ACADEMY OF
MEDICAL ROYAL
COLLEGES

ACADEMY OF MEDICAL ROYAL COLLEGES

Improving Assessment

JULY 2009

Designed and Typeset by
Millbank Media Ltd

CONTENTS

EXECUTIVE SUMMARY

The Academy of Medical Royal Colleges was asked by the four Chief Medical Officers (CMOs) to explore "a UK wide approach to assessment with common standards and common processes and instruments." This report comprises the outcome of that project.

The main recommendations are:

- The purposes of any assessment method and system needs to be clearly defined and communicated to all participants
- The main purpose of Workplace Based Assessment (WPBA) is to help trainees identify areas for improvement and is thus formative not summative
- The trainee's performance in the current formative WPBA must not be used to rank them for selection to specialty
- When an individual WPBA is to be used summatively then the purpose must be clearly defined to the trainee and assessors in advance, trained assessors must be used and an adequate number of assessments made by differing assessors to obtain a valid judgement
- At annual review, outcomes from a range of WPBAs should be one component of the evidence on which a judgement is made about the trainee's progress
- It is neither possible nor desirable to formally assess each individual competency at every stage of training
- Colleges, Faculties and the Foundation Programme should have a clear plan for the further development and implementation of WPBA and a strategy to achieve wider acceptance of this approach
- All assessors should be trained in order to improve the standards of WPBA delivery
- There needs to be consensus on the definition of the roles of supervision, and the nomenclature used to describe them, to avoid confusion and help identify training needs
- Employers should recognise the substantial contribution to the training and productivity of the future workforce made by Supervisors and Clinician Assessors. Adequate time to complete assessments effectively should be written into their job plans and it is recommended that Educational Supervisors receive at least 0.25 PAs /trainee and Clinical Supervisors 0.25 PAs/trainee. Training excellence should be recognised and rewarded
- There should be clear support mechanisms in place through local structures and the Deaneries to support supervisors in dealing with poorly performing trainees
- While differing WPBA instruments need to be used in differing specialties to test differing attributes, instruments must be consistent within training programmes and rating systems should be similar where possible across specialties
- Assessments should move from use of numerical values to the standard expected at the end of a period of training or that required for independent practice
- An ongoing Workplace Based Assessment Forum should be established with sharing of expertise between Colleges, Faculties, regulators and Deans to achieve the highest standards of assessment systems.

1. INTRODUCTION

1.1 Background

Training in medicine traditionally followed an apprenticeship model, and formal assessments were mainly directed towards the testing of knowledge. Clinical and practical skills were only assessed as a component of formal examinations. Any assessments of actual performance in the workplace were largely informal, often anecdotal and rarely documented. This situation still persists for many other professions, but in postgraduate medical training, there has been an increasing focus on assessing doctors in the workplace – workplace based assessment (WPBA).

Whilst knowledge-based testing is largely very well established and enjoys a substantial investment in infrastructure and resources from Colleges and Faculties, the same cannot be said for WPBA. The introduction of performance-based training in postgraduate medicine has highlighted a need to develop and introduce suitable new assessment methods. Although the profession would naturally desire to improve practice and promote excellence in assessing doctors, developing and introducing such profound changes takes both time and resources, and requires a substantial change in culture.

1.2 A crisis in assessment

The introduction of WPBA into postgraduate medical training in the UK was made for Foundation trainees under the umbrella of 'Modernising Medical Careers'[1] Unfortunately, unrealistic timescales together with a lack of resources and inadequate assessor training led to the hurried implementation of WPBA and the development of undesirable practices. This has resulted in widespread cynicism about WPBA within the profession, which is now increasing; some of the mistakes have been repeated in implementing WPBA for specialty training with similarly hurried implementation due to a tight timetable set by the Postgraduate Medical Education and Training Board (PMETB). The profession is rightly suspicious of the use of reductive 'tick-boxing' approaches to assess the complexities of professional behaviour, and widespread confusion exists regarding the standards, methods and goals of individual assessment methods.[2-5]

If the confidence of the profession is to be regained and WPBA introduced successfully and fully across all specialties and between hospitals, there needs to be a change of thinking. There must be a move away from the increasingly mechanistic approach that is currently being promoted, and a move back to the basic educational principles that have served well previously. Assessment is inseparable from learning, and at the heart of it is the relationship between the educational supervisor and the trainee. It is this relationship that must be fostered and encouraged. The primary purpose of WPBA must be to promote learning and inform this relationship. The role of the Educational Supervisor needs to be clarified and defined, with an emphasis on teacher development and support through constructive feedback. Whilst WPBA methods must be reliable and valid, above all, they must be feasible within the normal time and resource constraints of clinical practice. This will take both time and resources. Imposing the unrealistic expectations of enthusiasts will be a recipe for failure.

1.3 Towards commonality

In September 2006, Dr Jewell wrote to the Academy of Medical Royal Colleges on behalf of the four CMOs and the UK Modernising Medical Careers (MMC) Strategy Group. They were concerned that confusion could result from varying documentation, scoring systems and differing standards used in different environments. The Group 'supported a UK-wide approach to assessment with common standards and common processes and instruments. They acknowledged that there was a wide range of instruments that could be used, but supported working towards consistent methodologies and documentation across the UK. The AoMRC was suggested as being a suitable candidate for this role as it could provide a national overview across specialties and across the Foundation/specialty training divide. It was also considered well placed to co-ordinate the necessary input from postgraduate deans and other stakeholders'.

The aim of this report is to document the evidence base for the assessment methods currently in place, address the issues raised by the four CMOs and make recommendations to improve WPBA in UK postgraduate medicine.

2. MYTHS AND MISUNDERSTANDINGS

The introduction of new curricula and assessment systems has raised many concerns. Misunderstandings have arisen, partly due to a lack of clarity over some contentious issues, which this section aims to explore.

2.1 Did we really need a new assessment system?

Although some may have thought there was little wrong with previous assessment systems, they were far from perfect. Although most incorporated some clinical or practical component, the emphasis was largely on knowledge-based assessments, frequently in the absence of clearly laid out curricula. Clinical competence does not just involve knowledge, decision making, or procedural ability, but other factors such as interpersonal skills, professionalism, self appraisal and an ability to improve. Traditional knowledge based exams do not assess many of these areas.

Formalised WPBAs are relatively new for the medical profession, and there is limited experience of their use in medicine. Unfortunately, their over-hasty implementation has meant that there has been insufficient time or resources to fully develop the role of the educational supervisor in assessment and to provide robust training for the large numbers of clinicians who will necessarily be involved in assessment.

Existing evidence shows that supervisors are not always good at observing the clinical skills of trainees,[6] and may document that they have assessed procedures which they have never observed.[7] There is a tendency for doctors to give colleagues the benefit of the doubt, and reluctance to give negative feedback in face-to-face assessment.[8-10] Doctors are poor at self assessment, particularly assessment of professional behaviour and are reluctant to rate themselves.[11] In the interests of accountability and transparency and improving practice new assessments are needed. We need to observe trainees performing tasks they will be called upon to perform independently once they complete their training.

Assessment drives learning. If designed and implemented well WPBA can lead to trainees learning and developing the skills behaviours and attitudes we want in the junior members of our team and ultimately in our future colleagues.

2.2 Who is driving these changes?

The drivers for the development of new assessment systems arise both from the desire of the profession to improve and promote excellence and from regulatory change.

- The public expect the profession to be accountable and doctors to be able to demonstrate all aspects of their responsibilities

- Most doctors who come before the GMC because of doubts about their fitness to practice do so because of communication and interpersonal relationship problems, not lack of knowledge and these domains can best be assessed in the workplace

- With the introduction of the European Working Time directive and changes in working patterns, the profession can no longer assume that trainees will acquire adequate exposure to an appropriate range of experience simply through time spent in training

- Assessment needs to be aligned with the new competency curricula developed as part of Modernising Medical Careers (MMC). A comprehensive and defensible assessment system will also provide information and evidence when dealing with a struggling trainee.

2.3 Is it possible to fail a single WPBA?

The traditional model of assessment in medical training is one of intermittent high stakes testing. Junior doctors have, throughout their careers (school, undergraduate and postgraduate) been used to sitting exams for which they will receive a grade and either pass or fail. Much of this assessment is of knowledge with the addition of assessment of practical skills in the form of Objective Structured Clinical Examinations (OSCE) and the clinical parts of membership exams. The consequences of failing any individual exam will vary from minor embarrassment to potential career derailment. Feedback given following written exams will also vary, potentially losing an opportunity to provide the candidate with valuable information and guidance for improvement.

The tension between the use of the newer WPBAs for learning development (formative - assessment *for* learning) and for use in judgement (summative - assessment *of* learning) has yet to be resolved. The most vital aspect of WPBA is the ability to detect and then inform the developmental needs of a doctor. Unfortunately, this essential aspect is often overlooked by both trainees and by trainers because of a mindset arising from traditional knowledge based assessment practice. It is very difficult to adjust this mindset, and this has led to many of the assessments being seen as a 'mini-exam' by both trainees and supervisors, where receiving what is perceived to be a 'poor' score is seen as a failure. This should not be the case for a number of reasons. The assessments are designed to be formative, in that they provide an opportunity for the supervisor / assessor to observe the trainee in their day-to-day practice and give feedback on performance.[12] Indeed the main purpose of WPBA is improvement in performance resulting from the feedback.

Performance in an individual assessment such as mini-Clinical Evaluation Exercise (min-CEX) or Case based Discussion (CbD) is very context specific.[13,14] A trainee may do very well when examining a patient with ascites, but have problems assessing a patient with acute confusion.

In addition, assessors vary in their expectations; there are consultants who naturally tend to mark highly and those who are more stringent. It is therefore not possible to make a summative judgment (i.e. pass/fail) based on a single, or even a small number of observations. Multiple observations are required, preferably from a range of different assessors to improve reliability.[15]

The value of an individual assessment lies in the provision of guidance and suggestions for improvement, and in observing the doctor patient interaction which may otherwise happen rarely in routine clinical practice.

For these reasons WPBAs should be seen as a low stakes process which it is not possible to fail, with the emphasis on feedback and suggestions for improvement. The score on an individual assessment does not represent a trainee's overall ability.

2.4 Can one bad assessment have a negative impact on a doctor's career?

A further misunderstanding related to the 'mini-exam' concept is that doing badly in an assessment will have a negative impact on a doctor's career. For assessors, a common reason given for not using the lower part of the scale on WPBA, is a desire to avoid negative consequences for the trainee. Trainees are concerned about scoring low marks and do not like being told they are 'average'. In a competitive environment they want to score 'top marks' in all assessments. This has led to many assessments being completed towards the end of the year when trainees have had more experience and feel they are able to perform better. This negates the primary purpose of the assessments in guiding the development of the trainee.

Many of the descriptors used for the assessments specify the standard to be expected at the end of a particular period of training or for independent practice. It is not likely or anticipated that a junior trainee will attain this level early in the placement, and trainees should not be disheartened if they do not do so. One advantage of using scales which set the standard at the end of a period of training is that it allows a trainee to show progression through training. This may well be more valuable than attaining high marks throughout.

As previously discussed, performance is context specific and an individual assessment is not a reliable measure of performance and should not be viewed in isolation. Any serious concerns raised during an assessment may prompt further enquiry by the educational supervisor, but cannot, without supporting evidence form the basis for any remedial action.

Although WPBA is primarily for aiding development, nonetheless judgements of progress still need to be made. For the reasons outlined above, it is not usually feasible to have a completely separate set of assessments solely for the purpose of making judgements, and the same assessments will need to serve both purposes.

2.5 Uncertainty about how the outcomes of WPBA will be used

Assessment drives learning, and a well thought out and implemented system of performance assessment is an opportunity to encourage trainees to develop. Using WPBA can provide a structure to facilitate feedback and does encourage trainers to make recommendations about performance.[16,17] The existing WPBA have been predominantly designed to provide formative feedback, not to make high stakes, summative decisions about progress or promotion. However, these decisions do need to be made using all available information, including WPBAs.

