

Ottawa 2010

Preliminary report with draft Consensus Statements and Recommendations for the Performance Assessment Theme

Theme Group Members: Katharine Boursicot (UK) - lead; Luci Etheridge (UK); Jean Ker (Scotland); Elango Sambandam (Malaysia); Zeryab Setna (UK); Sydney Smee (Canada) Alison Sturrock (UK)

(c) Ottawa Conference: not to be used or reproduced without permission

THE THEME IN CONTEXT

“Whatever exists at all exists in some amount, to know it thoroughly involves knowing its quantity as well as its quality”

Thorndike 1904

Definition of the theme

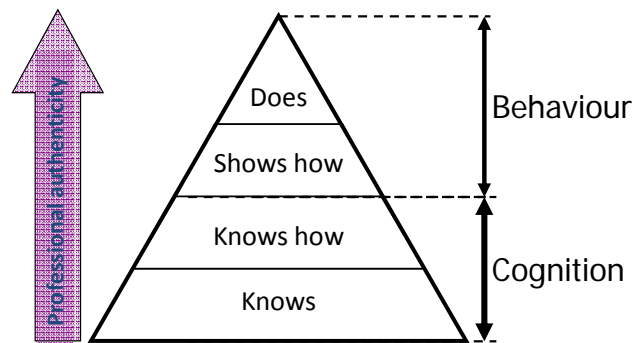
A modern approach to defining performance assessment in medical education requires recognition of the dynamic nature of the perspectives and definitions. These changes result from the variations in which clinical practice and education are delivered and a lack of clarity in defining competence and performance.¹

Competence describes what an individual is **able to do** in clinical practice, while performance should describe what an individual **actually does** in clinical practice. Clinical competence is the term being used most frequently by many of the professional regulatory bodies and in the educational literature.^{2, 3,4} There are several dimensions of medical competence including the scientific knowledge base and other professional practice elements; such as history taking, clinical examination skills, skills in practical procedures, doctor patient communication, problem solving ability, management skills, relationships with colleagues and ethical behaviour.^{5, 6, 7} In the last 50 years a wide range of assessment frameworks have been developed examining these different dimensions. Ensuring these are reliable and valid is, however, challenging.⁸

THEORETICAL BACKGROUND

Miller’s pyramid has been used over the last twenty years as a framework for assessing clinical competence.

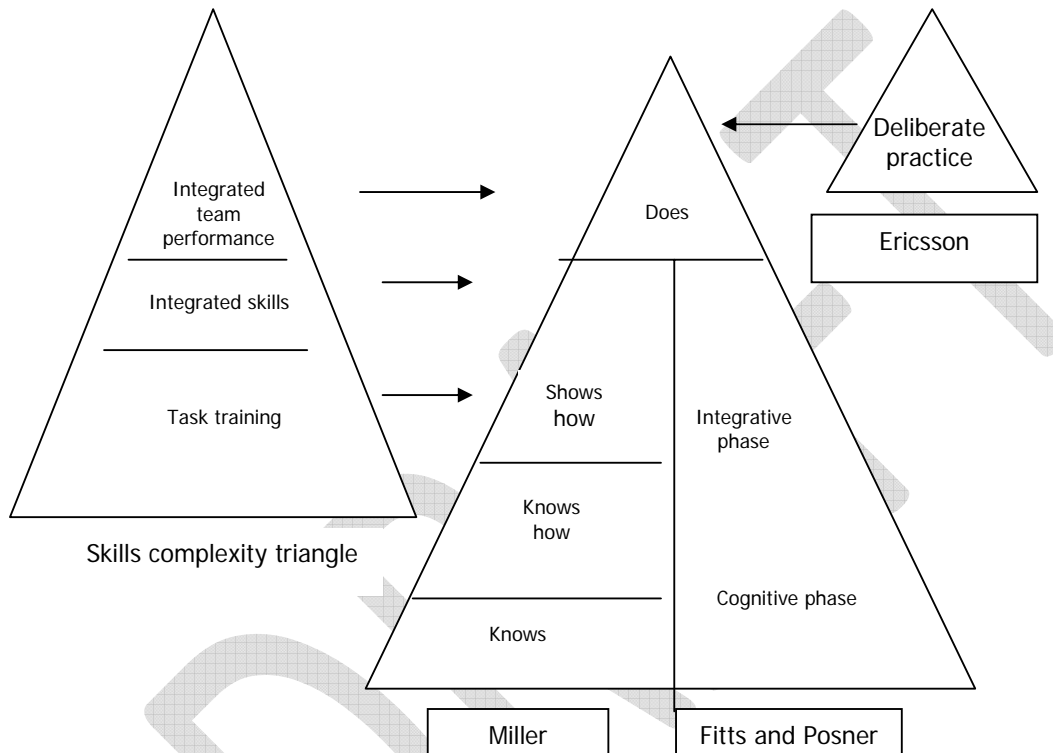
Miller’s model of clinical competence



Miller GE. The assessment of clinical skills/competence/performance. Academic Medicine (Supplement) 1990; 65: S63-S67.

This theme group addresses performance assessment, which we define as the **assessment of skills and behaviour**, both in academic and workplace settings. We recognise that there are many other perspectives which relate to standards required and credentialing. Performance assessment can be illustrated by building in a degree of complexity to Miller's pyramid, recognising both the development of performance expertise⁹ and the need for skills and behaviour maintenance through deliberate practice.¹⁰ This enhanced model recognises, through the skills complexity triangle, some of the contextual factors which impact on performance measures, both individual and systems related, including taking experience into account.

Miller's Model of Performance Assessment



We will examine the assessment tools designed to test the top two levels of Miller's pyramid; the 'shows how' and 'does' aspects.

Challenges for assessing the individual performance of health care professionals recognise that health care is increasingly being delivered in teams, so even if patients have the same medical condition the complexity of their care makes it difficult to compare performance.

