how to

ASSESS STUDENTS AND TRAINEES IN MEDICINE AND HEALTH

Edited by Olwyn M. R. Westwood, Ann Griffin and Frank C. Hay
How to Assess Students and Trainees in Medicine and Health
Companion website

Visit the companion website at:

www.wiley.com/go/assesshealth

On the website, you will find:
• Links to useful online resources
• Practical guidance
  • on running assessments and mark sheets
  • on standard setting
  • on running OSCEs
  • on simulation scenarios

This title is also available as an e-book. For more details, please see
www.wiley.com/buy/9780470670897
or scan this QR code:
How to Assess Students and Trainees in Medicine and Health

Edited by

Olwyn M. R. Westwood
Professor of Medical Education
Barts and the London School of Medicine and Dentistry
Queen Mary, University of London

Ann Griffin
Clinical Senior Lecturer in Medical Education and Deputy Director
UCL Medical School

Frank C. Hay
Emeritus Professor of Immunology
St George's University of London
Contents

Contributors, vii

Foreword, ix

Preface, xi

Chapter 1 Principles of assessment, 1
Olwyn M. R. Westwood and Ann Griffin

Chapter 2 Assessment of knowledge, 16
Steve Capey

Chapter 3 Development of skills competence assessment, 38
Alison Sturrock

Chapter 4 Assessment of performance, 59
Gavin Johnson and Allison Wiseman

Chapter 5 Assessment of professionalism, 76
Deborah Gill

Chapter 6 Setting the standard in assessments, 94
Steve Capey and Frank C. Hay

Chapter 7 Feedback and assessment, 114
Ann Griffin, Sophie Park and Catherine O’Keeffe

Chapter 8 Psychometrics in medical education assessment, 141
Katherine Woolf

Chapter 9 Examiners and examination boards, 167
Olwyn M. R. Westwood

Chapter 10 Future developments in assessment and feedback, 185
Jane Dacre

Index, 201
Contributors

Steve Capey
Director of Assessment
Swansea Medical School
University of Swansea

Jane Dacre
Professor of Medical Education
Director
UCL Medical School

Deborah Gill
Clinical Senior Lecturer in Medical Education and Deputy Director
UCL Medical School

Ann Griffin
Clinical Senior Lecturer in Medical Education and Deputy Director
UCL Medical School

Frank C. Hay
Emeritus Professor of Immunology
St George’s University of London

Gavin Johnson
Clinical Senior Lecturer in Medical Education
UCL Medical School

Sophie Park
GP and Principal Teaching Fellow
The Research Department of Primary Care and Population Health
University College London

Alison Sturrock
Clinical Senior Lecturer in Medical Education
UCL Medical School

Olwyn M. R. Westwood
Professor of Medical Education
Barts and the London School of Medicine and Dentistry
Queen Mary, University of London

Katherine Woolf
Lecturer in Medical Education
UCL Medical School

Allison Wiseman
Senior Tutor
School of Health and Social Care
University of Surrey

Catherine O’Keeffe
Senior Lecturer in Clinical Education
Institute of Education
University of London
Foreword

A historical perspective

Almost exactly 100 years ago Abraham Flexner published his famous analysis of medical education in the United States which generated a seismic change in the way doctors were trained internationally. Training in medicine has evolved from an apprenticeship model of teaching, where ‘wisdom’ was gained based on centuries of folk practice, to a regulated academic discipline, firmly founded on university-based scientific training.

The twentieth century saw an unprecedented expansion in our understanding of the basic sciences underlying medicine, and the location of medical education firmly in scholarly universities. To be sure, this may have led the ‘pendulum’ to swing too far away from the ‘art’ of medicine. However, recent developments in medical education have begun to redress that balance, such that there is now a widespread appreciation of the value of an integrated approach to care. The equilibrium between the art and science of medicine, although not yet completely established, undoubtedly offers today’s students a much better opportunity to develop into more effective and empathetic practitioners. Much of what is meant by the term ‘art’ in this context is actually no less scientific than modern genomics. What is meant by this term is the understanding of the complex interactions in a human being that encompasses not just the physical and natural but also the social sciences.

The watchword of many an attentive practitioner when confronted with a clinical problem is to ask for the evidence base on which this decision is to be made. While there are clearly problems with the over-zealous quest for the application of evidence, for example, in rare diseases. Nonetheless evidence-based medicine has undoubtedly improved clinical practice.

The evidence to support our practices in medical education needs to be better. This volume brings together some of those data to summarise our current understanding of medical assessment. The ability to select and assess healthcare professionals is a complex area and we need to continue to strive for better tools for assessing individuals and their performance. This will
only be achieved if we pay close attention to what we already know and, rather more importantly, what we do not know, so that we can fill these gaps.

The editors have brought together a significant body of talent to summarise the state-of-the-art methodologies in medical assessment and feedback. If we are to see the quality of medical education improve on a platform of a sound scientific evidence base, it is important that we take heed of this set of data and look forward to another century of growth and development – just as that which was built on the work of Abraham Flexner.

Professor Anthony Warrens
Dean for Education
Barts and The London School of Medicine and Dentistry
Queen Mary, University of London
Preface

Medical education as world leader in assessment

Medical education leads the world in its pioneering work on assessment. The international movement of medical graduates has certainly contributed to the perceived need for testing. The regulation of medical education and medical practice has likewise given rise to serious reflection on the assessment of progress and of attainment. Equally as significant, perhaps, is the complexity of medicine itself. The combination of knowledge, skills, problem solving and values has demanded a sophisticated approach to both curriculum design and assessment.

Added to this educational salmagundi is the diversity of sites of learning, the extensive range of teachers and the unpredictability of the clinical context, especially at postgraduate level. Scant surprise, then, that assessment has been seen as the dependable foundation on which to allow a learner to progress, to enter the profession and, in some countries, to stay within the profession.

So assessment in medicine is a serious business.

Reassuringly, throughout all the consequent research and technical development work that has been undertaken, the leading assessment theoreticians and practitioners have never disregarded their responsibility to communicate with medical teachers and students. And perhaps, even, with the public.

In more recent times, that communication has largely taken the form of technical academic papers, conference presentations and training courses. But there are few places where what is said and what is done are presented as one accessible argument.

This book addresses the deficit. One of its many strengths is that it represents a collaboration between two leading, and rather different, London medical schools. The practical and theoretical wisdom of the authors is apparent at every turn of the page. Not only have they presented the theoretical basis, but they have also set this within the context of practice. Not only do they know the literature, but they are also able to cast a light on the
experience of examiners and of the candidates. Their presentation of the field is accessible precisely because of their ability to reflect on both theory and practice. The two do not always perfectly coincide.

The informed honesty of this book is complemented by a writing style that is quite enticing. I did not feel that I was reading a dry academic text. I felt that I was involved in a personal conversation. And for that reason, I learned more and enjoyed more.

Janet Grant
Emerita Professor of Education in Medicine,
The Open University, UK
Honorary Professor, UCL Medical School
Chapter 1 Principles of assessment

Professor Olwyn M. R. Westwood1 and Dr Ann Griffin2
1Queen Mary, University of London
2UCL Medical School

OVERVIEW

Assessment is one of the most important aspects in education. It has a central role in the training of healthcare professionals and ensuring that professional standards are met. Assessment strongly influences what is taught and, more importantly, what is learnt. It is, therefore, rightly at the forefront of theoretical and practical developments because of its crucial role in teaching and learning. This chapter is divided into two sections. The first section introduces the reader to the basic principles of assessment; these are the core ideas and models upon which subsequent chapters will build. The what, why, and how of assessment will be addressed and it will be contextualised to the current healthcare environment in which assessments are practised. The second section will give an overview of the content covered in each of the book chapters, signposting the key points that they will address.

Introduction

Assessment is a vital and powerful force in teaching and learning. Assessments shape what individuals learn and what educators choose to teach. The feedback inherent in any form of assessment has a significant educational role shaping motivation and future learning. Assessment of the healthcare
How to assess students and trainees in medicine and health

professions is a field that has been the focus of educational research and has subsequently undergone significant improvements in recent years. New formats of assessments have been developed which have shifted the focus from factual recall to the application and synthesis of knowledge. Educationists increasingly strive to ensure that the tests they use have the prerequisite qualities that make them reliable, valid and acceptable, removing wherever possible subjective biases. Competence and performance have become central tenets, and assessment practices have been heavily influenced by socio-political concerns, ensuring that healthcare practitioners are fit to practise. This text provides the theory that underpins modern assessments and provides a solid foundation to the principles, practices and latest developments in assessment. This first chapter has two main aims. The initial section will cover the fundamental principles of assessment: key definitions, principles and models of assessment, and much of the work of subsequent chapters will be to develop these core themes and theories. The second section provides an overview of the entire book. The main subjects that each chapter addresses are highlighted, so that whilst the book can be read as a whole, individual chapters can also be selected according to individual requirements.

Trends in assessment

Assessment continues to develop and the advances made, to some degree, reflect the changing socio-political environment in which healthcare education takes place. Assessments are regarded as the device by which we can guarantee and regulate the calibre of our healthcare workforce and demonstrate to the wider public that the individuals under our tutelage or employ are fit to practise. Assessment is also regarded as a continuous process, one that happens throughout a professional’s career. This is a significant shift away from the once qualified, qualified for life view. This new perspective means that all healthcare practitioners need to be able to demonstrate their commitment to life long learning and continuous professional development (CPD). Appraisal, relies on multiple sources of evidence, which is reviewed by a peer. Success in a series of appraisals eventually leads to the individual practitioner being revalidated.

The necessity to qualify and licence healthcare practitioners that are fit for purpose has seen the introduction of competency-based assessments, many of which are now being carried out in the workplace, relying on observation of authentic clinical situations and, therefore, being more likely to reflect an individual’s actual performance. Simulation has been introduced in those areas where developing expert performance is associated with an inappropri-
ate degree of risk to patient care, for example the training and assessment of laparoscopic procedures in surgery. Simulation has been around in medical education for a while, objective structured clinical examinations (OSCE) have used simulated patients to role play clinical scenarios and models, and mannequins have been used in a variety of assessments of clinical skills – for example basic life support training – and model limbs for practising venepuncture. However, what has grown has been the use of high fidelity simulations that mimic the clinical setting as much as possible in an attempt to get the learner as fully immersed as possible in a ‘real world’ situation. This has led to the development of high-tech simulation suites throughout the UK that look exactly like clinical settings (e.g. mock wards and theatres).

The exponential growth in medical knowledge and advancements in treatment has changed what can, and should be taught: ‘need to know’ has firmly supplanted ‘nice to know’ and the notion of the ‘core curriculum’ has emerged. This has acted to concentrate educators’ minds about the important areas that any programme of study should contain. Having a core curriculum has resulted in a process called blueprinting. This is a method which ensures that the questions in the test match up with what has been taught, so that the test fairly reflects the curriculum. Assessment processes have shifted from ranking students within their cohorts – norm referencing – to focusing on whether or not an individual student has reached the desired level of achievement, and this is called criterion referencing. Norm referencing compares students within their cohorts and a decision is made about the proportion of candidates that will pass, however, even if all the candidates are brilliant some will still fail. Furthermore, this method ignores the exacting nature of the test, tests differ in their degree of difficulty and it is therefore not an appropriate method if you want to consistently guarantee that a healthcare practitioner is fit for purpose. The move to criterion referencing means that the assessment represents what a student can actually do, regardless of their place in the class, and is helpful in ensuring professional standards. Some tests are more difficult than others and some questions more challenging, and so if we want to say a student has reached a certain standard, as with criterion referencing, we therefore need to set the standard of each item in the assessment. Standard-setting is a process which ascribes a level of difficulty to questions; this allows them to be ‘weighted’ accordingly. There are a variety of ways this can be done, a modified Angoff and borderline group method being two commonly used practices. All these features and developments mean that assessment, the global assessment of performance, has become more complex and diverse. The next section will cover the basic principles, the nuts and bolts of assessment.
What is an assessment?

What is an assessment? It is a judgement, or appraisal, of someone’s ability and it allows the assessor to make a decision about their learner’s current level of knowledge, skill or behaviour. Assessments take a variety of different formats and each sort of assessment can be deployed to investigate a specific range of attributes; knowledge, skills, behaviour and professional attitudes. The use of more than one type of assessment, or a suite of assessments, has the ability to give a multi-faceted and more complete picture of an individual’s overall performance. The terms ‘assessment’ and ‘evaluation’ are not interchangeable. In the UK an assessment is a specific unit of appraisal, for example a multiple-choice questionnaire to assess someone’s level of knowledge. Evaluation, however, is a broader concept which typically would take account of a range of different sorts of assessments to give a much broader picture about somebody’s capability. Evaluation purposefully seeks a range of assessment data, looking for different attributes and giving a better overall picture of performance or capability. For example, evaluating someone’s capacity to practise will have to rely on a host of workplace-based assessments, written examinations and reflective portfolios. In the United States, the definitions of these words are reversed.

Why do we need to assess? To ensure our graduates and colleagues are fit to practise and able to provide a high standard of clinical care. Indeed, we have seen a raft of assessment methodologies developed that ensure that we are able to work effectively in our clinical context (see Chapter 4 on workplace-based assessments). But an assessment does more than just provide a guarantee that a practitioner is fit to practice; assessment, to a large degree, determines what is learnt. Learners will significantly alter what they do and learn in response to the sort of assessment that they are faced with. Assessment, therefore, can be used to motivate students to learn, to provide them with feedback about their performance and, additionally, provide feedback for educators about the progress of their learners. Assessment can be of learning as well as for learning.
Teaching, learning and assessment should all align with each other, this is something called the educational paradigm. This means that the purpose behind the teaching, its objectives or outcomes should relate to how it is taught and subsequently how it is assessed. For example, if you were teaching a clinical skill, taking a blood pressure reading for example, your objectives may be that by the end of your period of instruction the learner was independently able to use the equipment and reliably record blood pressure in a range of subjects. Your teaching methods would be likely to include a practical session, either in a clinical skills lab or in a clinical setting. An appropriate form of assessment would be an objective structured clinical examination (OSCE) or other form of observation, it would be unlikely to be a multiple-choice question. These three domains, the objectives or outcomes for the session, the teaching methods, and the assessment, should all align with each other.

What can we assess? There is an array of qualities that we can assess: communication skills, knowledge, clinical skills, professionalism and attitudes, our ability to lead and to work in teams.

However, not all knowledge, skills and behaviours are considered equal, there is a hierarchy; this was described by Bloom et al., (1956) and he called
Knowing a fact, whilst important, is not the same as understanding the principles underpinning it or being able to apply that piece of knowledge to differing contexts. Making judgements is regarded as the most exacting task in the knowledge domain as it is reliant on a widespread appraisal of all the relevant facts, a deep understanding of the context and a full evaluation of

Figure 1.2 The range of assessable attributes.
Figure 1.3 Hierarchy of knowledge, skills and performance based on Bloom’s taxonomy.
all possible choices. True/false multiple choice questions test at the lower levels of Bloom’s taxonomy, emphasising recall rather than application, and for this reason professional assessments tend to favour knowledge tests that test application, like single best answer (SBA) and situational judgement tests (SJT). Likewise, in a psychomotor domain, recognising that someone is having their blood pressure taken is different from describing the steps in taking a reading and very different from taking a blood pressure in an uncooperative patient in the middle of the night. Similarly, in the affective domain aptitude is demonstrated when the full complexity of similar or contradictory personal and professional values are successfully integrated into professional practice.

Tests of know and know how include: multiple-choice examinations, short answer questions, essays, vivas and other oral examinations.

‘Competence describes what an individual is able to do in clinical practice, while performance should describe what an individual actually does in clinical practice.’

Boursicot et al. 2011

**Competence and performance**

Competence and performance are complementary. Being competent relies on having the appropriate knowledge, skills and attitudes. Competence-based assessments measure against a clearly stated set of outcomes, their aim is to be able to describe, objectify and quantify what a person should be able to do and reflect that in the test criteria. Competences are activities that are genuinely needed for work and testing for them ensures those that pass the test have the prerequisite competence to practise. You can demonstrate you are competent in mock circumstances, like OSCEs, which control for many of the complexities and dynamics of the workplace. Tests of competence assess at the level of ‘show how’ (see Figure 1.4). Alternatively, if assessing in the workplace, contextualising competence to the clinical setting means you can now, if it has been done correctly, begin to assess performance; measuring what a practitioner actually ‘does’, i.e. an assessment in the top tier of Miller’s pyramid of clinical competence (Miller, 1990). Assessments taken outside of the workplace make a judgement on what a person can do (their competence) rather than what they really do in ‘real life’ (their performance). For example, a doctor may show that they are effective at breaking bad news using role play at an OSCE station, but may not be able to perform adequately in the middle of an over-running clinic or at 3 am in casualty. The
assessment of performance has led to a whole range of workplace-based assessments: case-based discussion (CBD), mini-clinical evaluation exercise (mini-CEX), direct observation of procedures (DOPs) and multi-sourced feedback (MSF) which are used in assessing professionals in medicine, nursing and allied health professionals. Workplace-based assessments (WPBA) gather information about doctors’ and students’ performance in their day-to-day practice. These assessments provide opportunities to make judgements about how well individuals work and perform in complex environments and how they apply their knowledge and skills on a day-to-day basis in practice.

**Professionalism**

Clinical specialities have historically tended to privilege assessment of knowledge and skills at the undergraduate and postgraduate level, but recent years have witnessed a growth in the assessment of professionalism. A range of high-profile healthcare scandals have propelled professionalism to centre stage and now a raft of new assessment tools have followed in its wake. Yet professionalism is a slippery concept that has escaped an accepted universal definition. Everyone knows when a colleague is or is not ‘professional’ but defining it is far more complicated. Being professional relies on a range of attributes and is demonstrated through practice (for a further discussion see Chapter 5). Assessment of professional practice relies on a broad range of different sorts of assessment, forming a wide range of sources in multiple

---

![Figure 1.4](image)
contexts. Multi Source Feedback (MSF) is a good example of using others to appraise professional practice by relying on peer assessment. Peer assessment has been around a long time in education, many universities mandate an annual observation of teaching by colleagues to support their quality assurance processes. Now this notion has been expanded into the clinical workplace and peer reviews of work, for example in appraisal meetings, validate evidence of maintaining professional practice.

Keeping logs, portfolios, and so on is seen as an important aspect of documenting the running commentary of learning and reflecting about work. The advantages of the portfolio mode of assessing practice is that it is flexible, allowing an individual to sequence and document learning activities that have arisen in an ad hoc way from the workplace. A well-crafted portfolio can certainly represent a meaningful reflection of someone’s work, but their role in high stakes assessments is not without controversy. The choice to include, or exclude, material in a portfolio is in effect a work of self-assessment. Self-assessment has to a large degree – and certainly until recent years – been ignored and overlooked. The evidence for its validity and reliability has often been lacking. However, this is changing and the role of self-assessment is becoming an important area of development. One particular area that has come to the fore is providing students with feedback on their progress and the use of self-assessment exercises which can provide learner’s with this insight. Feedback on assessments is an area that all higher education institutions strive to improve in response to National Student Surveys.

There are two main sorts of assessment: summative and formative. An example of a summative assessment is a multiple-choice paper, it aims to make a definitive judgement about whether somebody has reached a certain standard, whether someone has passed or failed. In contrast, a formative assessment is primarily an educational process. It provides feedback about an aspect of practice, not to pass or fail but to support development and advise the learner about where they must concentrate their efforts in order to improve. Some assessments, which will be covered later on (workplace-based assessments, appraisal, see Chapters 4 and 7 respectively), started off as formative processes and subsequently developed into assessments that have become summative, and this has introduced a range of complex issues for the healthcare professional to consider. There are also two main types of assessor, hawks and doves. A hawk is an examiner who is critical and tends to be harsher in their assessment whilst a dove is someone who is much more lenient with their marking and is more likely to mark higher. Having assessors that vary in this way is an issue that is a threat to the robustness of the test and one that disgruntles those being assessed (see also Chapter 9 which discusses examiner behaviours).
What makes a good assessment?

So, what makes a good assessment? There are a range of attributes that make an assessment a sound one. Good assessments are valid, reliable, cost-effective, acceptable and feasible. The van der Vleuten equation (1996) describes the usefulness of an assessment by the sum of these attributes.

Validity

A valid test is one that measures the attribute or performance that it sets out to measure. For example, a multiple-choice paper is a valid way of assessing somebody’s knowledge because it can ask a lot of questions that cover the full breadth of the topic. A single essay assessing somebody’s knowledge of, for example anatomy, would not be considered valid because the test would not be able to assess the completeness of somebody’s knowledge about this discipline. You may hear this being referred to as the ‘Ronseal test’, that is it does exactly what it says on the tin.

Reliability

A reliable test will give you the same results over and over again; it’s about how consistent the test is. In Chapter 8 you will read about the membership of the Royal College of Physicians examination, a high stake examination undergone by doctors who wish to become medical specialists. This examination has to consistently pass, or fail, individuals; each examination, despite having a different range of question items, has to be comparable and guarantee that only those who have achieved a certain standard will get through the exam.

Reliability is a mathematical estimate of how replicable and consistent a measurement is. Reliability is necessary but not sufficient for validity.

Validity is the evaluation of how well a test measures the theoretical attribute, or construct, that it purports to measure. Validity relies on a mathematic relationship between test results, but also on careful procedures and judgement. All validity is essentially construct validity.
Overview of the book

The subject matter within the chapters is aimed at medical and health professional educators with an interest in promoting best practice in assessment, student evaluation and feedback. The issues to be discussed include the methodologies used in criterion referencing that assure competence and fitness to practice in the areas of knowledge and skills. Having given an overview and a theoretical framework in Chapter 1 on which to build, Chapter 2 has been given over to exploring the utility of the different question formats, that is essay-type, short answer, single best answer and extended matching questions in the assessments for undergraduate and postgraduate healthcare professionals. The practicalities in their design for gaining content and face validity are discussed and an exploration of the advantages of the different formats is included, along with guidance on best practice and common errors to avoid when writing these questions. Likewise, research project and dissertation preparation and assessment are also discussed, with guidance on supervision and mentorship for proposals through to expectations for, and delivery of, the written product.

The issues discussed in Chapters 3 and 4 are very much aligned for they focus on competence assessment of clinical performance. Chapter 3 has captured the essence of the debate that continues over the uses of different forms of practical assessments of clinical methods, that is long case, objective structured long examination record and the various competence assessments. The practicalities of planning an objective structured clinical examination are articulated, not least the development and quality review of OSCE station design, with advice on training simulated patients and the examiners. A clear steer has also been given on OSCE scoring and the constant dilemma over providing feedback on an individual performance. The evidence in Chapter 4 goes further with a pragmatic consideration of the different modes for assessing performance in authentic settings, including simulated and clinical practice. The role of simulation suites in preparation for practice in acute settings have been advocated by the Patient Safety Agency. Johnson and Wiseman have provided a critique of their use as well as advice around writing scenarios and forward planning with technical support for dealing with the challenges of simulation in training. The different tools for in vivo assessment of competence, such as mini-clinical evaluation exercise (mini-CEX), directly observed procedural skills (DOPS) and case-banded discussions (CbD) are explained. The real concerns of assessment in the dynamic environment of the ‘world of work’, identifying busy clinical professionals and the conflict between different choices in knowledge and skills application for patient management are debated as well as the future of workplace-based assessments.
With professional development and the demonstration of professional behaviours being made explicit within medical and healthcare professional programmes, there is compelling evidence for developing robust assessment processes and assessment tools that measure attainment. Chapter 5 seeks to define the term ‘professionalism’, and gives a steer on its learning and assessment, effectively allowing these somewhat qualitative areas of practice to be quantified. Gill has also provided guidance on common spheres where professionalism can be assessed (formatively and summatively) in the practice setting with triangulation of evidence, in formal clinical assessments such as OSCE, and through the use of reflective writing.

The areas for discussion in Chapters 6, 7 and 8 are inextricably linked: in Chapter 6, an explanation is given for the use of the different types of assessment methodology in the standard setting for criterion referencing and ‘trustworthiness’ of the scores as an accurate reflection of candidate performance. Indeed this has been borne out by the testing to provide compelling evidence of competence for registration with the professional regulatory body. Thus this chapter describes the methodologies commonly used for defining the pass score and how to quality assure the assessment content through blueprinting against curricula outcomes. With increased use of virtual learning environments in assessment, the possibilities available for test construction and for evaluation of performance data for test items and post-test statistical analysis have been explained for enhanced test reliability.

Chapter 7 focuses on formative assessments and the role of feedback as an assessment for learning. It charts the rationale for why feedback has become so prominent in the assessment processes for undergraduate and postgraduate healthcare professionals. The context for giving feedback, the roles healthcare professions may have in enacting this duty, as well as the challenges to giving effective feedback are addressed, before moving on to look at the theoretical underpinnings for this practice. Practical suggestions and models of feedback are offered as a starting point for developing confidence and expertise in facilitating feedback conversations. The last part of this chapter looks at contemporary assessments which have feedback at their core, multi-sourced feedback and appraisal.

In Chapter 8 the complexities of psychometrics and assessment are discussed, and in particular the themes around reliability and validity, establishing and evaluating the evidence around these qualities. A helpful dialogue of classical test theory, item analysis and discrimination is also given. Likewise the key issues for establishing and evaluating test item validity and its impact on the assessment process are valuable. These issues, and that of the reliability of performance assessments in relation to inter-rater reliability and generalisability theory, provide a clear introduction to the areas under discussion in Chapter 9.
The various characteristics of examiners and candidates are discussed in Chapter 9, with practical approaches for identifying, and thus helping to avoid, assessment errors. The roles and behaviours of internal and external examiners are considered in agreeing assessment criteria, the quality assurance of the marking and the moderation process. From another perspective the multi-factorial basis as to why students fail is debated. That is, the extrinsic causes associated with the learning environment, and the intrinsic factors that may be one or more of the following: inadequate self-motivation, learning approaches that they find uninspiring and low levels of natural aptitude. In practical terms the possible mechanisms for identifying problems in those who seek, or are reluctant to seek, help and some ways of supporting them to reach their full potential, are recommended.

Finally Chapter 10 outlines possible future developments in higher education assessment and feedback. As mentioned in the introduction, it is not possible to predict the future, but the issues and challenges outlined are those that I believe are most likely to influence the lives of the higher education assessment faculties of the future. It is essential that, as an academic discipline, medical educators get better at developing their vision of the future, to ensure that as far as possible we are able to secure the highest quality of assessment and feedback for our students and trainees.

References


Further reading


Chapter 2  **Assessment of knowledge**

*Dr Steve Capey*

University of Swansea

---

**Introduction**

Schools that offer awards accredited by a professional regulator are publicly accountable, and must verify that the marks awarded to successful candidates equate with eligibility for registration in the chosen profession. It is the role of academics to construct quality assured assessments that test the published learning objectives precisely. Accordingly, the interpretation of results of reliable and valid tests should allow conclusions and legitimate recommendations to be made for student progression.

There are number of practical question formats to explore depth and breadth of understanding in a given discipline. For example, short written answers, longer essay-types, question ‘stems’ with a selection of answers from a number of options (i.e. multiple-choice questions or MCQs). It is therefore meaningful to explore how different question constructs are applicable to the nature of the knowledge and its synthesis. It is rare that a single question format can fulfil the requirements of testing both breadth and depth. Therefore it is advisable to use more than one test format to achieve a balanced approach for assessment of the various dimensions of knowledge acquisition – their configuration, utility and benefits are discussed throughout this chapter.

**Classification of written questions**

Written questions may be categorised on the basis of whether the answer is *constructed* or *selected*. Constructed answer questions call for an answer to
be given from memory or using the deductive reasoning processes appropriate to tackle the question. In contrast, selected answer questions require the candidate to choose the correct or most appropriate answer from a limited series of different items. Their response could be from a dichotomous choice (i.e. true or false), a single best answer, or extended matching questions where a collection of different responses must be matched to a set of questions or scenarios. Obviously the decision on the precise format should accord with its capacity to test a particular knowledge construct and to demonstrate face validity, that is the test appears to measure what it sets out to measure, each having its characteristic pros and cons.

Examination of pure or applied knowledge can be achieved effectively by selected responses, whereas the constructed answer format is more applicable to testing an ability to develop a reasoned argument around a given topic. Likewise, clinical vignettes for selected questions have transformed them from knowledge recall to knowledge application. Given limited time, a particular problem for long, detailed question formats is that of ‘case specificity’. While a candidate may demonstrate a superior and comprehensive knowledge relevant to the specific clinical specialism, their answer does not mean equal competence in a test question around another area. With the obligation to produce reliable and valid assessments, the in-depth questioning styles of modified essay questions and essays have been used less frequently owing to their relative inefficiency in sampling widely across a range of disciplines.

**Essay questions**

Essay-type questions are successful at testing comprehension, exploring a detailed knowledge and critique of a topic, and have their place for the development of academic writing as a transferable skill. The classic essay question generally includes a prompt on the expected content in the answer; this may cover a wide range of topics or explore salient features of a discipline. To avoid uncertainty in expectations for answer content, the question could include a short title that gives appropriate directives on the relevant information to include and how to structure an acceptable response. Hence, due diligence is needed in question construction to diminish the risk of ambiguity and to afford candidates the opportunity to excel (see Box 2.1); this also helps foil the exam-wise student who has a well prepared, wide ranging essay that they can slip into any too open ended question.

Marking is an isolating experience for the assessor; the hours set aside to mark essay questions are a substantial drain on faculty resources. Deriving precise and relevant criteria for the marking schedule is also challenging, and unlike short answer questions or selected answer formats, standardisation of
the assessors is complex, so marking can be subjective. Therefore, it is good practice for the criteria to be agreed in advance, by a group of assessors, who can standardise a comprehensive marking matrix that covers the expected answer content together with guidance on expectations around the award of marks within different grade boundaries (Jolly, 2010; see also Chapter 9 for a discussion on examiner behaviours, and the different assessment practices used to quality assure consistency in marking).

**Modified essay questions (MEQ)**
The MEQ was developed in the late 1960s to address the issue of ‘case specificity’ by increasing the knowledge sampling frequency above that of traditional essay questions (Jolly, 2010). An MEQ requires highly structured short paragraphs to answer a series of questions posed by a brief clinical scenario on a precise clinical area. Being shorter than an essay, this test format affords greater sampling of knowledge across the curriculum and, as
model answers are easier to articulate, the commensurate marks awarded can be more clearly defined. Test papers that include a structured marking scheme help candidates to prioritise their time.

The assessors need a comprehensive document outlining the intended and expected components of the model answer together with marks allotted, as well as a margin of flexibility in case candidates provide pertinent additional material. As with any essay format, it is prudent to have the answers marked by academics who are familiar with the material. The disadvantages of MEQs are again that the time allocated to mark them remains substantial, the marking schemes are sometimes hard to construct and it can be difficult to gain agreement on how to on award marks. While standardisation is very valuable it throws up differences of opinion among examiners that need to be considered when planning the marking scheme.

Although MEQs are reasonably reliable for use in high stakes assessments (Feletti, 1980), it is necessary to guard against testing only factual knowledge recall if clinical decision-making is to be tested as well. Intuitively many examiners feel that MEQs should be better at testing higher level cognitive skills but research has shown that it is so difficult to write good MEQs that, in practice, they are no better than MCQs, while MCQs are usually more reliable. (Palmer and Devitt, 2007; Palmer, et al., 2010).

---

**Box 2.2 Example of a modified essay question (10 marks)**

Geoffrey Appleby is a 73-year-old man with Stage IV lung cancer who has asked you about the possibilities of euthanasia should the pain became intolerable.

1. What ethical and legal principles should be included in your focused discussions with Mr Appleby? (6 marks)
2. What advice would you give Mr Appleby about treatment options for pain relief? (4 marks)

**Planning the mark scheme**

1. To include the legal and ethical issues related to end of life care and should be constructed in alignment with the teaching on the programme and the current regulatory and legal guidance.
2. To include the different management options for pain relief in Stage IV cancer therapeutics and holistic therapies.
Short answer questions (SAQs)
These are concise questions that require an answer in response to triggers and associated data, such as pathology reports and radiological images. Essentially, they permit the sampling of specific knowledge where the response might range from a single answer construct, a short paragraph of text, data interpretation or an explanatory or annotated diagram. The answer cueing effects that can occur with MCQ-type questions are not generally such a problem because the response is constructed (self-generated). The same diligence as with MEQs is needed to ensure all questions are unequivocal, because a disadvantage of SAQs is that they are open to interpretation by the candidates (and examiners) and, accordingly, post-assessment appeals.

Multiple-choice questions (MCQ)
Multiple-choice questions were introduced on a large scale for the ‘Alpha’ aptitude test, used by the US army, to assess World War One recruits. Leading the way, the US Medical Licensing Board implemented MCQs in the 1950s to replace their essay-style listening examinations, and they have since gained favour with many medical and healthcare professional curricula. MCQs can be adapted for several constructs and are wide ranging in their utility and purpose. Although initially True–false questions were common, more recently the single best answer (SBA) out of five options and extended matching questions (EMQ) are used more widely.

There are different scoring systems that have important implications for candidate behaviour, so the system used should be clearly articulated in the guidance document or Schedule of Assessment. Two types are frequently used:

• Right-scoring, where each correct answer is awarded a mark;
• Negative marking (formula scoring) where a mark is deducted for each wrong answer.

