are more appropriate for measuring the clinically relevant knowledge component.

3. Once the correct answers have been agreed upon, MCQs scoring is highly reliable.

3. The Essay

Nendas and Toklan (1999) stated that the advantage of essay exams are that they can more easily reflect the problem solving skills of students and provide a format that encourages students to integrate their knowledge.

In terms of face validity writing an essay question bears little resemblance to what a clinician does with patients. Proponents of essay examinations may argue, however, that physicians are professionals and, as such, should demonstrate that they can display their analysis of a problem in clear, logical and accurate prose. It is further argued that a good performance on an essay question demands the capacity to integrate information, to synthesise, to weigh arguments and justify choices, and perhaps even to be creative and original. While all these claims have some appeal, the evidence suggests that most of them have not been substantiated.

The nature of essay question suggests that their comprehensiveness is limited. Only a small number of topics can be discussed in a single examination, due to limitation of time. In general, it can be concluded that essay questions are poor on content validity.

Reliability is a major stumbling block in the use of essay questions. Different readers tend to give different grades to the same essay. A study was conducted to examine 300 essays written by college freshmen, in which 53 readers representing several professional fields rated them, using a nine-point scale. The result showed that none of the 300 essays received less than five of the nine possible ratings, 23 percent of the essays received seven different ratings, 37 percent received eight different rating and 34 percent received all possible ratings. However, there is some evidence that reasonable consistency among markers can be obtained if what the essay should contain is carefully specified.

The evidence that the essay question truly measures clinical competence is weak. The essay questions can only measure the component of knowledge. If not accompanied by other methods they fall short on many criteria of good assessment methods for measuring clinical competence.

4. MEQ

In 1975 Hodgkin and Knox developed the modified essay question (MEQ) as an examination tool for the Royal College of General Practitioners. "It examined the respondent’s ability to solve and manage clinical problems, and to assess problem related aspects of basic and clinical science, at least on paper. The features of the MEQ are that it is based on an actual case, is laid out in sequential fashion, and that it requires short written responses to open ended questions.

With respect to reliability, Felletti (1982) showed that the estimated reliability or internal consistency of VEQs lay between 0.57 and 0.91 using Cronbach’s alpha coefficient. Norman et al. (1987) reported that inter-rater reliability coefficients varying from 0.80 to 0.77. The magnitude of these coefficients is influenced not only by the quality of the criteria lists, but also by the test’s structure.

Newble et al. (1981) reported that the performance on free-response MEQ tests improves...
with the level of training. In terms of construct validity, correlations varying from 0.40 to 0.52 have been reported between MEQ and MQ scores. Correlation of approximately 0.42 have been reported between MEQ and medical bachelor clinical examination scores. Strasburger and Pierce-Pfen (1953) have reported a correlation of 0.64 between MEQ and clinical practice evaluation for physiotherapy students.

From the perspective of feasibility, the main drawback of MEQs is the requirement that they are hand marked, and they cannot be computer scored.

To sum up, although the MEQ appears to be a practical and credible written assessment, in terms of clinical competence it too only measures the knowledge component, and more work on its reliability and validity is still required.

5. Audit of case / patient records

Patient records have traditionally been used for providing documentation of a physician's assessment and management to serve as a reminder in the subsequent care of the patient.

The patient record has a number of advantages that make it suitable for both formative and summative evaluation. It is based on real clinical behaviour; it is experience-centered and relevant to the patient-care problems of the learner; the learner can have the opportunity to participate in the setting of the goals and the implementation of the program by involving in caring the patients; active participative learning is more effective than passive traditional techniques and record review allows a problem based approach that encourages active participation.

Audit of patient records or medical record review has been used to assess clinical competence in undergraduates, residents in postgraduate training, and clinicians practicing in the community.

The assessment is based on the patient record. In undergraduate education, record review is widely used on an informal basis by supervising residents and faculty for formative evaluation of students on clinical rotations, using the data from the student's write-up as a basis for discussion and feedback. The patient record has been implemented as the primary basis for evaluation in the clinical program of an English medical school on the premise that audit of problem-oriented records allows constant monitoring of learning objectives and immediate feedback.