Assessment tools for clinical competence include the Objective Structure Clinical Examination (OSCE), the Objective Long Case Examination Record (OSLER) and the Objective Structured Assessment of Technical Skills (OSATS). These are undertaken outside the 'real' clinical environment but have many aspects of realism of the workplace incorporated into them and are assessed at the "shows how" level of Miller's pyramid.

Workplace based assessments (WPBA) include the mini Clinical Evaluation eXercise (mini CEX), Direct Observation of Procedural Skills (DOPS) and Case Based Discussions (CbDs). WPBA tools assess at the "does" level of Miller's pyramid. Simulation exercises at this level are increasingly being considered as having a role in performance assessment.¹¹

Historical perspective

The assessment of clinical performance has historically involved the direct observation of assesses by professional colleagues. This stems from the traditional apprenticeship model which existed for hundreds, if not thousands, of years. Knowledge and skills in medicine were passed down from one person to another. The apprentice learnt from the master by observing and helping him treat patients.¹² The first medical schools in Greece and Southern Italy were formed by leading medical practitioners and their followers.¹³ Hippocrates, widely regarded as the father of medicine, was born in 460 BC on the island of Cos, Greece¹⁴ and is credited with suggesting changes in the way physicians practised medicine. They were encouraged to offer reasonable and logical explanations concerning the cause of a disease rather than explanations based on superstitious beliefs.¹³ The medical school in Alexandria established in 300 BC was considered a centre of medical excellence. Its best two teachers were Herophilus, an anatomist, and Erasistratus, who is considered by some to have founded physiology.¹⁵ Medical education in this school was based on the teaching of theory followed by practical apprenticeship under one of the physicians.¹³

In the United Kingdom medieval records identify that guilds maintained control of entry to the medical profession. Prior to the Medical Act of 1858, which established the medical professional regulatory body the General Medical Council (GMC), doctors in the UK were virtually autonomous practitioners and assessment was not the norm¹⁶. Aspiring doctors could pursue their studies at one of the three universities¹⁷ which offered undergraduate medical courses in the UK and would receive a university degree in medicine. Alternatively they could follow an apprenticeship with a senior practitioner and eventually join one of the licensing corporations; the Royal Colleges of Physicians and of Surgeons, which were established by royal charters in the mid-16th century, or the Society of Apothecaries, which was established in the early 17th century.¹⁸ There were no formal examinations for entry to the medical profession; apprentices were deemed satisfactory by their master and could then practice independently, or were awarded medical degrees within the universities.

The current situation in relation to performance assessment and national regulatory standards are that Canada, China and Japan have established national licensing examinations and the USA has national assessment for entry into postgraduate training. Several other countries are exploring the use of national licensing examinations e.g. Korea, Indonesia, Switzerland. There is currently no national licensing examination in the UK.

The importance of the theme

The current medical education world is dominated by discourse around accountability.^{19, 20} Along with the perceived loss of trust between society and professionals (including doctors)²¹, no one can be *assumed* to be competent. Clinical performance has to be proven by testing, measurement and recording. It also needs to be psychometrically acceptable and defensible. One of the major challenges is to ensure performance assessment is aligned to clinical teaching in the workplace and that it is feasible to deliver. A three stage model for assessing clinical performance has been suggested.²² This includes a screening test to identify doctors at risk who then undergo a more detailed assessment. Those who pass the screening test pursue a quality improvement pathway to enhance their performance, which may require different degrees of psychometric rigour. Hays et al suggested 3 domains of performance assessment which take account of experience and context; the doctor as manager of patient care, the doctor as a manager of their environment and the doctor as manager of themselves.²³

There is increasing acknowledgement of the need to explore the role of knowledge in competence and performance assessment. There is also a concern that competence and performance are not seen as two opposing entities but rather part of a spectrum along a

continuous scale. In order to achieve this there needs to be recognition of the complexities between performance and competence.²⁴

The increasing use of simulation is another area that requires debate and the development of supporting evidence. The level of simulation used in competency and performance assessment varies from the use of part task trainers and standardised patients in OSCEs to the use of covert simulated patients as 'mystery shoppers' in primary care in the Netherlands.²⁵ The international movement in quality improvement and patient safety has seen increasing use of simulation for performance assessment, led by anaesthesiology²⁶, but increasingly being used in other health care contexts. This has been led by the experience of simulation in other high reliability organisations, such as the aviation industry and the nuclear and oil industry.²⁷ Governments and regulatory authorities^{28,29} have developed national strategies in skills and simulation, recognising that consistent standards of practice enhance patient outcomes and experience of the health care services. Developing performance assessments using simulation may be the most defensible method of ensuring reliability and validity for individual practitioners' health care teams and organisations. The role of virtual reality and the use of technology will also need to be considered in performance assessment.

There is a need to develop internationally agreed standards of professional practice, and by developing standards in performance assessment we can begin to ensure consistency across national borders.

The remainder of this document provides an overview and background of some of the commonly used tools for the assessment of competence and performance.

SECTION A: COMPETENCE ASSESSMENT

Competence describes what an individual is *able to do* in clinical practice. The Association for the Study of Medical Education (ASME) booklet '*OSCEs and other Structured Assessments of Clinical Competence*'³⁰ was one of the background papers for this theme group, as it summarises the competence assessment tools which are being used. The following is based on this booklet. The tools described include the Objective Structured Long Case Examination Record (OSLER), the Objective Structured Clinical Examination (OSCE) and the Objective Structured Assessment of Technical Skills (OSATS).