In right-scoring, candidates benefit from answering all questions, even at random when they do not know the answer. Negative marking was introduced to dissuade guessing, but with it comes another variable of a negative psychometric effect. Candidates tend to deliberate more over each question where there is negative marking, and the less able ones may not complete the paper. Evidence for poor exam technique has been suggested as a reason for low scores in some instances (Hammond, McIndoe and Spargo, 1998). Conversely, good candidates are also often test-wise, so even with incomplete knowledge they may ‘guess’ the right answer. To discourage ‘guessing’, a ‘don’t know’ option has been introduced by some assessors (Muijtjens et al., 1999), there is after all the probity issue, that medical and healthcare students should recognise that they have knowledge gaps.
**True–false questions**

The dichotomous choice in True–false answer questions is the simplest form of MCQ that has been somewhat disregarded because they tend only to test knowledge recall rather than its application and synthesis. They are straightforward for the candidate, large numbers of items can be constructed relatively easily and they allow large areas of knowledge to be sampled in a short testing time. However, they have a number of innate flaws in that:

- The candidate who does not know the correct answer still has a 50% chance of a mark if they guess;
- There are few instances where a concept is unequivocally true or false in science and medicine, consequently the single best answer construct is more realistic and credible.

Typically a brief lead-in is followed by 4 or 5 statements each of which must be marked true or false. Any combination from all to none of the answers may be correct. As each statement must be considered and answered the student must evaluate far more material than with a single best answer.

True–false items tend not to be used as frequently in summative assessments, even so they have value for formative assessment of core concepts. But the prerequisite has to be that they are used in conjunction with immediate feedback in order to avoid any misleading ‘false’ material being retained by students as fact.

**Single best answer questions (SBAs)**

The dialogue continues as to whether to call them single ‘correct’ answers or single ‘best’ answers. Either way, the SBA requires a candidate to select an answer from a series of options or distracters, normally presented in a grid, with each distracter assigned a letter for simplicity in individualising the response. The number of distracters and their proximity to the correct answer depends on the nature of the test. Often the alternative distracters may be ‘reasonable’ answers, but not the *best* answer. With academic progression, the ‘distracters’ used may augment the level of complexity, and with it the knowledge application and the discriminatory function of the test.

When testing applied clinical knowledge, the SBA starts with a theme to provide context, has a clinical vignette or scenario known as the ‘stem’, followed by a ‘lead-in’ question for which the candidate must choose from a series of answers (see Box 2.3). They are one of the preferred question styles in medical and healthcare education as they exhibit high reliability for the number of hours of testing. In the debate around the number of optional responses that are most suitable; evidence suggests that perceived fairness increases with the number of options per question, and most test writers
How to assess students and trainees in medicine and health

would advocate five to reduce the chance of guessing (Haladyna and Downing, 2003; McCoubrie, 2004).

There are many advantages to using SBAs for assessment in the expansive curricula of medical and health professional awards, not least their functionality of:

• Sampling a wide range of knowledge in a relatively short time, which reduces the problem of case specificity;
• Assessing core knowledge and its application in one test;
• Having high face validity when combined with a clinical vignette;
• Being marked objectively and efficiently with optical marking equipment.

The main disadvantage of SBAs is that they are complex questions, which are time-consuming to write. So to produce a large enough question bank with items that are reliable and have face validity demands a large amount of faculty input and training for success. Likewise the problem of answer-cueing remains for a candidate has a one in five chance of selecting the correct response.

Extended matching questions (EMQ)

The EMQ was developed as an alternative to free response questions that would not have the same answer-cueing effects found with other MCQ

---

**Box 2.3 Single best answer**

**Theme:** Dermatology

**Stem:** A 48-year-old woman has been referred by her GP to the dermatology outpatient clinic with evidence of intensely itchy knees and elbows. She is known to be gluten sensitive and does not generally suffer from malabsorption as she is fastidious in keeping to a gluten-free diet. She has no history of allergic disorders.

**Lead in:** What is the most likely cause of the skin irritation?

**Answer options**

A. Atypical eczema
B. Dermatitis herpetiformis
C. Impetigo
D. Psoriasis
E. Scabies

**Answer = B**
Assessment of knowledge

formats (Case and Swanson, 1993, 1994). The EMQ item generally involves a number of clinical scenarios for which the candidate must select a response from a matrix of around 8 to 26 potential options. The emphasis must be on potential because although the upper limit of options can be extended, it is difficult to achieve this and ensure that options are reasonable responses. These items are different from other selected response formats as they have a substantial selection of possible answers available that are required as the responses to several questions with clinical scenarios. The most appropriate use for EMQs is as a test of applied knowledge rather than the more ‘descriptive’ approach in knowledge recall.

EMQ-type assessment papers may be difficult to construct, particularly for some of the more reflective disciplines, and require staff training. A large number of EMQs are required to sample a suitable breadth of knowledge, and the more item statements or scenarios included as part of an individual EMQ serve to increase the difficulty in construction. Nonetheless like SBAs they exhibit reliability when written effectively, and can be marked objectively and efficiently with optical marking equipment. Additional benefits of EMQs are that they:
• Allow the testing of a variety of clinical scenarios on a linked theme and so test knowledge in greater depth than a single MCQ;
• Avoid some of the cueing effects of MCQs because a larger matrix of answers is used;
• The quality of questions can be determined simply by the ‘cover up’ test, that is candidates should be able to select their response ‘without looking’ at the answer options – this also applies to SBAs.

A well-constructed EMQ set should include four components:
1. A theme,
2. A list of option/responses in a matrix,
3. A lead-in statement,
4. Item stems or clinical scenarios (see also Box 2.4).

Common flaws in question items using multiple-choice formats

Writing question items for examinations is one of the less favoured tasks – both for the academics who write them and the administrators who have to cajole their colleagues to prepare them. It is recognised that constructing well-written probing questions that assess precise learning objectives, at the appropriate level of study, is demanding, and diligence in their preparation is necessary (see Box 2.5). There are a number of ways to avoid flawed questions finding their way into a test paper – the obvious one being to have all questions peer reviewed with constructive critique. ‘Answer cueing’ is also a problem that frequently occurs unintentionally, but owing to the complexity
Box 2.4 Extended matching questions

An example of an EMQ that might be used as a Basic Clinical Pharmacology question.

**Theme:** Mechanisms of anti-microbial drugs.

**Lead in statement:** For the following methods of action for anti-infective chemotherapeutic agents select the most appropriate drug that uses this method.

**List of options:** Each option may be used once, more than once, or not at all.
A. Amphotericin B
B. Clarithromycin
C. Clofazimine
D. Doxycycline
E. Isoniazid
F. Mefoloquine
G. Penicillin V
H. Selegiline
I. Sulfadiazine
J. Trovafloxacin

**Items or stems**

1. Inhibition of peptidoglycan cell wall synthesis.  Correct answer: G
2. Inhibition of bacterial protein synthesis by binding to the 30S subunit of the bacterial ribosome.  Correct answer: D
3. Inhibition of bacterial protein synthesis by binding to the 50S subunit of the bacterial ribosome.  Correct answer: B
4. Act as a false substrate for p-aminobenzoic acid, leading to the inhibition of bacterial folic acid synthesis.  Correct answer: J
5. Interfere with the replication of bacterial DNA.  Correct answer: E

of writing test items the risk still occurs that the questions may prompt a test-wise candidate into selecting the correct response. Here are some potential pitfalls:

- The order in which the possible answers are written could inadvertently give away the answer; this is easily avoided by listing the distracters for
Box 2.5 Five steps to writing multiple-choice questions Adapted from Case and Swanson (2002)

**Step 1: Identify the topic for the MCQ:** The topic could be one of a number of areas within the curriculum, but it is good practice to group questions together that are around a similar theme or specialty.
- Patient treatment; for example drug treatments for hypertension;
- A feature of clinical treatment; for example management, diagnosis, investigations;
- Non-clinical studies, for example ethical issues.

**Step 2: Write the clinical vignette:**
The scenario or vignette should be concise, providing only essential information needed to answer the question, with loquacious and irrelevant information avoided.

**Step 3: Prepare the list of answers:**
List all possible responses as:
- Either a few words or short sentences;
- In alphabetical order.

**Step 4: Review the question and list of answers:**
Ensure that the following are reviewed as quality assurance regarding:
- The relevance of the featured question in the whole test item;
- There is only one ‘most appropriate’ answer for the question. All other answers should be ‘possible’ responses and relevant to the vignette, otherwise any overtly inappropriate responses would simply reduce the number of potential distracters and confer an answer-cueing effect. For example, if a candidate is asked to select a drug therapy, then all possible responses should be drugs.

**Step 5: Peer review:**
The final part of assessment preparation is the review of assessment items:
- By an experienced colleague with a credible knowledge base to critique the questions for content accuracy, technical construction quality and to check for any ambiguities;
- Any essential information required to answer the question is provided.

**Additional considerations for writing EMQs**

**Lead in:**
An essential component when writing the lead-in question is that it needs to:
• Indicate the relationship between the scenarios/vignettes and the options in the response matrix;
• Clarify the question posed for candidates.

**Item responses:**
Need to be in the same format to allow the majority of the options in the list to be reasonable distracters.

---

**Box 2.6 Example of a bad MCQ**

A 67-year-old woman with a 9-year history of Type 2 diabetes mellitus has come to the GP for her regular check-up, which of the following tests would be most appropriate for assessing her diabetes management?

A. DEXA scan
B. Full blood count
C. HbA1c
D. Serum alkaline phosphatase
E. Urinary calcium

In this rather obvious example, the test-wise student would reason that items A, B, D and E were not as relevant to diabetes, so could predict the correct answer to be C.

---

each question in alphabetical order, which provides uniformity of presentation that does not point to the answer;
• Avoiding the use of negative questioning;
• Grammatical errors are common, arising when the test writer has to provide 5 possible responses, with less attention given to the distracters than the correct response (see Box 2.6). A clear pointer might be using any one of the following:
  • Singular or plural terms;
  • A word or phrase that is given in both the stem and the correct response;
  • The correct response has more details than the other distracters;
  • When asked to identify the correct response, ALL are correct (see Box 2.7 for examples).

To reduce the incidence of errors and answer cueing in assessment writing requires critical peer review and agreement on the test items. It is essential to agree the purpose of the test so that the complexity of the test items is
Assessment of knowledge

appropriate to the level for academic progression of candidates. The Board of Examiners must ensure that the test blueprint represents a consensus on content and that it matches the precise learning objectives to be tested. It is acknowledged that tests are read by individuals who are under stress. Hence the peer review of question items should quality assure that they are in a clear-cut format, with the tasks required being unequivocal to the candidate taking the test.

Confidence assessment of multiple-response items

Guessing is an ever-present problem in assessments that use multiple-response type questions. One solution has been to ask students how confident they are in their answer. This has the added advantage that it trains students to think about how certain they are about their actions when in practice, ‘Am I certain I know what to do or should I look it up?’ (Gardner-Medwin, 2006). A system pioneered in London medical schools asked students to score each of their answers ‘1’ if unsure, ‘2’ if fairly sure and ‘3’ if very confident. If they answered the question correctly they were given their self-assigned confidence score as the mark, that is 1, 2 or 3. If they answered incorrectly they were given 0 if unsure, -2 if fairly sure and -6 if highly confident. The system tends to reward bright students who are confident in their

Box 2.7 Common errors when writing MCQs Adapted from Case and Swanson (2002)

- A question has more than one response and so last option is: ‘All of the above’ or ‘None of the above’;
- Too much information is included in the question;
- The candidate can predict the correct answer because a word or phrase is given both in the stem and the correct response;
- A summative assessment question is used as a teaching tool;
- Information in the question is inaccurate;
- Information in the question is ambiguous;
- Irrelevant information is included in the test item;
- ‘Trick’ questions are included that can confuse the candidates (for information in assessment of learning can also be a learning experience);
- Insignificant details are asked for, for example ‘What is the molecular weight of the alpha subunit of human insulin receptor?’ These questions are easy to construct, but they test recall rather than higher level cognitive skills.
knowledge but severely punishes poor students who are unaware of their ignorance, thus eliminating students likely to be dangerous in practice.

**Specific tests of clinical reasoning**

**Key feature problems (KFP)**

This test format has been used extensively in postgraduate examinations in Canada and Australia, and more recently in undergraduate examinations. It begins with a detailed clinical vignette of a patient problem, followed by a series of questions designed to probe the candidate’s ability to manage safely the specified clinical presentation. The remit of KFPs is to test the critical stages of clinical reasoning and decision-making skills (i.e. integration, interpretation of data and application of knowledge to make a clinical judgement), and may be employed to sample a wide range of acute or chronic scenarios according to the scope of practice. The answer system may be brief and specific answers may be generated by the candidate (see Box 2.8 Example) or chosen from a list of options (like an MCQ).

### Box 2.8 Example of key feature problems

**Clinical scenario:** David Thomas is a 75-year-old retired plumber who has been brought to the Accident and Emergency Department by ambulance, after his daughter found him in a confused and anxious. This was recent onset, as he had been his normal self when she visited two days ago. David Thomas is known to have Type 2 diabetes mellitus and hypertension. On examination he has a heart rate of 100 beats per minute and a blood pressure measurement of 100/70 mmHg. The attending physician found his abbreviated mental state score to be 3/10. He is currently being prescribed the following medication:

- **Metformin, 500 bd**
- **Bendrofluazide 2.5 mg one a day**
- **Aspirin 75 mg**
- **Simvisatatin 40 mg**

A sample of venous blood is taken for analysis and shows:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Normal Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Na</td>
<td>118 mmol/L</td>
<td>136–146 mmol/L</td>
</tr>
<tr>
<td>K</td>
<td>3.2 mmol/L</td>
<td>3.4–4.4 mmol/L</td>
</tr>
<tr>
<td>Urea</td>
<td>10 mmol/L</td>
<td>7.9–16.4 mmol/L</td>
</tr>
<tr>
<td>Cr</td>
<td>265 μmol/L</td>
<td>60–110 μmol/L</td>
</tr>
<tr>
<td>eGFR</td>
<td>45 ml/min/1.73 m²</td>
<td>100–130 ml/min/1.73 m²</td>
</tr>
</tbody>
</table>
Questions:
1. From the information you have what is the likely cause of David’s confused state? *List 2 only* (2 marks)
2. What further tests would assist you in your diagnosis? *List 2 only* (2 marks)
3. What are the major contra-indications for bendrofluazide? *List 3 only* (3 marks)
4. What are the main beneficial actions of metformin? *List 2 only* (2 marks)
5. What three (3) immediate interventions would you take to improve David’s confused state? *List 3 only* (3 marks)

Model answer and scoring plan:
Qu1. Dehydration (1 mark)
   - Hyponatremia (1 mark)
   - Hypoglycaemia (1 mark)
   - Hypokalemia (0.5 mark)
Qu2. Serum osmolality (1 mark)
   - Urine osmolality (0.5 marks)
   - Osmolality (0.5 marks)
   - Urine dipstick, BN stick (0.5 mark)
   - Blood glucose (1 mark)
   - Urine sodium (0.5 mark)
   - Urine FENA (0.5 mark) forced excretion of sodium
   - Full Blood Count (0 marks)
Qu3. Refractory hypokalaemia (1 mark)
   - Hyponatraemia (1 mark)
   - Hypercalcaemia (1 mark)
   - Hyperuricaemia (1 mark)
   - Addison’s disease (1 mark)
Qu4. Inhibition of hepatic glucogeonesis (1 mark)
   - Increase of peripheral glucose utilisation (1 mark)
   - Inhibits intestinal glucose absorption (1 mark)
   - Improves insulin production (0.5 mark)
Qu5. Stop bendroflurazide (1 mark)
   - Stop metformin (1 mark)
   - Rehydration therapy, Fluid resuscitation, intravenous infusion, IVI (1 mark)
   - Dextrose (1 mark)
How to assess students and trainees in medicine and health

As with all clinical assessments, their design and peer review are key elements for success so they are resource intensive. When constructing the clinical vignette it is important to provide the necessary information without either confusing the candidates, or betraying the answer. For an equitable assessment comprehensive marking guidance and criteria are essential, but are difficult to construct for free text response type questions. Then again, KFPs have two main advantages over other case-based question types:

• They allow a larger number of clinical scenarios to be tested in a short time frame, thereby increasing the reliability of the assessment and reducing problems with case specificity (Page and Bordage, 1995);
• Scoring the answers is generally easier because correct answers tend to be succinct.

The outcome for well-written KFPs is a question style that has been shown to demonstrate high reliability and validity (see also Box 2.9 for Eight steps to preparing key feature problems).

Script concordance items (SCI)

These are another form of assessment developed as a ‘Diagnosis script questionnaire’ to test clinical reasoning skills where there are elements of uncertainty in the patient presentation and management. With SCIs, candidates are questioned on a decision that is a crucial step within the clinical reasoning process (Fournier, Demeester and Charlin, 2008). The SCI marking grid is constructed thus:

• Test items are scored by the panel of experienced clinicians;
• Each option is awarded a score based on the number of experts that selected this option as the optimum solution.

The scores awarded reflect the level of agreement of the candidate decisions with those of a panel of experienced clinicians. When a candidate selects an option they are then awarded the score that relates to that option for that particular case presentation. A high degree of concordance with the panel equates to good practice in the use of information from the case presentation, and thus an indication of the clinical reasoning competence of the candidate (see Box 2.10). The SCI is known to demonstrate good face validity and is an effective test of the clinical reasoning process, having the capability to discriminate among candidates at different levels of experience (Charlin et al., 1998). Moreover, a significant number of cases can be reported on within a short time period, thereby giving a high sampling frequency. Again, faculty training and development sessions are essential to quality assure the assessments, both in writing the SCIs and in their delivery. Further, as students may be unfamiliar with the test format, the use of formative SCIs is advocated, prior to their use summatively. The number of questions required
Box 2.9 Eight steps to preparing key feature problems (adapted from Page, Bordage and Allen, 1995)

Step 1: Select a clinical problem. It is good practice to define the following:
• Age and gender of the patient;
• The setting of the clinical presentation;
• The appropriate clinical data, for example pathology reports, radiological images.

Step 2: ‘What are the critical steps in the resolution of this problem?’
• Identify the key steps that would be involved in the clinical decision making to manage or diagnose this clinical presentation.

Step 3: Think of different ways that patients present with this problem.
• The presenting complaint and/or reason for the clinical encounter, for example signs and symptoms and their duration.

Step 4: List the essential key features involved in the care of this group of patients.
• List as many of the essential features that are required for the resolution of the specific clinical presentation.

Step 5: Select a typical case presentation and write the clinical vignette for this particular presentation.
• The vignette should be as detailed as possible without containing unnecessary information that is irrelevant to solving the problem.

Step 6: Write the questions and construct scoring keys which test only the key features of the case presentation.
• The question format – short answers or a choice from a prepared list of responses;
• With short answers, the question format should be direct, for example ‘What is your provisional diagnosis?’
• Additional instruction should include the number of allowed responses,
  ○ list up to five . . . ;
  ○ select up to three . . . ;
  ○ when asked to select one response, it suggests a definitive answer is required.

Step 7: Scoring criteria.
• This is essential information for precise assessment which needs to be unambiguous and provides for any marginal difference in formulating the answers;
• Appropriate marks awarded for each element or answer to the questions;
• The relative weighting of responses should be an indication of the significance and potential consequences of the answer given in patient management and decision making;

**Step 8:** The final part of assessment preparation is the review of assessment items.
• By an experienced colleague with a credible knowledge base to critique the questions for content accuracy, technical construction quality and to check for any ambiguities;
• Check that any essential information required to answer the question is provided.

**Box 2.10 Example of a script concordance item (Fournier, Demeester and Charlin, 2008)**

An 86-year-old man suffering from acute chest pain and shortness of breath has been taken by ambulance from the nursing home where he lives, to the local Accident and Emergency Department.

**You were thinking of:** Angina pectoris.

The patient was administered Glyceryl trinitrate sublingually and his symptoms abated.

**What effect would this finding have on your diagnosis?**
+2 Almost ruled out
+1 Less probable
0 This finding has no effect on the diagnosis
−1 More probable
−2 Almost certain

to produce a reliable examination appears to be around 20 cases with each having 3 questions (Fournier, Demeester and Charlin, 2008) and, like other practical tests, the number of judges required is between 10 and 15 to create a robust item (Gagnon, *et al.*, 2005).

**Projects and dissertations**
Research projects are generally assessed by an extended written document which tests a number of essential generic and transferable skills:
Assessment of knowledge

- Project design;
- Research project delivery;
- Literature review;
- Data analysis and critique;
- Academic writing.

Guidance for successful delivery of projects and dissertations

Candidates are given criteria on which the content will be judged and guidance on expectations for written detail, such as word count, referencing style, layout and presentation. They are not only tested on these generic outcomes, but the specific focus of the project which might necessitate one or more of the following activities:

- A structured written report on research that follows the traditional layout (introduction, literature review, aims, methods, results, discussion, conclusions, reference list);
- A library-based project;
- A poster presentation – prepared alone or as a group activity;
- An oral presentation – alone or as a group activity.

Each of these activities has its value. When allowing students to complete original research, case study, literature reviews and audits, it is essential to provide guidance from an experienced supervisor. Often a relatively inexperienced person will actively oversee the project work, but they must be mentored by an experienced academic to ensure fairness in supervision and marking. Projects may be selected from an agreed list or be the student’s own design. With written projects, the local university registry normally has guidance on the word count which should take into account the discipline-specific needs, for example a reflective discipline such as the humanities having different criteria from scientific-based project write-up (see Table 2.1).

One concern with student-led research proposals is that their ideas may be rather ambitious with regards to the cost and what is achievable in the limited time, therefore a steer is essential. An initial approach is to ask candidates to deliver the project proposal as a verbal presentation to peers and potential supervisors. The very act of formulating their project for a ‘performance’ motivates them to focus strategically on potential problems in the design, so with constructive critique from academics and peers a deliverable project design should emerge. The projects permit students to explore in depth a subject area of interest and for those wishing to go on to postgraduate study, provide an insight into the positive, as well as the demanding and sometimes repetitive features of research.
Assessment of projects and dissertations

The important issues that exercise students and academics are:
• The level of guidance and supervision; this can be significant for undergraduates;
• The diversity of subjects covering narrow areas of content in-depth;
• The identification of assessors with the appropriate knowledge to judge them equitably.

Frequently the academic best placed to assess content will be the supervisor, but this is generally seen as a conflict of interest that could result in a skewing of the grades. Hence it is considered good practice for the project to be blind-double marked and the marks validated or adjusted as appropriate by the external examiners. Candidates and all examiners (local and external) require the assessment criteria in advance, and wherever possible the markers should be matched to the subject areas of the projects. In cases where new academics are introduced to these areas of activity, it is advisable for the
Assessment of knowledge

novice to shadow experienced assessors by ‘third’ marking projects followed by in depth discussions on the rationale for the scores awarded.

**Portfolios**

A portfolio is a compendium of documentary evidence, such as certificates of attendance, academic transcripts, log books, reflective practice, gathered either during the course of a programme of study and/or as proof of continuing professional development and lifelong learning. Portfolios have two distinct functions:

• As a learning portfolio – evidence of learning and reflective practice, part of which might be the log book used in undergraduate programmes or postgraduate training posts;
• As an assessment portfolio – if formative, as evidence of experience, for example workplace-based assessment (see Chapter 4), or may contribute to the summative assessment in a training programme.

When a manageable link between assessment and learning is established, then a high degree of face validity is achieved.

Portfolios facilitate the curriculum being learner-centred and the subject and materials included can be diverse, depending on their purpose. Therefore the rationale must be clearly defined, with guidance as to its contents, if the portfolio is to have an educational impact of worth for the student or trainee. The evidence provided is often subjective and personalised to allow flexibility and support the learner-centred approach (Van Tartwijk and Driessen, 2009). Without doubt, the increased acceptance of portfolios (paper-based and electronic) in academic programmes has been viewed as preparation for professional life. In vocational awards and post-registration, their utility has also been extended from being exclusively a learning tool to contributing to a personal development plan (PDP).

There are expectations for portfolios to include documents on reflective practice that contribute to academic progression, together with evidence of mentorship and an action plan for achieving the learning outcomes. Like other log book-type assignments, their immediate usefulness is not always acknowledged, and they can be seen as time consuming for students to prepare and the faculty to assess. Non-compliance, or at least procrastination over portfolio maintenance, is common, with a flurry of activity to amass proof of practice around the submission deadline. One concern that surrounds the portfolio-based assessment is the question of equivalence of marking the content, and to avoid any conflicts of interest the mentor for portfolio maintenance should not be the assessor for summative purposes.
It is worth mentioning that support for writing in a reflective style is essential, particularly for students from a predominantly ‘hard science’ education, to make sure that they have the skills to cope with the attitudinal and professionalism domains assessed via the portfolio. A more accurate longitudinal view of performance is afforded when portfolios are prepared systematically. An electronic format has the added dimension of flexibility of access for students and assessors and so ought to encourage concordance.

References


**Further reading**


Introduction

In undergraduate and postgraduate education of medical and health professionals, assessments usually include a combination of competence assessment and performance assessment. Competence assessments are to test what a professional is able to do in clinical practice, while performance assessments are used to test what they actually do in their clinical practice. Competence assessments are generally used in ‘high stakes’ assessments such as finals examinations or postgraduate assessment to gain membership of a Royal College. They are summative and convened at an allocated time in a contained environment. Accordingly, candidates are judged on their performance of the assigned task at a specific time.

Performance assessment, also called workplace-based assessments (WPBA) or structured learning events (SLE) test the professional in the clinical environment. These assessments are structured and continuous, unlike the tests that are used to make a judgement of competence. By completing repeated assessments, the candidate has the opportunity to accumulate documentary verification of their progress. Accordingly, this evidence can be used to identify any ‘gaps’ in practice which will allow the candidate to plan their individual development needs. By using a wide range of WPBA or SLE, the outcomes thereof can be evaluated for identifying personal strengths and weaknesses in distinct areas, such as technical skills, professional behaviours and team-working. (For more details about this form of assessment, see Chapter 4.)
Development of skills competence assessment

**Dimensions**

There are several dimensions to clinical competence that are the scientific knowledge base and other professional practice elements (see also Epstein and Hundert, 2002); these include:

- History taking, clinical method;
- Skills in practical procedures;
- Health professional–patient communication;
- Problem solving ability;
- Management skills;
- Relationships with colleagues;
- Ethical behaviour.

A range of assessments have been developed examining these different dimensions, however ensuring these are all reliable and valid is challenging.

**Test formats**

To assess clinical competence, there are a number of different test modalities or formats available and these include the

- Long case examination;
- Objective long case examination record (OSLER);
- Objective structured clinical examination (OSCE);
- Objective structured assessment of technical skills (OSATS).

The most commonly used format in both undergraduate and postgraduate assessments is the OSCE. The strengths and weakness of each of these formats are described in the next section.

**Competence assessments**

The criteria used to evaluate any assessment method include:

- Reliability;
- Validity;
- Educational impact;
- Cost efficiency;
- Acceptability;
- Responsiveness to change.

While all of these criteria are important for high stakes, the two fundamental issues are reliability and validity (see also Chapter 6; Schuwirth and Van der Vleuten, 2006). Reliability tests how good an assessment method is at measuring student performance, including discriminating between good and poor students. Validity asks whether it measures what it is supposed to measure, and as such also tests whether the assessment is linked to the curriculum. If any of
these criteria are not present then the results obtained by the assessment could be subject to challenge. Each format of assessment is discussed and the main advantages or disadvantages of each format highlighted in terms of these criteria.

**Long case examination**

In a long case examination the candidate has up to one hour to perform a history and examination on a real patient – this process is not observed. The candidate then presents their clinical findings to the assessors (normally two) and is then questioned on aspects of the case. Occasionally the candidate is taken back to the patient to demonstrate the clinical signs to the assessors.

**Advantages:** As the long case tests what a candidate does in actual practice, that is tests their ability to interact, assess and manage a real patient, it has high validity.

**Disadvantages:** The main disadvantage of the ‘long case’ as an assessment is that it has low reliability. One of the main factors affecting its reliability is lack of consistency between examiners, which is a consequence of the unstructured questioning format marking without anchor statements and the lack of agreement of the level of stringency before the examination (Norcini, 2001). Another factor is associated with the variability in degree and level of detail of information disclosure by the patient, which is influenced by their individual personality and health at the time of the assessment. Some illnesses are straightforward whereas others are complex and both categories of patient presentations traditionally have been included in these assessments. Candidates’ skills also vary according to the task being assessed, so examining a single patient may not provide an accurate estimate of their ability because using one patient tests a narrow scope of practice (Swanson *et al.*, 1995; Norcini *et al.*, 2001; Norcini, 2002).

Another disadvantage of the long case format is that there is no direct observation of the history-taking process so the candidate does not have to demonstrate good patient communication skills. All of these problems reduce the acceptability of this format of assessment for candidates.

**Objective structured long examination record (OSLER)**

In light of the criticism about the reliability of the long case, the OSLER was designed to examine in a more structured format whilst still investigating a real patient under examination conditions. In the OSLER, the candidate is given a fixed time period to perform a limited history and examination on a real patient before presenting their findings to the assessors (usually two). The assessors are in attendance throughout the examination and mark using a structured mark sheet (see Figure 3.1) which consists of ten items, four on
Figure 3.1 Example of an OSLER mark sheet.
How to assess students and trainees in medicine and health

history-taking, three on physical examination and three on management and clinical acumen (judgement). Each assessor marks the candidate independently then discusses their grades and the level of complexity of the case with their co-assessor and then a final grade is agreed. This is carried out for each item and also for the overall global grade (Gleeson, 1994, 1997).

• History-taking: The four items assessed include pace and clarity of presentation, communication skills process, systematic approach and establishment of the case facts;

• Physical examination: The three items include systematic approach, examination technique and establishment of the correct physical findings;

• Professionalism: During these activities professional behaviour is assessed. The remaining three items include construction of appropriate investigations in a logical sequence, appropriate management and finally clinical acumen.

Thus the overall ability of the candidate in being able to identify and integrate all aspects of the patient’s problems is considered together with their ability to establish an appropriate diagnosis or differential diagnoses, and discuss these in terms of investigations and overall management plan.

Advantages: The main advantages of the OSLER compared with the long case are that all the candidates are examined for same length of time using the same marking format making it a more objective assessment. The assessors are also present throughout so they can directly assess communication skills and examination fluency. As the assessors also discuss their grades and the level of difficulty of the case with their co-examiner there is a reduction in both the variability between assessors and the variability between cases. Overall this means that that OSLERs are more reliable than the standard long case and are more acceptable to both candidates and assessors.

Disadvantage: The main disadvantage of an OSLER is that in order to achieve good reliability 10 separate cases and 20 examiners are required, thus raising issues of the practicality and cost effectiveness of this type of assessment (Wass et al., 2001).

Objective structured clinical examination (OSCE)

Evolution of the OSCE
In addition to the long case, traditionally candidates were taken by two examiners to be observed in the examination and management of a number of patients with broadly differing conditions. They were then asked to examine a system or area, before presenting a differential diagnosis of their findings, demonstrating abnormal clinical signs or giving a spot diagnosis to their assessors. Students rarely saw the same set of patients and the cases
often differed greatly in their complexity. The assessment was not structured and the assessors were free to ask any questions they chose. Like the long case, there was no attempt to standardise the expected level of performance. This lack of consistency and fairness led to the development and implementation of the OSCE.

An OSCE comprises a circuit of short (usually 5–15 minutes) stations, in which each candidate is examined on a one-to-one basis by one or two assessors. Each station has a different assessor or pair of assessors and candidates move around sequentially to complete all stations on the OSCE circuit in a fixed time (Figure 3.2). At each station the candidate is asked to perform a specific task and the assessors mark the candidate using a structured mark sheet. Each station usually tests a combination of abilities, for example communication skills and clinical method, and can involve real patients, simulated patients or specific equipment, a video recording or interpretation of radiological image.

The basic structure of an OSCE may vary in the time allocated for each station, the use of checklists or global rating scales for scoring and the use of simulated patients or real patients. The underlying fundamental principle of each OSCE is that each candidate completes the same task in the same amount of time and is marked according to a structured schedule.

**Advantages:** The main advantage of an OSCE is that it has greater reliability than other formats of assessment. The main reasons for greater reliability are (Boursicot *et al.*, 2007):

![Figure 3.2 The OSCE circuit.](image-url)
How to assess students and trainees in medicine and health

- By increasing the number of assessors observing each candidate, individual assessor bias is reduced;
- The use of a structured marking schedules allow for more consistent scoring by assessors;
- The use of multiple stations means that many different skills can be tested. The more stations included in an OSCE, the more reliable the assessment is in terms of assessing overall competence.

The increased reliability of the OSCE over other formats and its perceived fairness by candidates have also helped to engender the widespread acceptability of OSCEs among test takers and testing bodies.

**Disadvantages**: OSCEs are very complex to organise; they require meticulous and detailed forward planning, engagement of a considerable numbers of assessors, real patients, simulated patients and administrative and technical staff to prepare and manage the examination. This obviously has an adverse impact on their cost effectiveness.

**Objective structured assessment of technical skills (OSATS)**

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

**Advantages**: OSATS have high face validity and strong construct validity, with significant correlation between surgical performance scores and level of experience. (Bodle et al., 2008)

**Disadvantage**: The data regarding the reliability of OSATS is limited and there is high reported inter-observer reliability (Goff et al., 2005; Fialkow et al., 2007)

**Practicalities to consider when planning an OSCE**

As previously discussed the OSCE is the most widely used format of competence assessment in both undergraduate and postgraduate education. A ‘good’ OSCE is very complicated to organise and it requires careful planning to ensure that the time, effort and cost of staging this assessment has the greatest educational reward. Factors to consider in this planning include:
Purpose of the assessment
At the start of planning an OSCE, one of the most important issues to clarify is the purpose of the assessment. Questions to consider include whether it is going to be formative or summative, when and where it is going to take place, what resources are available and what is going to be tested. It is obviously important to ensure that the content of the stations matches the content of the curriculum and that the stations are designed to reflect the educational objectives that the candidates have been set.