Participation in WPBA and the assessment outcomes may contribute to progress decisions made as part of the Annual Review of Competence Progression (ARCP) process and for eligibility for the award of a Certificate of Completion of Training (CCT). Those involved in making these decisions need to understand the properties of the individual assessment methods and the numbers needed for the results to be reliable. An individual assessment result should not be taken out of context.

Trainees need to understand how the results of assessments will be used and each specialty should have clear statements to this effect as part of the curriculum and introduction to the assessment system. If trainees feel they need to score highly throughout they will be reluctant to seek out the more challenging cases, from which they may learn more.

2.6 Can WPBA results be used for ranking?

One specific area of uncertainty and much controversy is the use of WPBA in recruitment and selection into specialty training posts. The collapse of the Medical Training Application Service (MTAS) process and associated problems have been well documented[3,18-21] and have led to concern and confusion about the use of WPBA for ranking and selection purposes.

The WPBA instruments are primarily formative; they were not developed to be used for selection and are not validated for this purpose. WPBA may be used to demonstrate that the required competencies have been met, and a well organised portfolio may reflect a committed and enthusiastic candidate but this needs further investigation. Neither are the WPBA instruments suitable to be used for ranking. Although the individual results of Multisource feedback can be presented alongside the means for the relevant cohort (medical school year, foundation trainees etc), this is to give the trainee an idea of his or her place amongst the group and guide areas for improvement rather than to provide an external ranking.

2.7 WPBA is frequently seen as a 'tick box' exercise.

WPBA has attracted much criticism, one of the main complaints being that it is a 'tick box' exercise which dumbs down medical training. This may be partly because of the association with competency-based training, and mistrust of the principles and implementation of Modernising Medical Careers as a whole.

Further factors which have contributed to this attitude include:

- The requirement for a minimum number of assessments. This reinforces the perception that WPBA is another hurdle to be jumped rather than an integral part of training which may be useful for personal development. There are good reasons for having a minimum number of assessments, including reliability, defensibility, transparency and to detect deterioration over time
- Rushed implementation has meant assessors are inadequately trained and trainees are not familiar with the assessment strategy
- Many of those completing the assessments do not understand the purpose of the assessment and the need for constructive feedback
- The assessments are not valued and are felt to be too simplistic. Seen as an imposed process by both trainees and trainers, they are therefore not completed effectively
- They are seen to assess mundane tasks at which the trainee is already competent, such as phlebotomy. This is particularly an issue for the more senior specialist trainees for whom WPBA has been introduced part of the way through their training. There have been reports of middle grades asking more junior doctors to complete Directly Observed Procedural Skills (DOPS) for such procedures.

2.8 Assessments are seen as separate from, and an addition to, the normal workload

The clinical day is already more than full, any addition to this workload is going to be difficult to achieve. Assessment does not take place in isolation; WPBAs are designed to assess aspects of what a doctor does as part of their daily workload. The value comes from assessing how doctors perform in the normal day to day setting. The more WPBA is integrated into routine practice the better the validity of the assessment. Further, the more the assessments are part of routine practice the greater the potential for realistic and useful feedback.

However, in order for assessment to be integrated into routine practice there needs to be a culture change. Ideally, in a training environment, all clinical encounters should provide an opportunity for feedback. When mini-CEX was originally developed, the cases used were not selected or prepared as it was designed to be part of routine practice, e.g. the first case of the day in outpatients.[22,23] This impromptu assessment can only be achieved with full integration into routine working and will only happen if the assessment and feedback is seen as useful by the trainees, moving away from the exam mentality. Integrating WPBA will help move away from the 'tick box' attitude that currently exists and prevent the problems associated with assessments being postponed to the end of the year.

2.9 Every component of the curriculum must be assessed

Recommendation:

- *It is neither possible nor desirable to formally assess each individual competency at every stage of training.*

Any assessment used as part of medical training needs to be reliable (i.e. reproducible – if the same candidate were to take the same assessment again, the same result would be obtained). For traditional assessments this may be achieved by increasing the number of test items, such as the number of multiple choice questions (MCQs) or the number of stations in an OSCE. As previously discussed performance assessment is both time and context specific with variability between cases, on the same case over time and between assessors.

This means that in order to attain reliable results for WPBA, multiple observations using multiple assessors are required. The assessment instruments have been evaluated in trials to identify the numbers required for results to be reliable. In reality this needs to be interpreted in a way which can be managed within the clinical environment. All assessment systems associated with the new specialty curricula are mapped to both the curriculum which is itself organised within the domains of Good Medical Practice (GMP) in a process known as blueprinting. This ensures that assessments sample across the domains of GMP and the areas covered by the curriculum. A more detailed description of blueprinting and reliability is given in the PMETB guide to developing and maintaining an assessment system.[24]

In practice this means that assessments need to cover a wide range of clinical topics and skills. In this way a representative picture can be developed without assessing every single clinical scenario. A vital role of the supervisor is to work with the trainee to ensure appropriate sampling across GMP and the relevant specialty curriculum.

2.10 Uncertainty among assessors of the standard required

The descriptors used for similar assessments developed for use at different grades and by different specialties vary. Some use the standard expected of a trainee at their level of training, some the standard expected at the end of a particular period of training and others that required for independent practice. In addition some scales use numerical values alongside

text descriptors whilst others have no numerical score. Consequently assessors are unclear about how to mark using the forms available.

Assessors are more comfortable making judgments with which they are familiar. Assessments that use scales with 'at the level of independent practice' as the outcome can be very reliable with a few observations.[25] There is a strong argument for moving away from numerical scales toward more specific text descriptors. The use of numerical scores may lead to an inappropriate statistical manipulation of assessment results, thus lending a spurious scientific validity to what are essentially subjective judgements of professional behaviour by competent experts.

3. DETERMINING THE PURPOSE OF ASSESSMENT

When designing an assessment system it is vital that the purpose of the assessment is clearly defined at the outset.[13,26] The purpose of the assessment system will determine:

- the choice of instruments used
- the way in which they are combined
- the number of observations required
- how the outputs are used to make decisions.

Instruments cannot be chosen purely for their reliability or ease of use.

For individuals to engage successfully with WPBA there must be a benefit perceived by trainees, assessors and trainers. Effective communication will help to ensure that participants in the assessment process understand the purpose of assessment and their role in achieving that goal (e.g. giving feedback). Any concerns about how assessment outcomes information will be used must also be addressed.

The purposes and benefits of assessment will be different for the individual being assessed, the training programme, and the employer/public.[27] Some of the purposes will be common, but where there is divergence there is potential for conflict, particularly between the formative and summative roles of assessment.

The majority of WPBA instruments are designed to be formative in nature, to provide an opportunity for constructive feedback and the development of learning objectives. It will still be necessary, however, to make summative judgements (e.g. at annual review), and there remains a lack of clarity about how data collected in a formative interaction can be used summatively. This leads to difficulty for supervisors who may have the dual role of teacher and assessor. Trainees are unlikely to embrace the assessments as an opportunity to challenge themselves if a less than perfect result could be used to inform summative decisions about their progression or selection.

Discussion at the WPBA forum identified the following possible purposes of assessment for three different stakeholders, the individual, the training programme and the employer/public.

Purpose and benefits of assessment for the individual:

- Providing feedback
- Providing motivation through positive feedback
- Encouraging an aspiration to mastery
- Promoting learning and informing learning objectives
- Providing evidence to inform the Annual Review of Competence Progression (ARCP)
- Demonstrating progression
  - over time within one post
  - across multiple posts on a rotation.

A clear pathway for progression helps clarify what is now expected of the trainee. The assessment process is a continuum, and the outcome of previous assessments should be taken forward to new posts to help inform the meeting with the educational supervisor. The supervisor can then build a better picture of the trainee and develop new learning objectives, reinforcing a positive perception of the relevance of WPBA.

One of the challenges in developing WPBA is how to assess the higher level aspects of performance, such as the integration of the many skills involved in making complex decisions.

Purpose and benefits of assessment for the training programme:

- Demonstrating trainee progression
- Contributing evidence for summative decision making at the ARCP, and providing evidence for appeals
- Identifying the doctor in difficulty
- Providing data to support decisions involving trainees in difficulty. Such decisions need to be defensible and made using robust information
- Providing information when designing remediation packages
- Identifying patterns of behaviour. The level of engagement with the assessment progress appears to be an index of commitment and possibly of ability. There are likely to be differences in behaviour between good and poor trainees - strong trainees will actively seek (presumably positive) feedback and the weak may not, only participating in the minimum number of assessments.

Purpose of assessment for the employer/public:

- Ensuring practitioners are competent
- Encouraging excellence.

Recommendations:

- *The purposes of any assessment method and system needs to be clearly defined and communicated to all participants*
- *The main purpose of WPBA is to help trainees identify areas for improvement and is thus formative not summative*
- *When an individual WPBA is to be used summatively then the purpose must be clearly defined to the trainee and assessors in advance, trained assessors must be used and an adequate number of assessments made by differing assessors to obtain a valid judgement.*

4. DESIGNING ASSESSMENT SYSTEMS

There has been increasing organisational and research interest in the assessment of performance of medical practitioners,[28] both in the UK and internationally. In the UK all specialties and the foundation programme have new curricula and assessment systems currently being followed by trainees. A great deal of work has gone into the development of the assessment systems and further work is ongoing to revise the assessment systems to meet all of the PMETB principles for 2010. In 2008, these principles, based on best practice and the available literature, were revised and incorporated into the PMETB standards.[29]

This section aims to discuss the steps required in the development of a reliable, well validated and defensible performance assessment system. The discussion is primarily aimed at Colleges and Faculties, but will be of interest to other stakeholders including Deaneries and Specialist Advisory Committees. There is currently a considerable variation in the provision and infrastructure of Colleges and Faculties for the development of WPBA.

Although the process described here represents an ideal that may not always be achievable, the basic steps need to be followed – it is not possible to pick a package of assessments off the shelf and expect them to work in a new situation or specialty. Adequate consideration needs to be given to all factors including the purpose of the assessments, the mechanisms of development and delivery, and how they align with the aims of the curriculum.

4.1 Structure

Developing and delivering curricula and assessment systems is a significant challenge, requiring adequate resources and infrastructure. Colleges already have an extensive and well developed structure for delivering membership exams; this is viewed as an important part of the role of the college in maintaining standards. The properties of performance assessment in the workplace are very distinct from those of the membership exams, and it is important that WPBA is addressed as a major part of an overall assessment package.

It is recommended that Colleges and specialties give consideration to the following factors:

- Leadership: there should be an identified individual in each College with specific responsibility for WPBA. This role needs to be clearly differentiated from exam support although there may be an overlap in personnel. A suggested person specification has been developed stressing the importance of appointing an individual with the seniority to lead and drive the process forward (Annex 1)
- Personnel: appropriate support will be required, both for the development process and, following implementation, for the maintenance of assessment programmes. This may include:
  - Clerical support, for the individuals developing and designing assessment, and dealing with queries from trainees
  - Data manager. WPBA generates large amounts of data which need to be managed
  - Psychometric and statistical support – again for the initial development then ongoing evaluation. Each College is unlikely to be able to resource this internally; sharing expertise through the AoMRC could help
  - IT support to ensure the incorporation of the assessments in an e-portfolio
- Policy: each College needs to consider their policy with respect to WPBA and the data generated on:
  - Data protection. Trainees need to have confidence in how their data will be used, and future litigation is a possibility
  - Data ownership
  - Evaluation and research – these are closely linked. Ongoing evaluation of the assessment programme is vital. The data may also be used for research, to contribute to the growing evidence base on the utility of WPBA in new contexts
  - Who is responsible for delivering assessments
  - How the results are to be used
  - Who is eligible to act as an assessor
- Links: It is important that the structure within each college has close links with other organisations involved in the development and quality assurance of assessment systems. The ideal is to share information, experience and best practice, and to promote a coordinated approach. The following organisations all have a role:
  - AoMRC
  - Deaneries and SHAs
  - PMETB
  - General Medical Council (GMC)
  - National Clinical Assessment Service (NCAS)
  - NHSMEE, NES and Northern Irish and Welsh equivalents
- Finance: Each College / Faculty will need to identify a budget stream to deliver their assessment responsibilities.