Objective Structured Long Case Examination Record (OSLER)

In the traditional long case examination assesses spend one hour with a patient, during which they are expected to take a full formal history and do a complete examination. The assessee is not observed. On completion of this task the assessee is questioned for 20 – 30 minutes on the case, usually by a pair of examiners, and is occasionally taken back to the patient to demonstrate clinical signs. Holistic appraisal of the assessee's ability to interact, assess and manage a real patient is a laudable goal of the long case. However, in recent years there has been much criticism of this approach related to issues of reliability, caused by examiner bias, variation in examiner stringency and unstructured questioning and global marking without anchor statements³¹ Measurement consistency in the long case encounter will also be diminished by variability in degree and detail of information disclosure by the patient, as well as variability in a patient's demeanour, comfort and health. Furthermore, some patients' illnesses may be straightforward whereas others may be extremely complex. Assessee's clinical skills also vary significantly across tasks³², so that assessing on one patient may not provide generalisable estimates of an assessee's overall ability.^{31, 33, 34}

While validity of the inferences from a long case examination is one of the strengths of the genre, inferring assessee's true clinical skills from a one hour long case encounter is also

debatable. Additionally, given the evidence of the importance of history taking in achieving a diagnosis³⁵, and the need for students to demonstrate good patient communication skills, the omission of direct observation of this process is a deficiency.

To address the shortcomings of the long case, whilst still attempting to retain the concept of seeing a 'new' patient in a holistic way, the OSLER was developed.³⁶

The OSLER has ten key features:

- It is a ten item structured record
- It has a structured approach – there is a prior agreement on what is to be examined
- All assesses are assessed on identical items
- Construct validity is recognised and assessed
- History process and product are assessed
- Communication skill assessment is emphasised
- Case difficulty is identified by the examiner
- Can be used for both criterion and norm referenced assessments
- A descriptive mark profile is available where marks are used
- It is a practical assessment with no need for extra time over the ordinary long case

The OSLER consists of ten items, including four on history, three on physical examination and three on management and clinical acumen. For any individual item examiners decide on their overall grade and mark for the assessee and then discuss this with their coexaminer, agreeing on a joint grade. This is done for each item and also for the overall grade and final mark. It is recommended that 30 minutes is allotted for this examination.³⁷

There is evidence that the OSLER is more reliable than the standard long case.³⁸ However, to achieve a predicted Cronbach's alpha of 0.84 required 10 separate cases and twenty examiners, and thus raises issues of practicality³⁹.

Objective Structured Clinical Examination (OSCE)

In addition to the long case, assessee classically undertook between 3 and 6 short cases. In these cases they were taken to a number of patients with widely differing conditions, asked to examine individual systems or areas and give differential diagnoses of their findings, demonstrate abnormal clinical signs or produce spot diagnoses. However, students rarely saw the same set of patients, cases often differed greatly in their complexity and the same two assessors were present at each case. The cases were not meant to assess communication skills but instead concentrated on clinical examination skills. The assessment was not structured and the assessors were free to ask any questions they chose. Like the long case, there was no attempt to standardise the expected level of performance. The lack of consistency and fairness led to the development and adoption of the OSCE.

In an OSCE the assessee rotates sequentially around a series of structured cases. At each OSCE station specific tasks have to be performed, usually involving clinical skills such as history taking or examination of a patient, or practical skills, and stations can include simulation. The marking scheme for each station is structured and determined in advance. There is a time limit for each station, after which the assessee has to move onto the next task.

The basic structure of an OSCE may be varied in the timing for each station, use of checklists or rating scales for scoring and the use of clinician or standardised patient as assessor. Assessee often encounter different fidelities of simulation, for example simulated patients, part task trainers, charts and results, resuscitation manikins or computer based simulations, where they can be tested on a range of psychomotor and communication skills.

High levels of reliability and validity can be achieved in these assessments.⁴⁰ The fundamental principle is that every assessee has to complete the same task in the same amount of time and is marked according to a structured marking schedule.

Terminology associated with the OSCE format can vary; in the undergraduate area they are more consistently referred to as OSCEs but in the postgraduate setting a variety of terminology exists. For example, in the UK the Royal College of Physicians' membership clinical examination is called the Practical Assessment of Clinical Examination Skills (PACES) while the Royal College of General Practitioners' membership examination is called the Clinical Skills Assessment (CSA).

The use of OSCEs in summative assessment has become widespread in the field of undergraduate and postgraduate medical education^{41,42,43,44,45} since they were originally described,⁴⁶ mainly due to the improved reliability of this assessment format. This has resulted in a fairer test of assessee's clinical abilities since the score is less dependent on who is examining and which patient is selected.

The criteria used to evaluate any assessment method are well described⁴⁷ and summarised in the ASME Understanding Medical Education booklet: *How to design a useful test: the principles of assessment*.⁸ These include reliability, validity, educational impact, cost efficiency and acceptability.

Reliability: OSCEs are more reliable than unstructured observations in four main ways:

- Structured marking schedules allow for more consistent scoring by assessors, according to predetermined criteria
- Wider sampling across different cases and skills results in a more reliable picture of overall competence. The more stations or cases each assessee has to complete, the more generalisable the test
- Increasing number and homogeneity of stations or cases increases the reliability of the overall test score
- Multiple independent observations are collated by different assessors at different stations. Individual assessor bias is thus attenuated.

The most important contribution to reliability is sampling across different cases; the more stations in an OSCE, the more reliable it will be. However increasing the number of OSCE stations has to be balanced with practicality. In practical terms, to enhance reliability it is better to have more stations with one assessor per station than fewer stations with two assessors per station.^{48,49}

Validity: Validity asks "what is the degree to which evidence supports the inference(s) made from the test results?" Each separate inference or conclusion from a test may require different supporting evidence. Note that it is the inferences that are validated, not the test itself.^{50, 51}

Inferences about ability to apply clinical knowledge to bedside data gathering and reasoning, and to effectively use interpersonal skills, are most relevant to the OSCE model. Inferences about knowledge are less well supported by this method than inferences about the clinically relevant application of knowledge, clinical and practical skills.⁵²

Types of validity evidence include *content validity* and *construct validity*. Content validity (sometimes referred to as *direct validity*) of an OSCE is determined by how well the sampling of skills matches the learning objectives of the course for which that OSCE is designed.^{52, 53} The sampling should be representative of the whole testable domain, and the best way to ensure an adequate spread of sampling is to use a blueprint.