Blueprinting (see also Chapter 6)
An important aspect in achieving good reliability and validity is to ensure that the content of the examination should be based on learning outcomes of the programmes. Moreover, the testing time should be proportional to the relative importance of topic and the sample size sufficient to test all areas. A simple way of ensuring this is to create a blueprint or test specification grid. An example of a blueprint is shown in Table 3.1.

This grid/blueprint provides an easy way of identifying the content of the test to ensure that the content is sufficiently broad to cover the whole curriculum and that the number of items in each cell is in proportion to the time spent in teaching and learning.

Developing OSCE stations
Meaningful OSCE stations are difficult to write and training is necessary for their construction to be reliable and valid. Ideally, the content of every station should be integrated, assessing more than one aspect of consultation or examination skills, for example a station testing communication skills, clinical knowledge and knowledge of ethics and law. All of these aspects should be reflected in the mark sheet.

All OSCE stations should include:
• Instructions to candidate;
• Simulated patient information (if applicable);
• Guidance for assessors;
• A mark sheet;
• An equipment list.
### Table 3.1 A test specification grid

<table>
<thead>
<tr>
<th>Simulated emergency</th>
<th>Short examination</th>
<th>Bedside test or chart</th>
<th>Practical procedure/clinical skills</th>
<th>Drugs</th>
<th>Consultation skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>CVS</td>
<td></td>
<td></td>
<td>SSO: CXR/CT chest</td>
<td></td>
<td>SSO: Prescription errors</td>
</tr>
<tr>
<td>RESP</td>
<td></td>
<td></td>
<td>LSO Station 2b</td>
<td></td>
<td>SSO: Abdo pain</td>
</tr>
<tr>
<td>GI</td>
<td></td>
<td>LSO Station 2c</td>
<td>SSO: Fluid balance</td>
<td></td>
<td>SSO: Urinalysis</td>
</tr>
<tr>
<td>RENAL</td>
<td></td>
<td></td>
<td>SSO: Headache</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENDOCRINE</td>
<td>SSO: DR ABCDE</td>
<td></td>
<td>SSO: Ophthalmology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEURO/OPHTH</td>
<td>SSO: Headache</td>
<td></td>
<td>LSO: Station 3a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSKELETAL</td>
<td>SSO: Joint examination</td>
<td>SSO: Bone scan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ONC/HAEM/ID</td>
<td></td>
<td></td>
<td></td>
<td>SSO: Explaining medication</td>
<td></td>
</tr>
<tr>
<td>SURGERY</td>
<td>LSO station 3c</td>
<td></td>
<td></td>
<td>SSO: Explaining procedure</td>
<td></td>
</tr>
</tbody>
</table>
The instructions to candidates should be as succinct as possible. They should include the role (for undergraduates) and setting of the candidate, the clinical scenario and the specific task that the candidate should complete and the time allotted for the task/activity. An example is shown in Figure 3.3.

Simulated patient information should include the name and occupation of the ‘patient’, the appropriate clinical information and social history, the ideas and expectations of the ‘patient’, a description of their personality, an opening statement and some appropriate questions for the patient to ask. The instructions should include the expected answers that the candidate will provide and instructions on how to address problems if the candidate strays from the task. As it is not possible for OSCE authors to predict all the questions that a candidate will ask, the instructions should also include how to deal with this.

Guidance for assessors should include information of the expected performance of the candidates and include definitions of the performance for each of the categories. An example is shown in Figure 3.4. If the station is based on evidence-based guidance, this should also be included.

The marking schedule should be tailored to the task that the candidate has been asked to perform and be consistent with information provided to candidates. Different types of mark schemes are available and should be chosen after consideration of the context of the assessment, that is who is being assessed and what the purpose of the assessment is (see next section for more details).
Station X: Examiner information

Students are taught how to examine a knee in their orthopaedic and rheumatology firms. They will also probably have seen it/performed it in their GP or Emergency medicine placements in final year. They have been asked to explain what they are doing and tell you their findings as they are completing the examination.

A mark sheet has been provided with detailed information on the criteria expected for classification of the candidate performance.

A good candidate would be expected to perform a well-structured and complete examination whilst demonstrating an appropriate level of communication skills.

A borderline candidate would perform an adequate examination (doing most parts stated in the mark sheet) however this may appear unstructured and less fluent than expected. Their communication skills should still be adequate.

A failing candidate would perform an incomplete and unstructured examination or demonstrate an inadequate level of communication/consultation skills.

If you have serious concerns regarding any aspect of the candidate performance, e.g. patient safety issues, professionalism, during the task assigned in the station, please fill in a Serious Concerns Form, outlining your concerns in detail.

The equipment list should include everything that is needed for the station. If specific props are required these should be included with this information.

After writing an OSCE station, it is important to pilot the station under non-assessment conditions to ensure that the details and tasks required are achievable in the time allocated – if possible, on a group of volunteer candidates. This is not always possible, however, to avoid any adverse incidents with candidates, it is essential to pilot the station with volunteers acting out the roles of the candidate and assessor.

Simulated patient and assessor training

While the main advantage of an OSCE is good reliability, this is attributable to its characteristic circuit structure. That is, every candidate completes the same stations with the same assessors, equipment and where appropriate,
the same patient (real or simulated), so has the same assessment. However, it relies on both the assessors and simulated patients demonstrating a consistent approach in their performance throughout the examination. In an ideal situation, both assessors and simulated patients attend a training session prior to the OSCE, but this is expensive in terms of the fees for employing the simulated patients (who are normally actors), and time for assessors who are generally clinicians with pressures of service delivery in addition to their education role.

To ensure a consistent approach with the simulated patient role, it is essential that individuals included in the Simulated Patient Bank have been given training on the expectations of their role in assessment. For the various patient scenarios to be enacted in any one circuit, obviously the patient characteristics, such as gender and age, must be aligned with the details of the scenario. Likewise, when composing the briefing notes, whilst being tailored to the resources available they need to also be as comprehensive as possible, covering:

- Medical and social history of the scenario;
- Details of how the patient would react to situations;
- Altered physical appearance – where appropriate, for example make-up used for a patient with jaundice.

In advance of a new station being used, the clinical scenario should be peer-reviewed for any errors to be put in order before they are used in a high stakes assessment. In addition to the briefing notes (ideally sent at least a week in advance of their 'performance'), some institutions provide simulated patients with a DVD of how the role is expected to be portrayed, others discuss it with the simulated patients on the day of the OSCE.

Ensuring consistency between assessors is another challenge and is one of the main complaints from candidates sitting a multi-centred OSCE (see also Chapter 9). Encouraging assessors to attend a formal training session is never easy and therefore one approach is to develop online guidance and expectations of the assessors that defines:

- The role of an assessor;
- The purpose of the OSCE;
- Sample station information and mark sheets;
- Video capture of performance examples of good, borderline and failing candidates.

Immediately prior to the OSCE circuit commencing it is useful have a compulsory attendance at an examiner briefing where the features and specific details are emphasised, and any questions answered or clarified. It is also beneficial to include the expected standard of candidates within the guidance to examiners in the station paperwork. Indeed, an example of a consistent
approach when there are multiple OSCE circuits (and thus multiple examiners for any one station) is to have all examiners for a named station guided through the expectations of their duties, with an opportunity for questions and answers as clarification.

**Feedback on performance**
The provision of feedback to candidates, assessors and simulated patients on their individual performance is constant dilemma and there is no ideal solution. The expectations of candidates are that they should be given feedback that is preferably individualised and detailed as an aid to their learning, so that they can improve for their next OSCE. Detailed feedback is particularly significant for borderline and failing students, albeit that it is requested by all candidates and, anecdotally, it is often the good students that are far more eager for their feedback than the less able ones. To provide a narrative on performance is rather time-consuming administratively, as well as monitoring the content to avoid an unfortunate comment from a ‘rogue’ assessor being disclosed. There is also the possibility that feedback could be used to challenge the scores awarded that could result in a rise in the number of appeal applications. As all examinations have limited resources, this is a legitimate burden to be considered in relation to the potential educational benefit of providing this information to candidates.

Although assessors do not currently expect feedback about their performance, it has been shown that providing assessors with their own stringency level (‘hawk’ or ‘dove’ score) reduces the number of outliers on either side of the continuum. This reduces inter-rater reliability and therefore increases the overall reliability of the OSCE (see also Chapter 9).

**‘Serious concerns’ documentation**
In addition to the individual station mark sheet, most OSCE assessments have a separate ‘Serious concerns’ form. This is a system used to report and feedback to the candidate any problems that may be difficult to address in full in the marking schedule, such as professionalism, problems with attitude or unsafe practice. Institutions that have adopted these forms use them in varying ways;

- Some schools have in the regulations a clause whereby if a student has been considered by more than a fixed number of examiners to be a cause for serious concern, then they are placed in the fail category irrespective of the marks awarded for the respective stations;
- Discuss the issues raised with the student and arrange remediation if required. More research on the predictive validity of these forms is required.
Current issues about competency assessment

There are several issues that are currently being deliberated and discussed in relation to competency assessment which are discussed separately. These include:

• OSCE scoring – which type and by whom?
• Which standard setting method is most appropriate?
• How do you define the borderline group?
• Sequential OSCEs;
• Should high fidelity simulation be used?
• Serious concerns cards.

OSCE scoring – which type and by whom?

There are two main issues being debated within the medical and health professional education environment regarding the type of scoring used in an OSCE. These are:

• Checklist rating vs global rating;
• Skills-based scoring vs station-based scoring.

Checklist rating vs global rating. There are a number of approaches to marking/scoring the performance in an OSCE assessment. Traditionally OSCE mark sheets have taken the complex elements of clinical performance and divided them into discrete, measurable tasks which are checked off on a list as to whether they are/or are not completed. However there has been a recent shift to the use of global rating scales which take a more holistic view of performance. This is particularly relevant when assessing more experienced candidates.

Studies using global rating scales in a postgraduate assessment have shown that they have higher inter-station reliability, better construct validity and better concurrent validity than checklists. More importantly, evidence shows that experienced clinicians may be penalised by the use of checklists, which neglect higher level components of clinical competence (Hodges and McIlroy, 2003). Comparisons of the two methods in postgraduate assessments of doctors have also shown that global rating scales are better able to distinguish the novices from the experts (Hodges et al., 1999). It is therefore important to consider the context of the examination before deciding which format is suitable.

Skills-based scoring vs station-based scoring. One of the criticisms of OSCEs is that the candidate can be incompetent in a particular skill, for example interpretation of clinical signs, but still pass the OSCE overall due to compensation from their performance in other stations (Cox, 2000). This has led to some assessments introducing an assessment in which the candidate needs to attain a minimum score in a series of skills as well as a
minimum total score for the assessment. In the MRCP PACES examination the skills are:

- Identifying physical signs;
- Clinical communication;
- Differential diagnosis;
- Clinical judgement;
- Managing patient concerns;
- Maintaining patient welfare.

At present there is no consensus as to which format of scoring should be used; it will again depend on the purpose of the assessment.

**Patient and simulated patient scoring.** Traditionally in the UK OSCEs were marked by clinicians, whereas in the USA and Canada simulated patients are frequently also the assessors for stations (Boulet et al., 2002). There is increasing evidence to suggest that including a simulated patient rating tests more humanistic aspects of an assessment such as empathy, rapport and attentiveness (Kilminster et al., 2007; Nestel et al., 2010). It has been shown that including such ratings in the assessor’s mark sheet highlights the importance to the candidate of a patient approach and also improves the psychometric reliability of the OSCE (Homer and Pell, 2009).

**Which standard setting method is most appropriate?**

In any summative examination it is important to set the pass mark or score that is appropriate and defensible – the process of deriving this mark is called standard setting (see also Chapter 6 for a more detailed discussion). There are two main types of standard setting: relative methods (norm referenced method) and absolute methods (criterion referenced method).

Relative standards (norm referenced method) is where the performance of candidates is compared and an agreed proportion of candidates fail, for example the bottom 20%. Conversely, absolute standards (criterion referenced methods) are based on how much the candidates know and candidates pass or fail depending on whether they meet specified criteria. For the majority of undergraduate and postgraduate assessments, the absolute standard (criterion referenced method) is used as most curricula are based on establishing that the candidate has achieved an appropriate level of competence. The most commonly used methods for standard setting an OSCE are the borderline regression method and the modified Angoff method.

**Borderline regression method:** In this method, in addition to completing the standard OSCE checklist score, assessors are also asked to complete a global score or grade for the station (see Figure 3.5). This overall grade varies, however it is generally made up of a 4 or 5 point global score rating, that is:
**Figure 3.5** Example of OSCE mark sheet for deciding the pass score via the borderline regression method.

- **Station Number:** 2a
- **Station Title:** Cardiovascular
- **Circuit:** &Circuit
- **Session:** &Session
- **Candidate Number:** &CandNo
- **Start Station:** &StartStation

**Patient Description:**

1. **Physical examination**
   - general inspection
   - checks pulses, notes blood pressure JVP, palpates carotids
   - inspects, palpates precordium, localises apex beat, auscultates valve areas with correct positioning
   - examines for peripheral pulses and ankle oedema when applicable

2. **Identification of physical signs**
   - identifies abnormal physical signs correctly
   - lists physical signs as checked by examiner

3. **Interpretation of physical signs differential diagnosis**
   - interprets physical signs sensibly
   - offers reasonable differential diagnosis and is able to defend it

**Overall grade (this will be used for calculating the pass mark only)**

<table>
<thead>
<tr>
<th>CP</th>
<th>P</th>
<th>BL</th>
<th>F</th>
<th>CF</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
</tr>
</tbody>
</table>

**Instructions:**

- Assess the candidate on each of the criteria, using one of the following categories: Clear Pass (CP); Pass (P); Borderline (BL); Fail (F); Clear Fail (CF).
- Please mark a horizontal line in the box to the right of your chosen category.

- **Data and examination:**
  - identifies abnormal physical signs correctly
  - lists physical signs as checked by examiner

If you have awarded a Clear Fail or a Fail for any of the criteria, please use this box for comment, incorporating the applicable anchor statement. If you have any serious concerns about this candidate’s fitness to practice, please complete a separate pink form.
Clear pass;
Pass;
Borderline pass;
Borderline fail;
Clear fail.

In this method, the overall scores of the candidates for a named station that have been graded as borderline pass and borderline fail are calculated and the average score derived to determine the pass mark for the station.

Examiners tend to favour this method of standard setting as it is less time consuming than the Angoff method, and it is quantifiable and based on actual observation rather than on the performance of a hypothetical borderline candidate. However this method is reliant on sufficient numbers of students being placed in the borderline groups for it to achieve reliability. It may therefore not be suitable for assessments with small numbers of candidates.

**Modified Angoff:** The Angoff method was originally designed for written items and has been adapted for OSCEs. It requires a minimum of eight assessor or judges that are ‘experts’ with reference to the content and standard expected for the assessment being scored. Initially the judges discuss the characteristics of a borderline candidate. Then each reviews every station individually along with the mark sheet and determines the percentage of items scored that they consider that ‘just passing’ candidates would achieve. The group then discusses the ‘borderline’ scores for each station in turn, for which there is a spread of numbers, before agreeing their individual scores. The pass mark for the named station is normally the average score of all the judges.

The advantage of this method is that it focuses attention on item content and it is relatively easy to use. However judges often have difficulty defining a ‘just passing’ candidate and often feel that they are ‘selecting numbers at random’. It is also a very time-consuming process (Kaufman et al., 2000).

**Sequential OSCEs**

OSCEs are complex to organise and expensive and as the majority of candidates pass, there have been recommendations to design an assessment that tests all candidates but targets the limited resources to assessing the borderline group (i.e. those just passing and just failing) – this has led to an interest in the use of the sequential OSCE where borderline and failing students identified using the borderline group methodology are given a second OSCE as a test of their level of competence.

**Defining the borderline group:** In any assessment, there is always a group of students that have ‘just failed’. This group are of significant concern to
exam boards and the subject of debate as to whether individual candidates have scored sufficient marks to fall within the reliability of the assessment. There has been no consensus within the education community as to how the problem should be addressed. This is becoming increasingly challenging throughout medical and health professional programmes as the number of student appeals are increasing. It has therefore been suggested that the more rigorous approach of the sequential OSCE should be adopted (Cookson et al., 2011; see also supplementary guidance from the GMC on assessment of undergraduate medical education).

The psychometric analysis of results (discussed in detail in Chapter 8) is employed to ensure fairness to the candidates and for quality assurance that the scores awarded are aligned precisely to their competence. When all candidates have completed the OSCE it is possible to calculate the standard error of measurement (SEm). Where a candidate has scored ‘two SEm’s below the pass mark, it is considered to be 97.5% certain that this is a ‘genuine’ fail. Therefore, for each OSCE, the score that was represented by 2SEms (or below) should be used to define the candidates that are ‘true’ fails from candidates that are a ‘false’ fail.

**Operationalising the sequential OSCE:** In this form of assessment all candidates take an OSCE which has fewer stations (e.g. 10 rather than 15). The borderline group of students is identified by deriving the SEm and calculating the cut score for those who have borderline or fail scores. These candidates (a reduced number) then take a second OSCE – the same number but different stations, and the extra information obtained about these candidates is taken into consideration when making a decision on whether they pass or fail. The advantage of this protocol for assessment is that it requires fewer examiner days as it focuses on the pass/fail candidates. However, it does not help discriminate among the higher achieving candidates.

**Should high fidelity simulation be used in an OSCE?**
Low fidelity simulation, such as venepuncture mannequins, has been used widely in competency assessments to allow individual competences to be measured reliably in an OSCE, such as procedural skills, history taking and physical examination. With high fidelity simulation (see also Chapter 4) there is good evidence of face validity, however the evidence for its reliability and predictive validity is not as convincing, which perhaps suggests that it is less appropriate for high stakes assessments (Hatala et al., 2011). Nonetheless, the potential benefits of high fidelity simulation are in the ability to assess the technical and non-technical skills as well as the team-working competences. However the drawback is the expense, therefore how it is used
needs to be established prior to the adoption of high fidelity simulation in the majority of competence assessments (see also Edler et al., 2009).

**Conclusions**

Assessment of competence is an integral part of both undergraduate and postgraduate medical training. There are many formats for these assessments, with OSCEs being mostly widely used as they have the best reliability. There are several different variants of OSCEs; the format used should be tailored according to the purpose of the OSCE and the resources available.

To ensure that a competence assessment has predictive validity (score awarded for the performance correlates with level of competence), it is necessary to sample multiple discrete areas of the curriculum and test a diverse range of clinical skills, such as communication skills, problem solving ability and professionalism. The integration of several skills within any one OSCE station also helps to increase the face validity of stations. As with all aspects of medical and health professional education, there are new developments emerging for the assessment of competence. However, additional evidence is required before any recommendations for widespread changes can be made.

**SUMMARY**

1. The most widely used competence assessment is an OSCE. This format has the best reliability.
2. Competence assessments should be developed after consideration of their purpose and context.
3. OSCEs are complex to organise and need careful planning to ensure that they are cost effective.
4. OSCE stations should integrate skills and these should be reflected in the marking schedule.
5. Training of assessors and simulated patients is required to ensure reliability.
6. Feedback should be provided to candidates and assessors.
7. The context of the assessment should be considered when deciding on the marking format.
8. The type of standard setting used depends on the number of candidates and resources available.
References


Further reading


General Medical Council (2009). *Tomorrow’s Doctors*. GMC.

General Medical Council (2011). *Assessment in Undergraduate Medical Education. Advice supplementary to Tomorrow’s Doctors GMC Additional Guidance*. GMC.

Chapter 4  Assessment of performance

Dr Gavin Johnson¹ and Allison Wiseman²
¹UCL Medical School
²University of Surrey

Introduction

Assessment of performance is the assessment of actual practice, what a practitioner actually does in clinical practice (Boursicot, 2011). It differs from the assessment of competence which is a demonstration of what the practitioner can do, that is completion of a task to an agreed standard, typically in staged in vitro settings (van der Vleuten, 1996). The assessment of performance is at the pinnacle of Miller’s pyramid (see Figure 4.1), because it is the nearest estimate of genuine practice, and indeed has been coined the ‘authenticity movement’ (van der Vleuten et al., 2000), This shift to evaluate real-life practice has seen two important innovations in medical education: the introduction of workplace-based assessments (WPBA) and simulation, particularly high fidelity approaches which attempt to provide legitimate contextualisation for assessment.

This chapter starts with the overarching issues in assessment of performance which may be considered as the closest level by which assessment can evidence professional practice. Multiple and continual episodes which capture the breadth and complexity of clinical work are needed to assure fitness to practise. The use of log books and portfolios are a mechanism by which many practitioners document this iterative process. Likewise, the use of simulation, in particular high fidelity simulation in the training and assessment of clinical and communication skills, has been included, as well as its utility in supporting the development of healthcare teams. A practical guide on how to plan and deliver a simulated session covers all the necessary steps to make the session run smoothly. The following section covers WPBAs

---

© 2013 John Wiley & Sons, Ltd. Published 2013 by John Wiley & Sons, Ltd.
in common use and then addresses the use of log books and portfolios in clinical education.

**Why assess performance?**

Assessments in ‘authentic’ settings focus on how a practitioner can combine knowledge and skills, judgements and attitudes and deal with the genuine challenges of the clinical setting. Continuing assessment of performance in day-to-day practice enables some skills to be tested that cannot be done authentically by other approaches: using a range of fundamental modalities it permits rigorous probing in areas such as professional behaviour, effective organisation and time management and communication with patients and colleagues in a multi-professional team. One of the real strengths of assessing performance lies in its efficacy as a formative assessment *for* learning rather than an assessment *of* learning. Both WPBA and high fidelity simulations have the capacity to provide feedback and assessment that is integral to, rather than separate from, working practices, involving genuine or extremely realistic interactions with patients and colleagues. They can help to develop higher order skills, like reflective practice and self awareness, and address the nuances of professional development. Accordingly, they offer opportunities for judgements on the application of newly acquired knowledge and skills in a complex and unpredictable environment (Hays *et al.*, 2002) (Figure 4.2).
Simulation

Simulation is the reproduction of a real world encounter or an aspect of it. Simulation can be low-tech or low fidelity, and a good example of this is the role play done by simulated patients in an OSCE. Here simulation is harnessed to test history-taking skills or communication skills. Clinical skills can also be assessed using simulation, mannequins support both training and assessment – such as breast examination and resuscitation skills – and increasingly sophisticated simulations, using computerised software and haptics, are used in medical education. Simulating the clinical setting can also be used to assess multi-disciplinary cooperation, teamwork and leadership. Simulation in general, but in particular high fidelity simulated practice, has been introduced to teach and test in settings which do not compromise patient safety, where errors made and their correction can form part of the educational event (Wiseman & Snell, 2008). This method has been modelled on its successful use in other industries that require high levels of reliability in areas of potentially hazardous and complex activities, such as aeronautics, air-sea rescue and nuclear power industries. Simulated clinical practice is an invaluable preliminary training ground, where scenarios and systems can be adjusted as part of the education process. These methods have proved so effective that ‘high-tech’ simulation training suites, (Trust-based and School-based facilities), designed to resemble clinical areas, are now common throughout the UK for undergraduate and postgraduate training. They
provide the opportunity for learning a wide range of skills which are endorsed by the World Health Organisation (2009). Likewise, the UK National Patients Safety Agency (2010) has advocated the role of simulation as a learning tool following serious incidents. Service users, carers and patients may also be included in sessions to enhance the realism or authenticity of the experience, and the professional context for inter-professional learning.

**High fidelity simulation**

High fidelity simulation uses complex clinical scenarios with computer-controlled mannequins, where vital signs can be adjusted to mimic an acute clinical setting, for example to provide visual (chest movements, papillary responses), auditory (vocal responses according to student interaction, breath heart and bowel sounds) and olfactory (odours and smells associated with medical/social conditions). Its use in teaching and assessment of technical skills, clinical decision-making and team-working is applicable across different levels of clinical experience and specialities, and evidence is emerging that it can improve outcomes related to patient safety (Issenberg et al., 2005). Trainee participation in the scenario may be captured on video for part of a debriefing session that includes self-assessment and immediate feedback on performance from the facilitator and peers.

The use of high fidelity patient simulators may be enhanced by creating a life-like practice setting; although it is advisable that the preliminary sessions are free of the distractions associated with practice (e.g. pagers, telephones, interruptions and other patients/healthcare professionals). As simulation aims to create the atmosphere of the actual workplace, it follows that the same principles of assessment and feedback of WPBA should apply to the assessment of simulated scenarios (Kneebone, 2010). Part of this realism is in the professional context of the learning experience; and while simulated practice might be seen as a gimmick that uses technology, it is a technique which endeavours to approximate experiences in a ‘real world’, and not all activities need to be in a simulator suite. For example, simulation can be undertaken in a low-tech environment when the learning objectives are human factors. For example, training that includes:

- A simulated patient acting out a scenario;
- A group of six to eight students who take turns in practising a communication skill, for example explanation of a clinical procedure;
- A tutor who facilities the skills learning and the feedback on performance from all in the session, that is patient, peers and tutor.

Accordingly, the development of clinical skills confidence and competence is progressive, with these pedagogic activities guiding the participant from novice to expert.
Some of the drawbacks to this technology are the expense of purchasing the equipment, the dedicated areas needed with in-house technical support and the regular maintenance contract with the manufacturer (see also Box 4.1).

**The students in simulated practice**

Part of the education philosophy behind the use of simulation is to develop the attributes of self-awareness and reflective practice, by both the individual and their colleagues. These sorts of behaviours can be applied to a range of training in clinical methods, problem-solving, error recognition and clinical decision-making. For simulation to be worthwhile, the sessions need to be customized to fit within the accepted scope of practice for the programme and training of the respective students groups, and the following areas need to be considered:

---

**Box 4.1 Advantages and disadvantages of clinical simulation**

**Advantages:**

1. Skills training for complex and acute scenarios undertaken in an environment that does not include a real patient.
2. Gives the trainee the opportunity to practise without compromising patient safety with a critique of performance.
3. Skills can be repeated to:
   - address possible gaps or incongruity of performance
   - at different levels of complexity for the students being trained.

**Disadvantages:**

1. The mannequin is not a real patient, so physiological and psychological responses cannot be recreated as effectively.
2. There can be simulation errors that occur owing to technical errors in computation that can impact on the scenario, giving inappropriate results.
3. Unlike an acute event in practice, the simulated environment does not have the dimension of ‘stress’.
4. Training of educators for:
   - preparation of authentic scenarios,
   - facilitation of the simulation sessions.
5. Resource implications of simulation:
   - purchase of the equipment.
   - dedicated space,
   - technical support and maintenance.
• Identify the learning objectives;
• Incorporate them into the scenario;
• Contextualise the scenario and provide guidance;
• Discuss how the exercise is assessed;
• Allocate time precisely
  - to hands-on orientation to the equipment and the technology prior to the scenario, to enhance efficiency of the simulation and reduce any potential performance anxiety levels;
  - pre- and post-scenario briefing, both individually and as a group.
Debriefing is a critical part of participating in simulated patient scenarios and should be led by educators who are competent to facilitate a group and who have a firm grasp of the students’ stage of training and level of understanding.

**Educators and expert practitioners**

Events that are a challenge in the clinical setting can be used to design simulation scenarios. These are often critical events which can involve a systematic breakdown in effective team-working and communication; they often involve more than one professional group and at their most extreme can contribute to patient morbidity and mortality. The fluid nature of modern healthcare teams is often cited as a reason for these sorts of failures, and simulation activities can be used to review such incidents and can include the examination of interprofessional working, thereby promoting a mutual understanding of these issues and power differentials within healthcare teams. The design of these sorts of multi-faceted scenario protocols can be onerous owing to competing priorities that make simulations across disciplines, organisations and cultures challenging (Wiseman and Horton, 2011). For a simulation event to be effective, specific key questions need to be answered for scenarios to be authentic and defensible (see Box 4.2). An agreed proforma and protocol for debriefing should also be written in to enable this vital part of the learning event to be instructive and constructive, to encourage self-reflection on strengths in the performance and opportunities for skills development.

**Planning a simulation episode**

The guiding principles are to identify the context of the learning experience and then the resources for its delivery, which include:
• Simulator fidelity – what can it do?
• Environment, for example clinical/non-clinical/outdoor;
• Additional staff, for example actors;
Box 4.2 Specification for scenario delivery (adapted from Jeffries, 2005)

1. Why develop a scenario?
2. Goals of the learning experience.
3. Assessment components.
5. Time allocated for the learning event:
   a. completing the scenario tasks,
   b. debriefing event.
6. The environment for the session to mimic a clinical setting.
7. Availability of audio-visual equipment and technical support.
8. Its application to the students:
   a. their current level of competence,
   b. their programme of study,
   c. the number of students,
   d. their prior experience of using simulators.
9. Faculty availability to facilitate the session.

- Clothing and special effects, for example make up;
- Health and safety for example, if a simulation exercise is happening outside:
  o Who do you need to notify?
  o What other equipment is required to meet health and safety obligations?
- Clinical documentation (anonymised) from practice;
- Monitoring equipment/medications/therapeutic equipment (oxygen masks/intravenous fluid/cannula);
- Facilitators – role identified and agreed;
- Adjunct faculty – roles.

Forward planning is crucial to identify the number of faculty required per session to teach a given scenario for an agreed number of learners, the impact the scenario might have on the learners, for example anxiety or distress, as well as technical issues such as a potential equipment malfunction. Therefore define roles for:
- A facilitator for the scenario with a script for guiding the scenario and debrief;
- Technical support for the HFPS/ audio-visual equipment;
How to assess students and trainees in medicine and health

• A runner to deal with adverse learner reactions;
• A facilitator to supervise and brief other students viewing the simulation remotely (Riley et al., 2011).

Simulation activities in healthcare education are not without their challenges for it is both an expensive and complex intervention. High quality and developmental scenarios need careful construction, with appropriate faculty development on how to prepare realistic scenarios with attention to detail of clinical and organisation contexts, including knowledge of evidence-based practice. The choice of learning activities, and how to assess, provide feedback and evaluation need to be reliable and valid (see Chapter 7) if the simulation technology is to be exploited for meaningful educational benefits and, ultimately, augment patient safety in clinical practice.

Workplace-based assessments

Previously, assessment of skills relied on in vitro tests, like the OSCE (van der Vleuten, 1996), where the test environment was standardised in order to maximise objectivity and to improve reliability. However these methods do not address the issues of competence in an ‘in vivo’ performance, that is in practice. For example, a doctor may achieve a global rating of ‘outstanding pass’ in an OSCE station that tests breaking bad news, but may be less empathetic when under the pressure of a hectic over-running outpatient clinic. Thus in-service training and assessments allow for multiple observations and decisions about performance in a variety of contexts which mark out day-to-day practice, and is a comprehensive and valid assessment of clinical performance. This section looks at the different instruments used in WPBAs. The list is by no means exhaustive, but represents some of the common assessment tools that are used widely.

Mini clinical evaluation exercise (mini-CEX)

The purpose of the mini-CEX is for an assessor to make a judgement on a ‘snapshot’ performance of a patient encounter in practice. This is generally organised by the trainee who selects the patient and the skill to be tested as an example of their workload. This may include any of the following:

• History-taking;
• Communication skills;
• Physical examination skills;
• Critical judgement;
• Professionalism;
• Organisation and efficiency;
• Overall clinical care.
The criteria for an assessor are that: (i) they are known to be competent in the skill being tested and (ii) they have been trained in mini-CEX observation and feedback. The details of the trainee encounter recorded must include:

- The complexity of the case in relation to their stage of training;
- The focus of the session, for example history-taking skills, the nature of the clinical problem, whether this is a new patient or a follow-up meeting;
- The clinical setting, for example outpatient clinic, ward.

The performance is scored using a Likert scale of 1–9 (where 1–3 = unsatisfactory, 4–6 = satisfactory, 7–9 = superior), with free text for constructive feedback that should be linked to a trainee’s personal development plan, that is areas of good practice and suggestions for improvement. Typically the exercise should take around 15 minutes followed by 5 minutes for feedback. For examples of mini-CEX assessment forms, see http://www.foundationprogramme.nhs.uk (accessed Nov 2012).

How many mini-CEXs are required to ascertain appropriate levels of performance? The answer varies, but a range of between 12–16 events has been estimated to give a reasonable level of reliability (reliability coefficient of 0.8) (Norcini et al., 2003; Wilkinson et al., 2008). Predictably, as the trainee gains more clinical experience, the mini-CEX scores generally increase, for example between the first and second half of the year, which would indicate validity (Davies et al., 2009), that is senior trainees achieved significantly higher scores than more junior contemporaries. There is general agreement that maximising the number of different assessors increases test reliability and validity.

**Directly observed procedural skill (DOPS)**

A DOPS is the observation of the trainee’s performance of a procedure that is judged using pre-defined standards – while the type of skill may vary depending on the speciality training, the same generic principles apply and the performance is rated using a Likert scale (1–6) where:

Score 1–2 Below expectations.
Score 3 Borderline.
Score 4 Met expectations.
Score 5–6 Above expectations.

There are also criteria (see Box 4.3) that form part of the judging of the competence in a given task, these include:

- Category of practitioner eligible to be the assessor, for example appropriately trained nurse, doctor;
- Level of complexity of the skill being tested (low, average, high);
How to assess students and trainees in medicine and health

Number of times the trainee has:
(i) completed the task,
(ii) seen the patient.

Feedback – free text comments on areas of strength and suggestions for development.