The dimensions of clinical competence that are involved in many studies include problem identification, data collection by history-taking and physical examination, investigation, data interpretation, problem formulation, refinement of problem, management of patient included therapy, consultation, patient education and follow-up.

There are two issues that relate to the face validity of record review. The first is the establishment of explicit criteria in relation to what is being assessed rather than implicit judgment. The second issue is related to the extent to which the chosen access influence the health of patients. Thus all criteria should be clearly relevant to therapeutic or prognostic decisions.

Tugwell & Dok (1953) suggested when evaluating clinical competence using a measure such as record review, three dimensions should be addressed: (1) completeness of recording; the extent to which record review accurately reflects the action of interest of the health professional; (2) relationship to patient outcomes; the demonstrator that patients who receive care from physicians judged more competent on the basis of evidence from the medical record have better outcomes; and (3) relationship to physician judgments.

The demonstration of a correlation between clinical competence assessed by record-keeping behaviour and patient outcomes would provide the best evidence that the record review assessment method is measuring an important component of clinical competence. Unfortunately the majority of studies involve only physicians in practice. No studies appear to have been carried out with undergraduate medical students.

For assessing clinical competence the record review has strengths and weaknesses. The main strength are: (1) the clinical record is an established component of clinical practice, (2) actual practice behaviour and complete episodes of care can be assessed, (3) it allows application of explicit criteria with high levels of consistency in abstracting the information, (4) relatively large numbers of cases with the same problem or diagnosis can be assessed, thus reducing the sampling problem that may result from basing conclusions on very few cases such as in oral examinations. The main
weaknesses are: (1) the inconsistency with which actions are recorded, (2) the difficulty in deciding who made the recorded decisions where more than one health professional discusses the patient before the record is written, (3) the failure to emphasize the evidence for the efficacy of the patient care being assessed, (4) using audit of medical record to assess clinical competence of undergraduate student will have many constraints, for the reason they are not allowed to treat patients. What was written in medical records does not represent what the medical undergraduate student actually does.

6. The Checklist and Rating Scale

A checklist consists of a list of specific behaviours, characteristics or activities and a place for checking whether each is present or absent. It may be used to assess a student performing a series of discrete steps in a procedure. A procedural checklist assesses whether a student follows the appropriate steps in a process or procedure. A product checklist focuses on the quality of the product. To create such checklist, a thorough understanding of the subject matter as well as the procedure or the product itself is needed. Without this knowledge it is difficult to identify critical steps, flaws in the product, and potential student errors.  

Rating scales, on the other hand, generally contain fewer items, and focus more on broader issues of performance. Rating typically use Likert-type scales, in which refers are required to score performance in a number of boxes, ranging from very poor to very good. There are several common errors that occur when a teacher rates students. A leniency error occurs when a teacher tends to make almost all ratings toward the high end of the scale, avoiding the low end of the scale. A severity error is when the teacher tends to make almost all rating toward the low end of the scale. A central tendency error occurs when a teacher hesitates to use extremes and uses the middle part of the scale only. This sometimes occurs when a teacher has to make strong inferences about a student, and tends to mark everyone as average. A halo effect occurs when a teacher lets her general impression of the student affect the rating on specific dimension. Personal bias occurs when a teacher has a general tendency to use inappropriate or irrelevant stereotypes to favour one category of student over another. A logical error occurs when a teacher gives similar rating on two or more dimensions of performance that the teacher believes are logically related but which are in fact unrelated.

Generally, the use of checklists is recommended over the use of rating scales.  

Two main arguments are used in this respect: checklists are more transparent (i.e. they provide a better definition of expectations for examinee and feedback on performance), and they are more objective (i.e. leading to better reliability of score). However, there are disadvantages associated with the use of checklists. First, some skills, e.g., communication and attitudes, are difficult to assess without trivializing the content. Second, checklists may have a negative impact on student learning. Van Luijk et al (1990)  

reported this negative impact when the students started to memorise previously used checklist, leading to a proficiency which is adequate on the surface, but in which deeper understanding of the skill appears to be missing. Moreover, the faculty raters, all experts in their field, can feel undervalued, because of the stringent nature of the checklist.