Construct validity (sometimes referred to as *indirect validity*) of an OSCE is the implication that those who performed better at this test had better clinical skills than those who did not perform as well. We can only make inferences about an assessee's clinical skills in actual practice, as the OSCE is an artificial situation.

To enhance the validity of inferences from an OSCE, the length of any station should be fitted to the task to achieve the best authenticity possible. For example, a station in which blood pressure is measured could authentically be achieved in five minutes whereas taking a history of chest pain or examining the neurological status of a patient's legs would be more authentically achievable in 10 minutes.⁵⁴

Educational impact: The impact on learning resulting from a testing process is sometimes referred to as *consequential validity*. The design of an assessment system can reinforce or augment learning, or undermine learning^{55, 56}. It is well-recognised that learners focus on assessments rather than the learning objectives of the course. By aligning explicit, clear, learning objectives with assessment content and format, learners are encouraged to learn the desired clinical competences. By contrast, an assessment system that measures learners' ability to answer multiple choice questions about clinical skills will encourage them to focus on knowledge acquisition. Neither approach is wrong – they simply demonstrate that assessment drives education and that assessment methods need to be thoughtfully applied. There is a danger in the use of detailed checklists as this may encourage assesseees to memorise the steps in a checklist rather than learn and practice the application of the skill in different contexts. Rating scale marking schedules encourage assesseees to learn and practice skills more holistically. Hodges and McIlroy identified the positive psychometric properties of using global rating scales in OSCEs, such as a higher internal consistency and construct validity than using checklists, but suggest the need to be explicit about the global ratings used as some are sensitive to the level of training of those being assessed⁵⁷.

OSCEs may be used for formative or summative assessment. When teaching and improvement is a major goal of an OSCE, time should be built into the schedule to allow the assessor to give feedback to the assessee on her/his performance, providing a very powerful opportunity for learning. For summative certification examinations, expected competences should be clearly communicated to the assesseees beforehand, so they have the opportunity to learn the skills prior to taking such assessments.

Cost efficiency: OSCEs can be very complex to organise. They require meticulous and detailed forward planning, engagement of considerable numbers of assessors, real patients, simulated patients and administrative and technical staff to prepare and manage the examination. It is, therefore, most cost effective to use OSCEs to test clinical competence and not knowledge, which could more efficiently be tested in a different format. Effective implementation of OSCEs requires thoughtful deployment of resources and logistics, with attention to production of assessment material, timing of sittings, suitability of facilities, catering and collating and processing of results. Other critical logistics include assessor and standardised patient recruitment and training. This is possible even in resource limited environments.⁵⁸

Acceptability: The increased reliability of the OSCE over other formats, and its perceived fairness by assesseees, has helped to engender the widespread acceptability of OSCEs among test takers and testing bodies.

Since Harden's original description⁴⁶ OSCEs have also been used to replace traditional interviews in selection processes in both undergraduate and postgraduate settings^{59, 60}; for example, for recruitment to general practice training schemes in the UK.

Objective Structured Assessment of Technical Skills (OSATS)

Another variation on this assessment tool is the OSATS. This was developed as a class room test for surgical skills by the Surgical Education Group at the University of Toronto.⁶¹ The OSATS assessment is designed to test a specific procedural skill, for example caesarean section, diagnostic hysteroscopy, cataract surgery. There are two parts to this: the first part assesses specificities of the procedure itself and the second part is a generic technical skill assessment, which includes judging competences such as knowledge and handling of instruments and documentation. OSATS are gaining popularity amongst surgical specialties in other countries.

Reliability: The data regarding the reliability of OSATS is limited. Many studies have been carried out in laboratory settings rather than in clinical settings. There is a high reported inter observer reliability for the checklist part of OSATS in gynaecological procedures and surgical procedures.^{62, 63}

Validity: OSATS have high face validity and strong construct validity, with significant correlation between surgical performance scores and level of experience.^{64, 65}

Acceptability: Most of the studies looking at OSATs have been done in simulated settings; therefore the evidence for acceptability is low. The majority of assessees and assessors report that the OSATS is a valuable tool which would improve a trainees' surgical skill and it should be part of the annual assessment for trainees.⁶⁴

Issues on competency assessment that merit discussion

There are several issues that merit further discussion or research in relation to competency assessment:

- the use of different systems of scoring
- who should be doing the scoring
- the use of different standard setting methods
- the development of national assessments of competence
- the use of simulation
- how to enhance training of assessors

The use of different systems of scoring

When OSCEs were first introduced extensive detailed checklists of each step of a clinical task were produced for each station. Examiners had to tick where an action was performed or not. Many examiners found this process unsatisfactory as deconstructing the performance to such a degree was felt to ignore the holistic aspect of professional competence. Checklists often focused on easily measured aspects of the clinical encounter and the more subtle but critical factors in clinical performance were overlooked or ignored.

While the objectivity of structured checklists would imply better reliability, checklists do not inherently provide more reliable scores. The use of rating scales to assess the performance of clinical skills has been shown to be reliable when used by expert examiners. Examiner training can improve their reliability further.

It may more effective to use checklists to assess technical skills in the earlier stages of learning (at the 'novice' end of the learning trajectory) and to use rating scales to assess more complex skills, especially with increasing levels of professional competence.

Who should be doing the scoring

There are two main approaches to scoring in OSCEs – examiners are either clinicians/physicians or non-clinician Standardised Patients (SPs). Physician/clinician examiners enhance the validity of the assessment because they can apply holistic judgements and integrate subdomains of sequence, logic, and other factors that might be difficult for a non-professional completing a binary checklist to capture.