4.2 Process

Those involved in the actual development of the assessment systems will need to consider a number of issues. There are many ways of looking at this process, here we describe a framework which divides the process into a number of steps.[13]

4.2.1 Step 1: Define the content to be assessed, and the focus of the assessment

The first step is to decide what is to be assessed. Assessment drives learning[30,31] and assessments must be designed to encourage trainees to learn and adopt desirable skills and behaviours. For WPBA, this raises a significant challenge in defining what makes a good doctor. In the UK the GMC has produced the document *'Good Medical Practice'* (GMP) which sets out the principles and values on which good practice is founded. These principles together describe medical professionalism in action and are divided into seven domains.[32] Good Medical Practice is the chosen anchor of PMETB which will provide a framework for all curricular and assessment systems.[24]

Defining a good doctor is a major and global challenge. The Royal College of Physicians and Surgeons of Canada have developed the CanMEDS framework[33] which identifies the different roles of a physician and the key competencies associated with each role. The Accreditation Council for Graduate Medical Education (ACGME) and the American Board of Medical Specialties (ABMS) in the USA have jointly focused on the outcome for their trainees: the "outcomes project."[34] Six domains of competence have been identified: patient care, medical knowledge, practice-based learning and improvement, interpersonal and communication skills, professionalism, and systems based practice.

For trainees the content to be assessed will also be directed by the curriculum, and in practice most specialty assessment systems are mapped to Good Medical Practice, as is the relevant curriculum. The new curricula are competency-based, and each College has been through a process of defining the competencies specific to its specialty. Although there are differences between specialties, much of what doctors do is common to all. The AoMRC and the NHS Institute for Innovation and Improvement has developed a Medical Leadership Framework and Competencies which may become common to all specialties and is also identifying generic components which need to be in all curricula. Assessing such non-clinical competencies presents a challenge.

Alongside these efforts, the role of clinicians as part of the healthcare team remains in evolution. Sir John Tooke's review into MMC 'Aspiring to Excellence', concludes that there needs to be 'common shared understanding of the roles of all doctors in the contemporary healthcare team.'[3]

4.2.2 Step2: Define the purpose of assessment

Once the content has been identified the next question to ask is "Why are we assessing the trainee on this topic/skill/behaviour?" or "What is the purpose of the assessment?" The specific purpose of any assessment will dictate the choice of assessment instrument and the way in which the results can be used. The purpose of assessment has been discussed in Chapter 3.

4.2.3 Step 3: The blue printing process

How a doctor delivers clinical care and functions within the healthcare team is highly complex and there is no single assessment method capable of assessing all facets of a clinicians ability.[14] It is necessary to use a combination of methods all designed to assess different aspects of performance.[35-37] For example, testing of knowledge can be achieved through written exams including MCQs, whilst assessment of more practical skills such as communication will require direct observation. The methods chosen may overlap and provide complementary evidence to a doctor's overall competence (triangulation).

It is a standard of PMETB that all assessment be referenced to GMP and the approved curriculum.[29] Blueprinting involves creating an assessment system that ensures all the attributes of GMP have been tested, along with important aspects of the curriculum. How each domain of good practice is to be assessed must be clear. The aim is to maximise the efficiency of the assessment package by reducing duplication, but making sure there is appropriate assessment over all of the domains of practice. Each domain must be assessed separately as they represent independent constructs, and improved performance in one domain does not necessarily mean improved performance in another.[38]

An effective way of presenting the assessment blueprint is in tabular form with the assessment methods mapped against GMP and the curriculum domains. A summary of all the currently approved blueprints are on the PMETB website. http://www.pmetb.org.uk/index.php?id=approvedcurricula

It is not realistic to assess every trainee on every aspect of the curriculum. The varieties of situations in which a clinician must work along with the range of clinical conditions make this impossible. It is therefore important to sample across the curriculum ensuring an appropriate mix of attributes has been tested and that all domains of GMP have been assessed. The weighting of each component of the assessment system in making an overall judgment about an individual trainee must be specified. The assessment system must be designed such that each trainee may be assessed on any component of the curriculum to encourage a wide breath of learning.

4.2.4 Step 4: Choose or develop methods

Part of the blueprinting process involves identifying the assessment methods which will make up the assessment system. There has been a proliferation of WPBA instruments being developed and implemented across postgraduate medical training in the UK. This includes variation both between, and within specialities at different levels of training. One example has been the geographical variation within the Foundation programme in the form of MSF in use. This made it difficult to transfer assessment outcomes across discipline and may have caused confusion for both trainees and assessors as trainees rotated through different specialties or geographies. Many of these methods are well established internationally, others have been recently developed specifically for training programmes in the UK.[17,39] In addition there are many instruments in development to fill gaps in the blueprints where there are no existing methods. Table 1 below, gives a summary of the WPBA methods, and the table at Annex 2 illustrates the methods which make up the assessment systems being used by all specialties and the Foundation Programme.

TABLE 1: WORKPLACE BASED ASSESSMENT METHODS

MULTI SOURCE FEEDBACK (MSF)
- Colleagues, clinical and non clinical
- Patient ratings

DIRECT OBSERVATION
- Videoed Consultations
- Directly Observed Procedural skills (DOPS)
- Procedure based assessment (PBA)
- Mini-clinical evaluation exercise (mini-CEX)

ROUTINELY GENERATED DATA
- Significant event reporting
- Clinical Audit
- Morbidity and mortality data
- Prescribing habits
- Case based discussion (CBD)
- Correspondence – SAIL
- Surgical reporting systems
- Critical incident review

COVERT SIMULATED PATIENTS

ORAL PRESENTATIONS

Colleges have often developed their own assessment instruments for specific purposes. The choice of assessment instrument will be driven by the pre-defined purpose and focus of the assessment. Individual assessments are context specific[13,14,37] and will need to be adapted to align with the curricular objectives and the needs of the individual specialty.[36,40] For example the same MSF instrument does not behave consistently across disciplines due to the varying mix of available raters.[41,42] The content to be assessed will vary, and by necessity so will the instruments chosen to make up the assessment system. There are differences between specialties in the attributes deemed important, for example procedural ability and decisiveness for Surgeons or report writing for Pathologists.

Differences exist between the performances expected of a doctor as they progress through training. For domains such as professionalism, different components are important at differing stages of a career,[11] with leadership and decision making being more relevant for the more senior trainee. Assessment instruments will need to be able to adapt to these requirements, or different assessment instruments may be required.

Even within a fairly narrow context, such as the assessment of procedural skills, the defined purpose of assessment may differ. For example the Royal College of Obstetrics and Gynaecology use assessment to decide when the trainee is ready to move to independent practice (five successful assessments for a specified procedure, with at least two different assessors one of whom must be a consultant). Whereas the assessment system designed by the Intercollegiate Surgical Curriculum Programme (ISCP) follows a trajectory through training. The trainee and the educational supervisor meet at the beginning of each attachment to review the portfolio and decide objectives and which procedural assessments should be met during that attachment.

Whilst it is necessary for specialties to devise and evaluate their own assessment systems there are potential areas to develop commonality. A number of possible advantages were identified during the joint Forum. These included:

- Cost efficiency
- Transparency
- Improving the training of assessors
- Working towards transferable assessments.

Resources for developing assessment systems are limited and using a collaborative approach would make the most efficient use of the resources available (both financial and human). There are cost benefits in jointly developing and trialling new instruments as well as in the ongoing administration and evaluation of existing programmes. The process for ensuring doctors are fit to practice and striving to improve needs to be transparent and understandable to the wider public, and this is easier to achieve with a simpler more standardised system.

However, for successful implementation of WPBA, development and implementation of instruments must

be specific to specialties and integrated into the curriculum. Desirable though standardisation may be, a top-down approach will not achieve results, and harmonisation of methods must develop in time through the sharing of experience.

Each assessment method has its own advantages and disadvantages. A useful framework for evaluating each method is the utility index described by Van Der Vleuten.[43] The index includes five components:

- Educational impact
- Validity
- Reliability
- Cost efficiency
- Acceptability.

This is further discussed in the PMETB document *'Developing and Maintaining an assessment system.'*[24] There will need to be a compromise between these components and the relative importance of each factor will depend on the purpose of the assessment. For high stakes test of knowledge such as membership exams ensuring high reliability is vital. However, for performance assessment the need for validity and authenticity may sometimes be more important than the reliability coefficient. PMETB have acknowledged the difficulties in producing accurate measures of reliability for WPBA and have encouraged the use of utility when assessing compliance with PMETB standards.[44] A range of different and complementary assessment methods will be required, with the further advantage of providing triangulation of evidence.

Feasibility is not specifically part of the original utility index but is implicit in the factors of cost efficiency and acceptability. Postgraduate training programmes in the UK involve a large number of individuals who will all need to contribute to the assessment process. In the context of implementing performance assessment in the workplace, feasibility is very important. Clinicians have many conflicting demands and limited time to spend on educational activities. The assessments need to be achievable within the time available, and those involved need to believe that the time is well spent. The overall assessment burden must be calculated. Methods that fit into normal routines and work patterns are more likely to be successfully implemented. The demands imposed upon the assessor and trainee by the assessment system must be realistic, otherwise there is a significant risk the assessments will be conducted poorly, the benefits will be lost and the system will fail.

A summary of the evidence for the most commonly used assessment methods is presented at the end of this report along with a discussion of their advantages and disadvantages (Appendices A to D).

All of these considerations will need to be taken into account when deciding which instruments to use as part of an assessment system. One attraction of using existing methods is that many are well characterised and fairly extensively evaluated. A note of caution however, the reliability and validity of any particular assessment method is not intrinsic to the instrument[40] but is but a product of subject and judge variation when applying the instrument or method in a particular context. Furthermore, we are interested in the reliability and validity of the judgment or decision made on the basis of information obtained as a result of the assessment, and not of the instrument itself.[45] This means that one cannot guarantee the same reliability with every implementation,[36] and each instrument will need to be re-evaluated when used in a new context. Reassuringly both MSF and mini-CEX have proved to be adaptable and have been used in a number of different clinical settings and specialties with similar results for reliability and validity.

From a practical perspective Colleges and Faculties need to make progress with developing comprehensive assessment systems and do not necessarily have the time or resources for detailed re-evaluation prior to implementation. It is recommended that the following diminishingly important set of four criteria is used when choosing their instruments. Further evaluation of reliability can then be undertaken as part of the quality management process.

- Does it measure what is intended? (as defined by the curriculum)
- Is it feasible to implement? (taking into consideration time and resources)
- Can it provide for all the intended purposes? (feedback profiles, etc)
- Has it undergone evaluation and proved satisfactory in another context?

Once chosen and implemented, every method/instrument requires quality management evaluation in two phases. First, is it feasible to use in this new context? If yes, does it achieve satisfactory reliability in this context?

It is possible to spend time and resources debating specific aspects of assessment design such as the relative merits of different rating scales, the number of items making up an assessment, and the wording of descriptors. In general, increasing the number of observations will have a greater impact on improving reliability than increasing the number of items.[46] The wording of descriptors is important as the standard expected will be implicit in the descriptors used.

Recommendation:

- *Assessments should move from use of numerical values to the standard expected at the end of a period of training or that required for independent practice.*

4.2.5 Step 5: Train Assessors

Training of all assessors is essential for the reliable use of WPBA. The training ideally should be face-to-face, but in some circumstances will need to be simplified to written or electronic media given the large number and variety of assessors used. No assessor should be allowed to grade a trainee's performance without having been trained in that assessment method and how to provide constructive feedback. Both the trainee and the assessor need to understand the purpose of WPBA as a formative instrument, the importance of the assessment as a learning opportunity and how the data generated will be used, including making progress decisions.

Recommendation:

- *All assessors should be trained to improve the standards of WPBA delivery.*

4.2.6 Step 6: Standard setting

Once the characteristics of a good doctor have been identified, and the blueprinting process is complete, standards need to be set against which to target assessment. Ultimately all attempts to assess standards for medical practice will involve making a judgement and the value of the assessment methods used lies in the defensibility of that judgment.[47]

There is much concern about setting standards for workplace based assessments. In trying to address these concerns it is first necessary to define the different types of standard.

- Quality standards
  - — of individual assessment methods
  - — of assessment system design
  - — of assessment delivery
- Performance standards.