Usually fewer observations are required for the DOPS than the mini-CEX for comparable reliability, and a general recommendation is that three different assessors should observe two different procedures (Wilkinson et al., 2008). Obviously some procedures take longer to complete than others which is an additional variable, but median times for the observation and feedback are again 10 and 5 minutes respectively (Davies et al., 2009).

Case-based discussion (CbD)
A case-based discussion is a structured interview exercise that explores the management and professional judgement skills and can be applied to both primary care and hospital-based practice. It is recognised that clinical practice is embedded in a complex and often changing environment. CbD probes the trainee’s performance in a specific patient case on issues such as their application of medical knowledge, clinical reasoning and decision-making skills. The trainee requires guidance in identifying suitable patient cases where there is either a measure of uncertainty and/or potential for controversy in their management. The CbD in this format has been in use since 2005; and has succeeded a similar type of assessment which was used: the

Box 4.3 Questions judged on a Likert scale 1–6
(or U/C = Unable to comment or not observed)
• Demonstrates an understanding of indications, relevant autonomy, the technique of the procedure.
• Obtained informed consent.
• Demonstrates the following:
  o appropriate pre-procedure preparation,
  o situational awareness,
  o technical ability,
  o aseptic technique,
  o seeking help where appropriate,
  o appropriate post-procedure management,
  o communication skills,
  o professionalism and consideration for the patient.
• Overall performance – global judgement.
random case analysis, or problem case analysis. Like the other forms of WPBAs, the formative elements should be used to enable the whole process to becoming a learning exercise.

**The acute care assessment tool (ACAT)**
The acute care assessment tool ([http://www.jrcptb.org.uk](http://www.jrcptb.org.uk): accessed Nov 2012) has been designed to assess clinical practice in an acute setting with immediate feedback, and is aimed at doctors-in-training. The assessor is normally a supervising doctor on an acute medical take who judges the trainee’s performance in the following domains:

- Clinical assessment;
- Investigations and referrals;
- Medical record keeping skills;
- Handover of patient care to another team;
- Management of critically ill patients;
- Organisational skills;
- Team-working;
- Leadership;
- Overall global judgement of performance.

Of the published data on the acute care assessment tool, it appears to have a good educational impact and face validity, achieving its goal of assessing trainee performance in the acute setting. But a word of caution to programme convenors: when incorporating WPBAs into a programme of study, it is advisable to consider the practicalities and areas that might affect the reliability and validity of the test, for example the trainee–assessor relationship (see also Chapters 8 and 9).

**Challenges of WPBA**
Assessment of competence to practice in the workplace is complex for two reasons. First, clinical practice is a dynamic environment involving multiple professional groups and even within singular disciplines assessors’ expertise will vary, and these factors can offer a different perspective on what is the ‘right’ choice in patient management. Therefore the assessment is likely to be influenced by the subjective variables of the assessor. Second, a written or observational evaluation of performance may be a valid approach for assessing application of knowledge and skills, but not a measure of their actual workplace performance. That is to say, as soon as a learner is observed the judgement is based on them ‘showing how’ (a competence), rather than ‘doing for real’ (performance) (Hays et al., 2002). Therefore, multiple assessments, with numerous assessors, in a variety of contexts and covering a diverse range of clinical cases, are necessary to maximise the reliability and
reduce potential sources of error in WPBAs. Similarly for a WPBA to have ‘face validity’, that is the test appears to measure what it sets out to measure, it must be appropriate to the level and skill of the trainee and reflect their current stage in training (see also Chapter 2). A further challenge is therefore to provide appropriate training and engagement so that WPBAs can be quality assured for reliability and validity. This training will need to include how to give constructive and timely feedback, for if a supervisor lacks confidence or aptitude in these areas (Cook et al., 2009), any tension may affect the trainee–supervisor relationship, and ultimately, the quality of the education. Realistically, these assessments require additional time to be performed well and they place additional time pressures on the supervisor/assessor; protected time is a necessary requirement.

**The future of the WPBA**

WPBAs were initially intended to be formative, their educational value arising from increased feedback from senior colleagues and embedding these discussions into clinical practice. WPBA are effective when taken seriously by all parties and put to appropriate use. A wide-ranging portfolio which encompasses the variety of WPBA (MSF, mini-CEX, DOPS, CbD) can be enabling for trainees and help them to achieve their educational objectives (Ryland et al., 2006). The UK Medical Schools are increasingly using adaptations of the WPBA in formative assessments and as part of the portfolio/log book systems used in clinical placements. However, ‘enthusiasts’ have championed their use as summative assessments and this has placed WPBA in a critical light. How can these tools developed for formative purposes be used in summative processes? How do the issues about inherent subjectivity and inconsistency sit with decisions about a trainee’s progression, particularly if it is contested? The GMC uses a range of ‘in-vivo’ performance assessments (case note review, third party interviews of colleagues, case-based discussions) in its Fitness to Practise (FtP) proceedings. Each assessment is not attributed a pass–fail criteria but is used by expert peer reviewers to shape a robust and legally defensible decision about capability. However the typical number of observations of practice is in the region of 1200 different assessments. What programme directors and supervisors can take from the GMC lead is that these reports need to be based on recorded observations that have been shared with the trainee in a feedback session and documented.

Even the current amount of WPBAs represent extra time for trainees and supervisors; some trainees are sceptical, viewing WPBAs as intrusive on training time (Pereira and Dean 2009; Johnson et al., 2009) and regarding the process as a ‘tick box’ exercise. This negative view is compounded when
the assessment forms are completed *en masse* in one meeting and the supervisor gives an overall impression of the trainee rather than linking comments to individual performances. Long delays in providing feedback also invalidate the assessment. The Academy of Medical Royal Colleges of the UK (2009) summarises the feeling of the medical profession from the findings of the evaluative studies of WPBA:

The profession is rightly suspicious of the use of reductive ‘tick-box’ approaches to assess the complexities of professional behaviour, and widespread confusion exists regarding the standards, methods and goals of individual assessment methods. This has resulted in widespread cynicism about WBA within the profession, which is now increasing.

So what can be done to improve the situation? Decreasing the assessment burden is another possible way forward. WPBAs have an unfavourable psychometric profile which is why so many of them need to be done to have an acceptable reliability (e.g. the number of mini-CEX ranging from 7 to 30). If each WBPA was more robust, then fewer assessments would be needed for a reliable judgement of a trainee’s performance. One way forward is to amend the standard rating scales used and align them with criteria descriptors for each levels of training – they would then be more realistic and reflect the developing clinical sophistication and independence of the trainee (a construct-aligned scale; Crossley *et al*., 2011). In a study that trialled these new formats, the number of assessors required to achieve a generalisability coefficient of 0.70 was reduced, with a concomitant decrease in the assessment burden for educational supervisors (see Box 4.4) and improved confidence in the trainees. Equally, changes to the format of questions in MSF questionnaires to make them one item–one question also improves reliability, that is it lowers ambiguity and increases the reliability (MacKillop *et al*., 2011).

**Log books for undergraduate training**

In clinical placements students may feel confounded by the move from university-based to workplace-based learning. Unfamiliar environments with their own working practices, as well as tutors whose dominant concern is patient care, can be a disorienting experience. Therefore log books can be a practical system for providing tangible advice to students on how to maximise their learning opportunities in the workplace. By defining the objectives they give some structure to the apprenticeship model of learning, particularly as clinical placements are often spread across large geographical areas, with
How to assess students and trainees in medicine and health

How to assess students and trainees in medicine and health

multiple tutors who have varying levels of engagement, this means log books can help support an equivalent clinical experience for all. An advantage of log books is that they are documentary evidence and, if scrutinized regularly, on-going experiential learning can be monitored and gaps can be identified and remedied. Again this process aims to encourage the continuing interface between practice-based mentors and students and to be successful it requires active engagement by both parties. Students need to be pro-active in completing their log books and Schools need to articulate the minimum expectations for learning assessment, feedback and remediation via the service level agreements with the placement education providers.

Personal development portfolios

Within the last decade, personal development plans and portfolios have become an integral part of professional education and training and a mechanism to structure and encourage lifelong learning. The professional regulators require professionals to demonstrate their professional learning through maintaining portfolio (see also Chapter 7) and beginning this in the undergraduate years is sound preparation for professional life. Increasingly, on-line portfolios are used and while students enjoy this format, their use in practice can cause problems owing to limited access to computers, time constraints in a busy clinical setting and the negative attitudes of some staff owing to less refined IT skills. Time allocation has always been an issue in clinical training, and the requirements of the European Working Time Directive have meant reduced working hours that has had led to a concomitant decrease in actual time spent in ‘apprenticeship’ training. Thus portfolios provide a structure and are a repository for:

• Documentary evidence, for example certificate of competence in intermediate life support.
• Signing off agreed skills and competences in designated phases of a curriculum in undergraduate/postgraduate training.

Box 4.4 Criteria-guided WPBA led to a reduction in the number required to increase assessment reliability

- 6 to 3 for the mini-CEX,
- 8 to 3 for the CbD,
- 10 to 9 for an ‘on-take’ ACAT,
- 30 to 12 for a ‘post-take’ ACAT.

(Crossley et al. 2011)
• Evidence of progression in development of complex skills acquisition in postgraduate training.
• Reflections on practice.

The portfolio, as a catalogue of reflective assignments on evidence and experiences in practice, is valuable for developing an insight into personal learning needs. However, the skill of reflective practice is one that some students find challenging, but is nonetheless necessary as preparation for their postgraduate training. A significant association has been found between low mean scores in workplace-based assessments and training difficulties (Mitchell et al., 2011). Therefore, portfolios should be used as a focus for progress meetings with an education supervisor, and for remediation in specific areas where a trainee is struggling. Even so, meetings are also a positive event for recognising excellence and practice development.

**Good practice in performance assessment**

This could comprise a consensus statement on performance assessment (Box 4.5) with recommendations based on best practice from the available literature; where there is little evidence then it is based on expert opinion.

**Conclusions**

• The rationale for formative assessment of performance in the workplace is to determine how the trainee performs in the genuine clinical setting, and with immediate and constructive feedback it can become a meaningful learning experience;

---

**Box 4.5 Consensus statement on performance assessment (Boursicot et al., 2011)**

(1) WPBAs assess the two behavioural levels at the top of Miller’s model of clinical competence, shows how and does, because the workplace provides a naturalistic setting for assessing performance over time.

(2) WPBAs can cover not only clinical skills but also aspects of skills such as professionalism, decision making and timekeeping.

(3) WPBAs, specifically the mini-CEX and multi-source feedback, can generate sufficiently reliable and valid results for formative purposes.

(4) WPBAs occur over multiple occasions and lead to documented evidence of developing clinical competence.

(5) WPBAs should provide the timely feedback that is essential to learning and enhancing its acceptability amongst users.
The distinct challenges for WPBAs are that any performance is a complex series of discrete tasks that are difficult to attribute to an individual practitioner, and a change in relationship dynamics emerges when supervisors become assessors, who in turn require training;

- The utility of an assessment must be considered in terms of its reliability, validity, practicality and educational impact, and when one of these aspects is given more importance, it is to the detriment of another;
- There is a contradictory evidence base for their feasibility and educational impact in practice owing to pressures of service that might lead to widespread misuse of the WPBA;
- There are limitations with the existing tools used in WPBA, not least their reliability, hence multiple assessments are needed for a reliable judgement of competence;
- The reliability of instruments can be improved markedly:
  - using rating scales that are based on the degree of independence and seniority of the trainee;
  - quality of questionnaire formats to ensure they contain non-compound straightforward items;
- In postgraduate training, performance assessment may be returning to the concept of a supervisors report. However this report should be based on the evidence of a doctor’s performance, and WPBAs may continue to provide opportunities for the observation and recording of performance.

References

Assessment of performance


Introduction

Professionalism has become a central concept in medical education, moving from the tacit to the stated; driven by challenges to, and changes in, the social contract of the medical profession with society. As ‘professionalism’ and ‘professional development’ are increasingly made explicit in healthcare education programmes, there is growing pressure to develop robust assessment processes and assessment tools that measure such attainment. However, professionalism is a complex and multi-dimensional construct and before we can attempt to assess learning or attainment in the domain of professionalism we need to answer some fundamental questions such as what is professionalism? And can it be taught and learnt?

There is no one definition of what constitutes professionalism in the health professions. This is important for educators because the discourse of professionalism, that is the way in which professionalism is understood and described, that is adopted by the profession or learning organisation will greatly affect what it believes should be assessed, in what way and for what purpose.

This chapter begins therefore with a brief introductory section on what professionalism might be, how it might be taught and how attainment of professionalism might be measured. It identifies the main discourses concerning professionalism, noting the link between what professionalism is conceived to be, and how it might be assessed. It then moves on to explore self, peer and faculty assessment of a range of competencies, attitudes and behaviours that can be described as ‘professionalism’ (see Chapter 7 on...
Feedback) highlighting that many of our traditional assessment tools that measure behaviours, skills and knowledge about what should be done, are only proxy measures for professionalism.

Building on the recommendations from the 2010 Ottawa Conference Assessment of Professionalism expert group, this chapter goes on to identify how individual, interpersonal and societal dimensions of professionalism assessment might be approached.

What is professionalism?

*If you wish to converse with me, define your terms.* – Voltaire

We all think we know what professionalism is without needing to explain it in great detail. We certainly think we know when someone is being unprofessional. However once we start to dissect this very tacit understanding in an attempt to identify *exactly* what professionalism is, and how we might teach and assess it, we start to run into problems with definitions. The literature, policy documents and curricula recommendations use inter-related but non-synonymous terms in descriptions, definitions and debates: they identify attitudes, values, attributes, behaviors, etc. Behind this inconsistent use of language, what remains unexamined is whether the profession, and the educators within the profession, conceptualise professionalism as a state or a trait. Is it:

- A set of behaviours based on a cognitive knowledge base?
- A collection of values and a process of identity formation?
- An approach to practice?
- Personal, interpersonal or societal?
- All these things?

Definitions of professionalism abound. Indeed some authors have suggested we cannot make any meaningful progress in learning, teaching and assessment without a clear definition (Swick, 2000). As these definitions have been generated by learned groups, researchers and policy makers there has been some commonality but also some significant differences. Some definitions hark back to an earlier age and a rather nostalgic view of the professions, some suggest there is nothing much teachers (and thus learners) can do as it is a way of being or a virtue, some focus on the interpersonal collaborative nature of modern professionalism, others focus on behaviour to be adopted to avoid getting into trouble either with the regulator or with your employer. So which is the most suitable definition to adopt? Sometimes, as educators, the definition is provided for us: many of the Royal Colleges and some licensing organisations have adopted their own definitions of professionalism. In
other circumstances educators are expected to create definitions for themselves within broad guidance or frameworks. Whatever definition is chosen or adopted by programme designers or learning organisations, that definition immediately begins to foreground some aspects of professionalism over others, thus influencing what is valued, what is learnt and ultimately, what is assessed.

It is worth identifying here some of the dominant discourses, or sets of meanings and understandings of professionalism, that are current in education for the health professions today. These can be categorised as:

- Professionalism as a set of learnable behaviours and orientations;
- Professionalism as a set of virtues: a way of being;
- Professionalism as compliance with the regulator (acting on behalf of society);
- Professionalism as being a good employee;
- Professionalism as an approach to work or practice.

It is important to identify what sort of discourse pervades in your own learning organisation because how professionalism is conceptualised will have an influence on pedagogical approaches (Mausley and Strivens, 2000) including the approach to assessment. It is also worth reflecting on one's own definition or understanding which may also be having an effect on what you feel it is important to learn and to assess. These definitions will affect: what you do in your own practice; what you expect others to do; what you consider is important for you as an educator to do to help others to develop it; and ultimately how you assess its presence or absence.

How do you teach or learn professionalism?

The definition of professionalism adopted will have an influence on a range of pedagogical approaches, including how we think professionalism develops and what sorts of education interventions might be used to aid this development. In addition, theories of learning also influence our approach to teaching in this complex domain. The competency culture that has pervaded healthcare education and assessment for many years tends to encourage a conceptualisation of learning as ‘acquisition’: an individual-centred, transmission activity. More contemporary ways of thinking about learning favour a ‘learning as transformation’ model where learning is a more social practice (for example Lave and Wenger, 1991). This model is compelling in education for the health professions, particularly workplace-based learning (see also Chapter 4), and may be an important element in understanding how professionalism develops.
There are some elements of professionalism that can be ‘acquired’ and therefore can be taught in a relatively didactic or classroom-based way, such as elements of the law, the structure and process of a patient-centred interview, what the GMC guidance is or how to complete a personal development plan. The majority of professional behaviours however tend to be learned through techniques such as immersion in practice, opportunities for legitimate participation and developing reflective practice. Because much of learning how to be professional happens outside a formal classroom setting, and is co-created with other learners, with patients or with other professionals, this creates particular challenges in how teachers might go about assessing whether such professional development has occurred.

**Approaches to assessing professionalism**

There are a range of approaches open to the educator concerned with assessing professionalism or professional development, and it is likely that a mixed approach, using a range of methods, is most likely to capture the many facets of professionalism. From what has been outlined so far, it is clear that those methods chosen will need to look at a range of knowledge, skills, behaviours and attitudes, and explore personal, interpersonal and societal dimensions of professionalism to truly assess this multi-faceted attribute.

A systematic review of approaches adopted in medical education by Wilkinson et al. (2006) revealed a range of approaches using a combination of some or all of the following: faculty or practitioner conducted observed activities; peer impressions such as multi-source feedback; user-focused assessment such as patients’ opinions; written tests and scenarios; simulation; proxy measures, such as review of research or teaching activities and self-assessments. The authors did suggest that although these are the common approaches, they might simply be commonly used because they represent, to some extent, the easier elements of professionalism to assess: they suggested other attributes still needed good quality assessment approaches such as measures of reflectiveness and approaches to personal development and practice – such as lifelong learning, seeking and responding to results of audits, dealing with uncertainty, balancing availability to others with care for oneself, and patient advocacy.

Assessments of professionalism, like assessments of other domains, can be considered within the framework of Miller’s Pyramid (Miller 1990). This is useful when designing or reviewing assessment of professionalism to ensure there is sufficient use of the most professionally authentic test methods. It is important not to be falsely reassured that tests of ‘knows’, ‘knows how’ or even ‘shows how’ truly represent the professionalism of the individual.
As with all assessment, a balance needs to be struck between the reliability, validity and feasibility of an assessment package. Some assessments will show greater reliability than others (important for high stakes assessments such as finals) others will show greater validity on one or more of the domains of professionalism. Many good assessments of professionalism create significant problems of feasibility: particularly those that require the cooperation of a significant number of assessors.

Assessment of professionalism should also utilise both formative and summative assessments. Professional development is a lifelong journey and it is difficult to see any assessment as truly summative: that is, that the individual has achieved ‘professionalism’. Instead, assessment in this domain should be seen as a continuous process with ‘check points’ along the way to ensure the individual is fit for their current level of practice. As professional development is a lifelong process, educators should also see their assessment as part of a continuum; paying attention to what has happened before and what will happen later in the learning and assessment of professionalism.

In the final part of this chapter, a range of assessment methods useful in the assessment of professionalism will be described, outlining their relative strengths and limitations. This is not to suggest a single approach will effectively and meaningfully assess professionalism: instead it will allow educators...
to choose a range of methods that test the domains of professionalism that their learning organisation feels are important and that are feasible in their own setting.

**Faculty/teacher–practitioner assessment: observation *in vitro* (simulation)**

‘In vitro’ testing includes a range of methods that test ‘shows how’. They are simulations of real life situations; some more authentic than others. Whilst they are relatively feasible and are usually safe to conduct, it must be remembered that these tests do not tell us about the professionalism of the individual; rather that they know what they *should* do and *are willing to do this when they are observed*. They give us limited insight into what the individual might really do in everyday practice.
Objective structured clinical examinations (OSCEs)

OSCE stations might include applying principles of the law and professional guidelines such as consent or communication with a third party. They may test communication with team members, often over mistakes or differences of opinion. They might present dilemmas that test probity or ‘doing the right thing’ in situations, such as error or patient safety issues, or they might address ethical principles, such as do not resuscitate orders or advanced directives.

Whilst coming up with a good scenario may be challenging, what is more challenging is coming up with the ‘right’ answers and responses. This brings us back to the point made at the beginning of the chapter: that we all have our own internalised idea of what is professional and what isn’t. It may be that the organisation that determines the curriculum being tested (e.g. the GMC, the Nursing and Midwifery Council) has defined objectives related to professionalism and has given clear indications of what is judged as professional in their understanding, but this is rarely the case. Therefore a discussion needs to take place amongst examiners and examination designers to identify the range of right answers, responses and behaviours that are acceptable: a consensus will never be reached. This sort of station is the most likely to show poor inter-rater reliability as examiners let their own judgement and internal compass override the mark sheet and guidance. This means very clear mark sheets and guidance, good examiner selection and robust examiner training with the scenario and marking is required.

Workplace-based assessments: Observations

Most workplace based assessments (WPBAs) contain some attempt to assess ‘professionalism’. Chapter 4 addresses WPBAs in more detail and so this ‘professionalism’ element only is addressed here. The clarity of what is meant by professionalism, or indeed lack of professionalism, is rarely spelt out in guidance on conducting WPBAs, very much leaving the judgement up to the assessor. This may seem acceptable for WPBA; who is a better judge after all of professionalism, or lack of professionalism, than a fellow practitioner? However, the consequences of getting a low score in the professionalism domain may be terribly upsetting and destabilising for the developing professional and lead to problematic remediation issues for the education supervisor, therefore good examiner training is required for those WPBAs conducted by individuals and clear written guidance will be required for WPBAs that involve multiple assessors.

Of the currently used observation-based WPBAs, the case-based discussion is particularly useful at assessing, to some degree, elements of practice that can be called professionalism. Using actual practice, and reflection on
and discussion about practice as a stimulus, is often a good test of many of the areas considered as professionalism, such as patient centredness, teamwork, ethical decision making etc. It is important for the examiner not to gloss over these aspects of the case: this may be a rare opportunity for supervisor and learner to talk about issues of professionalism.

Caution should be applied if the learner is able to choose their own cases for this type of assessment; they may only choose those that show them in a good light. Perhaps a mixture of cases chosen by the learner and cases chosen by the examiner is a good compromise.

**Faculty/teacher–practitioner assessment: Observation ‘in vivo’**

‘In vivo’ assessment of professionalism is probably the most authentic test of this domain of practice. *In vivo* tests aim to reach the ‘does’ level of assessment; examining what the learner does when they are not being directly observed; when they may be at their very best (because the stress of an assessment and an assessor is not there) or their very worst (when they think they can behave as they really want). Whilst these assessments are very valid their feasibility is often more problematic. They tend to be formative assessments of professionalism, but can have a role in a summative way, to make judgements.

**Assessment as part of practice**

This includes everyday assessment of everyday practice. It tends to be formative in nature and carried out by a clinical supervisor. The fact that the supervisor is present does tend to alter practice somewhat, although the more often the supervisor is present the more learners let down their guard and the more authentic the practice becomes. Each act of feedback however will reinforce this is ‘observed’ practice and so what is seen may not always represent typical practice.

Assessing professionalism in this way is particularly useful when it appears that professional behaviours and approaches are not the dominant aspect of activity taking place and the learner is not aware that it is professionalism that is being considered. For example: looking at professionalism when a trainee surgeon is performing a surgical procedure, when observing a team during a resuscitation attempt, when observing the delivery of a baby of a drug user, or when observing communication at handover. From experience, underlying and often ‘hidden’ attitudes tend to emerge in stressful situations: a learner may be very respectful of their colleagues in routine tasks but may reveal a lack of respect or understanding of roles in high pressure situations.
Critical incidents/significant incidents
Like case-based discussions, critical incidents and significant events capture moments of real practice. The discussion of, and reflection on, a significant event or a ‘near miss’ can uncover both individual and team breaches of best care, including aspects that can be considered professionalism, and can reveal system problems that create situations for individuals and teams where their enforced actions are not their preferred way of acting in what they conceive to be the most professional way. As a formative tool, when handled sensitively they can become rich sources of professional learning and socialisation into the norms of the profession and what is best practice. However the discussion of a significant event is often emotionally charged and can rapidly deteriorate into a blame scenario, losing its rich potential for learning and formative assessment.

Undeclared observation
Restaurants use undercover critics as diners and shops use ‘secret shoppers’ to investigate actual practice. This is a technique that is occasionally used in health professions assessment but has the potential for more extensive use in the future.

A review by Madden et al. (2009) of the use of ‘simulated client methods’ across 23 countries for over two decades suggested that these methods provide unique insights into the behaviour of practitioners. Under the Obama administration in the USA a $350,000 initiative is being planned for 2012 with the aim of uncovering fraud and discrimination against uninsured patients; and in the UK a recent hearing concerning a member of the British Association of Counselling and Psychotherapy who was attempting to ‘cure homosexuality’ involved the evidence of an undercover patient. A number of recent television documentaries have used undercover patients and undercover care givers to expose serious breaches in care and professionalism. Such assessment of professionalism may be fraught with ethical and practical issues but, with the introduction of revalidation and the increasing prominence of fitness to practise issues, it is likely to become an additional tool available for the assessment of professionalism.

Faculty/teacher–practitioner assessment:
Paper-based tests
Multiple-choice questions (MCQs)
MCQs are good tests of knowledge and application of that knowledge. They can test knowledge about expected professional behaviours, the law, professional guidance etc. They are of course not a test of what individuals do.
Assessment of professionalism

Situational judgement-style written tests (SJTs) extend beyond knowledge towards judgements, and may give better insights into thinking and typical behaviours. Like all MCQs and SJTs, producing well written paper-based tests of professionalism is a lot harder than it looks. Furthermore, agreeing what is the ‘best’ answer in this domain is even more challenging.

**Short answer questions and essays**

Although unreliable as a summative assessment tool, extended writing may be a more valid test of some aspects of professionalism, such as ethical dilemmas and applying the law and professional guidance. Again the assessment is of what people know is the right thing to do and think and what they say they would do, rather than any measure of reality in practice. Again, what constitutes the ‘right’ answer is problematic.

**Written reflections**

Reflecting on your own actions and inactions and those of the people around you is how adults learn. Reflective practice encourages learners to develop these skills further by undertaking activities of more critical reflection; by ‘stepping back’ from events and actions and exploring personal roles or interpretations. Reflective practice has become an accepted organising framework for professional preparation across the professions (Boud and Walker, 1998) not just healthcare. The term ‘reflection’, however, refers to a complex array of cognitively and philosophically distinct methods and attitudes (Van Manen, 1995) and as such it is hard to assess reflection and its relationship to underlying professionalism.

Reflective practice tends to take the form of either written reflections, such as essays or journals, or discussion groups. It tends to focus on everyday events where the individual can become still more mindful, or on significant events. When done well, with clear understanding of the purpose and supportive supervision and feedback, reflections have the potential to develop insight and create professional growth. They are probably more useful as a learning tool than as an assessment tool. Indeed the literature suggests that once reflections become open to scrutiny or assessment they become more superficial or formulaic; with learners writing what they think will score them good marks rather than their actual reflections.

**Scenarios**

Working through written scenarios provides a little more scope than MCQs or SJTs to assess professionalism as the length and subtleties of the scenario can be expanded to some extent and the options for potential responses can be unprompted. Like MCQs, SJTs and essays, producing well
written scenarios takes some considerable work: often the scenarios become too lengthy or too simplified or too formulaic. Using scenarios in virtual learning environments and other e-packages, where the response to one aspect of a scenario creates the next step or event (for example a decision to take a gift from a patient leads to a coercive request at the next consultation, or refusal of the gift leads to the patient becoming upset or angry), may enhance the use of scenarios as tests of professionalism.

**Service user assessment**

If we understand professionalism as concerning our social contract with society then assessment by service users, patients and clients, becomes the most important focus of the assessment of an individual’s or team’s professionalism. This can be done formally or informally. In medicine, it currently forms one of the key elements of appraisal and re-validation. There are often concerns raised about the reliability of service user feedback with claims that it amounts to a ‘popularity’ contest, or that service users don’t take a wider perspective when assessing their experiences. Having said that, most doctors would be reluctant to use a doctor or service themselves if user assessment was consistently poor.

**Patient satisfaction surveys**

Formally organised patient satisfaction surveys can be useful guides to the qualities and abilities of healthcare professionals and services. By exploring the ‘softer’ less tangible areas such as communication, empathy and shared decision making, they can make a meaningful contribution to the assessment of professionalism. These domains are of central importance to patients and are known to affect hard outcome measures, such as adherence to treatment plans. Because they matter to patients they are usually very willing to make comments as long as their anonymity is assured and they do not believe offering such an evaluation will adversely affect their own care. At times the patient experience is at odds with the way the individual intended the behaviour to be experienced, or indeed how they self-report about their actions and the perceived outcomes. Formally organised surveys can gather detailed information about a significant number of domains from a large number of patients. Results can usually be presented against norms for other practitioners in similar roles. A few cautions need to be applied. Patients may need some advice about what is useful feedback. Small numbers or surveys done on particular days may not reflect the breadth of opinion or the range of that individual’s practice. Sometimes the issue is with the system and practices of the organisation rather than the individual doctor. For example a
negative score on a doctor making themselves available to patients may reflect the set up of the service rather than the attitude of the doctor to availability. For these reasons results need to be reviewed in context.

**Ad hoc patient feedback**
This can also be a useful insight into an individual’s professionalism, whether it is thanks or criticism. It may range from a few words said to another practitioner to a thank you card or letter of complaint. Of course this represents one person’s experience or opinion but it can be useful in putting together a picture of professionalism when used with other measures.

**Peer assessment**
Our fellow professionals represent good judges of our professionalism. They have to work alongside the individual and deal with the consequences of their actions or inactions. If they are our colleagues in a different role, they can be affected by the consequences of our actions and inactions. If they are from the same professional group and are of similar seniority they also have a very clear understanding of the professional norms and expectations. Forms of multiple-peer assessment, such as the 360° assessment tools used in both workplace-based assessment of doctors in training and in practice and increasingly in undergraduate programmes (Cushing et al., 2011), represent the best balance of validity (if they ask the right questions), reliability (if they use enough assessors) and feasibility (if they do not make excessive demands of the assessors in terms of time to take part of complexity of the process). Like service user assessments, peers will often pick up strengths and weaknesses that we are ‘blind’ to, and are often brutally honest. Unlike service users it is often practicable to ask the peer to suggest how to improve, so the multiple peer review can be a powerful formative assessment tool. The reliability of using multiple raters is reflected in the central role peer review often takes in appraisal, progression, re-validation and fitness to practise issues. When well designed and conducted, peer review can explore complex behaviour and attitudes, describing effect rather than assuming intention, giving concrete examples, comparing to others and expected norms and suggesting ways of improving. It is best used in conjunction with self-assessment. Ideally it should be reviewed and discussed with an experienced facilitator or education supervisor to ensure it is constructive and leads to a plan of action for improvement.

**Self-assessment**
Self-assessment, if conducted in an environment of support, where insight and willingness to improve is valued over excellence, can be an excellent
assessment of professionalism. It can include the requirement to compare to peers and when used in combination with multi-source peer feedback it can be particularly powerful. Concern is often raised that individuals will over-rate their own performance. Whilst this might be the case in a minority who have poor insight into practice or who are dishonest, the majority under-rate their performance. As aspects of professionalism involve the internal – feelings, perceptions, attitudes etc. – it is only the owner of these that can truly assess them.

**Proxy measures of professionalism**

Proxy measures of professionalism are often used, by employers, admissions panels, medical schools, learned bodies and professional organisations. These can include: the seemingly behavioural, such as timekeeping, absenteeism, getting into trouble with the law enforcement agencies, involvement in voluntary activities; the practical, such as involvement in research or teaching; or the perceived attitudinal, such as undertaking continued professional development or taking part in patient participation groups.

In the contemporary world proxy measures of professionalism also include how professionals behave in relation to the social media and popular press. The use of proxy markers is widespread but is fraught with problems. Even simply trying to label these issues as ‘practical’, ‘attitudinal’ or ‘behavioural’, creates a rather simplistic view of professionalism. For example, poor time keeping may reflect a chaotic personal life, an unreliable transportation system or an attitude that places self-interest above the interests of others. Furthermore, the degree and frequency of the breach or attribute needs to be taken into account: does being rude once or getting one dangerous driving caution represent an inherently unprofessional individual? Does undertaking lots of CPD make you professional? It seems likely that proxy markers have their place in supporting other assessment evidence, particularly if they identify and persistent pattern of (un) professional behaviours.

**Looking towards the future: The recommendations of the Ottawa 2010 expert group**

The Ottawa Conference expert group represented a bringing together of researchers, writers, policy makers and practitioners from across the globe who concern themselves with professionalism and the assessment of professionalism. The recommendations of this group stress the importance of exploring the pervading discourses of professionalism and how these shape our understanding of what professionalism is, attempts at facilitating learn-
ing in professionalism and how it is assessed (Hodges et al., 2011). The expert group recommended a multi-dimensional, multi-paradigmatic approach to assessing professionalism and to ensuring that assessment of professionalism was not simply extended to the attributes of the individual but also to how that individual practised: incorporating inter-personal and societal–institutional dimensions into assessment practices (see Box 5.1). The authors argue that the assessment of professionalism may need to start at a diagnostic level, starting before selection onto undergraduate health practitioner courses, but warned that evidence was lacking in the predictability of behavioural traits for later professionalism or unprofessionalism. It needs to be used in early undergraduate training for its formative purposes and needs to extend across testing for both compliance behaviour and knowledge and towards the development of attitudes consistent with professionalism. It should use multiple tools on multiple occasions and with many assessors. It should include multiple sources of feedback including self, peers and service users. Extending into the postgraduate years, assessment of professionalism should continue to attempt to build up a complex picture of an individual’s

**Box 5.1 Recommendations from the Ottawa 2010 Conference expert group on the assessment of professionalism**

- A professional’s behaviour is a complex inter-relational phenomena and individual attributes of knowledge and values made manifest in particular, specific, contexts and tests are only an approximation; a proxy measure of authentic practice;
- Some aspects of professionalism may be related to an individual’s personality and therefore give credence to certain types of admissions testing, although research correlating these tests to future performance is still needed;
- The focus for the assessment of professionalism should be to build a portfolio of tools which contribute to a multi-faceted and robust means to assess this complex area;
- The sum of the parts is greater than the whole, that is the overall impression from a range of assessments is more important than just one measure and these assessments should be deployed in formative assessment and developing professional practice;
- The use of behavioural assessments of professionalism should be reliable and valid;
- Documenting negative professional behaviours should be considered;
- How different assessments/assessors/contexts of assessments support, or refute, each other is worthy of future research.
strengths and weakness, a body of evidence, both on a personal and interpersonal level, and extending to societal and institution-facing aspects of professionalism.