Especially in the USA, the use of trained Standardised Patients for scoring is the norm, as it has been shown that well-trained SPs produce similar scores to clinicians, at least in assessing general entry level practitioners. However, clinicians using holistic scoring models may have greater utility in capturing higher levels of expertise.

Currently in other countries, such as Canada and the United Kingdom, at both undergraduate and postgraduate levels, examiners are clinicians or other healthcare professionals. It would probably need a considerable shift in cultural acceptance to move to standardised patients as the sole assessors of clinical competence.

The use of different standard setting methods for OSCEs

Standard setting or establishing the pass mark is critical in the process of determining who passes and who fails any particular assessment of clinical competence. The standard or pass mark indicates the minimum score that every candidate has to reach to pass the OSCE. While it is difficult to quantify such a complex concept as clinical competence, the reality is that examinations such as OSCEs are used to discriminate between those who have sufficient clinical skills and those who do not, for a particular level or purpose.

The fundamental principle underlying all standard setting methods is to reach a consensus on professional values and standards. There are many standard setting methods described in the literature but all the traditional ones were developed for multiple choice questions. It is debatable whether it is appropriate for these methods to be used for complex performance-based examinations such as OSCEs.

While the Angoff method has been used quite widely for OSCEs, this has been superseded by the emergence of a standard setting method developed specifically for performance-based examinations such as OSCEs, where a clinician-examiner is present to observe and score the candidates.

The Borderline Group method (and its variations, such as the Contrasting Groups and the Borderline Regression) is very efficient in terms of using the clinician-examiners to set the pass mark and not having to convene a separate panel of judges. It is highly credible, as it utilises the expertise of the examiners in making judgements about professional standards, and collates the judgements over a large number of examiners. It does require some expertise in processing the data, it is more reliable if the examiners are trained but overall, it has become regarded as the 'gold standard' method of standard setting for OSCEs⁶⁷.

Developing national assessments of competence

In 1992 the Medical Council of Canada (MCC) added a standardised patient component to its national licensing examination because of the perception that important competences expected of licensed physicians were not being assessed.⁶⁶ Since inception, approximately 2500 assessees per year have been tested at multiple sites throughout the country at fixed periods of time during the year.

In 1998, the United States Educational Commission for Foreign Medical Graduates (ECFMG) instituted an assessment of clinical and communication skills expected of foreign medical graduates seeking to enter residency training programs. From 1998 to 2004, when it was incorporated into the United States Medical Licensing Examination (USMLE), there

were 43,642 administrations making it the largest high stakes clinical skills examination in the world.⁶⁷ The assessment consisted of a standardised format of eleven scored encounters with a trained simulated patient. Competence was evaluated by averaging scores across all encounters and determining the mean for the integrated clinical encounter and communication. Generalisability coefficients for the two conjunctively scored components were approximately 0.70 to 0.90.⁶⁸ In 2004, the USMLE adopted the ECFMG clinical skills assessment model (CSA) and began testing all United States medical graduates.

The use of high fidelity simulation

Simulation has been integral to competency assessment enabling individual competences of history taking examination and professional skills to be measured reliably in OSCE format.⁴⁰ One of the concerns in using high fidelity simulation is that it may engender abnormal risk taking behaviours in a risk and harm free environment. Some assessees may form a comfort zone within the simulated environment and use this as their normal standard in the face of a challenging clinical workplace -“simulation seeking behaviour”.

A systematic review of assessment tools for high fidelity patient simulation in anaesthesiology identified a growing number of studies published since 2000. While there is good evidence for face validity of high fidelity simulation there is less evidence for its reliability and predictive validity, which is important in high stakes assessments.⁶⁹

Potential benefits of high fidelity simulation in assessment include the ability to assess both technical and non technical skills and the ability to assess teams. The use of simulation is expensive but can be cost effective but may help reduce adverse events, thus providing long term cost effectiveness.

Assessor training

Despite extensive research into assessment formats, little research has been carried out to determine the qualities of a ‘good’ assessor. Trainers and assessors are usually the same people, although extensive understanding of a training program may not equate to the ability to use assessment tools fairly, objectively and in the manner intended. Inter rater reliability is known to be a major determinant of the reliability of assessments. However, while several approaches to assessor training have been suggested, there is little evidence of the impact of these on performance.³⁰

SECTION B: ASSESSMENT OF CLINICAL PERFORMANCE

Performance describes what an individual *actually does* in clinical practice. The ASME booklet, *Workplace Based Assessment in clinical training*,⁷⁰ was another key background paper.

WPBA is a form of authentic assessment testing of performance in the real environment facing doctors in their everyday clinical practice. It is structured and continuous, unlike the opportunistic observations previously used to form judgement on competence. By using repeated assessments, an assessor has the opportunity to collect documentary evidence of the progression of individual assessees. This evidence may then be used to identify ‘gaps’ in practice which will allow the assessor and assessee to mutually plan individual development needs. Using a wide range of WPBA tools helps identify strengths and weaknesses in different areas of practice, such as technical skills, professional behaviour and teamworking.

There are various types of WPBA tools currently in use. A summary is given below, including format, reliability, validity, acceptability and educational impact.

Mini Clinical Evaluation eXercise (Mini-CEX)

The mini-CEX is a method of clinical skills assessment developed by the American Board of Internal Medicine to assess resident's clinical skills.⁷¹ It was particularly designed to assess those skills that doctors most often use in real clinical encounters. It is now being increasingly used in undergraduate assessment as well.⁷² The mini-CEX involves direct observation by the assessor of an assessee's performance in 'real' clinical encounters in the work place. The assessee is judged in one or more of six clinical domains (history taking, clinical examination, communication, clinical judgement, professionalism, organisation/efficiency) and overall clinical care, using a seven point rating scale. The assessor then gives feedback. The mini-CEX is performed on multiple occasions with different patients and different assessors.⁷⁰

The average mini-CEX encounter takes around 15-25 minutes.⁷³ There is opportunity for immediate feedback,⁷⁴ which not only helps identify strengths and weaknesses but also helps improve skills.⁷⁵ The time reported in the literature for feedback can vary from 5⁷⁶ to 17⁷⁷ minutes.