4.2.6.1 Quality Standards

The standards set out in the PMETB document *'Standards for Curricula and Assessment systems'*[29] provide a model of quality standards for the development of Workplace Based Assessment systems. This includes defining the purpose of the assessments and blueprinting the system against the curricula and GMP. The overall system must be defensible, transparent and secure. Furthermore, there are quality standards for individual assessment methods; they must be reliable, valid, and fair. The document *'Developing and Maintaining an assessment system – a PMETB guide to good practice'*[24] gives a good overview of all of these concepts.

PMETB has a statutory role in quality assuring the assessment systems and all Colleges and Faculties are required to comply with the principles by 2010. However, the AoMRC also has a role in helping Colleges and Faculties achieve a consistent and high standard in the process of assessment system design and development. One mechanism for this is the institution of forum in WPBA with the aim of sharing expertise and providing practical and ongoing guidance and support for those involved in devising and maintaining assessment programmes. The first forum, jointly hosted by AoMRC and PMETB, was held in April 2008. Possible formats for such a forum might include the use of electronic resources and online discussion, though an element of face to face contact is needed to maximise potential gains.

Potential advantages of a forum include:

- Reducing the duplication of effort in developing assessment systems
- Facilitating the development of new assessment instruments
- Pooling experience and expertise in developing training programs for assessors
- Increasing the opportunity to utilise expert support
- Harmonising the quality of workplace based assessment
- Allowing peer review for quality assurance
- Providing a consistent approach to the purpose of assessment across specialties
- Helping to achieve a consistent high standard of assessment design and evaluation by providing advice and support
- Sharing good practice in the development of assessment systems
  - — Educational, Statistical, Psychometric.

Recommendation:

- *An ongoing Workplace Based Assessment Forum should be established to share expertise between Colleges, Faculties, regulators and Deans to achieve the highest standards of assessment systems.*

Even the most meticulously designed and validated assessment system or instrument loses value if it is not delivered appropriately. The formative impact is lost if no feedback is given, and a trainee is unable to demonstrate progression if all assessments are completed immediately prior to the ARCP. Currently there are significant problems with the implementation of the new WPBAs with many anecdotal examples of bad assessment practice reported by Foundation School leads, trainee surveys[4] and by delegates attending WPBA training sessions.

4.2.6.2 Performance Standards

Participants in assessments must achieve certain performance standards in order to pass the various assessment thresholds. It is unrealistic to attempt to set explicit criteria for performance standards for every competency identified in every specialty curriculum and at every level of training. This represents a reductive approach that is inappropriate when using the expert experience of the assessor in coming to a judgement on professional behaviour. Furthermore, the development of a single common standard across specialties is likely to remain a futile ambition, as the pre-defined purposes of assessment will vary across specialties and WPBAs are very context-specific and will behave differently.

Assessors should use rational and defensible methods to set performance standards, but there is rarely an empirical basis for standards-setting. When considering performance standards there are a number of issues to consider.

- Are the potential assessment outcomes linear (varying degrees of too poor, borderline, varying degrees of good) or diagnostic (a problem with insight, a problem with social and communication skills, no performance problems, etc.)?
- If linear, is it appropriate to aggregate the components of the assessment, or do they actually measure different aspects of performance as a profile?
- For each element in the final profile, is performance being judged against an intended absolute standard (whether or not this can be described in terms of clear criteria), or is it being judged relative to other participants?
- If relative, on what basis should the proportion of participants passing each threshold be defined?
- If absolute, is the standard based on assessment content (e.g. a borderline candidate will be able to answer question x or perform task y) or an observer response (that was better than borderline, and that wasn't)? The literature names and describes various systems for each.
- If absolute, will the assessment outcome correct for assessment precision (borderline +2 Standard Errors of Measurement (SEM), borderline -1 SEM etc.)?

There is a wealth of literature from knowledge and competence level testing on issues 4-6.[47-52] Issues 1-3, however, are the first and most important issues to address in workplace based performance assessment.

The purpose of the assessment instrument will determine whether the outcomes are linear or diagnostic. When combining these outputs of WPBA it is not appropriate to reduce the different measures to a single numerical pass or fail result, as the different domains assessed represent different constructs.[53]

Within each dimension the assessments should be referenced against absolute standards. In many cases, because of the highly complex and integrated nature of the tasks, these will have to be based on observer response standards that are difficult or impossible to express in terms of clear criteria. By deciding on the level of supervision required for a particular surgical procedure, in running the acute medical take, or reporting on CT scans, clinicians regularly make judgements about the ability of their trainees, and are comfortable in doing so. These expert judgements are based on previous experience of many trainees contributing to the development of the observer's own absolute standard of what is expected of a trainee at that particular point in training. 'Realistic criterion referenced standards are, in the end, normative'[50] and based on the experience of the assessor or standard setting panel.

4.2.7 Step 7: Reporting and review system

Once the assessment system is in place there needs to be a system for evaluation and review as part of the ongoing quality assurance process. Colleges need to demonstrate compliance with the quality indicators set out by PMETB.

5. THE ROLE OF THE SUPERVISOR IN WPBA

As WPBA becomes accepted as an essential part of the culture of quality training, the role of the trainer will require further development, with both the trainee and the trainer adopting a new and professional approach to the assessments. This section explores the relationships between the various roles involved in trainee supervision at the local level.

All doctors have a responsibility for teaching, training and assessing as outlined in Good Medical Practice,[32] and some will take on the additional role of supervisor. The process of WPBA is intended to be trainee-led and the relationship between trainee and supervisor a balanced one. For this to be successful a level of maturity is required of the trainee, who needs to seek out the trainer and arrange times to meet. This will involve an evolution in the traditional relationship between junior doctor and supervisor, from a parental approach to an adult/adult relationship, one in which juniors are viewed as colleagues. In order to ease the transition into postgraduate training trainees should be encouraged to develop a level of responsibility for their own learning during undergraduate education.

At present there is inconsistency in the nomenclature used for the roles and the structures of supervision, with different definitions used by some specialties and deaneries. This is a source of confusion, particularly at points of transition such as the start of specialty training or if a trainee moves between deaneries. The PMETB survey of trainers highlighted this inconsistency, finding many areas of overlap in the roles of those who considered themselves to be either educational or clinical supervisors.[54]

As part of the Quality Framework Operational Guide PMETB have redefined the roles of Clinical and Educational Supervisor (www.pmetb.org.uk):

- **Clinical Supervisor (CS)**
  A trainer who is selected and appropriately trained to be responsible for overseeing a specified trainee's clinical work and providing constructive feedback during a training placement. Some training schemes appoint an Educational Supervisor for each placement. The roles of Clinical and Educational Supervisor may then be merged

- **Educational Supervisor (ES)**
  A trainer who is selected and appropriately trained to be responsible for the overall supervision and management of a specified trainee's educational progress during a training placement or series of placements. The Educational Supervisor is responsible for the trainee's Educational Agreement.

For WPBA to succeed, it is important that appropriate levels of supervision are provided at each stage of development. There is considerable variation both in the clinical practice of different specialties and in the levels of supervision required at each stage. It is vital that there is a named individual at consultant level who is responsible for overseeing a trainee's educational progress.

Recommendation:

- *There needs to be consensus on the definition of the roles of supervision, and the nomenclature used to describe them, to avoid confusion and help identify training needs.*

The initial meeting between the trainee and the ES is very important, setting the tone for future interactions, letting the trainee know what is expected of them and that they are responsible for leading their own development and ensuring they complete the assessments.

However, although WPBA is designed to be trainee led, this does not absolve the trainer of responsibility. Neither does it absolve the College Tutor and Deanery.

Specifically in relation to WPBA, the ES has a role in:

- Ensuring that sampling is appropriate. For the assessments to be reliable and overcome problems with case specificity there must be multiple observations, by multiple observers, over time.[30,55] There is potential for poor trainees with insight – the conscious incompetent,[56] to choose cases and assessors likely to suit them. As part of interim meetings the ES should ensure that a variety of clinical situations have been assessed by a mix of raters including consultants

- Intervening if assessments are not happening (either due to lack of trainee engagement or organisational factors including a difficulty in finding people willing to assess). The ES may be able to help, and should seek to prevent trainees leaving assessments to the last minute

- Reviewing assessment outcomes and formulating development plans. This includes interpreting the results and investigating any concerns raised

- The outcomes of WPBA should inform the appraisal process, with clear learning goals being agreed, and a clear understanding of the purpose of assessment and the supervisor's dual role as teacher and assessor

- Providing feedback, including the reports of an MSF assessment. The trainee's response to feedback is dependent on the quality of that feedback. This is important if MSF is to stimulate a change in practice[57,58]

- Identifying and helping trainees in difficulty. This is one of the purposes of WPBA. The ES may provide initial help and advice for the trainee, and, if further action is needed, provide evidence for the Deanery

- Liaising with College Tutor/Foundation lead/Deanery, particularly when there are problems.

In order to fulfil this role effectively, Supervisors need to be trained in how to use the individual assessment instruments and be familiar with the purpose and characteristics of the assessment systems and the curricula they support.

The integrity of both assessors and of the process is also crucial, to prevent abuse of the system. A strategy for training the trainers and the development of common standards for training assessors may help improve the integrity of the process. Assessors and supervisors need training and support especially when there are concerns about performance requiring negative feedback. Mechanisms to provide such support should be in place both at hospital level and within the Deanery network. Consistency is required in making defensible judgements.

A common concern is the lack of resources and recognition for WPBA, which has significantly impacted on its development and implementation. This is supported by the PMETB trainers survey which found that over half of training programme directors have not been appraised for their educational duties.[54] The training and assessment of the medical workforce is important, and it must be adequately resourced and contributions rewarded. Resources are needed at both a national level for development and at a local level for successful implementation.

At the local level employers may not recognise the workload of educational roles such as educational supervisor, which should be have adequate time in the job plan. A lack of recognition for educational roles together with a de-motivated workforce is leading to reluctance to engage with the process of assessment and supervision, which in turn is seen as having been imposed without full evaluation.

Recommendations:

- *Employers should recognise the substantial contribution to the training and productivity of the workforce made by Supervisors and Clinician Assessors. Adequate time to complete assessments effectively should be written into their job plans and it is recommended that Educational Supervisors receive at least 0.25 PAs/trainee and Clinical Supervisors 0.25 PAs/trainee. Training excellence should be recognised and rewarded*

- *Colleges, Faculties and the Foundation Programme should have a clear plan for the further development and implementation of WPBA and a strategy to achieve wider acceptance of this approach*

- *There should be clear support mechanisms in place through local structures and the Deaneries to support supervisors in dealing with poorly performing trainees.*

ANNEX 1: SUGGESTED SPECIFICATION FOR WORKPLACE-BASED ASSESSMENT LEAD

Reforms to postgraduate training as a result of MMC have led to a much greater focus on work place based assessment. All Colleges and Faculties have now submitted their assessment systems for specialty trainees to PMETB for approval. A significant amount of time and effort has gone into this process; however, this is just the beginning of the road towards developing a well validated, reliable, defensible assessment system. PMETB has evaluated the assessment systems against principles one, two and five of the *'Principles for an assessment system for postgraduate medical training'* (published September 2004). Colleges and Faculties will need to comply with all nine principles by 2009. This is going to be a major challenge over the next few years and there is scope for greater co-operation between Colleges to help in the achievement of this goal.

A key step in the ongoing development of postgraduate assessment is for Colleges and Faculties to appoint an individual who has responsibility for leading and overseeing this process. This does not necessarily need to be a clinician, but be an individual who has the seniority and credibility required to drive the development process forward. Some Colleges already have such an individual and the AoMRC Assessment Committee would like to stress the importance of appointing a lead for WPBA, to coordinate the development of assessment strategy. Colleges will also need to give considerable attention and thought to the administrative and governance structure required to deliver a comprehensive, defensible assessment strategy.

Some possible considerations when developing a personal specification for a WPBA lead:

Essential attributes:

- Significant interest in postgraduate education
- A sound knowledge of the regulatory framework governing postgraduate medical education and the certification process for doctors completing higher specialist training
- Expertise in assessment methodology, particularly workplace-based assessments
- Familiarity with the College / Faculties' curricula and associated assessment strategy
- The ability to work closely with the Examination and Training departments or the equivalent within each College / Faculty.