Lower inter-rater reliabilities are generally found in measurement instruments using rating scales than those using checklists. Similar findings are reported for performance-based tests.  

Global ratings and checklists rank order students in a similar way. From this perspective, the global ratings are as objective as the checklists. This is even the case when only a one-item single rating is given for the overall quality of performance.  

Miller (1990) pointed out that:

"It is not simply a matter of checklists or rating scales, scoring by physicians or trained non-physician, but rather of reaching agreement on what aspects of the encounter to observe and how to combine and weight these observations to yield scores that reflect, in a meaningful manner, the adequacy of observed performance."  

Van der Vleuten and Swanson (1990b) supported the above warning. After reviewing many of the scoring forms currently in use (rating scales and checklists) they were moved to comment that the potential for omitting important items and including unimportant ones in checklists and rating scales is great. The former penalises examinees.
who take indicated actions that are not listed. The latter rewards examinees who are unjustifiably thorough.

7. Self-assessment

According to Boud (1995) the characteristic of self-assessment is the involvement of students in identifying standards and/or criteria to apply to their work and making judgements about the extent to which they have met these criteria and standards. People learn best when they are helped to define their own problems, acknowledge and accept their strength and weaknesses, decide on a course of action, and evaluate the consequences of their decisions.

Self-assessment is a critical component in medical practice. Various educators and researchers have indicated the necessity of introducing self-assessment to students in their undergraduate medical education.9 10 But although its importance appears to be widely recognised, very few medical schools have adopted a curriculum that enhances self-assessment behaviour among students.11

A large number of studies examining self-assessment conclude that individuals do not evaluate themselves accurately. Inaccuracies in self-assessment have been reported for family practice residents,6 clinical clerkships,6 and physiotherapists.7

Arnold et al. (1985)12 in a study of 221 students during internal medicine rotations during the last 4 years of medical school at the University of Missouri, found that the correlation between student and faculty rating decreased over the four years from r = 0.25 in year 3 to r = 0.16 in year 6.

On the other hand, another researcher has found that students have the ability to achieve levels of accuracy and validity in their self-assessment and can be good judges of themselves.13 By emphasising explicit self-assessment goals and training strategies, moderate to high validity outcomes or improvements over time were demonstrated.13

In terms of clinical competence, with predetermined criteria of clinical competencies, the self-assessment method can be as a tool to assist students in detecting their own strengths and weaknesses in mastering the clinical competence either in part of the knowledge, attitude or skills components.

8. Patient management problems (PMP)

In the 1970s the use of patient management problem (PMP) flourished. It became widely used for certification of competence at the undergraduate and postgraduate levels.14

A PMP usually begins with a clinical statement concerning the presenting problem of a patient with a brief history and examination data.15 It is structured in stages, first the student is asked to make a decision about the patient's management, which he/she considers appropriate to the situation. Once a decision is made, further hidden information is revealed indicating the result of the proposed action. He/she then bases his/her further decision on the result revealed. These three phases of providing information, proposed action and feedback are repeated many times before the end of a problem.

From the perspective of assessment, PMPs can be standardised and specific performance criteria can be defined. Furthermore, large numbers of learners can be evaluated simultaneously and economically with a broad sampling of performance in different content areas.16

By its design, the PMP demands performance in the four traditional categories of clinical competence, namely history taking, physical examination, laboratory investigation and management. It is presumed to demand cognitive or problem-solving processes similar to those occurring in the actual clinical setting. Experts who have reviewed PMPs and students and physicians who have worked through them report on systematic inquiry that the PMP closely approximates the clinical setting.17