Clarity was recommended: being clear to both the professional and the assessor where assessment is formative or summative and where it is supportive or regulatory. It was also suggested that rather than continue to develop new measures we should focus on evaluating and improving existing tools.

The assessment of professionalism as an individual phenomenon

1. Some component of professionalism may be related to inherent personality characteristics or traits. Assessment of traits (cognitive, personality, behavioural) prior to admissions may be relevant to later professionalism; however, use of such screening approaches requires that links between pre-admissions data, medical school performance, residency performance and professionalism-in-practice be demonstrated.

2. Professionalism may be understood as the external behavioural manifestations of the interaction of a complex set of cognitive and attitudinal elements and personality characteristics, mutually and with the environment. However behavioural assessments are proxy measures, resting on the assumption that observed behaviours are reflective of underlying dimensions. Research shows that this assumption is not always accurate. For this reason, documenting behaviours alone may be insufficient to capture a comprehensive construct of professionalism, which should also include knowledge, values, attitudes and the ability to employ professional behaviours in real practice settings.

3. Where behavioural assessments are used, instruments should be employed that have demonstrable reliability and can be used to support valid inferences. Both quantitative measures (e.g. numeric scores derived from observation-based survey instruments) and qualitative measures (e.g. narrative data from Dean’s letters) have been studied and may be employed in a defensible manner. A combination of methods over a period of time is likely to be needed.

4. Given the number of existing professionalism assessment tools, it may be more important to increase the depth and quality of the reliability and validity of a programme’s existing measures in various contexts than to continue developing new measures for single contexts.

5. Triangulation of multiple kinds of measures, by multiple observers, synthesised over time with data gathered in multiple, complex and challenging contexts is likely to be appropriate at all levels of analysis.
6. Identification and documentation of negative behaviours is likely to require a distinct system from one in which there is recognition, documentation and reinforcement of positive professionalism behaviours. Instrument design and validity research should be undertaken thoughtfully in such a way as to reflect this distinction.

7. The overall assessment programme is more important than the individual tools. The best programmes use a variety of tools in a safe climate, provide rich feedback, anonymity (when appropriate) and follow-up of behaviour change over time. Effective assessment and feedback programmes also incorporate faculty development.

The assessment of professionalism as an interpersonal phenomenon

1. In addition to its individual elements, professionalism also implies a set of behaviours and responses to situational and contextual phenomena that arise during learning and practice. The assessment of professionalism should therefore include assessment of the decisions, responses and behaviours of all actors in each context (perhaps using multi-source feedback), gathering longitudinal data from both teacher and student as well as other key players such as healthcare professionals, administrators, patients, etc.

2. Assessment of the learning/practice environment itself is also important. Inherent in this approach to assessment is feedback to improve the performance of teams (course faculty, clinical teaching teams, etc.) as well as to improve structural elements, be they organisational (e.g. policies that govern learning/work) or structural in an architectural sense.

3. Assessment of professionalism should include monitoring learning environments, student–student, teacher–student, student–health professional and student–patient relationships for problematic interpersonal phenomena. The concept of situation-specific professionalism challenges, dilemmas or lapses may be more useful than a global concept of unprofessionalism (characteristic or trait).

4. While a complete consensus on what are appropriate professional responses to complex problems and situations may not always be achieved, assessment and feedback should represent a collective perspective where possible.

The assessment of professionalism as an institutional/societal phenomenon

1. Professionalism can be understood in the context of the goals, aspirations and collective behaviours of healthcare and educational institutions and
How to assess students and trainees in medicine and health

of the profession itself. Assessment involves characterising societal expectations, through dialogue and meaningful input from public stakeholders, and measuring the degree to which the profession (be it a subgroup such as students, a whole medical school, a professional practice group or even the profession as a whole) meets these expectations. Accreditation requirements at every educational level require teaching and evaluating of professionalism. Effectiveness should be measured in terms of clear institutional/societal outcomes.

2. Assessment may involve critiquing the dominance of certain ways in which expectations and practices are framed or enforced (cultural, generational, gendered, hierarchical, etc.) and should lead to improved institutional and organisational climate and practice.

3. Professional lapses may arise from particular kinds of social interactions and problematic organisational and institutional settings and politics. Examining and making explicit the hidden curriculum and tacit problematic organisational or institutional norms is important in assessing and contextualising professional/unprofessional behaviours of students, teachers and institutions.

Summary

Professionalism has become a central concept in medical education and therefore robust ways of capturing and measuring professionalism or lack of professionalism are important. However, professionalism is a complex and multi-dimensional construct and there is no consensus, no one best way, to assess professionalism. There is no one accepted definition of what constitutes professionalism within the health profession, nor one way to understand how professionalism is learnt. Because of this, the way in which professionalism is understood and described within the learning organisation or programme significantly affects what aspects of professionalism are assessed, in what way and for what purpose.

As with all assessment, a balance needs to be struck between the reliability, validity and feasibility of assessments used. It is suggested that a range of assessment methods should be used in the assessment of professionalism; using faculty/practitioner–teacher, peers, service users, self-assessment and proxy markers to provide the best multi-dimensional assessment of professionalism. Assessments of professionalism should utilise both formative and summative assessments opportunities and tools. Professional development should be seen as a lifelong learning process and at any single point in time assessment of professionalism should be viewed as part of a continuum.
References


Further reading


Chapter 6  Setting the standard in assessments

Dr Steve Capey\textsuperscript{1} and Professor Frank C. Hay\textsuperscript{2}

\textsuperscript{1}University of Swansea
\textsuperscript{2}St George’s University of London

Introduction

The outcome of assessment in vocation-based training is perceived as recognition of competence for professional awards. There remains considerable debate as to the volume of different assessments in the training of healthcare professionals, with some maintaining they are over-assessed. The argument in favour is that the test results of applied knowledge, skills and professional attitudes provide compelling evidence of competence for registration with the professional regulatory body. As one might expect, tests prepared year-on-year are not going to be of equivalent complexity. Likewise, for reasons of test item security and curriculum alignment, one test cannot be reused repeatedly for a cohort or even subsequent cohorts as this would skew the test results and result in scores that did not reflect the skills and competence of the candidates. Therefore different tests must be prepared; but then the problem arises of quality assuring equivalence of standard between tests. This is achieved by deriving a justifiable pass score whereby one cohort has not been adversely prejudiced against another. The introduction of standard setting, that is criterion referencing, has provided an approach that allows examiners to construct assessment that (i) is aligned to the curriculum, (ii) equates to tests used for previous cohorts and (iii) agrees an objective pass score.

This chapter provides an explanation of the different types of assessments and methodologies for standard setting – how to use them, a discussion and measuring test reliability. This latter concept may be considered in terms of the ‘trustworthiness’ of the scores, that is did the score awarded reflect accurately the candidate performance?
Summative and formative assessments

The two main types of assessment used are summative and formative assessments, where the balance between the two formats is essential, with appropriate and timely feedback, to augment the learning experience throughout a programme of study. Summative assessments are often referred to as ‘high stakes’ assessments as they are used in the decision-making process about candidate suitability for academic progress to the next stage of training. Likewise, formative tests are generally viewed as ‘assessment for learning’ or ‘low stakes’ assessment, that test developing knowledge and skills, allowing students and faculty to discern areas of success and areas for further development.

Clearly it is necessary to strike an acceptable balance between the formative and summative elements. Excessive summative assessment can lead to examination fatigue owing to the amount of preparation needed to avert the adverse consequences of failure. The opposite may be true for formative assessments, when they are seen as having less of an influence on degree outcome, so do not generate much impetus to drive learning. Accordingly students prioritise the summative assessments at the expense of learning because they are recognised as the greater hurdles to negotiate. The educational literature has a wealth of evidence on how the ineffective management of the balance between these two assessment types has an impact on learning. When the student focus is on library-based study, it is to the detriment of the contact with patients which where they really learn their profession.

The educational influence of any scheduled assessment is significant – an effective approach to the support of learning is to have frequent, regular formative assessments (low stakes) within a programme that is clearly punctuated by summative assessments. Thus students are provided continuous feedback on performance, and faculty has an equitable level of evidence on the progress of individual students, without the pressure for both candidate and examiner when there is over-assessment.

Scores and standards in assessments

When defining scores, a School may opt to use a range of letters (A, B, C, D, E,) or a number to denote candidate performance relative to the agreed criteria for the assessment. But whatever the chosen system it must be robust if the assessment is to be of worth. Within a written multiple-choice test, the scores may be awarded for the number of correct responses. If a negative marking scheme is used, marks will be deducted for incorrect responses, a system generally used to deter candidates from guessing. Likewise on other
How to assess students and trainees in medicine and health

assessment formats where a skill is being tested, such as an OSCE, the mark sheet reflects the importance of defined areas of the performance which are weighted. For example, in an OSCE testing history-taking skills, the information retrieval section might be given a weighting of three (i.e. marks awarded then multiplied by three) compared with greeting the patient, where the marks have a weighting of one.

When considering standards for assessment, the pass or cut score must be defensible in terms of the marks awarded for the performance in knowledge or skill(s) being tested. The nature and construct of the assessment and how the standard is set may result in candidates that demonstrate competence, yet still fail (see Figure 6.1). Thus a cut score is at the margin between pass and fail, and assessment judges whether the candidate performance is at an appropriate standard for progression to the next academic level, for example academic year or graduation. The outcome of the examiners’ judgements for agreeing the cut score must be reliably defensible and be able to withstand scrutiny by all stakeholders, that is examiners (internal and external), candidates, professional regulators and the establishment (NHS).

When setting standards candidates may be competent to pass to the next stage of assessment. However as Figure 6.1 shows, some candidates’ test scores may not reflect their true abilities or ‘true scores’ and the need for setting robust and defensible standards is paramount.

**Standard setting**

There are various different standard setting methods used for deriving pass scores, but the level of competence required to gain a pass depends a good deal upon the nature of the knowledge or skill being assessed. For instance, for a basic competence in a skill that is an unequivocally/essential
requirement for registration, the pass score (quite reasonably) might be set at 100%, for example basic life support. Conversely, to expect a maximum score as a pass for a knowledge test may not be as appropriate, for in this situation the assessor may expect to test minimum competence to excellence of candidates. At first glance, the methods used to set standards may seem slightly bewildering; all report high levels of accuracy in setting pass scores but there has been no real consensus regarding the most appropriate methods to use for different test modalities, for example data analysis versus knowledge. The trend has moved towards the setting of an ‘absolute’ pass score that allows all candidates who meet the minimum competence of knowledge or skills (as agreed by the faculty), to achieve a pass grade.

Who sets the pass score?
An important variable in the standard setting process is the panel of judges, that is the number of judges, their specialist knowledge and the appropriate gathering, is fundamental to the credibility of the work. Judgements in a standard setting process are generally considered to be rather subjective. The individual decisions made are frequently adjusted judiciously during panel discussions. But in order that their conclusions stand up to scrutiny, what qualities are needed in a panel, and realistically how many judges should be present in order that the system is unbiased, methodical and reproducible?

Understandably, a formative assessment which is part of a learning session is not ‘high stakes’, that is an academic hurdle to progression, so one faculty member organising the test is adequate. But to uphold good practice in summative assessments, so that the results are meaningful, the panel of judges should fulfil the following criteria:

- A minimum of six colleagues, representing a breadth of expertise in the subjects being assessed, that is specialists (e.g. consultants), generalists, clinical and non-clinical faculty (Brennan and Lockwood, 1980);
- A range of personal attributes, that is age, gender and race, to avert any real or perceived conflicts of interest;
- All to have a thorough understanding of the curriculum being assessed;
- All to have practised and demonstrated proficiency in the agreed method of standard setting.

The meeting itself requires administrative support as panel members invariably have demanding work schedules that necessitate adequate forward planning (Brennan and Lockwood, 1980; Norcini and Shea, 1997).

Setting the pass score
The decision made by a panel of judges is to formulate a pass score that represents minimum competence in a candidate performance. Standard
setting is the process used for making decisions on where the boundary lies between candidate scores that demonstrate competence to pass, and those who fall below that order or specification. A definition of standard setting that is offered is: ‘Standard setting refers to the process of establishing one or more cut scores on examinations. The cut scores divide the distribution of examinees’ test performances into two or more categories’ (Cizek and Bunch, 2007). Nearly as many standard setting methods have been devised as the number of different test modalities that use them. It is reasonable to state that no individual method is a ‘gold standard’ for all test modalities, the important factor to consider is that an appropriate method is applied to the test (Schuwirth and van der Vleuten, 2010).

Methods for setting the pass score or cut score can be (roughly) divided into the two major categories: relative and absolute – the choice is dependent on the specific objectives of the test. The relative methods are founded on the principle that candidates’ scores are used to derive the cut score (pass score) for the test. These methods are generally used to classify a group of candidates into a ranking order, and where only an agreed certain proportion of candidates go on to pass the assessment – this is known as ‘norm referencing’ (Bandaranayake, 2008). Deriving the cut scores in this instance is a relatively simple calculation, and is based on setting a precise and agreed percentile. A candidate that achieves a score within those percentiles would be deemed to have passed. The main drawback of the technique is that the outcome of the assessment is dependent on the performance of peers rather than an individual’s competence. Whilst this may be acceptable in formative assessments, or when candidates need to be ranked against each other, it would seem inappropriate for high stakes assessments where a judgement is being made on candidates and their fitness to practise safely (Bandaranayake, 2008).

The absolute methods are also referred to as ‘criterion referencing’ standards. These are more complex and typically they require a panel of judges to make a decision on the pass or cut score. The methods used in these calculations are rather more involved and require prior training for the judges. But the major advantage of absolute standard techniques is that the cut scores agreed are more defensible in high stakes assessments, and the pass/fail decisions are made against predetermined criteria, that is the standard of minimum competence in the test.

A further category of methods has emerged, known as compromise methods, that share some of the characteristics of absolute and relative methods. As the term suggests these methods occupy the middle ground between the two extremes, and commonly use a norm referenced element in conjunction with a series of tolerance levels as assigned by the panel of judges that determine the cut score. Compromise methods have gained
popularity for their simplicity of use and appear to provide a reasonable level of acceptability to candidates. However, such methods should be used with caution for high stakes examinations when decisions about competency to practise are being made.

**Methods used for standard setting**

The type of standard setting method to be used and the procedure employed depends on the purpose of the test and this can have a large amount of relevance to the justification for choosing a certain methodology. A simple six-step process for standard setting was described by Norcini (2003) that provides a reliable framework to ensure that appropriate criteria are determined (see Box 6.1).

The first two steps are the most crucial to setting a defensible standard for a test and these will be explained in greater detail. Selecting the type of standard to be used is critical as it is the basis for the activities in the later steps. The purpose of the test is an initial question to be asked in this stage, for if it is to be used as a ‘high stakes’ assessment, for example a licensing exam, then an absolute method is the standard of choice. However if there are a limited number of positions available, then it is pragmatic to consider the best of the candidates. Accordingly it would be reasonable to rank them in order of competence, by using a relative method. Once the type of standard has been decided, the decision on the most appropriate method relies on a number of parameters, those being (i) the desirable characteristics of the chosen method, such the ease of use, (ii) the availability of judges to meet as a group and (iii) the utility of the method.

**Relative methods**

Relative methods are where an agreed percentage of candidates are permitted to pass the assessment irrespective of their competence. Thus there could be
a situation where all candidates within a cohort are competent, but only those that fall into the top percentiles make the grade. It is the simplest standard setting method in that it merely requires the judges to define that percentage. The procedure is as follows:

- Each judge is asked to estimate the percentage pass rate for all candidates taking the test;
- The panel of judges discuss their opinions on the pass rate and are free to change their decision following the discussion;
- The estimated percentages from all the judges are collated and an average pass rate is calculated and applied to the test.

For example: a fixed percentage calculation for five judges would be as follows:

Judge 1 = 60%
Judge 2 = 75%
Judge 3 = 65%
Judge 4 = 70%
Judge 5 = 89%

The percentage of candidates to pass would be: \[
\frac{60 + 75 + 65 + 70 + 89}{5}
\]

= 71.8%.

The average pass rate can be used repeatedly for subsequent tests as it is recognised as allowing the highest achievers in a cohort to pass – a method that has been used in professional examinations to ensure a limited number pass at each sitting.

An alternative form of the ‘relative method’ is when a sub-set of the cohort is used as the reference population to set the standard for the whole of the cohort, that is the reference group method. By defining the reference group and their characteristics, the judges then describe the pass rate for this subgroup using the same approach as for the fixed percentage method. The reference group method is valuable when the judges are familiar with the capability of the candidates and where the large cohort is spread over multiple sites to take the same test. Conversely, the major disadvantage is that the reference group may not be truly representative of the whole cohort (Norcini and McKinley, 2009).

**Absolute methods**

**Nedelsky method:** One of the first absolute methods was the Nedelsky method (Nedelsky, 1954) that’s main focus was introducing the concept of
the ‘borderline group’ of students. The notion of a borderline group was described as candidates who were minimally competent – it is a model that has been applied in a number of subsequently formulated standard setting methods. In all such methods the critical feature is for judges to define precisely the concept of borderline in order to determine the appropriate standards.

The system used in the Nedelsky method is relatively straightforward for multiple-choice questions:

• Judges give a score to each question which is derived by adding up the number of incorrect answers that the borderline group of students would be able to rule out as not being the correct answer;
• The score for that question is the reciprocal of the number of answer options with the number of incorrect answers identified subtracted. For example, in a multiple-choice question with five options and where three answers were identified as being incorrect for a borderline student, the Nedelsky score for that question by that judge would be: \( \frac{1}{(5 - 3)} \) or \( \frac{1}{2} = \) Nedelsky score of 0.5.
• The Nedelsky scores for all judges are taken and then an average is derived for each question;
• The sum of the mean item scores is the Nedelsky passing score for the whole paper.

The Nedelsky method has since undergone several modifications, with discussion of individual items and an agreed/consensus score reached, rather than means being reported, to improve agreement between judges. Another modification has been the introduction of multiple rounds of scoring, with judges discussing item scores between rounds and being able to change their scores if they feel it is appropriate.

Angoff method: The Angoff method continues to be one of the more widely used absolute methods for setting standards (Angoff, 1971). It applies the same concept of a borderline group of students that was introduced by Nedelsky, but with Angoff scoring the borderline concept is described and agreed ahead of the standard setting session. The original Angoff method used the concept of 100 minimally competent candidates to define this group. Thus:

• Judges rate each question independently, estimating the proportion of the ‘100 borderline students’ that would give the answer correctly, for example a score of 0.75 for a question by a judge who considered that that 75 of the borderline group would answer the question correctly;
• When all questions have been rated independently by the judges, they discuss their scores and may alter their individual scores depending on the outcomes of the discussion;
Each judge’s scores are compiled for all questions;
A pass score is derived from the sum of the item Angoff scores, divided by the number of questions and expressing it as a percentage.

For example: the Angoff calculation for one judge on a five item test would be as follows:

Q1 0.75
Q2 0.15
Q3 0.40
Q4 0.60
Q5 1.00

Clearly the judge considered Question 2 to be far more difficult than Question 5

The Angoff score for the 5-item test would be:

\[
\frac{(0.75 + 0.15 + 0.40 + 0.60 + 1.0) \times 100}{5} = 58\%
\]

Using this method the final cut score is determined by calculating the mean of the whole test Angoff scores for all judges. There have been some modifications to the Angoff method since it was first introduced. Most of them have considered the fact that judges generally find it difficult to envisage a group of 100 borderline students. Therefore judges now may be asked to return the probability that a single borderline student would answer the question correctly.

A further modification described is somewhat simpler: this is the ‘yes/no method’ of Impara and Plake (1997):

• Judges are asked to score a question with a one if they believe that a borderline student would answer it correctly, and zero if they believe they would give an incorrect answer;
• The cut scores suggested by each judge are calculated by totalling the number of scores for the test;
• Judges are then provided with ‘impact information’ which is essentially the number of students predicted to pass or fail based on their predicted cut scores;
• Judges undertake a second round of standard setting where they discuss each question and their reasons for attributing the scores;
• Finally, the judges’ scores are collated and the average calculated to provide the final agreed cut score.

Compromise methods
Compromise methods can prove useful, particularly when a School or Faculty first introduces standard setting procedures as part of the assessment
strategy. Whilst there might be enthusiasm for such developments, the level of expertise may be limited within the panel of judges. Moreover, colleagues may be reticent about answering putative questions around the arbitrary borderline student. But if the panel know that more tangible techniques are also being applied as a safeguard for agreeing cut scores set, it is likely to encourage confidence that these standards are appropriate.

**Hofstee method:** This method, first described in 1973, was updated in 1983 and is a system that combines elements of norm-referencing and criterion-referencing (Hofstee, 1983). The Hofstee method of standard setting requires the panel of judges to answer four strategic questions which are:

1. What is the highest percentage cut score that would be acceptable to you even if every examinee achieves this score?
2. What is the lowest percentage cut score that would be acceptable to you even if every examinee achieves this score?
3. What is the maximum acceptable failure rate, for this cohort in this exam, expressed as a percentage?
4. What is the minimum acceptable failure rate, for this cohort in this exam, expressed as a percentage?

In this scheme for setting the cut score:
- Judges can answer these four questions after reading the test items/questions, then reaching an opinion on the putative score that the ‘borderline student’ would achieve in this examination;
- The opinions of all judges are collated and the average percentage score calculated for each question.

The average scores are used to define the corners of a box that is plotted on a graph as shown in Figure 6.2. A diagonal line is drawn from the top left corner of the box to the bottom right corner. The observed scores from the examinees are plotted onto the graph as a cumulative Ogive and the point where the observed score crosses the diagonal is used to define the cut score. Some problems can exist with this method as the observed score may not pass through the box and so will not cross the diagonal, multiple solutions have been used to deal with this issue from using the minimum or maximum pass rate as the cut score (Norcini and McKinley, 2009) to extrapolating the lines on the box or the diagonal to intersect the observed data. Unfortunately these measures do not often provide a suitable or acceptable solution and this is the major weakness of the method.

**Cohen method:** The Cohen method is a more recent method that uses the scores of the highest performing candidates to delineate the cut score for those taking the assessment. The cut score is defined by the highest scoring candidates or the score at a specific percentile – in this case the 95th
How to assess students and trainees in medicine and health

Figure 6.2 Hofstee plot. Adapted from Cizek and Bunch (2007).

<table>
<thead>
<tr>
<th>Judge</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max %</td>
<td>55</td>
<td>60</td>
<td>58</td>
<td>50</td>
<td>58</td>
<td>60</td>
<td>50</td>
<td>55</td>
</tr>
<tr>
<td>Min %</td>
<td>45</td>
<td>50</td>
<td>38</td>
<td>40</td>
<td>45</td>
<td>45</td>
<td>35</td>
<td>42</td>
</tr>
<tr>
<td>Max fail</td>
<td>10</td>
<td>25</td>
<td>30</td>
<td>15</td>
<td>3</td>
<td>70</td>
<td>15</td>
<td>24</td>
</tr>
<tr>
<td>Min fail</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

Average scores from seven judges that create the corners of the Hofstee box.

Figure 6.2 Hofstee plot. Adapted from Cizek and Bunch (2007).
percentile (Cohen-Schotanus and Van der Vleuten, 2010), thus the best performing students are used as the point of reference.

Using this method, the pass mark (PM) for an exam is calculated using the following equation:

\[
PM = C + 0.60(P - C).
\]

Where the pass mark (PM) is calculated by taking the score expected by just guessing (C) and the score of the student at the 95th percentile (P). Some criticism has been indicated in that, as the passing score is corrected for guessing, why is a factor value of 0.60 used? Undoubtedly, the 0.60 factor could be altered to change the level of difficulty. Nonetheless when cut scores derived using the Cohen method and an alternative absolute method are compared, Schools might want to consider working with the retrospective data, and making adjustments using a 'reverse engineered' locally derived Cohen factor, that is to derive a precise value by assessment modality and year of study to produce cut scores of equivalent standard in the local assessments (see also Taylor, 2011).

A final thought on standard setting . . .
Standard setting methods have both advantages and drawbacks that need to be considered when deciding on which method to apply for a given assessment modality. The opportunity is there to use multiple methods to make a decision on cut scores. While this might be relatively straightforward to do, it could lead to a greater dilemma if the two different methods gave markedly dissimilar scores: what score should be used? Probably the most practical solution is to select an appropriate standard setting tool that is based on: (i) what is known about the methods, (ii) resolving the issue on the purpose of the assessment, (iii) what is necessary to develop expertise within the faculty so that there is confidence in the decisions made by the panel of judges and (iv) a method so that the candidates feel that the assessment has been unambiguous and fair.

Blueprinting of assessment
The description of how the intended learning outcomes of a programme or module of study will be assessed is known as the blueprint of assessment. The production of a blueprint is an essential document for students and faculty to understand the scope and range of the examinations. The simplest representation of an assessment blueprint is to produce a matrix or grid, with the modalities of assessment on one axis and the intended learning
outcomes for the program on the other (Sales et al., 2010). The assessments that are intended to measure a specific learning outcome are indicated on the individual ‘squares’ of the matrix (see Figure 6.3). The blueprint can be developed further to assist in the setting of assessments by representing multiple aspects of the construction of the test in the blueprint; medical speciality and common tasks carried out by a doctor can be used as part of the blueprint to ensure an even coverage of assessment items across the range of specialities; other demographics such as age, ethnicity and gender can all be utilised in the blueprint to ensure a reasonable representation across all these variables. In an integrated assessment system it is important to ensure that all of the assessment modalities are represented across the different areas of the blueprint.

Figure 6.3 illustrates a simple and integrated blueprint that is based on medical speciality and common tasks or areas of activity undertaken by clinicians. Three assessment modalities have been included: multiple-choice questions (MCQ), short answer questions (SAQ) and the objective structured clinical examination (OSCE). The assessments are distributed throughout the matrix so that even coverage is achieved. The blueprint for each year should be consulted in subsequent years so that all areas represented in the blueprint can be demonstrated to have been assessed.

The central activity in blueprinting is ensuring enough flexibility in the blueprint to allow the setting of test items that truly reflect the curriculum so that constructive alignments with the learning outcomes of the course can be achieved. The volume of work that must be covered and tested in any healthcare education programme usually means that it is not possible to test competence in all areas of knowledge and skills. Therefore sampling across the learning outcomes is required if assessment fatigue in the students is to be avoided. However, the problem that exists with the sampling approach to assessments is the issue of case-specificity. The fact that a candidate may be able to solve a complex patient problem in one medical context is not a reliable predictor of their performance in other patient problems in a different medical context. Hence the sampling of knowledge needs to be across a range of medical specialities and contexts. The most appropriate system for achieving this is to blueprint across the programme outcomes to guide the selection of test items for a broad sampling of candidate knowledge and skills (Geoffrey et al., 2006).

**Computer-based assessments**

The development of computer technology has opened up many opportunities for test delivery that were previously unavailable in standard pen-and-
Figure 6.3 Integrated assessment blueprint. Adapted from Amin et al. (2006).

<table>
<thead>
<tr>
<th>TASK</th>
<th>Cardiovascular</th>
<th>Renal</th>
<th>GIT</th>
<th>CNS</th>
<th>Respiratory</th>
<th>Reproductive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis</td>
<td>MCQ</td>
<td></td>
<td>OSCE</td>
<td>SAQ</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Management</td>
<td>MCQ</td>
<td>SAQ</td>
<td>OSCE</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>History</td>
<td>OSCE</td>
<td>SAQ</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Examination</td>
<td>OSCE</td>
<td></td>
<td>MCQ</td>
<td>OSCE</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Investigations</td>
<td>SAQ</td>
<td>MCQ</td>
<td>OSCE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethics</td>
<td>SAQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public health</td>
<td>SAQ</td>
<td>MCQ</td>
<td>MCQ</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of OSCE: 6
Number of MCQ: 6
Number of SAQ: 6
paper formats. The statement by Thomas Oakland that ‘Computer use is shaping the ways tests are constructed, normed, validated, administered, and scored. The reach of this technology on test practices is broad and international’ describes the impact that computer technology and the Internet have had on assessments (Bartram and Hambleton, 2006).

With the advent of the personal computer came the development of computer-delivered versions of paper and pencil tests that did little more than deliver the test in the same manner but through an electronic medium. However the capabilities that computers offer allow innovation to take place within this new medium. The type of innovation that computer based testing (CBT) allows comes in two main areas: (i) content and (ii) delivery.

The advances in content are reasonably straightforward, with the use of sound and video to enhance the supplementary materials being used in the test. It is evident that sound can also open up new opportunities in assessments that were previously only used in a practical setting, for example heart sounds heard via auscultation. Video data can also provide a simple method for testing visual characteristics, such as skin lesions, movement disorders and radiological images. Indeed, some may be accessible for use in ‘real time’ such as ultrasound or echocardiography.

Simulation is a further development of content that utilises more of the capabilities of computer-delivered content, and attempts to utilise realistic patient-based vignettes to simulate consultations (as used in OSCEs) or other complex assessment situations have been successful (see also Chapter 4). Examples of simulated practice have been developed in many fields of healthcare education as a formative approach in testing patient–professional consultations, such as exploring socio-cultural issues, which has potential for use in summative assessments (Perron et al., 2009). Simulations have also been developed further to cover surgical procedures and even virtual ward environments that would all be valuable tools for formative assessment (Brown et al., 2012).

Delivery of assessments has been revolutionised by computer-based technology where adaptive testing methodologies can be utilised to improve test efficiency. Computer assisted test construction has facilitated the development of question banks. Likewise, the requirements for delivery of papers to geographically diverse locations can be easily overcome, which is of significance for licensing exams that can be delivered to test centres. However the utilisation of electronic delivery of tests does have some considerable item security issues associated with it (Shepherd, 2003). Hence the availability of software that can synchronise individual test items in a specific order that is unique to the individual candidate, which can ameliorate some of the concerns around test security.
The possibilities for evaluating test items against candidate responses using statistical item response theory (IRT) has been made more straightforward and rapid with the introduction of on-line assessments (see also Chapter 8). IRT is possible when there is a significantly large enough question bank available that has specified the complex characteristics of the individual test items. Therefore it is quite possible for assessments to be delivered in real-time that will adapt to candidates’ abilities.

An interesting development described by Lai, Gierl and Turner (2012) of the University of Alberta has been a computer-based item generator (IGOR) that allows a large number of multiple-choice questions to be prepared based on one clinical scenario. The potential benefits of the time outlay in programming the scenario could be realised by the return of a large number of questions in a specific subject area to be used in subsequent examinations.

### Assessment management

**Administration of assessments**

The administration of on-line assessments has the benefit of large amounts of processing power, made feasible by software designed to allow the creation of assessments that utilise most of the principles that have been described earlier. The development of tailored software packages for test construction and their analysis allows automation for rapid marking, with an individualised feedback facility for candidates. Moreover, the introduction of these easy-to-use software packages has allowed a much greater analysis of the reliability of assessments to be carried out than was feasible to conduct prior to this. There are a number of assessment data management software packages available that will deliver most of the assessment tasks from blueprinting to assessment item analysis. Nonetheless, a change in assessment modality requires considerable organisation change and faculty development for it to work effectively, as well as software packages to link between curriculum changes and the assessment item database. Likewise, technical support is essential to reduce the incidence of errors and intrusive and fraudulent behaviours that can compromise the security of the examination process.

There are greater efficiency advantages with on-line assessment. The test analysis has improved significantly with a level of analysis, previously incredibly labour-intensive, which is now straightforward and fast. Statistical analysis of assessments has been greatly simplified by the use of customised software, making it possible to calculate the reliability estimations for complex tests, and allowing the move forwards from classical test theory to the more elegant generalisability theory, Rasch modelling and IRT (see also Chapter 8). These advances have offered meaningful item analysis, thus
enabling test-writers to construct more robust tests with the possibility of being able to predict the characteristics of the test before the candidates sit it. Such detailed analysis has certainly assisted in guiding the principles of test design and in identifying poorly performing test items.

**Blueprinting management**

The on-line blueprinting applications require continual attention to ensure that item-tagging information allows the blueprint created to be updated and aligned with any changes in curriculum, its delivery and the possible advances in medical knowledge that may invalidate some assessment items. The analysis sections of these packages provide a robust interface that is user-friendly and can produce high quality assessment reports that permit assessment reliability to be explored rigorously. Their potential for developing joint curriculum and assessment management packages that allow the curriculum learning outcomes – mapping them against the requirements of the regulatory bodies – is also being realised which affords a seamless link for using the assessment management software in curricula quality assurance.