Principle responsibilities:

- To continue to develop an assessment strategy with particular reference to WPBA
- To ensure that the assessment strategy meets all the PMETB's Principles for Assessments
- To advise on projects piloting new methods of assessment to establish an evidence base
- To set standards through competency based frameworks for assessment
- To liaise with PMETB, MMC, the Academy of Medical Royal Colleges and other organisations involved in training and assessment
- To be responsible for the development of processes for the training and performance monitoring of assessors with reference to work place based assessment
- To ensure that appropriate measures are in place for the quality assurance of the WPBA tools used
- To ensure integration of WPBA with other assessment methods.

ANNEX 2: TABLE OF ASSESSMENTS USED BY EACH SPECIALITY INCLUDING THE RECOMMENDED NUMBERS.

	ANAESTHETICS	ITU	GENERAL PRACTICE	EMERGENCY MEDICINE	PUBLIC HEALTH
DOPS	Yes 12	Yes	Yes As appropriate	Yes 4-6	Direct Observation
OTHER ASSESSMENT OF PROCEDURE					
MINI CEX	Anaes CEX 8	Yes	Yes (in hospital posts) 6	Yes 4-6	
CBD	Yes 4	Yes	Yes 6	Yes 4-6	Yes
MSF	Yes 2/Year	Yes	Yes ≈1/year	Based on SPRAT ≥1/year	Yes 1 in total
SAIL					
PATIENT SATISFACTION QUESTIONNAIRE			Yes 2-3 in total		
PORTFOLIO	Yes		Yes	Yes	Yes – and log book
TEACHING				Yes - tools not clear	
OTHER	Annual assessment of attitude and behaviour (AAAB)	Plus primary specialty AAAB	Consultation observation tool (COT) if in primary care 6	Observed clinical care in the workplace Audit Piloting 4 WPBA tools	Written reports
EXAM	FRCA	Diploma optional	AKT and CSA	MCEM FCEM	MFPH - OSPHE
TRAINERS REPORT	Yes	Yes	Clinical supervisors structured report (CSSR) 3 in total	Yes	Yes

	OCCUPATIONAL MEDICINE	OPHTHALMOLOGY	PATHOLOGY, MED MICRO	PATHOLOGY CHEM PATH	PATHOLOGY HISTOPATH
DOPS	Yes – for visits and meetings 4	Yes 2 for each of 22 procedures in total	Yes ≥6	Yes ≥6 in ST1 & 2	Yes ≥6
OTHER ASSESSMENT OF PROCEDURE		OSATS 2 for each of 13 procedures in total			
MINI CEX	Yes 4	No (CRS)		Yes ≥6	
CBD	Yes - including external assessors 14	Yes 1/month Total of 60	Yes ≥6	Yes ≥6	Yes ≥6
MSF	Yes 1/year	Yes 1/year	ePATH SPRAT 3 in 5 years	Yes 3 in 5 years	Yes 3 in 5 years
SAIL	Yes 8				
PATIENT SATISFACTION QUESTIONNAIRE					
PORTFOLIO	Logbook	Yes Reflective portfolios	Training and learning record	Yes	Yes
TEACHING					
OTHER	Dissertation	Clinical rating scales (CRS) 22 in ST1	Year 1 assessment	Year 1 assessment Evaluation of Clinical Events (ECE)	Year 1 assessment Evaluation of Clinical Events (ECE)
EXAM	Yes	FRCOpth	MRCPath	MRCPath	MRCPath
TRAINERS REPORT	Yes	Yes	Yes	Yes	Yes

ANNEX 2: TABLE OF ASSESSMENTS USED BY EACH SPECIALITY INCLUDING THE RECOMMENDED NUMBERS. CONTINUED.

	PAEDIATRICS	PSYCHIATRY	O&G	CLINICAL RADIOLOGY	CLINICAL ONCOLOGY
DOPS	Yes 1 for each procedure	Yes		Yes 3-6	Yes
OTHER ASSESSMENT OF PROCEDURE			OSATS– used to assesprogression to independent practice. 5 for each procedure		
MINI CEX	Paed mini CEX 4-6	Mini ACE 4 in ST1 & 2	Yes Total 10 for each competency	Yes 2	Yes 4
CBD	Paed CbD 4-8	Yes 4 in ST1 & 2	Yes – including case reports	Yes 3-6	Yes 4
MSF	eSPRAT 1/year	Mini PAT Team assessment of behaviour (TAB) 1/year	Team observation exercise 2/year	SPRAT 1/year	Yes 3 in 5 year
SAIL	Yes Total 7 in ST4-7				
PATIENT SATISFACTION QUESTIONNAIRE	SHEFFPAT 1 in total	PAT		Yes 1/year	Yes 3 in 5 years
PORTFOLIO	Yes	Log book	Yes		Yes – and log book
TEACHING	Under development	Assessment of Teaching (AoT) 2			
OTHER	Plan to extend SPRAT to cover handover	ACE 8 total Case Presentation (CP) 4 Journal Club Presentation (JCP) 4		Direct assessment of diagnostic radiological skills (DARDS) Research 6	Research Assessment of IRMER regulations and knowledge of chemotherapy
EXAM	MRCPCH	MRCPsych Integrated with WPBA	MRCOG	FRCR	MRCP FRCR
TRAINERS REPORT	Yes	Yes	Yes	Supervisors report	Supervisors report

	GENERAL INTERNAL MEDICINE	SURGERY	FOUNDATION PROGRAMME
DOPS	Yes - procedure specific for StRs Until independence	Yes – basic procedures 6	Yes ≥ 4-6 List of 14 procedures
OTHER ASSESSMENT OF PROCEDURE		Procedure based assessment (PBA) As per learning agreement	
MINI CEX	Yes ≥4	Yes 6 in ST1 & 2	Yes 6
CBD	Yes ≥4	Yes 6 in ST1 & 2	Yes 6
MSF	Yes 1/year	Mini PAT 3 in 7 years	TAB Mini PAT MSF 1/year
SAIL			
PATIENT SATISFACTION QUESTIONNAIRE	Under development		
PORTFOLIO	Yes	Yes Procedural log	Yes
TEACHING	Under development		
OTHER	Acute care assessment tool (ACAT) 6	Learning agreement	
EXAM	MRCP	MRCS	
TRAINERS REPORT	Yes	Yes	Yes

APPENDIX A
DIRECT OBSERVATION OF PROCEDURAL SKILLS (DOPS)

Where does DOPS come from?

Direct observation of procedural skills was developed by the Royal College of Physicians of London, and initially piloted with Specialist Registrars (SpRs). The method is a variation on mini-CEX, and was designed to assess practical skills.[59]

Previously the ability to undertake procedures had been documented through the use of log books and a record of complication rates. DOPS improves on logbooks by being designed to be a more objective assessment of competence than simply counting numbers of procedures, and by allowing supportive feedback.

What is DOPS?

A DOPS assessment involves the observation of a clinical procedure performed by the trainee. The procedure chosen will vary with the specialty and level of experience of the trainee. The foundation programme has a list of recommended procedures, and many specialties have lists of required procedures e.g. surgical DOPS for ST1 and ST2 trainees. As with mini-CEX both trainee and assessor agree in advance that the encounter is to be assessed, and the period of observation is followed by an opportunity for feedback and suggestions for improvement.

In addition to a global rating the DOPS form includes ratings of a number of possible components of clinical competence including; consent, analgesia, aseptic technique, post procedure management and communication. The assessment of the less procedure specific skills such as communication, approach to the patient and analgesia are in many ways more important than the technical skill in performing the individual procedure.

The duration of a DOPS assessment varies with the procedure observed, feedback may take an additional 20%-30% of the procedure time.[60]

What is the evidence for the utility of DOPS?

DOPS is a relatively new instrument and there is limited published data on utility. As a variation on the mini-CEX studies showing the validity of mini-CEX may apply to DOPS as well.[61]

Reliability
The pilot study by the Royal College of Physicians London with SpRs found a generalisability coefficient of 0.89 (SEM 0.27) with six encounters or raters. They recommend that for each procedure a trainee should be observed by at least three assessors observing two procedures each to achieve adequate reliability.[60]

Validity
DOPS has high face validity as it involves the observation of a real clinical encounter. The RCPL pilot study showed evidence for construct validity as more senior trainees received higher scores.[60] Procedure type had no significant effect on ratings.[60]

Other instruments for assessing procedural skills

Surgical specialties have developed instruments for the assessment of more complex procedures. The rating scales mostly use descriptors based on the level of supervision required by the trainee for that procedure.

Procedure based assessment (PBA) is being used by higher surgical trainees, based on previous work to assess operative skills[62,63] there are procedure specific forms for all surgical specialties. The choice of procedure to be assessed is driven by the objectives set at the meeting with the Educational Supervisor.

Obstetrics and Gynaecology have piloted Objectively Structured Assessment of Technical Skills (OSATS) in ST1 trainees. OSATS are used to assess 10 key procedures; each procedure requires five assessments at the level of independent practice to be signed off.

Evidence for the utility of other instruments for assessing procedural skills

OSATS were developed in Canada and the USA for assessing Surgical and Obstetrics and Gynaecology trainees.[64-66] The original studies used either pigs or lifelike models in a multiple station exam format. This method is reliable (Cronbach's alpha 0.89-0.95), has good inter-rater reliability and can distinguish between levels of experience of the trainee.[65-67] There is a suggestion that global ratings are more reliable than checklists.[64]

Direct observation or videoing of real procedures using structured check lists based on OSATS can demonstrate high inter-rater reliability and test-retest reproducibility,[68] and experienced trainees receive higher scores.[48]

Cautions

Individual DOPS assessments are not designed to be a sign off for independent practice. The instruments become reliable through the use of multiple observations with multiple observers.

The PMETB Survey suggests that only around 40% of foundation trainees found the feedback from DOPS helpful.[4] This may reflect a lack of assessor training and time available for assessment.

Examples of how DOPS is used

A generic DOPS is used by all foundation trainees and procedure specific forms have been developed by most specialties. The Foundation instrument uses 11 items and a six point scale, those used in specialist training vary from 10-12 items with a 4-9 point scale.

Adaptability

Instruments for the direct observation of surgical skills can be adapted to use in sub specialties such as ENT and remain highly reliable with good construct validity.[69]

APPENDIX B
CASE BASED DISCUSSION (CBD)

Where does CbD come from?

Case-based Discussion is a variation of chart stimulated recall (CSR), which was developed by the American Board of Emergency Medicine.[61] CSR provides information about how a trainee makes decisions on diagnosis, investigations, and treatment.[70] CSR was further adapted for use as part of the GMC fitness to practice procedures,[71] and is used as part of the Physician Review and Enhancement Program (PREP), by the College of Physicians and Surgeons of Ontario.[72]

What is CbD?

Case based discussion is designed to assess clinical decision-making and the application or use of medical knowledge. A CbD assessment uses case notes as the focus for discussion. The trainee will identify relevant patient notes and the assessor should choose a suitable case. The discussion may focus on a particular aspect of the case such as deciding on a management plan, or ethical issues raised. The assessor should probe for the reasoning behind any decisions made.

What is the evidence for the utility of CbD?

There is very little published evidence for the reliability and validity of CbD. Some specialties have run pilots prior to implementing their assessment systems in 2007. Anecdotal reports would suggest that CbD is the WPBA instrument clinicians are most comfortable with and the easiest to integrate into everyday practice. The 2007 PMETB survey found it was the WPBA instrument from which the trainees reported most helpful feedback.[4]

Evidence for the utility of CSR

Reliability
The initial studies of CSR demonstrated good inter-rater reliability of 0.85, subsequent studies have shown a reliability coefficient of 0.54 with three cases[73] and of 0.74 with two cases.[70]

Validity
CSR correlates with measures of performance using standardised patients and there is a weak correlation with previous certification scores.[73] CSR can differentiate between candidates of varying levels of experience.[70]

Cautions

Data from some studies suggest that performance on CSR may be case specific requiring multiple observations to be reliable.