Studies assessing construct and concurrent validity of PMPs have been undertaken. Hubbard (1978)9 reported that the performance of students at the third year was significantly lower than performance at the level of internship. McKenzie et al. (1972)22 compared performance on PMPs and multiple-choice questions at all levels, medical student, resident, and certification exam candidate. The findings demonstrated uniformly low correlations, ranging from 0.20 to 0.40, and did not exceed 0.50 even when the tests were deemed highly reliable because of their length. It is concluded that the tests are measuring different
competencies. Goran et al. (1973) compared a clinical team's performance or, PMP with clinical performance assessed by chart audit. While exact cases were not compared, the chart audit and PMPs assessed similar clinical problems. The basis for comparison was a list of ten items of history, six items of physical examination and a laboratory test on urine culture. The clinical importance of each item was established by a literature review. The finding showed that the ten items of history were requested far more frequently on the PMP than they were recorded on the chart. Moreover, the critical decision to order a urine culture was also requested twice as frequently on the PMP. Differences in each comparison achieved statistical significance. The author added that the subjects performed consistently better on the PMP than in the clinical situation and that those performing best on the PMP did not necessarily perform best in the real life setting. These findings raise doubts regarding the concurrent validity of PMPs. Norman and Feighner (1981) compared the performance of two similar groups of clinical clerks on four cases, each in one written PMP and in a live clinical simulation format. The study found that students elicited up to twice as much information on history, physical, laboratory, and management in the PMP formats. The findings are consistent with the study conducted by Goran et al. (1973), which raise questions about performance on PMPs. Internal consistency reliability has been reported to be at level of 0.80 to 0.85 using the Kuder-Richardson Formula 20. Levy and McGuire (1966) used an alternate approach to measure the reliability of branching PMPs, and showed the coefficient of reliability ranging from 0.75 to 0.85 for a single but lengthy problem. For tests of two or three long problems in one particular content area, reliability estimates range from 0.80 to 0.90.

According to Miller (1990), the problem with PMPs is the difficulty of gaining consensus among independent judges on the positive or negative weights to be assigned each possible intervention and even to the optimal path that should be about equally reliable.

Other problems that have arisen for the PMP are that the scores have been found to be case specific and that it still only accesses declarative knowledge. In fact, the PMP has been shown to tap the same knowledge as a well-constructed MCQ, with no improvement in validity, reliability or efficiency.

To test clinical reasoning and problem solving by case simulation, patient management problems were used extensively until the late 1980s when evidence of psychometric flaws led to their partial discontinue.

9. Portfolios

Portfolios are not new, but their use in initial and continuing professional development in medicine is still in its infancy. Snadden and Thomas (1998) defined the portfolio as a collection of evidence that learning has taken place. This articulates what has been learned by the students.

In this definition a portfolio is not merely a collection of events seen or experienced, but involves a critical reflection process about the relation between the events and the professional tasks and the learning that has taken place. Moreover, the event itself should be a key event or in other words it should be meaningful for one's professional life.

Paulson et al. (1991) gave a wider definition. A portfolio is a purposeful collection of student work that exhibits the student's efforts, progress and achievements in one or more areas. The collection must include student participation in selecting contents, the criteria for selection, the criteria for judging merit and evidence of self-reflection.

In this definition a portfolio is more than a mechanism to promote reflection; this becomes personal collections of educational experiences over a period of time. There have, therefore, been attempts to develop it into some system of assessment, on the assumption that it provides a more equitable and sensitive portrait of what students know, and are able to do, than do traditional assessments.

It is important to emphasise the difference between portfolio and logbooks that have been introduced into some higher specialist training programs. Such logbooks are collections of tasks.
carried-out; they do not contain critical reflections and may be seen as chores rather than a way of stimulating learning.

Redman (1994) described a series of key aspects that are generally recorded in the portfolio. The first is the experience. It infers what has happened, what has been done, seen, written, or made. The second is the learning discovery that what has been recalled has significance for doing or changing things in the future. The third is the evidence that demonstrates how the learning is being applied in an appropriate context. The fourth is learning needs that are identification of where it would be appropriate to go next. The last is learning opportunities that mean an educational action plan identifying ways in which learning needs might be met.