**Computer-assisted marking of assessments**

The use of optical marking processes for selected answer tests such as MCQ examinations and OSCE mark sheets, have increased the efficiency and fidelity of the assessment process, particularly in one of the more error-prone areas, that is collating test information in a short time frame that can result in marking and transcription errors. Accordingly, the high fidelity optical markers provide rapid and efficient processing times for large item MCQ tests and performance-based marking checklists. However, a word of caution: care must be taken that the marking protocols are robust, and checks are in place to ensure the validity of data returned by these methods. Minor errors in database or spreadsheet construction can have disastrous consequences for the candidates whose marks are returned incorrectly, especially in any high stakes, summative examinations.

A recent development in the delivery of tests is real-time computerised capture of assessment data. The availability of small and relatively cheap computers has made it possible to create large assessment centres where candidates register and complete the assessment, which can reduce the incidence of transcribing errors. The on-line capture of performance based assessments is also being explored that will allow assessors to complete an electronic checklist, for example by using an iPad, that then ensures all items have been scored before data is transmitted to a central computer for immediate collation. Thus wireless Internet access has made it possible to collect
test data from multiple sites, for example OSCEs are synchronised at several NHS Trust sites, which has sped up this type of data collection, and its return and collation by a central administration. The introduction of electronic portfolios as part of workplace-based assessments (see Chapter 4) has also made these documents accessible to colleagues involved in appraisal, allowing a more streamlined and distributed process around these types of assessment activities (Van Tartwijk and Driessen, 2009).

Recent innovations have been rather ambitious in the developing of computer- assisted free text marking. The challenges here are much more considerable than those faced when marking selected answer formats. For this type of marking to work, the computer must be programmed with very robust marking criteria for the question to be answered. Currently the software identifies key words in the response, and matches them to the marking criteria. Several systems have been developed that can offer a wide range of free text marking, such as the ‘e-rater’ system developed by Educational Testing Services (ETS; (Attali and Burstein 2006) that will mark constructed responses that correlate well with human scoring. A second system has been produced by Intelligent Assessment Technologies (IAT) that has been used successfully to score short answer questions of medical knowledge. The main problems that have been encountered have been in the handling of complex answers produced by candidates, spelling errors and discriminating the context in which the answer has been written in. Another more human problem is that of trust, for candidates and even faculty have questioned the use and reliability of this technology in high stakes assessments. However, in short answer questions the methodology has been proved to be as reliable as hand marking by experienced faculty members.

Computerised assessments have greatly increased the possibilities available for test construction and for evaluation of the performance data for test items. Such developments in post-test statistical analysis have the potential to enhance the reliability of tests and to identify methodologies for appropriate test delivery. Although still in a relatively early phase, with the right support and infrastructure they can augment the assessment for, and assessment of, learning. Likewise, the potential for simulation-based assessments with virtual patients is a very promising area of innovation that may provide a significant contribution to the types of assessment available in the future.

References


**Further reading**


Chapter 7  **Feedback and assessment**

Dr Ann Griffin¹, Dr Sophie Park² and Dr Catherine O’Keeffe³

¹UCL Medical School
²University College London
³Institute of Education

**Introduction**

Feedback is an essential element in the facilitation of effective teaching and learning. Timely and well-crafted feedback enhances the performance of learners by highlighting to them the abilities they already possess, as well as giving constructive advice about how to develop those skills which have scope for improvement. Feedback is a key method which can promote reflection within the learner and foster practitioners that are more critically aware of their own performance. In recent years we have seen significant progress in embedding feedback into clinical practice. Giving feedback is no longer seen as the exclusive domain of the educator, rather an expected responsibility which is shared by the whole healthcare team and patients. We have seen a range of new devices, for example the 360° multi-source feedback tools (Davies and Archer 2005; Davies *et al.*, 2008), e-portfolios and workplace-based assessments. Feedback practices are based upon a diverse range of educational theory and the research proves that quality feedback is effective in enhancing performance. Despite the undoubted benefits of feedback in teaching and learning there are important issues to consider. Education establishments and clinical placements are often criticised for the lack of feedback they give learners, particularly in relation to feedback on formal assessments. Some would suggest that students fail to recognise when feedback is being given, perhaps because teachers are not making the activity
explicit enough. The practice of feedback is, however, frequently perceived by educators as one of their most challenging roles, largely because of the interpersonal issues that are involved. We do know that giving honest and developmental feedback is a highly nuanced skill that not all teachers feel confident about and this concern may produce some reluctance to provide open and meaningful feedback for learners, particularly where there is concern about its reception. Feedback is increasingly used in formal assessments, changing the nature of the process from formative to summative. This introduces additional complexities and tensions which may impact on the authenticity and quality of the feedback dialogue and its potential impact upon learning.

In this chapter, we will help to address some of these concerns and offer practical guidance for those who have taken on the task of being the facilitator of feedback conversations. First we will set the scene: the context in which feedback is given and received in healthcare settings. Next we will consider the feedback exchange from the different perspectives of feedback receiver and facilitator. We will also cover some of the theory underpinning feedback; provide you with some principles and models of feedback; and discuss some developments in feedback, both in terms of new tools and sources, focusing on multi-source feedback and appraisal. Students value feedback on assessment and often report that this is inadequately done, therefore we will conclude this chapter with a few ideas that can be developed to provide more effective feedback on summative assessments. This text is written primarily for those who are going to be responsible for orchestrating feedback conversations: the facilitators. We are, however, aware that readers are also likely to be in a position of also receiving feedback on their own performance and hope that this chapter will also provide some additional support to make sense of these feedback conversations.

**What is feedback?**

Feedback is essentially a dialogue about performance. Eraut (2006) defines feedback as ‘any communication that gives some access to other people’s opinions, feelings, thoughts or judgements about one’s own performance’. Its aim is to provide learners with insight about their own abilities. The analogy of the mirror is often used to illustrate how feedback works. Holding up a mirror allows a learner to see something for themselves, revealing a particular, hitherto unseen, aspect of their behaviour. This new insight allows them the opportunity to reflect on their current practice and raises their own awareness about how to expand to develop their skills, including those of self-assessment. As Winne and Butler (1994) suggest, feedback
enables individuals to refine and reshape their thoughts and actions, and to change. Feedback can be an assessment for learning, that is a formative assessment made by someone about something they have observed or reviewed, but it can also be an assessment of learning, when it may take on a more formal or summative role.

‘Feedback is information with which a learner can confirm, add to, overwrite, tune, or restructure information in memory, whether that information is domain knowledge, meta-cognitive knowledge, beliefs about self and tasks, or cognitive tactics and strategies.’

Winne and Butler, 1994

Skilled feedback works. In the clinical setting, Veloski et al. (2007) in their systematic review showed that feedback had a positive effect on a physician’s performance. More broadly, Black and William’s meta-analysis (1998) demonstrated that high quality feedback is effective for improving learning and performance, proving particularly beneficial for struggling students. There are other benefits of feedback too. Feedback conversations create a formal space which acts to demand the attention and time of senior, respected colleagues to discuss developing professional practice, providing an opportunity for promoting meaningful contact. The shift to embed feedback into the everyday practices of the workplace is important and offers great potential for increasing developmental exchanges, which can enhance individual learning as well as support quality enhancement in clinical practice.

The increasing importance of feedback in healthcare education

There have been many recent developments formalising and expanding opportunities for feedback within clinical training and continuous professional development (CPD) and a growing emphasis on providing learners with increased amounts of feedback on their performance.

‘In clinical education, the importance of feedback extends beyond pedagogy. The goal of clinical training is expertise in the care of patients. Without feedback, mistakes go uncorrected, good performance is not reinforced, and clinical competence is achieved empirically or not at all.’

Ende, 1983
Ende’s statement is still true today; however, healthcare education has frequently been criticised for its deficiency in observing and feeding back to its learners. Even when observations or assessments have been made, data about clinical practice has not always been shared with the learner, denying them the opportunity to shape their own performance. We have seen many changes in the education and training of the healthcare professional and these developments have been driven by shifting societal expectations about what should define a competent, contemporary healthcare practitioner. They have resulted in significant structural modifications to training, which have attempted to standardise the curriculum and pathway to expert status. Ongoing observation, the ‘currency of feedback’ (Ende, 1983), and feedback on performance, have become core requirements of clinical education, assuring the contemporary practitioner is ‘fit to practise’. This has not been confined solely to undergraduate education and postgraduate training, it has also extended into continuing professional development (CPD), in the form of appraisal.

Feedback has been embedded in a variety of new tools, including workplace-based assessments, structured reports about progress and portfolios as means to guarantee closer and more frequent observations of the learner. The use of these tools has seen a concomitant increase in the amount of written feedback requiring supervisors to provide documentation about an individual’s growth. These permanent and formal texts are increasingly drawn upon to verify an individual’s competence and are used as an important source of certification to enable career progression. This process produces a ‘paper trail’ of assessments, potentially available for further layers of scrutiny in the future.

Feedback is now sought from a widening range of individuals including other healthcare professionals and learners, educators, peers, those involved in administration and management of healthcare and patients. Any of these stakeholders may now be invited to comment on an individual’s performance and remark about an increasing range of professional behaviours, knowledge, skills and attitudes (see Figure 7.1). Providers of feedback have a range of ‘official’ names including assessor, manager or supervisor. In this chapter we will use the term ‘facilitator’ to refer to the person who is essentially in control of orchestrating feedback conversations. The choice of terminology is purposeful as it attempts to emphasise the importance of trust, equality and mutual respect within the learner–facilitator feedback dialogue.

**Roles and challenges in facilitating feedback**

Many individuals involved in training, in any healthcare domain and at both undergraduate and postgraduate level, will have a variety of roles...
Much of the time feedback is immediate, in response to a particular circumstance that has been observed: this could be a clinical teaching session with an undergraduate; peer observation of a colleague’s lecture; or watching a colleague consult with a patient. These sorts of circumstance are commonly what we think about in relation to the term ‘feedback’, often regarded as a relatively informal act. However, there is a growing range of more formal feedback roles which include educational supervision, clinical supervision, appraisal, mentoring, coaching and being a ‘trainer’. This role diversification represents a growing recognition that supporting professional practice, through developmental feedback conversations, should be embedded in teaching and training; operate across the healthcare practitioner’s career; and should assure clinical practice which supports and enhances patient care (see Figure 7.2).

There are a host of reasons attributed to the ‘vanishing’ or difficult nature of feedback, influencing why feedback might be of poor quality, or not ‘given’ at all. The facilitator of feedback, for example, might be concerned that the information may do more harm than good, be damaging or upsetting or even hinder development. Indeed, Black and William (1998) found that poor quality feedback could be detrimental. Teachers also express concern that delivering challenging feedback to students may have unfavourable effects.

**Figure 7.1** The widening sources of healthcare feedback.
on their own popularity, resulting in poorer evaluations of their teaching. Subjectivity is often cited as an issue particularly when it’s ‘just one person’s opinion’ and differing perspectives can lead to disagreement and conflict. Increasingly, however, facilitators are working with the accounts of others and a growing amount of feedback data from a variety of sources, which can help them to present a broader and possibly more ‘objective’ perspective. Coordinating an array of feedback documents, sometimes containing diverse and at times opposing views yet presenting a unified and developmental overview to the individual, is a significant skill the modern educator now needs to possess.

Lastly, the shift in feedback from an assessment for learning to an assessment of learning produces a number of important questions about the purpose of feedback as either a tool for individual learning and development, or a tool for monitoring employees’ adherence to stipulated professional activities. Many of the assessments currently used have a mixed purpose, with a blurring of the divide between formative and summative. While this may still provide some useful learning for the individual and additional benefits for the institution in terms of producing evidence about practice and subsequent control, it provides a number of complex challenges for learner and facilitator. Is the facilitator, for example, acting in the best interests of the individual, or the institution? Are they monitoring this individual’s adherence to institutional policy and practice, or supporting wider definitions of professionalism and the individual’s perspective? Is this feedback activity for the broader development of the individual, or a tick-box exercise to provide evidence of an interaction for institutional or performance monitoring purposes? The answer is often yes to both formative and summative expectations – providing some difficult tensions for both facilitator and learner to manage, or even resolve.

Recent developments in feedback have tended to focus on contained and measurable areas of change, becoming increasingly behavioural, summative
and monitorable, or ‘performative’, in nature. This has a number of potential effects, determining the extent to which individual players are prepared to engage in the feedback process. Feedback for externally imposed or summative monitoring may continue to meet a number of learner needs by enabling them to produce structured, accurate and accessible records about aspects of their practice. The external structure may, however, contrast with formative feedback methods aimed at the learner’s own development, which are more likely to be aimed at supporting a learner to change in order to meet their own desired outcome. While these external and internal goals may at times align, an externally imposed form of feedback is likely to feel less genuine or meaningful to the learner and alternative behaviours of disengagement and resistance, or even strategic compliance may arise.

Not only do we need to think about the facilitator’s role and purpose in the feedback process we also, very importantly, need to consider the individual who is the attention of the feedback conversation (Figure 7.3). It is eminently clear that learners value feedback, but they frequently comment that they do not get enough, do not recognise it when it is offered or feel that it is unclear, lacking in guidance for them to be able to develop. It is important to remember that ‘delivering’ feedback does not ensure its digestion, as Eraut (2006) identifies ‘the feedback given is not the same as the feedback received’. Learners will often require a significant amount of support, or even training, to help them both accept and utilise feedback to enhance future performance and the evidence suggests this guidance is generally lacking.

In addition to a lack of training in how to interpret feedback, there are a host of individual factors, including past experiences, that will either enable

![Figure 7.3 Factors involved in effective feedback: the individual, language and power.](image-url)
Feedback and assessment

or restrict an individual’s motivation to change. Trope et al. (2001) demonstrated that an individual’s ability to accept feedback about how to improve their performance depended upon their mood, the diagnostic nature of feedback being offered and the importance of the task being assessed. Having a positive mood also affected their ability to remember advice and to subsequently adapt their practices. Likewise, confidence, self-esteem and readiness are also individual factors which impact on motivation to change practice. If learners have questions about the suitability of the provider of feedback, perceiving them, for example, to be a poor role model, lacking in essential prerequisite skills or the authority to be able to provide what they feel they need, then feedback is likely not to be accepted. Learners in difficulty tend to avoid receiving feedback altogether. Johari’s window (see Figure 7.4) illustrates the importance of the relationship between facilitator and learner in

**Figure 7.4** Diagram to illustrate Johari’s window (Park, 2011) adapted from Luft, 1955.
controlling the boundaries for effective feedback. The degree of their rapport influences the extent to which the learner is prepared to ‘reveal’ insights about themselves. Their respect for the facilitator will influence the extent to which they permit and accept the facilitator’s comments. If mutual trust is established, the process is likely to extend awareness of both teacher and learner, thereby maximising the potential for meaningful learning.

Language, power and emotions play an important role in making feedback effective: it is often not a conversation between equals. Our workplaces are emotionally charged environments (Menzies-Lyth, 1988). Learning in this context, particularly about oneself, cannot be separated from our feelings and yet there is often a ‘professional silence’ which acts to minimise talk about sensitive and emotive issues. Leadership has an important influence on the workplace, shaping whether the fundamental role of humans’ feelings or emotions in learning and feedback are accepted and acknowledged. The healthcare environment is still, despite significant moves towards teamwork and distributed power, one of hierarchy. Furthermore, within the tight communities in which medicine is practised, it would not be unusual for a senior figure to reappear in future encounters as an assessor or indeed the person who impacts upon a trainee’s future career path. If we envisage feedback as a dialogic process, we can see that the influences of seniority are still likely to exert an undue force upon the communication, affecting the boundaries of the talk. Feedback, as healthcare professionals, is likely to include discussions about how we interact with patients, how good we are at providing up-to-date, evidence-based care and so on. Comments about these sorts of subjects go to the core of our professional identity and, if not handled with sensitivity, can have devastating effects.

All these factors are likely to influence the way and extent to which an individual engages and retains elements for development from the feedback experience. They may, for example, engage purely at a strategic and functional level in order to fulfil assessment criteria and produce the required repository of work ‘evidence’. In contrast, however, given opportunities for personal and meaningful learning, they may make significant leaps in their growth as individual learners and practitioners.

‘Relationships rely on emotions. Our understanding of workplace learning cannot afford to sidestep this critical connection’.

Theories of feedback

Despite its importance in the learning process and an array of empirical enquiry into its delivery through various mechanisms, the theoretical underpinnings of feedback are generally diverse and under-conceptualised.

The range of somewhat eclectic perspectives about the feedback ‘process’ are demonstrated in Figure 7.5. The original concept of feedback loops was derived from the disciplines of mathematics and engineering with initial mechanical applications such as improving performance in rocket science. In this context, data about performance was fed back into a mathematical/scientific programme which enabled subsequent rocket launches to become more effective. The concept of the feedback loop has since then been applied widely by other disciplines. In industry, for example, the feedback loop forms the basis of performance management strategies designed to improve employee productivity. The concept of the feedback loop has also now been extensively embraced and developed as a pedagogic, or educational, device.

Luft (1984) used a psychodynamic perspective to explore this area. He conceptualised the feedback process in terms of a child-parent or adult-adult interaction. A feedback exercise, for example, which requires submissive and/or conforming behaviours of the learner, is likely to develop adherence, dependency, or complete rejection. At its most extreme, the learner is treated as a child in relation to an adult monitor and is, therefore, likely to behave as such. If, in contrast, the learner is engaged within an adult–adult relationship and has developed trust and respect towards their teacher, they are more

Figure 7.5 The range of theoretical underpinnings of feedback.
able to openly discuss areas which they perceive as weak, accept guidance or criticism and make meaningful changes to the ways in which they make sense of experience.

A sociological perspective of learning holds that the learner is constantly constructing and performing their own self to develop a professional identity in relation to their external environment and those around them. The individual will, then, imitate and adapt elements of witnessed professional practice or role-modelling, through their socio-cultural participation in the day-to-day activities of everyday ‘work’. From a psychological perspective, the strongest motivator for an individual to change is, of course, if the learner themselves can explicitly see a need to adapt their own learning. Malcolm Knowles (1984), in his writing about how adults learn (also known as ‘andragogy’), emphasises that the internal drivers to change and learn are frequently the most important and are crucial to the process of facilitating feedback and harnessing an individual’s own desire to develop.

Schon (1983) writes about the reflective practitioner, who constantly self-monitors their own performance, understanding acceptable professional boundaries within which they can conform or adapt their own approach. Neighbour (2005) refers to this process as developing a ‘second head’ which comments upon practice as it evolves. A novice learner will require a great deal of initial feedback in order to establish new frameworks and definitions of professional expertise. As this learning consolidates, the learner’s need becomes less constant and the responsibility for the feedback process becomes more internalised as the learner becomes more expert. Developing this complex internal feedback and sense of responsibility is itself a learning process. It may at times be challenging, but is likely to result in the development of responsible, engaged and autonomous professionals.

**Feedback for effective learning: key principles**

What learners often seek from facilitators of feedback is an emphasis providing positive, specific and clear guidance on how to improve. There are many things that the facilitator can do in order to increase the acceptability and effectiveness of feedback, and what follows is a description of these key principles.

Feedback is a conversation involving mutual dialogue about someone else’s professional performance. The terminology often used in relation to feedback, about giving and receiving it, infers that it is a linear, one-way process, but in our view it is a shared endeavour, an integral part of the working roles and responsibilities of a healthcare community. Feedback should be balanced, identifying the good, as well as constructive guidance
about the aspects of performance that need improvement. If agreed by both parties, feedback should ideally be embedded within any sort of professional encounter so that it is expected and explicit; not coming as a surprise, but expected by learners as an everyday, ordinary activity. Facilitating feedback does not have to be time-consuming. In fact, the more it is built into a regular routine, the easier conversations become, developing relationships of trust and an expectation of continual learning. It can, however, be very challenging to establish, requiring a marked shift in cultural approach and expectations. Giving challenging feedback is never easy and can easily be ‘put off’. Constructing regular routines, however, may mean that conversations can be much more frequent and involve much more bite sized and manageable chunks of material, making the process feel less daunting.

Facilitating successful and effective feedback can only be achieved by understanding thoroughly both sides of the operation and ‘how to build bridges between the two’ (Sadler, 1998). The facilitator first needs to understand the context and constraints placed upon the learner. Facilitators need to understand the body of knowledge and skills to be learned, the aims/goals/standards expected, as well as the more personal attributes of their learner. Second, a feedback dialogue must clearly identify any gaps between current and desired performance as well as acknowledge performances which have reached the required standards. Learners therefore need to possess a concept of the goal/standard or reference level being aimed for. A skilled facilitator will make that gap felt by the learner so that they are internally motivated to bridge the gap (Nicol and MacFarlane, 2004). The discussion then needs to extend to support not only the learner’s awareness of their ‘goals’ and ‘the gap’, but what is actually necessary to be done to achieve its closure (see Box 7.1).

**Box 7.1 Seven principles of feedback**

2. Encourages teacher and peer dialogue around learning.
3. Helps clarify and refine what ‘good’ performance is.
4. Provides opportunities to close the gap between current and desired performance.
5. Delivers high quality information to students about their learning.
7. Provides information to teachers that can be used to help shape the teaching.

Nicol and MacFarlane-Dick, 2004
The principles of verbal feedback and written feedback are the same but, naturally, there is a greater time delay in providing written feedback. For both, just giving praise is ineffective. The skill of giving feedback, particularly when it is timely or immediate, is to prioritise the most significant areas to cover and to develop a language which describes the behaviour identified for discussion. The great asset of immediate feedback is the potential opportunity for a re-performance which incorporates agreed suggestions for improvement. Sometimes feedback can take place at an interval after the observed episode has occurred. This can be helpful as it gives both individuals the opportunity to reflect on the situation and to identify which aspects should form the focus of their discussions (particularly if it is about a sensitive situation) allowing the facilitator the opportunity for rehearsal, practising the words and phrases that will facilitate the process best. These delayed feedback conversations should be mutually planned, ideally prior to the observation of practice, to agree the ‘rules of engagement’. Either way, if learners know and agree in advance that they are going to receive feedback, they are more likely to be receptive than if it is sprung upon them after the event.

Feedback may involve a particular activity, or may be much more conceptual and over-arching in its aims to improve an individual’s learning or performance. Colleagues can offer detailed and specific feedback on an element of practice. They may also convey wider messages or opinions about the learner’s level and achievements as well as role model a number of qualities revealing their own perceptions about quality and ‘good’ practice. Part of the feedback process should help students to interpret academic and professional guidance, within the broader professional context. These meta-feedback skills will enable the learner to place other people’s reflections on their performance within a wider perspective about their own development and enable them to take greater control and direction. The later section in this chapter on feedback on assessment builds upon on this idea and contains some practical ideas for setting up discussions, interpreting feedback and helping learners to understand its meaning.

Practical suggestions for effective feedback

Setting the climate for successful feedback is the first step. There are a number of important and practical considerations. Healthcare environments are busy and stressful places and feedback about clinical work is a very powerful and important learning experience; it is important to be mindful of the sensitivities invoked. Efforts to ensure that the feedback session is within ‘protected time’ can be very worthwhile. Ensuring confidentiality is vital and finding an appropriate place is important; interruptions from bleeps and phones should
be avoided if at all possible. The individuals should be positioned to maximise eye contact and avoid marked embodiments of power relations (such as standing over them). If completing a record of the feedback conversation, consider positioning learner and facilitator together so that you can ‘co-author’ any written comments.

A useful framework for approaching any feedback conversation is illustrated in Figure 7.6. Consider the participant’s readiness for feedback and set the right climate. Make sure you have the necessary material, be it observational or in written format, to be able to help shape a constructive conversation. Listening is the most important thing a facilitator can do. Listening will allow the learner to set their agenda, own the areas of practice that they want to raise, both good and not so good, and will avoid unnecessary interruptions and comments. In any feedback conversation the talking should be done primarily by the person whose performance is being reviewed. Ask questions to encourage the learner’s reflection and allow the learner to identify their own needs, rather than telling them. This will encourage the learner to feel the gap between their current and the desired performance and develop motivation to change (Grant 2002). While it is tempting to deliver a list of recommendations for change, feedback is likely to be more effective if you allow the learner the space to reflect upon on areas they wish to address and possible solutions, complemented by your own comments.

After these stages have been accomplished there may be particular areas you wish to reinforce and give further guidance on. It is more likely at this stage, having established your interest and learner-focused approach, that

---

**Figure 7.6** A schema for feedback conversations.
your suggestions will be taken on board. Feedback conversations are often easier if individuals are first encouraged to reflect and articulate their own insights about the strengths and weaknesses of their own performance.

In order to help those with a responsibility for feedback conversations, a range of tools, or models, have been developed to provide a structure to the flow of the dialogue. There are many different models for feedback and we will discuss a couple of common examples (see Figure 7.7). Essentially, they all suggest ways in which to order the conversation. Pendleton’s rules are well known and helpful for getting the balance of feedback right, emphasising positive attributes (see Pendleton et al., 1984). A more outcome-based way to facilitate feedback can be the Cambridge Calgary model (Silverman et al., 1997). This asks the learner what they were aiming to achieve; what they would like specific help with; what went well; what could be improved with alternative strategies; and then provides an opportunity for role play or rehearsal of alternatives.

Facilitators and learners are sometimes frustrated by the constraints of following one particular model, reporting them as formulaic and artificial, forcing their conversations to prioritise and conform to the model, rather than the immediate needs of the learner. There needs to be, like any consultation model, a degree of flexibility with the application of any of these models. It may be useful to think about the principles behind each model, rather than the specific orders and behaviours. The GROW(N) model (see Figure 7.7) comes from a coaching and mentoring background. This type of conversation may offer the opportunity for the learner to take more control within the feedback process (see Whitmore, 1992) and help support reflective and self-directed practice.

No matter which model you employ to structure the way your conversation proceeds, the facilitator needs to phrase their constructive advice in a way that can be easily understood and palatable. ‘What do you think happened when you mentioned Mrs Smith’s weight problems?’ is very different from saying, ‘She was very angry when you said she was fat’, the latter representing negative feedback about a negative performance. Pendleton (2003) illustrates the use of positive feedback for both a good performance: ‘Your body language, your encouragement and nodding, I think really helped Mrs Smith talk about and address her weight problem’, and a performance that needs to be improved: ‘I could see that you realised you had hit a sensitive nerve when Mrs Smith got angry and you began to change your consultation style to address this’. This type of talk also acknowledges that the facilitator’s feedback is opinion emphasising the importance of ownership.

The following sections discuss approaches that have been developed to support the feedback process: multi-source feedback (MSF) tools and appraisal.
Feedback and assessment

Figure 7.7 Three models for facilitating feedback. (Pendleton et al., 1984; Silverman et al., 1997; Whitmore 1992).

Pendelton’s Rules

- Teacher discusses what they should improve.
- Learner discusses what they did well.
- Teacher discusses what went well and what could be improved with alternative strategies.
- Provide an opportunity for role-play or rehearsal.

Agenda-led outcome

- Clarify purpose of meeting.
- Learner discusses what they did well.
- Teacher discusses what they should improve.

GROW Model

- G - Goal setting
- R - Reality checking
- O - Options
- W - What next, when, and by whom
- N - Next steps
Multi-source feedback

Multi-source feedback tools (MSFs) can be defined as mechanisms by which multiple individuals representing a range of groups provide feedback to a recipient. MSFs follow similar processes but may vary according to the number of raters, their professional relationship with the individual being assessed, the degree of anonymity of raters, the content of the assessment, the extent to which verbatim feedback is shared and the use of a mentor or coach to help interpret feedback. MSFs may be constructed as questionnaires with Likert scales only, while some will also provide opportunities for free text comments. Most have opportunities for self-assessment.

MSFs were originally developed by the military and have been applied extensively in management and industry. In the past 20 years MSFs have become a commonly used resource in healthcare settings. Drivers promoting the use of MSFs include a shift from traditional examination-based curricula to the introduction of competency, workplace-based assessments. Recognition that such tools can provide important learning opportunities that may facilitate the development of insight into the nuances of professional practice have also further promoted their use.

For trainees, students and practitioners to engage effectively with MSFs, it is important that the tools themselves are seen to be credible. A key factor promoting credibility is the extent to which the MSF tool is perceived to be based upon observation of practice. Establishing the content validity by ensuring that questions are clear, specific and relate directly to observable practice is therefore crucial. By ensuring that specific colleagues, students or patients who have close contact with the individual in their professional capacity are selected, content validity can be further reinforced. A common concern however is the potential for bias in the process of selection of raters. MSF tools can use self-nomination of raters leading to concerns that the health professional may, for example, choose students, colleagues or patients to assess them who they feel will rate them positively. Findings from a study by Ramsey et al. (1993) in the USA suggest that peer ratings are not significantly influenced by the method of selection or relationships between the peer and raters. While further research is needed to confirm this finding in different settings, this study suggests that self-selection does not promote bias. An advantage of self-selection is that the professional concerned may find the process less threatening and may therefore engage more fully and actively in the associated learning opportunities.

Due to the close professional relationships involved, raters who are colleagues, students or patients may be reluctant to share their true observations on performance, especially if such observations are critical. While positive
feedback can reassure colleagues, lack of constructive criticism providing specific suggestions with regard to how performance can be improved may undermine the perceived value of such tools. Approaches to promoting honesty could include providing clear guidance on how to write criticism constructively, such as keeping comments factual, in instructions for use with tools that have options for free text comments. Berk (2009) however suggests that the organisational context and culture within which MSFs are used can greatly impact on their utility. To be effective an atmosphere of ‘... openness, mutual trust, and honesty, plus a genuine interest in and desire for performance improvement must exist . . . ’. While such ideal situations may not always be possible in practice, consideration can be given to how to create a supportive learning environment when using MSFs. Opportunities to discuss feedback from the MSF with a tutor or mentor may help to create a supportive environment and could be very valuable in enabling individuals to interpret and learn from the experience of using this type of tool.

Peer assessment and feedback from multiple sources has also been used successfully with medical students using a formative OSCE format which was not only to provide feedback on improving performance, but also to learn how to give constructive feedback (Cushing et al., 2011). Finally, the main advantage of MSF is its multi-source component rather than the potential for bias from one source of the evaluation. Collated evidence from several assessors is likely to provide a more precise, reliable and equitable judgement.

Patient satisfaction questionnaires
Patient opinions on their interaction with health professionals have become a focus in the quality assurance of undergraduate and postgraduate training. While a number of questionnaires have been developed, there is a relative paucity of data available on their use. Although subjective, these assessment tools are part of our consumer society, where patients are the ‘clients’ with greater access to healthcare knowledge from the various media. Their perceptions of a trainee’s performance are often as the ‘external lay reviewer’ in assessment of professional and education progression. But the degree of success as a learning tool is pivotal on the trainee attitude and willingness to accept feedback from a non-specialist, that is their personal insight and humility to accept the critique.

As for the number of patients’ opinions necessary for attaining assessment reliability, the data is rather inconsistent. Various studies have quoted numbers ranging from 16, 25 or even 66 patients to approach a generalisability coefficient of 0.8. Equally the category of patient is an added variable, for instance, children may be too idiosyncratic to give a reproducible result (e.g.
15 paediatric patients, generalisability coefficient of 0.36; Crossley et al., 2005). Moreover the precise educational impact of patient satisfaction questionnaires for changing trainee practice and behaviours remains to be thoroughly evaluated, for example the CARE questionnaire (Mercer et al., 2004).

**Doctors’ interpersonal skills questionnaire (DISQ)**

The doctors’ interpersonal skills questionnaire (DISQ) is another example that consists of 12 items exploring the patient–doctor interaction to be used in formative feedback; around 25 patients are thought adequate for test reliability. The reluctance of trainees to engage with the process owing to anxiety is common with novices in all professions. Therefore supervisors have an obligation to be sympathetic in their management of trainees’ self-confidence issues. From the patients’ standpoint, they too may be disinclined to give feedback to their doctors, resulting in the recurring problem of finding sufficient numbers of willing patients to take part. Here are some practical points to facilitate the use of patient feedback, these include:

- To engage colleagues of all grades in the area of practice so the trainee does not feel singled out.
- Availability of patients to give feedback – in practice areas where cognitive competence is a problem, it is useful to gain feedback from carers.
- To have agreed exclusion criteria for selecting patients.
- To explain to patients the reasons for obtaining feedback in order to gain their trust and engagement with the process.
- To ensure adequate security in the access to questionnaires and subsequent storage of data.

Flexibility in the organisation and clinical contexts for patient feedback is needed for its success as well as reassuring the trainee that this is a formative process to inform practice of all professionals and grades.

**Appraisal**

Appraisal is defined as ‘the action of assessing’ and ‘a formal assessment of an employee’s performance’ (OED, 2006). It has become a widespread practice in healthcare education. Appraisal is typically less frequent than some of the other types of feedback discussed earlier. aiming to provide a bird’s eye view of the achievements an individual has made and helping to set goals for the following interval of time, yearly in some contexts. Appraisal should ideally be a holistic review of an individual’s professional, educational and personal needs.

Appraisal conversations are based on a slightly different format to the other feedback conversations discussed so far. The process begins with a
Feedback and assessment

Purposeful self-assessment by the learner; the individual appraises their own performance and are frequently required to write a reflective statement to illustrate their thoughts. The self-assessment exercise usually reflects the range of the individual’s professional responsibilities, educational activities and their progress against any pre-determined goals. A substantial part of the appraisal is about the appraisee’s own healthcare practice and therefore their documentation tends to reflect this: feedback from patients, colleagues and students being an increasingly integral component. The next step is to hold a meeting, typically one-to-one, adopting the principles and guidance that have been discussed earlier in the chapter. The output of the appraisal is to construct a mutually agreed, personal development plan (PDP) which allows the appraisee to consolidate and develop their practice.