Reliability: Although there is good evidence of the reliability of the mini-CEX in the literature, a large proportion of this data comes from studies that employ the mini-CEX in experimental settings rather than in naturalistic settings. Experimental settings make it possible for a performance to be assessed by multiple assessors, a situation that is not practicable in real clinical settings. The reported inter-rater reliability of the mini-CEX is variable even among same assessor groups.⁷⁸ In addition, there is variation in scoring across different levels of assessors, with studies showing that residents tend to score mini-CEX encounters more leniently as compared to consultants.⁷² Inter-rater variations in marking can be reduced with more assessors rating fewer encounters, rather than few assessors rating multiple encounters⁷⁹ Between 10 and 14 encounters are needed to show good reliability, if assessed by different raters and on multiple patients.⁷³

In practical terms the number of mini-CEX encounters possible in real clinical practice settings needs to be balanced against the need for reliability. One possible method of reducing assessor variation in scoring is through formal assessor training in using these tools. The evidence for reducing this variation through assessor training is, however, variable, with some studies showing that training makes little difference to scoring consistency⁸⁰ whereas others report more stringent marking and greater confidence while marking following assessor training.⁸¹

Validity: Strong concurrent validity has been reported between mini-CEX and Clinical Skills Assessment (CSA) in the USMLE.⁷⁸ Any assessment tool that measures clinical performance over a wide range of clinical complexity and level of training must be able to discriminate between junior and senior doctors. In addition, senior doctors should be performing more complex procedures and performing more procedures independently. The mini-CEX is able to discriminate, with the senior doctors attaining higher clinical and global competence scores.⁸² The domains of history taking, physical examination and clinical judgment within the mini-CEX correlate highly with the similar domains of the American Board of Internal Medicine (ABIM) evaluation form.⁸³

Acceptability: One measure of acceptability is to record the uptake of forms by both assessors and assessees, with the assumption that higher uptakes will reflect greater acceptability for the assessment method. Some studies report a high uptake of mini CEX⁷² whereas other studies report a lower uptake.⁷⁷ A reduced uptake of forms may be related to time constraints and lack of motivation to complete these forms in busy clinical settings. Assessee may find completing WPBAs time consuming, difficult to schedule or even stressful and unrealistic.⁸⁴ However, both assessors and assessees have reported high

satisfaction rates with regard to the ability of the forms to provide opportunities for structured training and feedback.⁷⁹

Educational Impact: The perceived educational impact of the mini-CEX is related to the formative use of the assessment to monitor progress and identify educational needs. Appropriate and timely feedback allows assessees to correct their weaknesses and to mature professionally.⁸⁵ Assesseees find the mini-CEX beneficial as it reassures them of satisfactory performance and increases their interaction with senior doctors.⁸⁴ However, it is important to train assessors to give both good quality observation and feedback.

Direct Observation of Procedural Skills (DOPS)

DOPS is a method of assessment developed by the Royal College of Physicians in the UK specifically for assessing practical skills.⁸⁶ Although mainly used for postgraduate doctors, at present many medical schools are planning to use it as part of their undergraduate assessments. It requires the assessor to:

- Directly observe the assessee undertaking the procedure
- Make judgements about specific components of the procedure
- Grade the assessee's performance

Reliability: As with the mini-CEX, DOPS needs to be repeated on several occasions for it to be a reliable measure. Subjective assessment of competences done at the end of a rotation has poor reliability and unknown validity.⁸⁷ There is a wide variety of skills that can be assessed with DOPS, from simple procedures such as venepuncture to more complex procedures such as endoscopic retrograde cholangiopancreatography . The DOPS specifically focuses on procedural skills and pre/post procedure counselling carried out on actual patients.⁷⁷

Validity: The majority of assesseees feel that DOPS is a fair method of assessing procedural skills.⁸⁶ It has also been established that DOPS scores increase between the first and second half of the year, indicating validity.⁸⁸

Acceptability: The majority of assesseees feel that the DOPS is practical.⁸⁶ The mean observation time for DOPS varies according to the procedure assessed; on average feedback time took an additional 20–30% of the procedure observation time. Davies et al have demonstrated a mean of 6.2 DOPS were undertaken by first year foundation trainees in the UK. The median times for observation and feedback were 10 and 5 minutes respectively.⁸⁸

Educational impact: There is little current research on the educational impact of DOPS, although the opportunity for timely feedback gives it the potential for high educational value.

Case-based discussion (CbD)/ Chart Stimulated Recall (CSR)

These are essentially case reviews, in which the assessee discusses particular aspects of a case in which they have been involved to explore underlying reasoning, ethical issues and decision making. Repeated encounters are required to obtain a valid picture of an assessee's level of development.