The portfolio can be used in a number of ways. Some portfolios may be developed in order to demonstrate the process of progression, a method of personal development and a way of tracking progress while others will be assessed against specific targets of achievements. Some will be essentially private document, for personal review only, while others will need to be open to public scrutiny.

Portfolio-based learning has been introduced successfully into a range of educational and professional learning contexts. This may have been prompted largely in the UK by the move towards competence-based assessment and the introduction of National Vocational Qualifications (NVQs), where the emphasis is on evidence of achievement, rather than the educational processes undertaken by the learner. Within this framework the portfolio is designed to demonstrate how the learner has met the outcomes that form the basis of the qualification. However, the use of portfolio as a tool for development in initial and continuing professional development, taking into account process as well as outcome, is growing.

Portfolios are attractive as assessment tools as they appear to be one way of assessing performance in practice over a period of time. In other words they assess the application of theory and the performance of the student or doctor. Jasper (1993) described the portfolio as being successfully used for both formative and summative assessment in several nurse and midwifery training institution. Snadden and Thomas (1998) found a different result when using portfolio with general practice registrars. There is negative impact of formal assessment on the material. College portfolio for registrars was unlikely to collect in a reflective journal incidents which have not gone well. Such incidents are a rich source of learning and do give insight into the development in learning, but it is understandable that young doctors would not wish to allow such material to be used for assessment purposes as the traditional view of assessment in medicine is that of a competitive examination that seeks excellence.

The portfolio has a huge potential for assessment in the contexts of adults and self-directed ideas about learning. But the concept of portfolio, which means student-selected non-standardised work, means that the present dependence of assessment on notions of measurement and validity is challenged. This in essence means that the effective assessment of portfolios will continue to be difficult if we remain trapped within our traditional view of assessment. In other words, assessment based on comparing students with each other and with target grades or marks does not fit easily with portfolios with are essentially non-standardised. The portfolio will remain difficult to assess until new non-comparative assessment methods are developed. While the emphasis on grading, excellence and comparison between students and doctors remains in assessment and medicine, it is likely that portfolios will have a greater role as a learning tool than as a summative assessment tool, but as a learning tool they may have considerable influence on learner’s performance in summative assessments. If, however, they can be grasped to develop innovative assessments they may have the potential to influence the type of learning of reflective practitioners.

Portfolios will normally be directly related to a personal or professional learning plan. This constitutes the framework within which portfolio development takes place, and provides a statement of the outcomes that the portfolio seeks to demonstrate. There are many benefit that arise from the use of portfolio-based learning that may not be encompassed in other forms of educational activity:

1. It recognises and encourages the autonomous and reflective learning that is an integral part of professional education and development.
2. It is based on the real experience of the learner, and so enables the consolidation of the connection between theory and practice.
3. It allows a range of learning styles to be used according to the preferences of the learner.
4. It enables assessment within a framework of transparent and declared criteria and learning objectives.
5. It can accommodate evidence of learning from a range of different contexts.
6. It provides a process for both formative and summative assessments, based on either personally derived or externally governed learning objectives.
7. It provides a model for lifelong learning and continuing professional development.

Mitchell (1994) has commented, however, that there has been little critical appraisal in the literature on the use of portfolios and highlighted several key areas of concern:
1. There may be conflict for portfolios serving both the purposes of individual learning/assessment and large-scale assessment.
2. Can portfolio be assessed if the work is not standardized?
3. What criteria should be involved in making judgements about students' attitudes, feeling and behaviour?
4. What about issues of privacy and confidentiality as some portfolios contain highly personal material?

The portfolio may offer the breadth of information about the learners' progress and achievement to be considered as appropriate means of assessment.

10. The Objective Structured Clinical Examination (OSCE)

The OSCE (Objective Structured Clinical Examination) was developed in 1979 by Harden and Gleeson. The OSCE has been recognized globally, as a useful instrument for the assessment of clinical skills. In the 1980s, its use became widespread, and it is a feature of most medical schools in the developed world.