The purpose and form of appraisal has evolved over the years. The initial expectation, particularly in medicine, was that this process would be predominantly formative, to support reflective professional learning. Although elements of this remain, an increasing emphasis has been placed upon the summative purpose of the process, using appraisal to guarantee that a practitioner is providing ‘good’ and up-to-date care. For the medical profession, appraisal is now the route to re-validation and a requisite for the GMC licence to practise (GMC, 2010). This explicit shift from formative to summative purpose is likely to shape what doctors perceive as relevant to completion of appraisal. There is an important distinction between the goals set by an individual for ‘learning’, perhaps following constructive discussion with their appraiser, and ‘performance goals’ which are likely to be recorded by individuals when their performance is to be ‘assessed’. As with any assessment, once predominantly summative, the individual involved is likely to produce a strategic performance to best fit the expected outcomes in order to achieve ‘success’ and gain favourable judgement about their practice or competence. This contrasts with the potential for more sensitive, but informative, discussions which might occur between peers within the boundaries of a formative conversation. In this latter context, the individual is likely to be prepared to set goals which produce a greater challenge to their learning and support formation of new knowledge and understandings (Eraut, 2006), risking failure and negative judgement. After reading the other chapters in this book which cover what makes a good assessment, that is reliability, validity, cost effectiveness, and considering the research evidence about self-assessment (Gordon, 1991; Tracey et al., 1997), you might wish to draw your own conclusions about the authenticity of appraisal as an assessment method, particularly in a high stakes process. Whilst evidence supports its formative role in facilitating professional development, some question its combination with summative purposes and the effectiveness of blending of the two roles.
Feedback on assessment

One very powerful and desirable form of feedback for learners is in response to their completion of assessment, either formative or summative. Those involved in education will hear the plea for this particular feedback time and time again and a range of trainee surveys hold testament to this fact. Providing feedback on assessment is not an easy area for many reasons, maintenance of academic standards and integrity of exam banks being commonly cited issues. However, there have been a range of interesting new developments in this area and this penultimate section will briefly give the reader some ideas about how they might develop this aspect of feedback for their own learners.

Making feedback an iterative cycle of assessment, beginning with formative and moving on to summative, is a helpful process for the learner. Initial, formative feedback is provided in an attempt to develop the learners’ work, giving them an opportunity to address areas where shortcomings have been identified and then re-submitting for a summative assessment. Opportunities integrated into any teaching and learning program which involve talking

Box 7.2 Ideas for conversations about feedback

Task: encouraging ‘feedback dialogue’

Students can be involved in reviewing their feedback comments and learn how to understand the feedback they have received. This can be done in a variety of ways:

- Peer-led discussions offer opportunities for students to learn from each other and share each other’s insights. Peers understand the student predicament best of all and they are often very good at providing a sensitive and well-tuned commentary.
- Self-assessment or learning loops (Petty, 2011) are performed when students identify a weakness in their work and set a target about what to achieve on the next assignment, and can be rolled out indefinitely.
- Tutor-led discussions about feedback comments will let both learner and teacher construct a better understanding of comments and why people sometimes say different things.

All these sources allow for a bigger picture, a wider range of feedback data to work with and comparison against standards.

All of these activities will facilitate self-assessment/meta-feedback skills.

Petty, 2011
Box 7.3 Structuring feedback on assessment.

Structured feedback

- Overall impression and grade.
- Feedback on areas you’d asked us to look at.
- Our feedback about what worked well.
- Our feedback about what lost you marks.
- What you might want to do next time.

Developments in information technology have also been harnessed effectively in providing feedback on assessments. Pod casts giving students individualised feedback, quizzes with standard answers and assessments with individualised profiles about where a student has done well or poorly, for example multiple-choice questions covering a range of topics, can be broken down into specific areas, such as anatomy, physiology, use of medicines and so on, allowing students to receive feedback about their performance in each particular domain. Online peer- or tutor-led feedback discussions, blogs and wikis have all been used with good effect. What follows next is a specific example of an online assessment and feedback used in medical education.

Case of the month

Case of the month is an example of an assessment with both instantaneous and delayed feedback on assessment. The modules are online, hosted by a virtual learning environment, and deal with authentic clinical cases that students are likely to encounter once they are qualified (see Figures 7.8 and 7.9 for an example). They require students to answer multiple-choice questions about the scenario as well as to write a free-text response about the
Case 1 | Part 2

You are a Foundation Year (FY1) doctor working in Medicine for the Elderly in hospital. Shortly after you arrive for your day, you review the list of patients under your Consultant. One of the new inpatients under your care is a 91-year-old man who was admitted from a nursing home at 07.45h.

A photograph of Mr Smith taken just before his 90th birthday

Below is the letter that accompanied him on admission:

Dear Dr X

Re: Mr Alfred Smith, D.O.B 14/05/1918

Thank you for admitting this 91-year-old man with vascular dementia, previous CVA, and signs of a chest infection, who currently lives in Fairtree Nursing Home.

I was asked to review him after he became generally unwell and developed a temperature of 38.8°C and a productive cough. On examination, he is confused and has bronchial breathing throughout the lung fields with widespread crackles. He has some dependent oedema. I suspect that he has either pneumonia or heart failure.

Many thanks.

Yours sincerely

Dr O. O. Hours

Figure 7.8 Case of the month scenario 1. With thanks to Deborah Gill, Alison Sturrock and Steve Rowett for permission to reproduce this case.

diagnosis and management plan of the case. Both get an instant response, for the MCQs a pass mark is given, and there is a model answer for the text response which appears by the side of the student’s submission. The free text comments are subsequently assessed by an online tutor who then provides feedback and clarification about any outstanding practical, professional and theoretical issues that the student’s answer warrants and provides a holistic approach to feedback.
Case 4 | Part 7

Mrs Peters remains unintelligible and is very distressed. George, who is managing the case, is concerned by the tenderness in her abdomen, as since arriving in resus 60 minutes ago her abdomen has become rigid. Her pulse is now 110 bpm and BP is 100/60mmHg. He makes a tentative diagnosis of internal organ damage and arranges for a focused abdominal sonography for trauma (FAST scan – click for more info on FAST scans).

Click here for an article on investigating abdominal trauma: http://www.bmj.com/cgi/content/full/336/7650/938?view=long&pmid=18436949

The FAST scan shows intraperitoneal fluid and so an urgent abdominal CT is arranged. Below is one of the images:

A verbal report from the on-call radiologist says: "As you can see in this image it appears there is a tear in the spleen and although it is not ruptured it is bleeding quite substantially. There is no other obvious source of bleeding”.

Activity 7

It has now become obvious that this woman needs a laparoscopy and/or laparotomy. What are the issues to do with the risk of emergency surgery in this case?

Your answer:

Figure 7.9 Case of the month scenario 2. With thanks to Deborah Gill, Alison Sturrock and Steve Rowlett for permission to use this case.

Conversations can shape how individuals set their own goals

In this chapter, we have discussed some of the theories and contextual factors that influence feedback. In particular we have discussed some of the tensions that can occur with its use in healthcare settings, the context in which we give feedback in healthcare situations and the main considerations for both the learner and those responsible for initiating feedback conversations. We have discussed theory; principles and models of feedback; and then dealt with two specific sorts of feedback: 360° multi-source feedback and appraisal as illustrations of methods gaining significant importance in clinical education. As you reflect on this chapter and its relevance to your own roles and working contexts, Figure 7.10 may provide a useful summary of the key points to consider when giving feedback.
We hope that these principles, practical suggestions and models will give you a framework for developing your own feedback, or even *feed-forward*, conversations. Like all skills, practice makes perfect.

**References**


GMC Revalidation – the way ahead General Medical Council London 2010


How to assess students and trainees in medicine and health


Further reading


Chapter 8  **Psychometrics in medical education assessment**

*Dr Katherine Woolf*

UCL Medical School

**Introduction**

By adulthood, most of us have been assessed many times, whether it is a spelling test in primary school, the long jump on sport’s day, a school examination such as a GCSE, a driving test, or a specialist postgraduate examination such as the Membership of the Royal College of Physicians (UK). Testing plays a hugely important part in our lives, but the impact of the same test can vary massively depending on how it is used. For example, doing a number puzzle on the commute in to work is for pleasure, whether or not you solve the puzzle is relatively unimportant and is unlikely to have a significant effect on your life. However, if that same puzzle were used to measure numeracy in a job selection process, the consequences of failing would be more significant. Being turned down for a job because you did not solve the puzzle might lead to concerns about the quality and fairness of that as a test, and might lead you to question whether:

- You had been given sufficient opportunity to prepare;
- Stress adversely affected your performance;
- Other candidates had prior notice of the testing;
- The test had any errors in it;
- The test was overly complex or difficult;
- Another test would have been better at demonstrating your level of numeracy.

This chapter looks at how psychometrics can be used to assess whether a test is a ‘good’ test; what elements of a test can make it flawed; and what can be done to maximise ‘goodness’ and minimise ‘badness’. We will
concentrate on summative tests, where the way in which the test results are interpreted and used has significance to the candidate, to those responsible for its administration, and to society in general.

**Psychometrics in educational assessment**

The word, ‘psychometrics’ makes most people think of the long personality profiling questionnaires that are used to select recruits for the police, the civil service or other large organisations. Personnel selection is one important, and hugely profitable, application of psychometrics. But more generally, psychometric theory is applied to understand and evaluate any test that measures the working of the human mind, including educational tests and assessments (Furr and Bacharach, 2008; Rust and Golombok, 2009). An understanding of the concepts and theories that underpin psychometrics is therefore necessary to understand how to design and evaluate educational assessments and how to minimise the problems that arise in testing.

**What is psychometrics?**

To understand modern psychometric theory, it is useful to know the context in which psychometrics was first established. The use of tests to evaluate human ability and potential for personnel selection is thought to have originated in ancient China over 2000 years ago, but in the 19th century, rapid advances in science led to the advent of the scientific measurement of human thought, psychology, and to its close relation, psychometrics.

It was Sir Francis Galton who coined the word ‘psychometry’ in an article published in a 1879 edition of the journal *Brain*:

> Psychometry, it is hardly necessary to say, means the art of imposing measurement and number upon operations of the mind. (Psychometric Experiments, 1879: p.149)

Strange though it may seem now, the idea that one might quantifiably measure the ‘operations of the mind’ was groundbreaking in 1879. In 1879 psychology was barely even recognised as a discipline. Sigmund Freud was a 23-year-old medical student in Vienna, and Wilhelm Wundt had just set up the world’s first experimental psychology laboratory in Leipzig. Although scientists were starting to measure human reaction times, measuring thoughts and thinking was yet to be done in a scientific manner. But Galton was obsessed with measuring everything, especially to do with the human body and mind, and devised innovative and sometimes eccentric methods
to do so (Brookes, 2004). Although the methods Galton\(^1\) describes in *Psychometric Experiments* do not stand up to modern scrutiny, the idea underlying those methods – that by measuring human behaviour we can understand the human mind – is still the basis of most scientific psychology and, in many ways, the basis of educational assessment.

Galton’s use of the word psychometry (the process of measuring the mind), differs to the definition of psychometrics I gave in the first paragraph of this section (psychometrics is the science of evaluating psychological and educational tests). This is because the word psychometrics has both meanings. For example, the occupational and clinical psychologist Coaley (2010), suggests that ‘psychometrics are designed to do measurement; in fact, the terms is an abbreviation for “psychological measurement”’. While academic psychologists Furr and Bacharach (2008) write: ‘psychometrics is concerned with evaluating the attributes of psychological tests’. In this chapter, ‘psychometric’ or ‘psychometric testing’ refers to tests that measure mental processes or attributes, and the use of those tests, respectively. Thus, ‘psychometric’ or ‘psychometrics’ refers to the features or properties of those tests and to the evaluation of their use. The word ‘test’ is also commonly used ambiguously. In this chapter, it refers to any tool used to measure mental processes or attributes, including examinations and educational assessments. An ‘item’ is a test question and both terms are used interchangeably.

---

*A psychometric test* measures an aspect of human thinking or human psychology, and produces a quantifiable output. Most educational tests and assessments can be considered psychometric tests.

*Psychometrics* can be defined as BOTH:

The process of administering a psychometric test, including an educational assessment, and interpreting its results

AND

The science of evaluating psychometric tests, including educational assessments.

---

\(^1\)Francis Galton is considered the father of psychometrics but his important legacy has an unfortunately nasty flavour. Set alight by his cousin, Charles Darwin’s, work on evolution, Galton became an enthusiastic eugenicist, championing the use of psychological and physical tests to make decisions about controlling human reproduction to ‘improve’ the human race.
Are educational assessments psychometric tests?

Traditionally, psychometric tests were used to measure personality or intelligence (IQ). Personality tests usually consist of pages of statements that candidates rate on a 5 point (Likert) scale to indicate how much they agree or disagree with them. IQ tests usually consist of pages of short questions with multiple-choice answers, and candidates have to select the one they think is correct. It is easy to see how this type of test is similar in many ways to the ‘single best answer’ (SBA) or ‘best of five’ multiple-choice questionnaires used to measure knowledge at medical school or in postgraduate examinations.

But what about tests of performance? Can a practical examination like an objective structured clinical examination (OSCE) in which a candidate is marked on their ability to break bad news sensitively, or a 360 degree appraisal where colleagues rate one another’s competence and attitudes, be considered psychometric tests?

The answer is yes, they can. If a test is measuring an aspect of human behaviour that relates in some way to the workings of the mind (i.e. to psychology), and the test has a quantitative numerical output or score, the test can be considered psychometric, using Coaley’s (2010) definition, or, using Furr and Bacharach’s (2008) definition, the test can be subject to psychometric evaluation. Educational tests or assessments tend to be psychological in as much as they measure a person’s ability to recall, recognise, analyse or synthesise information; to prioritise, make decisions, act on those decisions and so on. As such, any educational test with a quantitative outcome can be considered a psychometric test, or can be subject to evaluation using psychometric methods.

Psychometric validity and reliability in assessment

Tests provide measurements, but a measurement is only useful if it relates to something interesting in a predictable manner. Unlike physical attributes such as height or weight, psychological or mental attributes are theoretical and can only be inferred from test results (Raykov and Marcoulides, 2011). The challenge of providing accurate and consistent measurements is therefore particularly great for educational tests, and the goal of psychometrics is to evaluate how well this is achieved by measuring the reliability and validity of the result of that test. This section explains in detail what is meant by reliability and validity, and outlines the various ways in which they are evaluated. It is followed by a section looking at threats to the validity of a test result, and ways in which those threats can be minimised.
What is reliability?
The reliability of a test result is an estimation of its reproducibility or consistency. Reliability is one of the cornerstones of the scientific method (Downing, 2004). Technically, reliability is the statistical similarity of one measurement compared to all possible measurements of the same thing; the greater the similarity, the greater the reliability (Crossley, Humphris and Jolly, 2002). To be able to demonstrate reliability of educational assessments is extremely important when the stakes are high. For example, Part I of the Membership of the Royal College of Physicians (UK) (MRCP (UK)) examination is an entry qualification for training as a physician (internal medicine) in the UK. The fees to take the examination are around several hundred pounds, and only those who pass can continue in that career – so it is high stakes. It is also tough: only about 50% of candidates pass first time (see http://www.mrcpuk.org/Results/Pages/ExamPassRates.aspx [accessed Nov 2012]). The examination has several sittings (diets) per year. To prevent candidates who sit the exam twice or more having an advantage, and to prevent cheating, most of the questions in each diet are different from those in the previous diet, with about a third of questions coming from a range of past examinations. Imagine a candidate who has failed Part I MRCP(UK) by 5 marks at the first diet; they might wonder whether they would have passed if they had sat the next diet of the exam and perhaps had an easier set of questions. To ensure equivalence across diets and counter such challenges, MRCP(UK) uses robust statistical methods. (see http://www.mrcpuk.org/SiteCollectionDocuments/EquatingPart%202.pdf [accessed Nov 2012]).

Reliability is necessary for validity, in that a test result cannot be considered valid if it is not reliable. However, reliability is not sufficient for validity. It is possible for a test to reliably give you a particular result, but that result to have no meaning in the context you are using it in, and therefore be lacking in validity. For example if students with A level English always achieve higher marks in Biology essays, this may be because they are better at writing essays rather than because they know more Biology. The test result would be reliable because you would get it consistently, but not valid because you are trying to test knowledge of Biology, not English writing ability.

What is validity?
A test result is considered to be valid if it reflects what it is designed to measure. This simple statement hides a considerable amount of complexity in practice. We saw at the beginning of this chapter with the number puzzle example that many questions can be asked about a test to assess whether it
is measuring what it is supposed to. The answers to these questions, and to related questions, provide evidence of validity.

Many introductory textbooks talk about different ‘types’ of validity, such as content, concurrent or face validity; however it is now commonly accepted that there is only one ‘type’ of validity, and that is construct validity, and the other ‘types’ are just ways of establishing construct validity. Psychometrics comes from psychology, and in psychological terms, a construct is a specific theoretical attribute of the human mind (Furr and Bacharach, 2008) such as intelligence, personality or depression. Because they are theoretical, constructs cannot be directly measured in the way that physical factors such as height and length can, so instead must be inferred from measurements of overt behaviour. Construct validity of a test can be defined as an estimate of how well that test measures the theoretical construct it is intended to measure.

Messick (1995) has been extremely influential in validity studies, has emphasised the fact that establishing validity is actually about gathering evidence, and defines validity as:

An evaluative summary of both the evidence for and the actual – as well as potential – consequences of score interpretation and use.

So although you will often hear people say ‘this test is valid’, in fact it is test results rather than tests that are more or less valid, and validity depends heavily on the context in which the test is taken, who it is taken by, and how the results are used. The validation of a test is a constantly changing and evolving process. One must gather evidence to make a judgement about whether the test result is valid enough to be acceptable and defensible to use in a particular situation for a particular purpose (Messick, 1995). Because validity involves judgement, when the testing stakes are very high there is great pressure to provide large amounts of high quality evidence as to the validity of the results.

**Reliability** is a mathematical estimate of how replicable and consistent a measurement is. Reliability is necessary but not sufficient for validity.

**Validity** is the evaluation of how well a test measures the theoretical attribute, or construct, that it purports to measure. Validity relies on the mathematic relationship between test results, but also on careful procedures and judgement. All validity is essentially construct validity.
Establishing and evaluating reliability
The reliability coefficient is used to establish the reliability of a test, and calculating it which requires data. A new test should be piloted on a sample from the population you wish ultimately to test (it is important to select a pilot group who are as representative of the population of interest as possible). By calculating the item statistics, you can identify the test items that improve the test’s reliability and those which lower it. The poorly performing items can be replaced or re-written to improve the overall test reliability. The same process can be used when creating a new test using some old items.

Reliability coefficients and item statistics are generally calculated using methods based on classical test theory and the reliability of test items can be calculated using item response theory. Whilst it is not necessary to go into the details of the equations used because calculations are almost invariably done by computer, it is important to understand and be able to interpret the test statistics.

Classical test theory
One of the first things we learn in science class at school is how to take a measurement. For example, to estimate accurately the length of a table, we are taught to measure the table several times and take the average of those measurements. The more measurements are taken, the more accurate the average measurement, that is the higher its reliability. The principle underlying that method is the same principle underlying classical test theory (CTT) and its close associate, the theory of true scores. Classical test theory (CTT) is very commonly used in educational assessment and, despite its limitations, it remains a useful tool (Coaley, 2010; Kline, 2005).

CTT indicates that no measurement can ever be perfect and thus all measurement contains error. The theory of true scores gives us an equation for this statement:

\[ X = T + E \]

where \( X \) is the observed score, \( T \) is the true score and \( E \) is the error term.\(^2\)

Crucially, the error term in this equation is random. It is caused by inconsequential factors, including distractions, small administration errors and so

\(^2\)It should be noted here that the concept of a true score is firmly rooted in the positivist scientific paradigm. Many social scientists believe that there is no such thing as an objective truth. They believe truths are socially constructed and are therefore relative to their social context. By contrast, psychometrics is a quantitative discipline based on the idea that there is an objective truth, which we strive to measure using the most accurate tools possible.
on, which are unrepeatable and non-systematic (Raykov and Marcoulides, 2011).

Because measurement error is random, it can lead to an observed score that is either randomly higher or lower than the true score. If a person takes the same test an infinite number of times, sometimes the error will inflate their score and sometimes it will deflate it, but on average the error will be null with all of the measured error terms being normally distributed around that average (null) error score. The theory of true score equation can thus be solved by replacing E with zero, resulting in an average observed score that is equal to the average true score. So taking a test an infinite number of times, according to CTT, means the average test result will accurately reflect the average true score on whatever construct the test is measuring.

In practice it is not possible to take a test an infinite number of times, but you might want to take the test a number of times to gain an accurate estimate of the true score. But how many times is enough? Two, five, ten, one hundred, one thousand? Often it is not even practical to take a test more than once, particularly with a short enough gap between testing to counter any other changes such as learning which may alter a candidate’s true score. Luckily, as Kline (2005) points out, because the error in the theory of true scores is random, the mathematics works just as well if a single person takes a test an infinite number of times as if an infinite number of people take the same test once. In the latter case, however, the equation changes, so it is the variance in the observed scores, true scores and error that is used rather than the mean score (the variance being a measure of the spread of scores).

\[ \text{VAR}(X) = \text{VAR}(T) + \text{VAR}(E) \]

Therefore, if a test is given to a large number of people, then according to CTT, their average observed score will give a good estimation of the true score on that test. This is very useful because it is much more practical to do this than to have one candidate take a single test several hundred times (Kline, 2005).

**Assessing the reliability of a written test result**

Although textbooks often refer to different ‘types’ of reliability, in fact these are usually just different methods of obtaining data and calculating reliability coefficients. In classical test theory, a reliability coefficient is an estimation of the similarity between the observed score and the true score, or more technically, the proportion of the observed score variance that is shared by the true score variance (de Champlain, 2010; Kline, 2005). Because a reliability coefficient represents a proportion, it ranges between 0 (the test result is no different to a random test result) and 1 (exactly the same result would be
obtained with repeated testing). The nearer a reliability coefficient is to 1, the more reliable the test result. Typically, tests that yield a reliability coefficient of lower than 0.7 are considered too unreliable to use in summative assessments; however, as with all psychometrics, the context is all-important. In high stakes medical education assessments, a coefficient of 0.9 is often required for the test to be sufficiently reliable to know that those who failed the exam were not just unlucky and those who passed were not simply fortunate (Downing, 2004; Tighe et al., 2010).

One feature of test reliability is that the more items you have in a test, the larger the reliability coefficient and the more reliable the test. However, the longer a test, the more tired the candidates will become, or the more they will rush through items. This may affect the validity of the test because even candidates with a lot of knowledge may perform poorly. This shows how reliability is not sufficient for validity, and how test designers need to strike a balance between competing elements. Various methods of obtaining reliability coefficients are outlined below. These are for use with written tests, as performance tests require different type of reliability calculations and we will come to these later.

**Test re-test**

This is where the same test is given to the same group of individuals after a short interval. The two test results may be correlated by using an intra-class correlation coefficient (Bland and Altman, 1996). The test re-test method can be used to assess the reliability of personality tests and other measurements of constructs that do not change significantly in the short to medium term. However, use of the test re-test method in educational assessment is impractical.³ It would be like assessing the reliability of a tape measure by

---

³ Some educational tests are designed to be administered several times. There are many situations in which this can occur, for example when the number of students is such that it requires the test to be administered to two groups of the students on separate occasions. It may be because the purpose of the test is to measure change in learning over time. For example, some medical schools use progress testing, that is the same test is given to students in every year – they are expected to perform fairly badly at the start of medical school, but as their learning increases, so does their test result. Another similar example is when students are given a test pre- and post-teaching to measure how much they have learned. Another situation in which tests can be taken several times is when candidates are allowed to sit the same test until they pass it. This is the case with Royal College Examinations and the driving test. A large proportion of candidates fail the first time they sit the Membership of the Royal College of Physicians examination, and they are allowed to take the test again, when hopefully their knowledge will have increased. The important point here is that in all of these situations, multiple administrations of the same test are not being conducted to measure the reliability of the instrument, but are being used for some other purpose.
repeat measurement of the height of a growing child a month apart – a difference in the two measurements is more likely to be because the child had grown than because the tape measure was giving an unreliable result.

In educational terms, with the same test given a month apart, it is likely that candidates will achieve a higher mark the second time, either because they have learned how to take the test, or because they have learned more about whatever construct the test is measuring. It is therefore not feasible to assess the reliability of educational tests using the test re-test method.

**Parallel forms**
In the parallel forms method two different but equivalent forms of the same test are created that measure the same construct, but use slightly different questions. Although this can reduce problems due to candidates rote learning the answers of specific test items (or obtaining the answers via cheating), the parallel forms method does not get over the problem that candidates could have learned more about the construct by the time they take the second test, not to mention the administrative complications and cost of administering a test twice. Luckily it is possible to assess reliability using methods that only rely on a single test administration.

**Split-half**
The split-half method of assessing reliability is effectively the parallel forms method, but instead of giving two tests at two different times, two equivalent tests that are half as long are combined into a single test, or to put it another way, a single test in which all of the items are measuring the same construct is split into two equivalent halves. The correlation between the two halves of the test is calculated (to give its reliability) and then, using the Spearman–Brown formula, it is extended to give an estimate of the reliability of the whole test. This method gets around the practical and theoretical problems associated with two administrations of the same test.

**Internal consistency**
In the split-half method, a single test is split into two tests. A more sophisticated method is to calculate the relationship between the results from all possible ways of splitting the test into two halves. In effect, this is how a test’s internal consistency is calculated (Downing, 2004). Internal consistency is the most commonly used measure of reliability in educational assessment. It is a measure of how similar the items are to one another: if a test has high
internal consistency it is presumed that the items are measuring the same underlying construct reliably (although see Schmitt, 1996, for a clarification of the difference between homogeneity and consistency).

The most commonly calculated internal consistency statistic is Cronbach’s alpha. Cronbach’s alpha is a generalisation of the Kuder–Richardson formula 20 (KR 20) for dichotomous items. So on a test of true–false items, the KR 20 rather than Cronbach’s alpha will be calculated. However, they are very similar to all intents and purposes. A good assessment will sample different aspects of the same trait. Cronbach’s alpha’s coefficients range from 0 to 1, a higher the score indicating a more reliable test. The statistical derivation of these methods is basic common sense and works on the test re-test concept: if a cohort of students is given a test, and then the same or equivalent test is re-administered, then provided the students have not in the interim gained or forgotten information pertinent to the test, the scores attained at the later sitting should be about the same as the first.

**Standard error of measurement**

The standard error of measurement (SEm) is a useful psychometric statistic that allows us to calculate confidence intervals around a test score, and thus estimate how much a person’s observed test score represents their true score. SEm is calculated using the following formula:

\[
SEm = \text{Standard Deviation} \times \sqrt{1 - \text{Reliability}}.
\]

where the standard deviation is the spread of candidate scores, and Cronbach’s alpha value is typically the index of reliability. SEm is used to estimate the accuracy of the pass mark, although there is evidence that it may also be a better measure of test reliability than alpha under certain conditions (Tighe, McManus, Dewhurst & Mucklow, 2010).

**Item analysis**

Since the reliability of a written test depends on how the items relate statistically to one another, a low reliability coefficient will usually be caused by poorly-performing items. Item analysis allows the items that are decreasing test validity to be identified so they can be removed or re-written. The most sophisticated way of doing this is by using item response theory; however it is possible to calculate useful statistics based on classical test theory.
Item difficulty

An important statistic is the item difficulty or p-value, which is the proportion of candidates who achieved a correct score on an item (Kline, 2005). For example, if 360 people attempted an item and 200 of those answered it correctly, the p-value for that item would be $200/360 = 0.55$ (this p-value should not be confused with the $p =$ value that tells you whether a statistical test result is significant). Multiply a p-value by 100 and you get the percentage of correct answers. As such, p-values range from 0 to 1. Harder items have p-values of closer to 0 and easier items are closer to 1. Items with a p-value of 0.5 are the best at discriminating between candidates, as half get the right answer and half get the wrong answer.

If items with a p-value of 0.5 are best at discriminating between people, one might intuitively think that all test items should have a p-value of 0.5. However, as we have seen, reliable tests have items that correlate with one another. Kline (2005) argues that if all items in a test are highly correlated and all items have p-values of 0.5, the whole of the test will not be any more discriminatory than a single item, and 50% of people will pass. As such, it is useful to have a series of items with different p-values, but with an average p-value of 0.5, in order to differentiate effectively between candidates. This also enables test designers to start their tests with easier items to make candidates feel confident, reduce their test anxiety, and thus improve validity.

Another situation in which p-values of 0.5 might not be so desirable is when the test is criterion-marked, in other words, the pass–fail decision is based on the candidate meeting a pre-defined cut-score (e.g. achieving a score of 70/100 to pass). In those situations, it may be better to select items whose average p-value is near the cut-score (e.g. 0.7) (de Champlain, 2010). Therefore the decision to be made is whether to have a criterion-referenced test (as is the case in the driving test, for example where there is a required standard that you pass) or a norm-referenced test (where an agreed percentage of the group pass) or some combination of the two. This is a complex matter than cannot be fully covered here (see also Chapter 6 on standard setting).

Item discrimination

If it is shown that candidates who do well on most items (who have a high overall test score) tend to perform badly on an item, but lower scorers tend to do well on that item there may be a problem with the item. To estimate the ability of each test item to discriminate between those who achieve high and low scores overall, it is necessary to calculate the correlation between the item score and the overall test score excluding that item. If the items are
dichotomous (binary), as they are in multiple-choice questions when an item is either correctly (1) or incorrectly (0) answered, the point-biserial correlation coefficient is calculated. If the items give a continuous score, such as in a personality test where each item is scored on a 5-point scale, the product-moment item-to-total correlation coefficient is calculated, which gives similar output.

Just as in any correlation, the coefficient ranges from $-1$ to $1$, with zero indicating no relationship between the item score and the test score. On average, one would expect high test scorers to achieve high scores on each item. This would be indicated by a positive correlation of 0.2 or greater (de Champlain, 2010). A negative correlation coefficient for a particular item would indicate high test scorers tend to do badly on that item. This may mean there is an error, either in the scoring system, (for example if the correct answer is marked as B when in fact it is C) or in the way the item is written.

**Reliability of performance assessments: inter-rater reliability and generalisability theory**

Communication skills, clinical judgement, the interpretation of evidence and professionalism are all important attributes or skills that a doctor needs to have. They are also difficult to assess in written examinations and are instead assessed by having candidates’ clinical performance rated by one or more examiners. Performance assessments can be with real patients in real clinical settings, such as in the mini-CEX (mini clinical evaluation exercise), or in much more structured simulated environments, such as in the OSCE.

OSCEs pose particular problems for reliability analyses, which we will come to in a moment. However, in general, all assessments that rely on a person observing and rating another person’s behaviour are less reliable than examinations that are marked by computer. This is because human perception is a top-down process, influenced partly by what the person is expecting to see. It is also because in assessments of real clinical encounters, there is not always a clear right or wrong answer, and the examiner or rater must use their judgement, which is, by necessity, subjective and varies from rater to rater.

In performance assessments, the main source of measurement error is in the differences between raters (Downing, 2003). This error can be estimated in several ways to create a reliability coefficient for the overall assessment. The simplest way is to estimate the percentage agreement between raters, but this does not account for chance agreement, and so will over-estimate the
reliability. The kappa statistic is a good alternative, as is the intra-class correlation coefficient, but Downing (2004) considers the best way of estimating the reliability of rater judgements is to use generalisability theory (GT). Generalisability theory is an extension of CTT (Downing, 2003). It uses the statistical analysis of a variance test to estimate how much of the overall test result is accounted for by various different factors, including the raters. From this, a generalisability coefficient can be calculated, which estimates how well the raters represent all possible raters and thus how reliable the measurement is.

OSCEs are more standardised than mini-CEXs, and therefore the error due to variance between standardised patients, and differences in marking schemes is, in theory, lower. However, OSCEs are problematic for reliability analyses because the unit of analysis, instead of being the item (question), is the case. As we have seen, the more items there are in a test, the more reliable it is. In an OSCE there are usually between 12–25 cases, which is very few compared to a three-hour multiple-choice examination with 300 items, and therefore the OSCE reliability will be much lower. The reliability of an OSCE is often called case specificity, which is how much each specific case tested relates to the universe of possible cases that could be tested. For instance, a candidate may perform well in an OSCE station measuring advanced life support, but does that have any bearing on their ability to break bad news, or interpret a chest x-ray? There is some debate about how much variability in an OSCE is due to case specificity (and indeed exactly what case specificity means – whether it due to context or content specificity (Norman et al., 2006)); however, using generalisability theory, one can estimate how much of the measurement error in an OSCE comes from various different aspects of the test, including cases, items, standardised patients and raters (Downing, 2003; Norman, Bordage and Page, 2006) – see also Chapter 9 for an exploration of the effect of examiner bias.

Problems with classical test theory
One of the problems with classical test theory is that it is sample dependent. Imagine you give a test to a cohort of Year 3 students at your medical school. Then colleagues at a different medical school asks if they might give your test to their Year 3 students. If for some reason the average ability of the Year 3 students in their school is lower than in your school (or they have not covered the necessary material), and the test results are analysed separately, the same item on the same test would have a higher p-value in your school (indicating the item is easier) than in their school.

Another problem with CTT is, because the error is random, it is assumed to be identical for all scores, whereas in fact very high or very low scores are
less accurately measured because in a normally distributed sample of scores, there are fewer high and low scores than there are scores in the middle of the distribution (De Champlain, 2010). Although in practice CTT is still a useful tool (Coaley, 2010), item-response theory is starting to be used to get around those problems.