The CSR⁸⁹ was developed for use by the American Board of Emergency Medicine. The CbD tool is a UK variation. These tools can be used in a variety of clinical settings, for example clinics, wards and assessment units. Different clinical problems can be discussed, including critical incidents. The CbD assesses seven clinical domains; medical record keeping, clinical assessment, investigations and referrals, treatment, follow up and future planning,

professionalism and overall clinical judgement.⁷⁵ The assessee discusses with an assessor cases they have recently seen or treated. It is expected they will select cases of varying complexity. A Cbd should take 15-20 minutes and 5-10 minutes for feedback.⁸⁸ The number of cases suggested during the first two years of training is a minimum of six per year.⁷⁵

Reliability: There is data available for the CSR which shows good reliability.⁹⁰

Validity: A study using different assessment methods has shown that assessment carried out using CSR is able to differentiate between doctors in good standing and those identified as poorly performing and correlates with other forms of assessment.⁹⁰

Acceptability: This has not been extensively reported in the literature. As with all WBPAs, the evidence regarding the true costs of these assessments is scarce. Costs include training of all assessors as assessor bias reduces the validity of WPBA. There are also costs associated with assessor and assessee time in theatres, clinics and wards. It has been calculated that the median time taken to do an assessment and provide feedback is about 1 hour per month for one trainee.⁸⁸

Educational impact: Again, there is little published research on the educational impact of these methods. However, as with DOPS, the opportunity for feedback on clinical reasoning and decision making is thought to be valuable in helping learners progress.

Mini Peer Assessment Tool (Mini-PAT)

The mini-PAT is a modified version of the Sheffield Peer Review Assessment Tool (SPRAT), which is a validated assessment with known reliability and feasibility.⁹¹ The mini-PAT is used by the Foundation Assessment Programme and various Royal Colleges in the UK.

There are a number of methods to collate the judgements of peers, the most important aspect of which is systematic and broad sampling across different individuals who are in a legitimate position to make judgements. The mini-Pat is one example of the objective, systematic collection and feedback of performance data and is useful for assessing behaviours and attitudes such as communication, leadership, team working, punctuality and reliability. It asks people that the assessee has worked with to complete a structured questionnaire consisting of 15 questions on the individual doctor's performance.⁹² This information is collated, so that all the 'raters' remain anonymous, and fed back to the trainee. A minimum of eight⁹⁴ and up to twelve assessors⁹⁴ can be nominated. The time taken to complete a mini-PAT assessment is between 1-50 minutes with a mean of 7 minutes.⁸⁸

Reliability: Good inter-item correlation in the mini-PAT has been reported but correlation between assessments is lower.⁸⁸ Inter-rater variance has been reported, with consultants tending to give a lower score. However, the longer the consultants have known the assessee, the more likely they are to score them higher.⁹²

Validity: The 15 questions within the mini-PAT assessment were modified and mapped against the UK Foundation Assessment Programme curriculum and the General Medical Council (GMC) guidelines for Good Medical Practice, to ensure content validity. Strong concurrent validity has been reported between the mini-PAT, mini-CEX and Cbd. There is evidence of construct validity, with senior doctors achieving a small but statistically significantly higher overall mean score at mini-PAT compared to more junior doctors.^{88, 92}

Acceptability: As discussed earlier, acceptability may be inferred from the number of completed assessment forms. A high response rate of 67% was achieved in one study.⁹² One of the main advantages of such assessments is that they can be kept anonymous, encouraging assessors to be more honest about their views. The main criticisms are that

feedback is often delayed and that it is difficult to attribute unsatisfactory performance to specific clinical placements due to anonymous feedback.⁹⁵

Educational Impact: The mini-PAT has been used primarily as a formative form of assessment rather than summative.⁹⁵ This allows the assessee to reflect on the feedback they receive in order to improve their clinical performance.

This method of assessment has been used in industry for approximately fifty years and in medicine within the last ten years.⁹⁶ The aim is to provide an honest and balanced view of the person being assessed from various people they have worked with. This can include colleagues, nurses, residents, administrative staff but also medical students and patients.⁷⁵ However this method has some disadvantages, which include risk of victimisation and potentially damaging harsh feedback.⁸⁴

Issues that merit further discussion on performance assessment

There are several issues that merit further discussion or research in relation to performance assessment:

- reliability of workplace based assessment tools and their alignment to learning outcomes
- the feasibility of tools in contemporary clinical workplaces
- the use of simulation in transferring performance to the workplace
- the use of feedback in performance assessment
- the development of tools to assess teamwork

Reliability, feasibility and predictive validity

As discussed, there is good evidence for individual WPBA tools but less evidence regarding how these fit together into a coherent program of ongoing assessment. A major concern of clinicians is the impact the administration of these assessments will have on clinical workload for both assessors and assessee. In addition, less is understood about how the results of these assessments predict the future clinical performance of individuals.

Simulation and transfer

There is initial evidence that using simulation to train novices in surgical skills, such as laparoscopy, results in improved psychomotor skills during real clinical performance. However, the complexity of real clinical performance is often overlooked and a simulation program needs to be part of a comprehensive learning package⁹⁷ in order that high stakes decisions can be made.⁶⁹

Feedback in performance assessment

There has been a tendency to move towards multisource feedback, as evidence of performance in the workplace.⁹¹ Peer group feedback can give learners a realistic perspective about standards of performance. In addition feedback can be given by simulators by collecting data of events and interactions and from attached monitors. Tutors and facilitators can provide feedback where the main focus is on striving for better professional performance.

Assessing teamwork

While teamwork and professionalism are emphasised in many curricula documents,⁵ the assessment of these 'soft' skills is problematic. Formative assessments, such as multisource feedback, often incorporate these elements. However, there is a lack of validated summative assessment tools.

DRAFT CONSENSUS STATEMENTS AND RECOMMENDATIONS

Widespread use of competency assessments such as the OSCE, and of performance assessments such as the mini-CEX, grew out of assumptions and principles that are incorporated into the consensus statements and recommendations that follow.