The OSCE is an approach to the assessment of clinical competence in which the components of competence are assessed in a planned or structured way with attention being paid to the objectivity of the examination. It is not an examination technique per se but represents a format within which a variety of techniques can be employed. The basic concept of the OSCE is that each component of clinical competence is tested uniformly and objectively in all the students taking up a particular examination. It is a multi-station test of clinical skills that uses sater and direct observation. The student is assessed at a series of stations with one or two aspects of competence being tested at each station. The OSCE combines the reality of live clinical interactions, the standardization of problems, and the use of multiple observations of each student.

The advantages of the OSCE are that it forces multi-component testing, allows the use of multiple methods, utilizes use of limited resources, implies objectivity, simulates situations, allows profiling and feedback. Van der Vleuten (1990) added that the use of clinically relevant tasks tailored to the skills to be assessed, controlled and standardized testing situations, and predefined grading criteria are thought to provide a significant advantage over traditional unstructured ratings of performance in clinical training. It enables an increase in the sample of clinical behaviour that might be evaluated in a reasonable period of time, using facilities and resources generally available in most medical schools.

Van der Vleuten and Swanson (1990b) in their comprehensive review stated that there are three main factors that influence the reliability of the OSCE.

First are rater-related factors. Although different indices of reliability and validity vary from study to study, inter-rater agreement generally appears to be adequate. Rater agreement for history-taking and physical examination skills is generally better than for communication and patient education skills. There is no systematic difference between physician and non-physician raters; adequate agreement can be achieved with non-physician raters, in particular use of standardized patients as raters seems feasible.

Second is a standardized patient-related factor; it relates to the accuracy of portrayal. If
standardised patients are inaccurate in portraying patient roles, this may have an impact on examinee performance and on the reproducibility of scores, particularly if multiple patients are trained to play the same role. The use of multiple standardised patients to portray the same patient role results in a relatively small decrease in the reproducibility of scores, as long as examinees are randomly assigned to the standardised patient playing the same role.

The third are station-related factors. Content specificity is the largest source of measurement error in OSCEs. The performance of an examinee at one station is not a good predictor of performance at other stations. As a result, fairly long tests are required to obtain reproducible scores. Furthermore testing time per station has limited impact on the reproducibility of total test scores. Total testing time appears to be more important than duration of individual stations.

Furthermore Van der Vleuten and Swanson (1990b) stated that assessment of history taking, physical examination, and communication skills is the most reproducible. Assessment of laboratory utilisation, differential diagnosis, and treatment generally reduces the reproducibility of scores and increases testing time requirements.

In an OSCE, skills are tested in very small units and stations, therefore, last only 5 minutes. Harden mentioned as a disadvantage of an OSCE that skills are isolated without any clinical context. In addition many skills cannot be tested because they last more than 5 minutes. This could impair the validity of OSCE in measuring clinical competency.

Construct validity has been demonstrated by Petrusa et al (1987) who compared the level of performance on the five major clinical skills for 60 house officers in their first year of postgraduate training (PGIs), 14 house officers in their second year (PGII) and 283 third year medical students. The result showed that the house officers’ performance were significantly better than the third year medical students. It is reported that there is considerable improvement in OSCE scores among foreign-trained medical graduates at the end of a one-year clinical clerkship. Stillman et al (1991) showed that mean OSCE scores increased in medical students from year one through year four. Many other studies support both construct and concurrent validity of OSCEs in assessing clinical competence.

As OSCEs involve many assessment methods, there might be a danger that some of the problems we now attribute to MCQs test might appear in the OSCE. We must be quite clear what skills can be appropriately measured by this approach and more importantly, what skills cannot be measured.

SUMMARY

Various methods to measure clinical competences have been discussed. Each has strengths and weaknesses. The triangulation principle should be employed to provide more clear picture and fair judgment. The psychometric properties of the whole assessment methods employed should be taken into account when designing an assessment system.

REFERENCES


20. Van Der Vleuten CPJ, Swanston DB. Assessment of clinical skills with standardized patients: state of the art. Teaching and learning in medicine, 1980; 2: 58-76.