Clinicians are comfortable with the CbD assessments as they commonly listen to their trainees present cases and then discuss management plans. To make best use of the potential of CbD as a formative instrument and to assess decision making skills, the assessor will need to use more probing questions than would be necessary on a routine ward round.

Since there is weak published evidence of the reliability and validity of CbD, decisions about number of assessments required have been extrapolated from data for similar instruments.

Examples of how CbD is used

Case based discussion is used by all specialties and the foundation programme. The domains are mostly consistent across specialties with a few specialty specific modifications. The Foundation instrument uses seven items and a six point scale, those used in specialist training vary from 7-10 items with a 5-6 point scale.

APPENDIX C
THE MINI-CLINICAL EVALUATION EXERCISE (MINI-CEX)

Where does mini-CEX come from?

The mini-CEX was designed by John Norcini to be used in US Internal Medicine postgraduate training programmes. It was derived to overcome many of the problems with the Clinical Evaluation Exercise (CEX) then recommended by the American Board of Internal Medicine.[23] A full CEX assessment consists of a long case evaluated by single examiner outside of routine practice, in a period of up to two hours, with a reliability coefficient of less than <0.3.

The mini-CEX is designed to overcome these issues of poor reliability by using multiple observations with multiple assessors over a period of time. This also addresses problems with case specificity. It is designed to be a formative instrument, to provide an opportunity for feedback.

What is mini-CEX?

A mini-CEX assessment involves the direct observation of a trainee / patient interaction. This can be in any setting (inpatient, outpatient, A&E) and the focus may be one part of the clinical encounter such as examination skills or breaking bad news. The clinical interaction should reflect everyday practice and both the assessor and assessee need to know in advance that the encounter is to be assessed. Following the period of observation there is an opportunity for discussion and feedback for the trainee on his/her performance. The assessor completes the assessment form that includes identifying areas for improvement and formulating an action plan. The original mini-CEX form included ratings of four components of competence: history-taking skill, physical examination skill, clinical judgment and synthesis, and humanistic qualities along with an overall rating on a nine point scale.[23]

For each encounter the assessor records the complexity of the case, the clinical environment and the context of the visit. The assessment is repeated with different cases, in different settings and with multiple assessors over the course of an attachment.

A mini-CEX assessment takes between 15 and 25 minutes to complete including around five minutes for feedback.[23,60]

What is the evidence for the utility of mini-CEX?

Reliability
The mini-CEX is reliable with between eight and 14 raters, the original studies showed a reliability of 0.8 with 12-14 raters,[23] more recent studies have shown reliable results with as few as eight raters.[74,75] In these studies the results show a narrow standard error of measurement (SEM) suggesting that those who have high (or low) scores initially may need as few as four encounters and further assessment can be focused on trainees with borderline results.

A study carried out by the Royal College of Physicians of London into the use of the mini-CEX for specialist registrars demonstrated a generalisability coefficient of 0.83 with six encounters or raters.[60] However, given the wide confidence intervals they recommend a trainee should be observed by at least eight different assessors observing at least two encounters each.

Norcini demonstrated good inter-rater reliability, with no large differences in ratings between examiners and across settings.[76]

Validity
Mini-CEX has good face validity as it involves the observation of a real patient encounter in a real clinical environment. Mini-CEX is able to differentiate between levels of experience. Scripted videoed performances demonstrated faculty are able to differentiate between satisfactory and unsatisfactory performance.[77] Scores do improve over time[22,74] and more experienced trainees receive higher ratings.[23,60] Mini-CEX correlates with other measures of performance.[22,74,75,78]

Educational impact

There is some evidence that mini-CEX promotes deep learning,[79] and encourages self reflection.[78]

Cautions

There may be a significant halo effect with a high correlation between scores achieved on individual competencies.[22,80] Care needs to be taken when interpreting the results of a mini-CEX instrument which attempts to assess multiple distinct domains of performance.

Assessors tend to give higher ratings for more complex cases.[22,60]

Studies of student mini-CEX in the USA suggest that faculty ratings are lower than resident's ratings of students.[74,81] This difference could be important as many of the Foundation assessments are completed by registrars.

Mini-CEX is not intended for use in high stakes assessments and should not be used to rank or compare candidates.[61]

Assessors need to be trained in the use of the mini-CEX assessment method. The primary purpose of a mini-CEX is to provide an opportunity to observe the trainee's clinical skills (which otherwise happens rarely) and give constructive feedback. For this to happen effectively both the assessor and the trainee need to be familiar with the assessment instrument and the assessor needs to be both trained and competent in the procedure/skill they are assessing (in order to be able to make a judgement) and trained in how to give feedback.[82]

Examples of how mini-CEX is used

The mini-CEX has been adapted for use in many different settings. It is used by all foundation trainees in the UK and most (but not all) specialist training programmes. The Foundation instrument uses seven items and a six point scale, those used in specialist training vary from 7-9 items with a 5-9 point scale. In the USA a significant minority of Internal Medicine core clerkships use mini-CEX, mostly in a formative fashion.[16]

Adaptability

The mini-CEX can be adapted to assess different domains and different levels of seniority of practitioner whilst remaining reliable e.g., the professionalism CEX[78] which uses a 24 item four point scale (reliability 0.8 with 10-12 raters), and an instrument for use in palliative care[83] (no reliability data). Kogan et al have shown mini-CEX to be feasible, reproducible and have concurrent and construct validity when used with students.[74] It is also feasible to adapt mini-CEX to be used on a PDA.[81]

APPENDIX D
MULTI SOURCE FEEDBACK (MSF)

Where does MSF come from?

Multisource feedback is widely used in the commercial sector[84] where it was developed to support employee decision making and quality improvement.[85] Within medicine variations on MSF are increasingly used to assess attitudes and behaviour in North America and in the UK at all levels of training from medical students through to consultant and GP Principal.

What is MSF?

Multi source feedback is a way of obtaining feedback from a wide variety of sources on everyday performance in the workplace. There are a number of types of MSF used in medicine depending on the groups asked for their views. Peer review MSF surveys clinical colleagues though they may not necessarily be the same grade or specialty. Both MSF and 360° assessment survey allied professionals and administrative staff in addition to clinicians. There are also instruments specifically designed to elicit feedback from patients.

A straightforward questionnaire is distributed to a selection of colleagues who have been able to observe the individual being assessed; this may be done on paper or via email. The forms are completed anonymously and either returned to the educational supervisor / appraiser to be collated or collated electronically and the report sent to the supervisor. The assessee also completes a self assessment. Most instruments include a small number of statements to be graded on a rating scale along with a global question asking if there are any reasons for concern and a space for free text comments. The results are then fed back during the appraisal and any plans for further development made.

What is the evidence for the utility of MSF?

Multiple studies from industry demonstrate that if assessees like the assessment system and find it motivating, MSF can lead to organisational benefits.[84] However, managers are often poorly trained in giving constructive feedback, and they also require organisational support if the feedback is to be accepted.[86]

Reliability
The various MSF instruments available require different numbers of raters for the instruments to be reliable and depend on the mix of raters and the number of domains to be assessed.

SENIOR CLINICIANS: Early studies of peer ratings (by physicians and nurses) of practicing internists showed acceptable reliability with 11 raters in the domains of 'cognitive and clinical management skills', and 'humanistic qualities and management of psychosocial aspects of illness.'[84,87] Subsequent studies of senior clinicians/consultants have shown a peer assessment instrument to be reliable; reliability coefficient 0.7 with 11 raters,[88] 0.98 with 7-8 peer raters,[89] generalisability coefficient of 0.61 with 10 raters.[90] MSF instruments have been shown to have a reliability coefficient >0.85 with eight raters,[91] and to be reliable with nine raters.[41] Studies of patient survey instruments have shown: reliability coefficient of >0.85 with 22 raters,[91] to be reliable with 15 raters,[41] a generalisability coefficient of 0.67 with 25 ratings.[90] The RCPL pilot of MSF for consultants showed an overall generalisability coefficient of 0.71 although this did vary across specialty.[42]

TRAINEES: A study of MSF for medical SpRs recommends 12 raters to achieve adequate reliability.[60] See Table 2 for comparison of miniPAT and TAB.[92-96]

STUDENTS: Student peer review can be reliable (0.7) with six raters and independently rate the two domains of work habits and interpersonal skills.[97]

Validity
MSF has high face and content validity as demonstrated in most of the studies referenced.

There is limited evidence from industry that MSF reports can predict future performance.[84]

A review found several studies which demonstrated that peer ratings correlate with grades given by faculty and written examination performance.[98]

MSF does reflect speciality differences. Factor analysis of an instrument used across three specialties identified the same four factors. Construct validity is supported by the fact that they accounted for a different amount of the variance[89] (with communication being important for psychiatry and patient management for internal medicine).

A few studies have demonstrated consequential validity with 61%,[99,100] 71%[101] and 42%[90] of assessees contemplating or initiating change as a result of MSF feedback. Subsequent work found a lower proportion of assessees initiating change. The feedback needs to be specific and congruent with other sources.[57]

Educational impact

The correlation between self ratings and ratings from others is often low;[85,90] this provides an opportunity for the appraiser to discuss any discrepancy.

Choice of rater

Feedback following the introduction of the foundation programme and core specialist training suggests that trainees are choosing colleagues to complete their MSF assessments that they think will mark them highly. The evidence from the literature is contradictory.

Early studies suggested that peer ratings are not biased substantially by the method of selection of the peers or the relationship between the rater and the subject.[87] Further, a more recent study of 360° assessment (TAB) of UK trainees mentioned concern that trainees would select raters who would give favourable scores. However, the data partially refute this by demonstrating that some trainees were still rated as having cause for concern.[94]

A focus group study of family physicians who had participated in MSF suggested that those who deliberately did not choose reviewers who knew them well were disappointed by the results.[102]

The assessee response to the results of a MSF assessment depends on the perceived objectivity of the process and of the credibility of the raters, who must be able to observe physicians to rate them.[99,102] However, assessors who are more familiar with the assessee give better ratings.[90,99]

The mix of raters is important, with several studies showing small but significant differences in mean ratings from medical staff and allied professionals.[41,42,60,92] Others have shown differences with 3.4%-7% of variance due to the occupation of rater, the length of working relationship and working environment.[92,93]

Clerical or secretarial staff may not feel able to complete all parts of the MSF as they concern behaviours they have not observed,[96] and other aspects of performance such as teaching may not be observed by all raters.[91]

Overall the published literature suggests that any differences in ratings attributable to the choice of rater are small. There are advantages to the assessee nominating some of his or her own assessees; the choice of rater cannot be challenged, although the credibility of the raters needs to be as high as possible. For MSF that includes a mix of roles, there needs to be a mechanism for ensuring the balance of roles is appropriate.

Cautions

Most studies of MSF consistently demonstrate a marked halo effect with factor analysis showing two distinct domains – 'humanistic or interpersonal skills (which accounts for most of the variance) and 'technical or cognitive factors.'[84,88,92,93,103] This suggests that raters are unable to distinguish between more than two domains of performance and care needs to be taken before using an instrument attempting to do so. Cognitive factors may be best assessed using other methods, and MSF is primarily used to assess interpersonal and communication skills.

There are discrepancies in the assessment of humanistic behaviour by medical staff and nursing staff.[11,84] This does not necessarily mean ratings by either group are less reliable but may reflect differences in the opportunities for assessment, and experience of the trainee, giving an alternative (and possibly complementary) perspective. There is also mixed overlap between ratings provided by colleagues and those provided by patients suggesting they may reflect different aspects of performance.[90,91]

Anecdotal reports suggest that there is widespread uncertainty as to how the results of MSF will be used, and reluctance to make comments or give ratings that would have a negative impact on a trainee's career. There is published data to suggest that in practice faculty are unwilling to report incidents of unprofessional conduct, despite stating in interview that they thought they would.[104] In addition one study found that those identified as outliers have fewer questionnaires returned,[91] this could reflect an unwillingness to express adverse views.