Section A: Competency Assessments (OSCEs)

Consensus:

1. Competency assessments are needed to assess at the “shows how” level of Miller’s pyramid.
2. Competency assessments should be designed and developed within a theoretical framework of clinical expertise.
3. Competency assessments can generate scores with sufficient reliability and validity, which can be conflicting for high stakes decision making and feedback.
4. Competency assessments can provide documented evidence of learning progress and readiness for practice.
5. Competency assessments should be designed using a wide range of methodologies

Recommendations for individuals (including individual committees)

1. Articulate clearly the purpose of the competency assessment. Ensure that the purpose and the blueprinting criteria are congruent and that they reflect curriculum objectives.
2. Develop assessments that measure more than basic clinical skills; for example, assessment of clinical competence in the context of complex patient presentations, inter-professional and team skills, ethical reasoning, and professionalism.
3. Use scoring formats appropriate to the content and training level being assessed; no single scoring format is best for all.
4. Employ criterion referenced methods for setting standards for decision making; methods like contrasting groups, borderline group, or borderline regression.
5. Identify how competency assessment can be best linked to a remediation framework

Recommendations for institutions

1. Implement competency assessments, especially when summative, as a component of an overall assessment plan for the curriculum.
2. Engage in collaborations within and across institutions where common approaches can enhance standards of practice.
3. Provide appropriate levels of support to the faculty members who create and administer competency assessments. These assessments require support staff, equipment, space, and funds. Creating content, designing scoring instruments, setting standards, recruiting and training simulated or standardised patients (SPs) and examiners, plus organising the assessment itself, all require time and team work to a degree that is frequently underestimated and underappreciated.
4. Formalise the recognition of scholarly input to competency assessment.
5. Engage faculty who participate in competence assessments as content developers and examiners, with faculty development. Recognise their participation.

Outstanding Issues

1. Establishing a common language and criteria regarding scoring instruments to clarify what is meant by global rating (versus, for example, generic rating), checklists, and grids
2. Ensuring all aspects of competence are assessed, including ‘softer’ competences of leadership, professionalism etc
3. Ensuring assessment of competence but promotion of excellence
4. Creating a wide range of scoring tools for scoring assessments based on complex content and tasks. Moving beyond a dichotomous discussion of checklists versus global

rating scales to development of scoring formats keyed to the content and educational level of a specific assessment.

5. Ensuring consensus around use and abuse of terminology
6. Promoting better understanding and use of existing standard setting methods.
7. Developing feasible as well as psychometrically sound standard setting methods for small cohorts.
8. Articulating guidelines and rationales regarding who should score assessments, including standardised patients, non physician and physician raters; as well as further establishing training requirements and methods for each.
9. Exploring integration of simulation and SP based methodologies for assessment
10. Gathering further information about the predictive validity of use of simulation in high stakes assessments
11. Developing and promoting more feasible equating methods for assessments
12. Establishing the role of the OSCE as a 'gatekeeper' for progression
13. Engaging stakeholders in identifying key areas from assessment e.g., patient groups
14. Ensuring constituency of assessors by training and observation

Section B: Performance Assessments (WPBA)

Consensus:

1. Workplace based assessments assess the two behavioural levels at the top of Miller's model of clinical competence, shows how and does, because the workplace provides a naturalistic setting for assessing performance over time
2. Workplace based assessments can cover not only clinical skills but also aspects of skills such as professionalism, decision making and timekeeping.
3. Workplace based assessments, specifically the mini CEX and multisource feedback, can generate sufficiently reliable and valid results for formative purposes
4. Workplace based assessments occur over multiple occasions and lead to documented evidence of developing clinical competence
5. Workplace based assessments should provide the timely feedback that is essential to learning and enhancing its acceptability amongst users.

Recommendations for individuals (including committees)

1. Ensure assessments go hand-in-hand with learning and are not isolated exercises
2. Encourage assesseees to engage in assessments regularly and throughout the year rather than clustering them just as their final assessment is due
3. Articulate and disseminate the purpose of the assessment clearly. Ensure that each assessment is appropriately driven by learning objectives, practice standards and assessment criteria
4. Create workplace based assessments within the framework of a program of assessment that incorporates the use of multiple methodologies, occasions and assessors. This can reside within a portfolio

Recommendations for institutions

1. Ensure time and costs for training and assessments are included in workplace planning; for example, included in job plans, clinic schedules, and budgets
2. Select methodologies which are appropriate for the intended purpose (summative or formative)
3. Use results in accordance with the intended purpose to protect assessment processes. For example, using formative outcomes for summative decision making pushes assesseees to delay assessment occasions till the end of the assessment period, thereby undermining the documentation of their development over time
4. Educate assessors and assesseees about the assessment method and the use of the rating tools to enhance the quality of feedback, reduce stress of both groups, and to preserve the naturalistic setting

5. Promote strong reliability by minimising inter-rater variance (e.g., increasing number of assessors), providing clear assessment criteria and training to all assessors and minimising case effect (e.g., drive wider sampling, take into account assessee level of expertise, and set case selection criteria)
6. Enrich validity by incorporating use of external assessors to offset the effect of assessees choosing their own assessors
7. Combine workplace based assessment with other forms of performance assessment to provide more comprehensive evaluations of competence, as occurs with a portfolio
8. Individualise assessments based on assessee performance; for instance, individuals in difficulty may need more frequent assessments and feedback than those who are performing well
9. Identify individuals in difficulty as early as possible with robust systems of assessment

Outstanding Issues

1. Establishing the reliability and validity of outcomes from Direct Observation of Procedural Skills (DOPS) and Case based Discussions (CbD)/ Chart Stimulated Recall (CSR) in workplace (naturalistic) settings.
2. Estimating the total number of DOPS/CbD/CSR required to achieve sufficient reliability and validity for decision making (guiding remediation) within specified contexts (e.g., specialty-based)
3. Exploring how various workplace assessment tools, when combined, facilitate learning and meet the overall objectives of training or practice
4. Exploring the role of workplace based assessments in summative assessment, including approaches to reporting outcomes (e.g. scores, profiles)
5. Researching further the concept of transfer between simulation and workplace based performance
6. Developing more feasible approaches to workplace based assessments and how these relate to portfolio assessment processes
7. Conducting predictive validity studies on the impact of workplace based assessments on long term performance
8. Developing workplace based assessments for team working and professionalism