Overall, most mean scores on MSF assessments are skewed towards favourable results.[90,91] This may reflect the fact that the subjects tend to be volunteers. Care must be taken when using these instruments to identify outliers. It may be necessary to look at dis-aggregated scores and investigate further any individual poor ratings or comments of concern.[41]

There may be gender and ethnicity difference in the ratings received by individuals. A study of MSF for consultants in the UK demonstrated that colleagues rated UK trained doctors more highly than non UK trained ($P<0.05$). Patients rate female doctors significantly more highly for their relational skills.[41,90]

A review of studies of peer ratings that used rigorous criteria for inclusion found only three instruments that met all criteria for development and validity. Most of the studies referenced in this report would not have met these criteria.[105] The same group reviewed the literature for patient surveys and found six instruments with little validity data.[106] They recommend that further evidence for validity and reliability is required prior to the use of either assessment for summative purposes.

Examples of how MSF is used

The American Board of Internal Medicine (ABIM) have developed a patient and physician peer assessment module which can be used to gain credits towards recertification.[107] The patient instrument focuses on professionalism and interpersonal and communication skills, whilst the peer instrument covers professionalism, medical knowledge and patient care.[90] The College of Physicians and Surgeons of Alberta Canada have developed instruments for peers, co-workers and patients, with the goal of formative review and quality improvement.[85] A form of MSF called 'Physician assessment' may be used to contribute to the maintenance of professional standards programme of the Royal Australasian College of Physicians.[108]

In the UK all foundation trainees and all specialty training programmes are using a form of MSF. It is already used by groups of consultants and GPs and is likely to become part of the revalidation process. A few of the specialties and some foundation trainees are using a instrument based on TAB with four domains and a three point scale, others are variations on miniPAT with up to 27 items and a six point scale. Overall the instruments vary from 9-27 items with a 3-9 point scale.

Adaptability

Multi source feedback or peer review has proved to be feasible in many different settings and in many ways is the most valid way of assessing performance in everyday practice. Given the variations in the instruments used it is important that the characteristics of the assessment system are well understood, that there is clarity of purpose, and a clear statement of how the results will be used. This is vital for MSF to be accepted and useful.[98,102,109,110]

Table 2: Comparison of MSF Instruments[92-96]

	SPRAT	MINIPAT	TAB
PURPOSE	Inform quality assurance process. Contribute to quality improvement	Derived from miniPAT	Identify trainees whose behaviour does not meet GMC requirements, so that appropriate action may be taken
GROUP IN WHICH PILOTED	Paediatric trainees	Foundation pilot	Initially O&G trainees then SHO's across specialty
NUMBER OF QUESTIONS	24 questions covering five domains	15 questions and global rating	Four questions covering four distinct domains
RATING SCALE	6 point scale:1,2 below expectations; 3 borderline; 4 meets expectations; 5,6 above expectations	6 point scale: 1,2 below expectations; 3 borderline; 4 meets expectations; 5,6 above expectations	Three point scale (pass/ borderline/fail changed to no concern/some concern/ major concern)
DOMAINS IDENTIFIED BY FACTOR ANALYSIS		'humanistic' and 'clinical care'	Initial study suggest 76% variance due to 'relationship' factor
RELIABILITY	83% of doctors only need four raters to be reliably place either side of the borderline category	74% of F2s would need eight raters to achieve a small enough SEM to place them accurately above or below satisfactory (eight raters would be enough for 53% of F1s)	Generalisability coefficient of 0.8 for nine raters. 10 assessments are needed to give a 0.85 probability of picking up a poorly performing trainee (0.96 with 15 raters)
CONTENT VALIDITY	GMP	GMP and foundation curriculum	GMP
CONSTRUCT VALIDITY	SpRs rated higher than SHOs ($p<0.001$)	F2s rated higher than F1s ($p<0.001$) The 'Humanistic' domain was rated higher than that of 'clinical care'	
IDENTIFYING POOR PERFORMERS		5.2% of F2s assessed as below expectations	12% received at least one concern rating
OTHER			Test-retest reliability Both assessors and trainees are positive about the process

REFERENCES

1 MMC. Modernising Medical Careers. www mmc nhs uk [2009 Available from: URL:www.mmc.nhs.uk

2 Multiple. Report of the AoMRC and PMETB WPBA forum. 28-4-2008. Academy of Medical Royal Colleges (available on request).

Ref Type: Report

3 Tooke J. Aspiring to Excellence: Findings and Recommendations of the independent inquiry into Modernising Medical Careers. 2007.

Ref Type: Report

4 Smith D, Riley S, Kazmierczak A, Aitken M, Paice E, Le Rolland P. National Survey of Trainees 2007. Postgraduate Medical Education and Training Board [2008 [cited 2008 July 28]; Available from: URL:http://www.pmetb.org.uk/fileadmin/user/QA/Trainee_Survey/National_Survey_of_Trainees_2007_Summary Report_20080723-Final.pdf

5 Riley S, Smith D, Le Rolland P. National Survey of Trainers 2007. Postgraduate Medical Education and Training Board [2008 [cited 2008 July 28]; Available from: URL:http://www.pmetb.org.uk/fileadmin/user/QA/Trainer_Survey/National_Survey_of_Trainers_2007_Summary_Report_20080723-Final.pdf

6 Holmboe ES. Faculty and the observation of trainees' clinical skills: problems and opportunities. Acad Med 2004; 79(1):16-22.

7 Paisley AM, Baldwin P, Paterson S. Accuracy of medical staff assessment of trainees' operative performance. Medical Teacher 2005; 27(7):634-638.

8 Colletti LM. Difficulty with negative feedback: face-to-face evaluation of junior medical student clinical performance results in grade inflation. J Surg Res 2000; 90(1):82-87.

9 Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. [Review] [121 refs]. Teaching & Learning in Medicine 2003; 15(4):270-292.

10 Pfeiffer CA, Kosowicz LY, Holmboe E, Wang Y. Face-to-face clinical skills feedback: lessons from the analysis of standardized patient's work. Teaching & Learning in Medicine 2005; 17(3):254-256.

11 Arnold L. Assessing professional behavior: yesterday, today, and tomorrow. [Review] [173 refs]. Acad Med 2002; 77(6):502-515.

12 Carr S. The Foundation Programme assessment tools: an opportunity to enhance feedback to trainees?. [Review] [29 refs]. Postgrad Med J 2006; 82(971):576-579.

13 Hays RB, Davies HA, Beard JD, Caldon LJ, Farmer EA, Finucane PM et al. Selecting performance assessment methods for experienced physicians. Med Educ 2002; 36(10):910-917.

14 Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990; 65(9:Suppl):S63-S67.

15 Downing SM. Reliability: on the reproducibility of assessment data. Med Educ 2004; 38(9):1006-1012.

16 Kogan JR, Hauer KE. Brief report: Use of the mini-clinical evaluation exercise in internal medicine core clerkships. J Gen Intern Med 2006; 21(5):501-502.

17 Beard J, Strachan A, Davies H, Patterson F, Stark P, Ball S et al. Developing an education and assessment framework for the Foundation Programme. Med Educ 2005; 39(8):841-851.

18 Health Select Committee. Select Committee on Health: Third Report: Modernising Medical Careers. House of Commons Publications [2008 [cited 2008 May 21]; Available from: URL:http://www.publications.parliament.uk/pa/cm200708/cmselect/cmhealth/25/2504.htm

19 Black C, Maxwell P, Marshall M, Rees M, Dolphin T. MTAS: which way now? Interview by Rebecca Coombes. BMJ 2007; 334(7607):1300.

20 Everington S, Fielden J, Rees M, Hilborne J, Meldrum H, Spencer JC. MTAS: Response from the BMA. BMJ 2007; 334(7607):1288.

21 Lydall GJ, Malik A, Bhugra D. MTAS: Mental health of applicants seems to be deteriorating. BMJ 2007; 334(7608):1335.

22 Norcini J, Blank L, Duffy F, Fortn G. The mini-CEX: a method for assessing clinical skills. Ann Intern Med 2003; 138(6):476-481.

23 Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. Ann Intern Med 1995; 123(10):795-799.

24 PMETB. Developing and Maintaining an assessment system - a PMETB guide to good practice. Davies H, Joshi H, Holsgrove G, Rowley D, editors. 2007.

25 Beard J. Evaluation of Procedure Based Assessments. 28-4-2008.

Ref Type: Personal Communication

26 Hays R. Assessment in medical education: roles for clinical teachers. The Clinical Teacher 2008; 5(1):23-27.

27 Epstein RM, Hundert EM. Defining and assessing professional competence. [References]. JAMA: Journal of the American Medical Association 2002; 287(2):226-235.

28 Norman G. Research in medical education: Three decades of progress. Br Med J 2002; 324(7353):1560-1562.

29 PMETB. Standards for Curricula and Assessment Systems. 2008. Postgraduate Medical Education and Training Board.

Ref Type: Report

30 Wilkinson TJ. Assessment of clinical performance: gathering evidence. [Review] [48 refs]. Internal Medicine Journal 2007; 37(9):631-636.

31 Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. [Review] [34 refs]. Lancet 2001; 357(9260):945-949.

32 GMC. Good Medical Practice. 2006. The General Medical Council.

REFERENCES

Ref Type: Report

33 The Royal College of Physicians and Surgeons of Canada. CanMeds 2005 Framework. The Royal College of Physicians and Surgeons of Canada [2008 [cited 2008 Mar. 9]; Available from: URL:http://meds.queensu.ca/medicine/obgyn/pdf/CanMEDS_2005_Framework.pdf

34 ACGME. Outcomes project. Accrediation Council for Graduate Medical Education [2007 [cited 2007]; Available from: URL:www.acgme.org/Outcome

35 Epstein RM, Dannefer EF, Nofziger AC, Hansen JT, Schultz SH, Jospe N et al. Comprehensive assessment of professional competence: the Rochester experiment. Teaching & Learning in Medicine 2004; 16(2):186-196.

36 Schuwirth LW, Van Der Vleuten CPM. Changing education, changing assessment, changing research?[see comment]. Med Educ 2004; 38(8):805-812.

37 Eva KW. What every teacher needs to know about clinical reasoning.[erratum appears in Med Educ. 2005 Jul;39(7):753]. [Review] [47 refs]. Med Educ 2005; 39(1):98-106.

38 West CP, Huntington JL, Huschka MM, Novotny PJ, Sloan JA, Kolars JC et al. A prospective study of the relationship between medical knowledge and professionalism among internal medicine residents. Acad Med 2007; 82(6): 587-592.

39 Wilkinson J, Benjamin A, Wade W. Assessing the performance of doctors in training. BMJ 2003; 327(7416):s91-s92.

40 Dauphinee WD, Blackmore DE. Assessing the assessors' assessment. [comment]. Med Educ 2001; 35(4):317-318.

41 Crossley J, McDonnell J, Cooper C, McAvoy P, Archer J, Davies H. Can a district hospital assess its doctors for re-licensure? Med Educ 2008; 42(4):359-363.

42 Mackillop L. Multi source feedback and revalidation: can one generic instrument work? 2007.

43 Van Der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. Advances in Health Sciences Education 1996; 1(1):41-67.

44 Holsgrove G. Reliability issues in the assessment of small cohorts. 2009. Postgraduate Medical Education and Training Board.

Ref Type: Report

45 Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ 2003; 37(9):830-837.

46 Williams RG, Verhulst S, Colliver JA, Dunnington GL. Assuring the reliability of resident performance appraisals: more items or more observations? Surgery 2005; 137(2):141-147.

47 Southgate L, Hays RB, Norcini J, Mulholland H, Ayers B, Woolliscroft J et al. Setting performance standards for medical practice: a theoretical framework. Med Educ 2001; 35(5):474-481.

48 Beard JD. Setting Standards for the Assessment of Operative Competence. Eur J Vasc Endovasc Surg 2005; 30(2):215-218.

49 Cusimano MD, Rothman AI. Consistency of standards and stability of pass/fail decisions with examinee-based standard-setting methods in a small-scale objective structured clinical examination. Acad Med 2004; 79(10:Suppl):Suppl-7.

50 Cusimano MD, Rothman AI. The effect of incorporating normative data into a criterion-referenced standard setting in medical education. Acad Med 2003; 78(10:Suppl):Suppl-90.

51 Norcini JJ. Setting standards on educational tests. Med Educ 2003; 37(5):464-469.

52 Searle J. Defining competency - the role of standard setting. Med Educ 2000; 34(5):363-366.

53 Schuwirth LW, Van Der Vleuten CPM. A plea for new psychometric models in educational assessment.[see comment]. Med Educ 2006; 40(4):296-300.

54 Riley S, Smith D, Le Rolland P. National Survey of Trainers 2007. Postgraduate Medical Education and Training Board [2007 [cited 2008 July 28]; Available from: URL:http://www.pmetb.org.uk/fileadmin/user/QA/Trainer_Survey/National_Survey_of_Trainers_2007_Summary_Report_20080723-Final.pdf

55 Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment.[erratum appears in Med Educ. 2003 Jun;37(6):574]. Med Educ 2002; 36(10):972-978.

56 businessballs.com. Conscious Competence Learning Model. www businessballs com/ [2008 [cited 2008 July 29]; Available from: URL:http://www.businessballs.com/consciouscompetencelearningmodel.htm

57 Sargeant JM, Mann K, Sinclair D, Van der Vleuten C, Metsemakers J. Challenges in multisource feedback: intended and unintended outcomes. Med Educ 2007; 41(6):583-591.

58 Sargeant J, Mann K, Sinclair D, Van der Vleuten C, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. Advances in Health Sciences Education 2006.

59 Wragg A, Wade W, Fuller G, Cowan G, Mills P. Assessing the performance of specialist registrars. Clinical Medicine 2003; 3(2):131-134.

60 Wilkinson JR, Crossley JGM, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. Med Educ 2008; 42(4):364-373.

61 Norcini J. Workplace-based assessment in clinical training. London: Association for the study of medical education; 2007.

62 Thornton M, Donlon M, Beard JD. The operative skills of higher surgical trainees: measuring competence achieved rather than experience undertaken. Bulletin of The Royal College of Surgeons of England 2003; 85(6):190-218.

REFERENCES

63 Burt CG, Chambers E, Maxted M, Grant JR, Markham N, Watts H et al. The evaluation of a new method of operative competence assessment for surgical trainees. Bulletin of The Royal College of Surgeons of England 2003; 85(5): 152-155.

64 Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 1997; 84(2):273-278.

65 Goff BA, Lentz GM, Lee D, Houmard B, Mandel LS. Development of an objective structured assessment of technical skills for obstetric and gynecology residents. acogjnl 2000; 96(1):146-150.

66 Goff BA, Nielsen PE, Lentz GM, Chow GE, Chalmers RW, Fenner D et al. Surgical skills assessment: a blinded examination of obstetrics and gynecology residents. American Journal of Obstetrics & Gynecology 2002; 186(4):613-617.

67 Swift SE, Carter JF. Institution and validation of an observed structured assessment of technical skills (OSATS) for obstetrics and gynecology residents and faculty. American Journal of Obstetrics & Gynecology 2006; 195(2):617-621.

68 Beard JD, Jolly BC, Newble DI, Thomas WE, Donnelly J, Southgate LJ. Assessing the technical skills of surgical trainees. Br J Surg 2005; 92(6):778-782.

69 Roberson DW, Kentala E, Forbes P. Development and validation of an objective instrument to measure surgical performance at tonsillectomy. Laryngoscope 2005; 115(12):2127-2137.

70 Oandasan IF, Byrne N, Davis D, Shafir MS, Malik R, Waters I et al. Developing competency-assessment tools to measure the family physician's ability to respond to the needs of the community. Acad Med 2001; 76(10:Suppl):Suppl-3.

71 Southgate L, Cox J, David T, Hatch D, Howes A, Johnson N et al. The General Medical Council's Performance Procedures: peer review of performance in the workplace. Med Educ 2001; 35 Suppl 1:9-19, 2001 Dec.:9-19.

72 Cunnington JP, Hanna E, Turnhbull J, Kaigas TB, Norman GR. Defensible assessment of the competency of the practicing physician. Acad Med 1997; 72(1):9-12.

73 Solomon DJ, Reinhart MA, Bridgham RG, Munger BS, Starnaman S. An assessment of an oral examination format for evaluating clinical competence in emergency medicine. Acad Med 1990; 65(9:Suppl):Suppl-4.

74 Kogan JR, Bellini LM, Shea JA. Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mCEX) in a medicine core clerkship. Acad Med 2003; 78(10:Suppl):Suppl-5.

75 Durning SJ, Cation LJ, Markert RJ, Pangaro LN. Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. Acad Med 2002; 77(9):900-904.

76 Norcini J, Blank L, Arnold G, Kimball H. Examiner Differences in the Mini-Cex. Advances in Health Sciences Education 1997; 2(1):27-33.

77 Holmboe E, Huot S, Chung J, Norcini J, Hawkins R. Construct validity of the miniclinical evaluation exercise (MiniCEX). Acad Med 2003; 78(8):826-830.

78 Cruess R, McIlroy JH, Cruess S, Ginsburg S, Steinert Y. The Professionalism Mini-evaluation Exercise: a preliminary investigation. Acad Med 2006; 81(10:Suppl):S74-S788.

79 ves de LA, Henquin R, Thierer J, Paulin J, Lamari S, Belcastro F et al. A qualitative study of the impact on learning of the mini clinical evaluation exercise in postgraduate training. Medical Teacher 2005; 27(1):46-52.

80 Margolis MJ, Clauser BE, Cuddy MM, Ciccone A, Mee J, Harik P et al. Use of the mini-clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: a validity study. Acad Med 2006; 81(10:Suppl):Suppl-60.

81 Torre DM, Simpson DE, Elnicki DM, Sebastian JL, Holmboe ES. Feasibility, reliability and user satisfaction with a PDA-based mini-CEX to evaluate the clinical skills of third-year medical students. Teaching & Learning in Medicine 2007; 19(3):271-277.

82 Holmboe ES, Yepes M, Williams F, Huot SJ. Feedback and the mini clinical evaluation exercise. J Gen Intern Med 2004; 19(5:Pt 2):t-61.

83 Han PK, Keranen LB, Lescisin DA, Arnold RM. The palliative care clinical evaluation exercise (CEX): an experience-based intervention for teaching end-of-life communication skills. Acad Med 2005; 80(7):669-676.

84 Wood L, Hassell A, Whitehouse A, Bullock A, Wall D. A literature review of multi-source feedback systems within and without health services, leading to 10 tips for their successful design. [Review] [81 refs]. Med Teach 2006; 28(7):e185-e191.

85 Lockyer J. Multisource feedback in the assessment of physician competencies. [Review] [22 refs]. J Contin Educ Health Prof 2003; 23(1):4-12.

86 Overeem K, Faber M, Arah OA, Elwyn G, Lombarts KMJM, Wollersheim HC et al. Doctor performance assessment in daily practise: does it help doctors or not? A systematic review. Med Educ 2007; 41(11):1039-1049.

87 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance.[see comment]. JAMA 1993; 269(13):1655-1660.

88 Ramsey PG, Carline JD, Blank LL, Wenrich MD. Feasibility of hospital-based use of peer ratings to evaluate the performances of practicing physicians. Acad Med 1996; 71(4):364-370.

89 Lockyer JM, Violato C. An examination of the appropriateness of using a common peer assessment instrument to assess physician skills across specialties. Acad Med 2004; 79(10:Suppl):Suppl-8.

90 Lipner RS, Blank LL, Leas BF, Fortna GS. The value of patient and peer ratings in recertification. Acad Med 2002; 77(10:Suppl):S64-S66.

91 Campbell JL, Richards SH, Dickens A, Greco M, Narayanan A, Brearley S. Assessing the professional performance of UK doctors: an evaluation of the utility of the General Medical Council patient and colleague questionnaires. Qual Saf Health Care 2008; 17(3):187-193.

REFERENCES

92 Archer JC, Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training.[see comment]. BMJ 2005; 330(7502):1251-1253.

93 Archer J, Norcini J, Southgate L, Heard S, Davies H. mini-PAT (Peer Assessment Tool): A Valid Component of a National Assessment Programme in the UK? Advances in Health Sciences Education 2008; 13(2):181-192.

94 Whitehouse A, Hassell A, Bullock A, Wood L, Wall D. 360 degree assessment (multisource feedback) of UK trainee doctors: Field testing of team assessment of behaviours (TAB). Med Teach 2007; 29(2):171-176.

95 Whitehouse A, Walzman M, Wall D. Pilot study of 360 degrees assessment of personal skills to inform record of in training assessments for senior house officers. Hospital Medicine (London) 2002; 63(3):172-175.

96 Whitehouse A, Hassell A, Wood L, Wall D, Walzman M, Campbell I. Development and reliability testing of TAB a form for 360 degrees assessment of Senior House Officers' professional behaviour, as specified by the General Medical Council. Med Teach 2005; 27(3):252-258.

97 Dannefer EF, Henson LC, Bierer SB, Grady WT, Meldrum S, Nofziger AC et al. Peer assessment of professional competence. Med Educ 2005; 39(7):713-722.

98 Norcini JJ. Peer assessment of competence. Med Educ 2003; 37(6):539-543.

99 Sargeant JM, Mann KV, Ferrier SN, Langille DB, Muirhead PD, Hayes VM et al. Responses of rural family physicians and their colleague and coworker raters to a multi-source feedback process: a pilot study. Acad Med 2003; 78(10:Suppl):Suppl-4.

100 Hall W, Violato C, Lewkonia R, Lockyer J, Fidler H, Toews J et al. Assessment of physician performance in Alberta: the Physician Achievement Review. Can Med Assoc J 1999; 161(1):52-57.

101 Violato C, Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. BMJ 2003; 326(7388):546-548.

102 Sargeant J, Mann K, Ferrier S. original research Exploring family physicians' reactions to multisource feedback: perceptions of credibility and usefulness. Med Educ 2005; 39(5):497-504.

103 Wood L. 360 degree assessment: encouraging results of a 6 year study. Annual Meeting of the Association for the Study of Medical Education . 2004.

Ref Type: Abstract

104 Ginsburg S, Regehr G, Hatala R, McNaughton N, Frohna A, Hodges B et al. Context, conflict, and resolution: a new conceptual framework for evaluating professionalism. [Review] [61 refs]. Acad Med 2000; 75(10:Suppl):Suppl-S11.

105 Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians.[see comment]. [Review] [26 refs]. BMJ 2004; 328(7450):1240.

106 Evans RG, Edwards A, Evans S, Elwyn B, Elwyn G. Assessing the practising physician using patient surveys: a systematic review of instruments and feedback methods. Fam Pract 2007; 24(2):117-127.

107 Americal Bourd of Internal Medicine. www abim org [2008 [cited 2008 Jan. 7]; Available from: URL:http://www.abim.org/pims/choose/module/patient-and-physician-peer-assessment.aspx

108 Newble D, Paget N, McLaren B. Revalidation in Australia and New Zealand: approach of Royal Australasian College of Physicians.[see comment]. Review] [5 refs]. BMJ 1999; 319(7218):1185-1188.

109 Arnold L, Shue CK, Kritt B, Ginsburg S, Stern DT. Medical students' views on peer assessment of professionalism.[see comment]. J Gen Intern Med 2005; 20(9):819-824.

110 Rees C, Shepherd M. professionalism The acceptability of 360-degree judgements as a method of assessing undergraduate medical students' personal and professional behaviours. Med Educ 2005; 39(1):49-57.

ACKNOWLEDGEMENTS

This report was written by Dr Anne Collett, Educational Fellow at RCPL, with help from

Professor Sir Neil Douglas, RCPEd/AoMRC

Professor Alastair McGowan, CEM/COPMeD

Dr Keith Myerson, RCoA/Academy Assessment Committee

Ms Winnie Wade, RCPL

The following served on the working groups, led by Professor McGowan and Dr Myerson

Professor Roger Barton, RCPL

Dr Colin Campbell, RCPCH

Dr Jim Crossley, RCPCH/Psychometrician

Mr Robert Gillies, RCSEng

Professor Arthur Hibble, COGPeD

Dr Has Joshi, PMETB

Dr Ian Kestin, RCoA

Ms Winnie Wade, RCPL

With thanks to

Rosie Carlow

Claire Coomber

James Taylor

Academy of Medical Royal Colleges
70 Wimpole Street
London W1G 8AX

Registered Charity
Number 1056565