**Item response theory**

In contrast to CTT, item response theory (IRT) is concerned with items rather than the overall test score. Using IRT it is possible to estimate reliability based on the difficulty of the item, the ability of the individual (similar to the true score in CTT, except it is a property of the candidate rather than the test), the power of the item to discriminate between high and low ability candidates, and other factors such as the likelihood of guessing the correct answer (Baker, 2001).

IRT has advantages over CTT in that predictions about item performance do not depend on a particular sample of candidates, and a candidate’s predicted ability is not dependent on the particular items the test is made up of. This means that performance on, say, Royal College examinations can be compared across cohorts, so long as all tests contain a sample of common ‘anchor’ questions. However IRT often assumes that performance is predicted by a single construct – candidate ability – which may not be the case in OSCEs for example. Both De Champlain (2010) and Downing (2003) have written useful briefings on IRT, how it compares to classical test theory, and how both are used in medical education, which interested readers might want to refer to.

---

*Classical test theory* is the idea that no measurement can be perfect, and thus has random error associated with it. Any observed test result consists of the true underlying result (the candidate’s true ability or true score) plus random measurement error. This is reflected in the theory of true scores equation $X = T + E$, where $X$ is the observed score, $T$ is the true score and $E$ is the random error of measurement.

*A reliability coefficient* is the statistical estimation of the amount of overlap between a candidate’s observed score and their true score, that is an estimation of the random error of measurement associated with a test. The larger the coefficient, the greater the overlap, the less error and the more consistent the test result.
Generalisability coefficients are calculated using generalisability (G) theory, which is an extension of classical test theory. G theory allows the estimation of error variance due to various facets or aspects of assessment that can be especially problematic in performance testing, for example, inter-rater reliability.

*Item response theory*, like classical test theory, is based on the idea that individuals have a true score on the construct the items in a test are measuring, but in IRT the true score is called the latent ability and it is a property of each candidate, rather than a property of the test. Unlike CTT, IRT is concerned with items rather than the overall test score. Using IRT one can estimate reliability based on the difficulty of the item, the ability of the individual, the power of the item to discriminate between high and low ability candidates, and other factors such as the likelihood of guessing the correct answer.

**Establishing and evaluating validity**

Assessing validity can be extremely complex. Downing (2003) provides a useful summary of the main sources of evidence to assess the validity of a medical educational test result (see also Cook and Beckman, 2006). In brief, this involves asking questions about:

1. **The content the test items are covering**, for example medical schools finals are designed to test whether a candidate has the knowledge and skills necessary to perform on day one of their first clinical job as a doctor. Does the test measure performance in all of those knowledge and skills? Or only a sub-set?
2. **How candidates respond to the test** (the response process) for example what evidence is there about the quality assurance process that ensures the fairness and accuracy of the marking process? Is the correct scoring key being used? Have examiners been properly trained?
3. **How the test items relate to one another and the overall test result** (internal structure) for example how difficult or easy are the items for candidates who generally score high or low? How does performance on one item relate to performance on a similar item, or a different item?
4. **How the test results relate to other variables**, for example how similar (or different) are the test result to the results of other similar (or different)
tests, either those that are taken at the same time, or that have previously been taken, or will be taken in the future?

5. The consequences of testing for the individual candidate as well as for society, for example are the results being used in a fair and equitable manner?

To answer many of Downing’s questions, test designers and administrators need to think carefully about the processes involved in writing and delivering their test, and to keep accurate records of those processes. However, psychometrics is a scientific discipline, and it demands numerical evidence as to the validity and reliability of a test outcome. This section will look at number four in the list above to explain how and why we measure the statistical relationship between test scores and other scores to establish predictive and concurrent validity. We will talk more about the other aspects of validity in the section about threats to validity.

Evidence for validity of a test – predictive, convergent and discriminant validity

To gain a place at medical school, applicants are required to take all sorts of tests. They need to take the right school examinations and a special medical aptitude test such as the BioMedical Admissions Test (BMAT) or the UK Medical Admissions Test (UKCAT), they have to show they have experience of working with people, and many also have to sit through an interview. Medical schools put their applicants through such a rigorous selection procedure because doctors have a privileged and responsible position in society, and it’s important to select people who are going to be able to fulfil that role. The idea is then that by measuring performance on school leaving examinations, aptitude tests, interviews and so on, we can predict whether or a candidate will be able to learn the knowledge, skills and attitudes required to be a doctor. How well each of those measures predicts future performance is the predictive validity of those measures.

The usual method of establishing the predictive validity of a test is to calculate the correlation between performance on the test and performance on measures of the same construct taken at a later date. For example, to know the predictive validity of GCSE examinations for medical students, you might measure the GCSE grades of a cohort of medical school applicants, and then correlate their GCSE performance with their first year medical school performance. Obviously, this will only express how well GCSE grades predict first year medical school examination grades, not how well they predict final year grades, or communication skills or professionalism, or any other factor that might be of interest. But then why should GCSE grades predict communication skills? Just because a candidate can pass a written
examination does not mean they can talk to patients. As with any assessment of validity, when assessing the predictive validity of a test, clarity on the precise construct the test is measuring is essential. Generally a simple correlation of $r = 0.4$ to $0.6$ indicates a moderate statistical relationship between two measures and provides evidence of predictive validity (assuming that the two measures are thought to measure the same construct).

Another way of assessing validity is by measuring the relationship between scores on tests that measure dissimilar constructs. This is called divergent (or discriminant) validity. For example, the results of a newly-designed written test of biochemistry knowledge might have correlated highly (converged) with results of other tests of biology and chemistry knowledge, but less well with results of an OSCE measuring communication skills. In an early and influential paper, Campbell and Fiske (1959) proposed a sophisticated method for calculating convergent and discriminant validity of a test called the multi-trait multi-method approach. They argued that, because measurement methods create error, it is necessary to gather measurements of the same trait (construct) using different methods, and each of the methods must also be compared by measuring several traits (constructs). The resulting correlations between traits measured using different methods is viewed in a multi-trait multi-method matrix, which clearly highlights which traits and methods have high correlations (indicating convergent validity) and which have low or even negative correlations (indicating discriminant validity).

**Threats to validity: systematic bias**

Imagine that a medical student who always scores in the top 5% of the year, achieves a low mark on the final medical school knowledge test. Why is this? Because there was something wrong with the student? Or because there was something wrong with the test? This section will cover various factors that can call into question the validity of a test result, and explains what can be done to reduce their negative impact.

Classical test theory explains that no measurement can ever be perfect, so every educational test result will contain some measurement error. This measurement error is random – sometimes it increases scores and sometimes it decreases scores, but on average, scores are unaffected. However, sometimes test scores are skewed and do not reflect the true ability, knowledge, skills, or aptitude of candidates. There are many possible reasons for this. For example, on an end-of-term history test, if we increase the number of questions about facts that were not covered in the syllabus, students will achieve lower marks. Importantly, those low marks will not accurately reflect the knowledge students have learned on the course because the test is measuring knowledge of information they were not required to learn. This type
of error leads to unwanted bias, which is unrelated to the construct being measured by the test (knowledge of history covered in the syllabus). The error is also systematic in that it is predictable (as opposed to the random error of measurement). Systematic error can be caused by a number of factors in test design and administration (Raykov and Marcoulides, 2011).

**Construct under-representation**
If a test does not cover everything it is supposed to measure, this can lead to biased results. For example, MRCP (UK) is designed to measure various aspects of competence related to internal medicine. If it only asked questions about gastroenterology, the construct (competence in internal medicine) would be under-represented by those items. This is what Messick (1995) calls *construct under-representation*. Details of how to create tests that cover all aspects of constructs is outside the scope of this chapter, although interested readers are referred to Crossley, Humphris and Jolly (2002).

**Construct-irrelevant variance**
If an examination is scheduled for 2am then students’ scores are likely to be lower than their true ability. This bias is systematic and predictable, and the source of error is unrelated to the construct being measured but contributes to the observed score. It is therefore called *construct-irrelevant variance* (Messick, 1995; Downing and Haladyna, 2004).

Construct-irrelevant variance can be person-specific or group-specific (Haladyna and Downing, 2004). The example refers to person-specific construct irrelevant variance, because it depends on something about the individual taking the test. Another example would be if non-native speakers perform more poorly on a mathematics examination because the questions are written in a way that requires a good command of English. An example of group-specific construct-irrelevant variance would be if students’ essays are marked by one of two examiners. One of the examiners is a very harsh marker, so everyone in that group gets a lower mark. Another example is if candidates sit one of two papers, which are supposed to be of equivalent difficulty, but in fact are not, so that the candidates taking the harder paper achieve lower marks (Downing and Haladyna, 2004).

Obviously the first step to minimising construct-irrelevant variance is to have a clear definition of the construct that the test is designed to measure (Rust and Golombok, 2009) because without defining the construct, it is impossible to know what is construct-relevant and what is construct-irrelevant. A proper definition of the construct is also required to minimise construct under-representation, as described earlier. Once the construct has been defined, there are a great many factors that can cause bias due to
construct-irrelevant variance, and test designers as well as those administering tests have to be wary to minimise them. Downing and Haladyna (2004) have produced a useful taxonomy of construct-irrelevance error sources. I have reproduced the main areas that relate to medical education, have included ways to minimise bias associated with these, and go into further details about many of them later in the text.

**Exam stress**
This is also referred to as test anxiety and can be a considerable problem. To minimise it, students can be given examples of past exam papers to practise, and tips on anxiety reduction in exam preparation. In practical examinations, assessors can be trained in the importance of putting candidates at their ease. Although medicine can be a stressful job, unless the examination is designed to measure ability at dealing with stress, and the examiner reacts the same way to all candidates, stress caused by examiners will lead to construct-irrelevant bias in exam results.

**Errors in exam papers**
To err is human, and it is always possible for an exam to contain errors. In 2011 Ofqual (England’s regulator of qualifications, examinations and assessments) investigated errors that appeared in school examinations. These included a geography exam with a graph showing the velocity of a river as 0.5 instead of 0.05, a biology exam in which all the answers to a multiple-choice question were wrong, and a maths exam in which students were asked to solve an equation without being given the information needed to do so (see http://www.guardian.co.uk/education/2011/jun/09/exam-board-watchdog-questions-errors [accessed Nov 2012]). Careful proofreading of questions is vital to guard against such errors.

**Badly written examination questions**
Badly written questions are a source of construct-irrelevant difficulty, in that they tend to be harder than well written questions and cause more students to fail (Haladyna and Downing, 2004). As any question writer will know, it can be extremely difficult to write a good question (see also Table 8.1). The details of good question writing are too lengthy to go into here, however, interested readers are referred to the National Board of Medical Examiners guidance by Case and Swanson (2002), who have provided guidance for clinical item writing; and Haladyna, Downing and Rodriguez (2002), who reviewed the literature for evidence that supported or refuted their taxonomy of multiple-choice item writing guidelines for use in general educational
settings. Schuwirth and van der Vleuten (2004) have also written about which formats are appropriate in which medical educational contexts. Whichever items are chosen for a test, their quality and effectiveness should be assessed using item analysis.

**Guessing**

Much of the construct-irrelevant variance we have talked about thus far is systematic error that causes under-performance, that is the bias is negative. Positive systematic error, or *construct-irrelevant easiness* (Messick, 1995), can occur if a test that is written in a way that makes it easy to guess the right answer, because it will tend to give people artificially inflated scores.

Item statistics can be used to identify items that might be subject to guessing – for example if a candidate who generally scores poorly on a test (to put it in CTT terms), or someone of low ability (to put it in IRT terms), achieves a high score on a particular item, it may be because it is easy to guess the right answer. This may be because of “cueing”, where something in the way the question is written gives away the answer, for example:

- The words in the question stem are replicated in the correct answer or;
- One of the answer options contains many more words than all the others.

Confidence-based marked is sometimes used as a way to try to prevent guessing. Candidates give their answer and then rate how confident they were that the answer is correct. Correct scores with high confidence ratings score the highest and incorrect scores with high confidence ratings score the lowest. There are concerns that negative marking leads to poorer performance (Betts *et al*., 2009) and strategic test taking, despite random guessing not really being much of a problem because it tends not to lead to high scores (Downing, 2003). Corrections for guessing have therefore rather fallen from favour in medical education.

Positive systematic error may also occur if a test is has a very socially desirable answer. For example, in a survey that asks how much you donate to charity, a respondent might report that she gives more than she really does, making her observed score on that item higher than her true score. The Marlow–Crowne social desirability scale is designed specifically to measure socially-desirable responding. It includes items such as “I’m always willing to admit it when I make a mistake” (the socially desirable answer being “true”) and “I never resent being asked to return a favour” (the socially desirable answer being “false”) (Crowne and Marlowe, 1960). If scores on the scale are high, scores on other test items given at the same time should be interpreted with care as they may be artificially inflated.
Table 8.1 Sources of construct-irrelevant variance in assessment and ways of minimising them (from Haladyna and Downing, 2004)

<table>
<thead>
<tr>
<th>Category</th>
<th>Instances</th>
<th>Minimisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test preparation</td>
<td>Whether students get test preparation and, if so, its extensiveness</td>
<td>Give students sufficient time to prepare, providing examples and past papers, teaching exam technique. This should be done ethically, i.e. without giving too many clues as to the correct answers (see cheating).</td>
</tr>
<tr>
<td>Test development</td>
<td>Item quality</td>
<td>Train question/item writers. Perform item analyses and remove poorly performing questions/items.</td>
</tr>
<tr>
<td>Test item format</td>
<td></td>
<td>Use formats that have evidence to support their validity in the context they are being used, e.g. when designing multiple-choice questions in tests of medical knowledge use single-best-answer questions rather than multiple true–false questions.</td>
</tr>
<tr>
<td>Differential item functioning</td>
<td></td>
<td>Assess item performance in different groups that are presumed to be equal in ability, and reject items that discriminate between the groups.</td>
</tr>
<tr>
<td>Test administration</td>
<td>Location of test site</td>
<td>Ensure the site is accessible for those taking the test.</td>
</tr>
<tr>
<td></td>
<td>Altering the administration</td>
<td>Write standard operating procedures for administering tests to ensure uniformity. Provide sufficient time to ensure weaker students are not disadvantaged.</td>
</tr>
<tr>
<td>Test scoring</td>
<td>Scoring errors</td>
<td>Use machine marking when possible. Always check item performance.</td>
</tr>
<tr>
<td>Test form comparability</td>
<td></td>
<td>Use statistical test equating procedures. Create an item/question bank that includes statistics on the performance of those items in different groups and use to create tests from items that perform similarly and are measuring the same construct.</td>
</tr>
<tr>
<td>Rater severity</td>
<td></td>
<td>Train raters, including giving them feedback on how harshly or leniently they are scoring. Monitor inter-rater reliability when possible. Standardise marking forms.</td>
</tr>
<tr>
<td>Accuracy of passing scores</td>
<td></td>
<td>Use an evidence-based method for standard setting the pass mark.</td>
</tr>
</tbody>
</table>
**Cheating**

There is likely to be at least some cheating in all assessments, including those at medical school and in postgraduate medicine. Cheating can include copying another candidate’s answers during an exam, accessing unauthorised information during an exam (e.g. bringing in notes), gaining access to exam papers prior to the exam, candidates being given too much information by teaching about the test prior to the exam, and plagiarising another’s work in a submitted assignment. In fact, one of the difficulties with cheating is in defining it (Franklyn-Stokes and Newstead, 1995).

According to a survey study of 31 US medical schools, medical schools’ codes of honour have a small effect on reducing cheating behaviour, although they do not prevent it (Baldwin et al., 1996). In their August 2010 newsletter, the Royal College of Physicians (UK) explain how they are combating cheating in their written examinations by increasing the ratio of invigilators to candidates, giving invigilators specific zones of the room to focus on, and

---

**Table 8.1 (Continued)**

<table>
<thead>
<tr>
<th>Category</th>
<th>Instances</th>
<th>Minimisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students</td>
<td>Influence of verbal abilities on test performance</td>
<td>Train item writers to write plainly. Review items for comprehensibility and re-write if necessary.</td>
</tr>
<tr>
<td></td>
<td>Test anxiety, motivation, fatigue</td>
<td>Try to reduce student anxiety during test administration. Give students practice examinations. Give students breaks between tests.</td>
</tr>
<tr>
<td></td>
<td>Accommodation for special student populations</td>
<td>Ask students in advance whether they need special provision for e.g. dyslexia. Ensure provisions are available and encourage students to use these if required. Train test administrators or examiners in how to respond to students with special needs.</td>
</tr>
<tr>
<td>Cheating</td>
<td>Individual and institutional</td>
<td>Keep test materials secure. Produce statements about acceptable and unacceptable behaviour in relation to testing. Encourage reporting of suspect behaviour. Use software to detect cheating and plagiarism.</td>
</tr>
</tbody>
</table>
increasing training to help invigilators report incidents that might be cheating (http://www.mrcpuk.org/SiteCollectionDocuments/Examiner_13.pdf [accessed Nov 2012]).

Computer software is increasingly being used to detect cheating, and the knowledge that such software is being used may act as a preventative measure. Many universities now use Turnitin, which compares submitted written work with ‘17+ billion web pages, 200+ million student papers and leading library databases and publications’ (https://turnitin.com/static/index.php [accessed Nov 2012]).

McManus, Lissauer and Williams (2005) developed ‘Acinonyx’ software to detect cheating in written multiple-choice examinations. The software is now used in various high stakes postgraduate medical assessments. In 2011, a doctor was struck off the UK’s medical register after being caught cheating by this software.

**Summary**

Educational assessment and testing is hugely important and particularly in medical and healthcare education, where the stakes can be very high. Therefore it is crucial that the tests used in medical education are of high quality, and that those involved in designing and administering those tests understand how to assess and provide evidence for test quality. Although assessors may be skilled in interpreting and assessing the quality of clinical test results, they are less able to assess the quality of educational assessments, and frequently misinterpret and misapply information about the quality of assessments (Cook & Beckman, 2006).

This chapter has provided an introduction to psychometrics in assessment with a focus on medical education. Psychometrics is the science that enables educators to assess whether assessments are useful and defensible. This chapter provided the historical and theoretical context for understand psychometrics. It explained how classical test theory – the idea that all measures have a degree of random error associated with them – is used to calculate the statistics that show whether a test produces consistent, reliable results and whether its results are valid in the context in which they are used. An overview of the more recently-developed item response theory was also given. The various methods used to calculate reliability coefficients and item statistics are explained, and readers were shown how to interpret these statistics. Based on the work of Messick (1995) and Downing (2003), readers were given five aspects of test construction to question (test content, candidate response, internal test structure, relationship to other tests and consequences of testing), to help assess the validity of a test result. Further details were given about assessing the predictive, convergent and discriminant valid-
Psychometrics in medical education assessment

ity of a test result. Finally, some major threats to test validity were explained together with tips on how to reduce their impact.

References


Case, S.M., Swanson, D.B. (2002). Constructing Written Test Questions For the Basic and Clinical Sciences, 3rd Edn. BME.


**Further reading**

Chapter 9 Examiners and examination boards

Professor Olwyn M. R. Westwood
Queen Mary, University of London

Introduction

The psychometrics of assessments (see Chapter 8) and ultimately the marks awarded denote the judgement by examiners on competence, for which in turn the candidates must be assured of their expertise and complete integrity. However the interpretation of assessment criteria is not always clear-cut, therefore procedures are needed for the standardisation of examiner practice. Without these it could be entirely feasible that candidates demonstrating the same competence in an equivalent performance, for a specific task, might receive very different marks. In this chapter, the various issues around the characteristics of an examination board, the examiners and the candidates are discussed along with practical approaches for identifying where errors might occur, the reasons why students may fail and how to support them to reach their full potential in assessments.

Academic quality indicators

Universities have local standard operating frameworks for examination boards that incorporate programme-specific requirements, but the generic principles of assessment apply. Transparency around the defined assessment criteria that equates to a measureable result is essential and must be shared with all stakeholders, that includes students, academics and examiners (local and external). But these definitions are among the more complex educational tasks and even the most straightforward descriptors need explanation and a discussion on the limitations. Therefore, clear instructions should be provided in advance to avoid an erratic performance by the examiners and
thus reduce potential marking errors. A common problem is in dealing with the personal (and sometimes unrealistic) expectations of standards, particularly when there are several examiners, for example with a clinical examination that is delivered and synchronised for multiple sites. Box 9.1 provides some characteristics for designing stable assessment criteria.

To obtain a reliable judgement of competence in a practical assessment, data needs to be collected from multiple sources (see also Chapter 4 on workplace-based assessments). Likewise, the marking criteria and rating scales need to be described precisely, and be related to the different assessment items, corresponding scores and domains. The implications for how the criteria relate to the assessment will vary according to the modality and the context, such as year of study, practical versus written. Thus any judgement and scores awarded require reference to:

- The areas of knowledge or skills rated by the examiner;
- How the final score is derived – some areas of the assessments have a higher weighting than others, for example:
  - the written examination score versus the in-course assessment;
  - in an OSCE station, the domains to test the chest examination skills of the candidate may be weighted higher than the initial introduction and request for permission to examine the patient;
- How the final score correlates with the level of competence measured;
- How the individual scores are used and contribute to calculating the assessment reliability.

**Marking and moderation of assessments**

Universities are expected to abide by the Quality Assurance Agency Code of Practice for Assessment of Students (see QAA 2006, 2011). There are various

---

**Box 9.1 Features of assessment criteria**

- Mapped to the learning outcomes
- Written in unambiguous style
- Appropriate for the assessment modality
- Easily accessible to assessors (and candidates without compromise of test security)
- Able to stand up to external scrutiny
- Quantifiable
- Reliable
- Valid
- Responsive to change
methods so clarity in details of the procedures used with supporting documentation must be available both to advise examiners and for scrutiny by external assessors, auditors and the students. This includes, for instance, specifics on how an institution expects their examiners to function in their role in double, blind double, and anonymised, marking.

Double marking is where the work is marked by two markers and blind double marking is where they mark independently and if there is a discrepancy between scores, there is a discussion to agree a final mark. Anonymised marking has become the norm in many institutions, where a candidate numbering system is used so their identity is hidden from the assessor. While it might seem like an equitable process, there have been reservations expressed by students and examiners alike. One of the drawbacks of anonymity is that it precludes personalised feedback that is afforded by prior knowledge of the student and their progress (or lack of it). In any case, such a system only works with certain written assignments for it is almost impossible to have an anonymised scheme when practical skills are being assessed.

Moderation is a regular practice, where a second marker is tasked with double marking samples of work to check for consistency of practice and diligent adherence to the assessment criteria. Usually the moderator will assess independently between 15 and 25% of the total number of assignments from across the ranges of pass marks, and all those deemed to fail. It is possible for there to be systematic differences of opinion between assessors where one marking consistently higher or lower than the moderator. While it is almost unachievable for the two to always agree marks within precision of less than 2%, discrepancies that amount to a difference in grade or degree classification require discussion. Where an agreement cannot be reached the work should be referred to a third internal marker or the external examiner for their opinion. Where the moderator sees only a sample of the work it is not advisable to re-grade any individual piece of work alone, as students whose work was not moderated may be unfairly discriminated against. Significant discrepancy necessitating re-grading should apply to all students sitting the assessment. If there is inconsistent marking between different students, this will require all the papers to be remarked by a senior examiner.

Compensation, condoned fails and mark capping
The QAA advises on assessment policy for the award of degrees and these regulations must be clearly written and accessible to all stakeholders. Most universities have adopted a credit accumulation system for their awards, with many using the Bologna Declaration recommendations, that is European Credit Transfer and Accumulation (ECTS) credits. For students who are on
the borderline or just below the pass score, there are three options available to examiners: compensation, condoned fail or referral. (The methods for agreeing a pass score were dealt with in detail in Chapter 6.) A student may be given a condoned fail when the failed element is not considered in determining academic progression – institutional regulations may permit only an agreed minimum number of condoned fails in an academic year or award. A compensated mark is where a student might have fallen just below the pass score (usually up to 5%), but when the mark is included with scores of other assessments, the final mean score is deemed a pass and so academic progression is granted.

In instances where the student has been referred and because the score was an outright fail, regulations normally permit a second attempt. However, pass marks awarded for the re-sit are generally capped to the minimum pass score. This regulation is to deter students from ‘failing’ strategically some modules in order to gain a better score at the second attempt. Re-assessment carries its own problems for assuring equivalence of assessment with smaller numbers of students and while, with some awards, candidates may only expect to re-sit the elements they failed, others may have in their regulations the expectation that all elements should be re-taken, for example if they fail the written examination, students may or may not be expected re-take the OSCE and the written assignments. In the case of postgraduate awards, there may be stringent regulations for progression to the research component or dissertation. These might be the time frame between passing the taught modules of the programme and/or level of academic attainment in these elements. For example those to do not perform well in the taught component may be offered a postgraduate diploma rather than attempting independent research.

**Mitigating circumstances, complaints and appeals**

Occasionally an event might adversely affect a candidate’s ability to complete an assessment to their full potential, or they may require an extension to the assignment deadline. The mitigating circumstance could be an acute illness, a relapse of a known chronic disorder, a bereavement or personal tragedy which should be explained in the local documentation that is submitted for consideration by the Examination Board. If a student is taken ill during an examination then obviously this is grounds for mitigation. However, where institutions have a ‘fit-to-sit’ policy, a candidate who is attending is signifying they are capable and competent to take the examination, deciding later they were not on top form is not grounds for mitigation at a later date. Therefore students who feel unwell should be advised not to attend the examination.
In any examination period students are particularly stressed and are more likely to submit a complaint or grievance if they feel their performance has been adversely affected by the content of the assessment or environment in which it was taken. The quality assurance of examination content is easily addressed by:

- Coding all curriculum learning outcomes and cross-referencing them to the respective test items and blueprint (see also Chapters 3 and 6);
- Thorough proofreading of the papers in advance.

Conversely the environment is a more volatile commodity and some incidents may be more difficult to control, such as a fire alarm, temperature of the examination hall or level of noise in an OSCE circuit. Issues that are outside the control of the university also require consideration in an action plan for dealing with potential problems, such as a breakdown in the transport network or adverse weather conditions. Likewise, the impact of the changing demographics of the student population require thought – many mature students have dependants that make demands on them, and in turn they compel institutions to make suitable adjustments for their learning and assessment needs.

There are some instances when students might wish to appeal against the Examination Board decision; this may be disagreement over the outcome of the assessment in an individual module, or the decision regarding academic progression. The more common reasons for appeal are against judgements of:

- Being deregistered from the programme;
- Being advised to re-take the whole academic year;
- Disputes over the final degree classification.

**Role of the external examiner**

An external examiner (EE) is an academic from another institution whose role is to provide an objective view of the assessment marking procedures, and to assure fairness of standards in their delivery. Their function is central for verifying:

- Equivalence in academic standards of assessment in relation to the quality of student performance;
- The compliance of a programme, where appropriate, with the national qualification frameworks and professional regulatory bodies.

The nomination of an EE is a multi-layered process that involves identifying an academic with experience and expertise in the discipline, and confirming no conflict of interest prior to authorising their tenure. These are university appointments, normally for three years with the possible extension of an additional year, which affords continuity between the out-going and in-coming EE, and mentorship of new appointees.
Universities vary in assessment methods, so an explanation of the rationale for their use in the context of the programme and learning delivery will avert any misunderstanding prior to the EE receiving work to scrutinise. Some institutions offer annual EE training seminars, while others simply provide a mass of paper-based information in advance. The expectations of the EE vary, but all require generic local information on the programme-specific examination regulations.

Effective communication between the EE and programme administrators is pivotal to a smooth and effective process, as the duties are likely to occur at an equally busy time in their own institution. Obviously contextualising the assessments and signposting busy periods of EE work is a basic obligation, that is:

- Agreeing examination paper content/coursework titles.
- Diarising important dates for:
  - examination board dates;
  - receiving work to scrutinise and to turn around their marking;
  - delivery of EE reports.

Their role is often as a moderator to scrutinise examples of coursework and examination scripts of all pass grades and all deemed as a ‘fail’ as quality assurance of local examiners’ marking, that is opinions gained from an independent judge. The role of an EE is critical, particularly for the case where a student appeals against the Board decision. They should also be viewed as an accessible adviser on issues relating to curriculum delivery and assessment. Their dialogue with local examiners, either formally (or informally at the examiners’ dinner), need not be an ‘us-and-them’ approach, for we all want to deliver best practice. Equally the EE meeting with students should also be seen as constructive for enabling academic developments.

**National steer in the UK for the external examiner role**

With the expansion in the number of UK universities, public accountability has resulted in the EE function coming under increased scrutiny. *Universities UK* with *GuildHE* have recommended key principles in support of national criteria for EE appointments, with minimum expectations for their induction, lines of communication and confidentiality issues. Some of the proposals are obvious, such as appointment of first time EEs to an existing team for mentorship and clear action plans for dealing with areas of concern. The *Quality Assurance Agency* (QAA) is also unambiguous in its steer over academic rigour and EE’s responsibilities. A system is being introduced whereby EEs who believe their concerns have been persistently ignored by the institu-
tion can alert the QAA. Such action is not considered a justifiable reason for terminating their appointment. From academic year 2012–13, all EE reports are expected to include a judgement on student learning opportunities and its enhancement. It has also been mooted that a newer model for having an independent EE panel might replace the current structure, for appointing academics who are ‘known’ may not necessarily be perceived as impartial.

**Examiner behaviours**

The principal decision made by an Examination Board on a candidate reaching the final examinations of a professional award is: ‘Do they have enough specialised knowledge and skills to be fit to practise in the chosen profession?’ This has to extend beyond conceptual knowledge and skills (‘knowing that’) into competence in skills and knowledge synthesis (‘knowing how’) for safe clinical practice. These academic decisions are prone to variation, but where the consequences are disciplinary action, it is even more important that equivalence and the appropriate procedures are seen to be applied. It is freely acknowledged that ‘hawk and dove’ activities present, and examiners may or may not be aware as to which category they belong to. This is not a new phenomenon, Osler (1913) wrote in the *Lancet* (as cited by Harasym *et al.*, 2008) that some examiners were reluctant to fail any student, however dismal (i.e. Molluscoids), but others were equally reluctant to pass any student, (i.e. the Metallics). Even in Schools which endeavour to maintain best practice in their statistical accountability and reporting of assessment metrics, disparity in examiner behaviours can still go unnoticed.

**Examiner judgement process**

Knowledge and skills competences are tested using a number of modalities, and, generally, universities have a policy for anonymised marking. But with an oral or practical skills assessment anonymity is nigh on impossible. Central to an impartial process for awarding marks on performance, is a robust faculty development system to help diminish the risk elements of examiner bias. Indisputably this is a delicate area to negotiate with one’s colleagues. Therefore it is politic to assume best practice, and to include with thanks for their contributions feedback that is, (i) embedded and individualised for examiner development and (ii) generic on the overall examination process.

The global score (overall opinion) of an expert examiner on a candidate performance is a function of the methodologies used for gauging the appropriate pass score. In the traditional OSCE station, assessors score a
series of discrete and specified activities against the marking schedule to provide an objective and a precise written record of the perceived performance. The circuit of different OSCE stations with multiple assessors is considered more reliable than the traditional ‘long case’ and oral examinations of a previous generation (hence the powerful reason for no longer using the latter). To assure satisfactory reliability and validity, a minimum number of stations are required, which de facto means using multiple examiners and sampling different skills in order to obtain a collective view on a candidate’s competence (Swanson, 1987; Newble and Swanson, 1988).

Some more senior examiners believe the check-list system of OSCE marking to be not making full use of their expert judgement. Indeed, they have a point, for the global judgement over and above the analytical raw scores derived from the sum of the individual items is believed to be a valid process for deciding the cut score (see also Chapters 6 and 8). Similarly, standardisation of examination materials should be a given for maintaining a consistent examiner performance. Nonetheless, how expert judges reach their conclusions about a performance is not necessarily consistent.

Comparable levels of concurrent validity (correlation with previous tests) and reliability are obtained from the analytical and global scores. Some assessors may tend towards greater stringency, particularly with ‘lower achieving’ or borderline candidates. One view is that perhaps certain expert examiners have unrealistic expectations of performance, particularly in the early years of a programme. In contrast, a novice examiner who is less self-assured may ask for guidance, hence an unambiguous explanation of assessment criteria is fundamental to help reduce these sources of error.

Workplace-based assessments are useful formative exercises on ‘real’ rather than standardised patient interaction, but enough time should be allocated for the activity to be observed and immediate and focused feedback provided on the performance (van der Vleuten and Schuwirth, 2005). It is possible that scores given may be open to interpretation owing to a reluctance to fail on account of familiarity. As with summative assessments preparedness is the guiding principle, with candidates assessed across a wide range of scenario complexities and using multiple assessors. In the evaluation of the mini-CEX, Norcini et al., (1997) agreed that the issue of familiarity could be a problem, but attested that as an adaptable assessment tool, across a breadth of different clinical settings, with multiple examiners (in various different scenarios), generally scores given were comparable (see also Chapter 4).

**Examiner severity**

A persistent struggle for anyone tasked with coordinating examinations is recruiting examiners with the appropriate training and experience to deliver
them. Although examiners can distinguish between different levels of competence, the raw scores given generally lean towards their personal levels of stringency, for example severe examiners awarding lower scores, and lenient examiners higher scores. In reality, knowing the candidate could have a positive or negative influence if scores of the performance in real time are adjusted because of their prior knowledge in a different context.

Examiner confidence and accuracy are not always well correlated with the borderline candidates. In the one-to-one interaction of an OSCE, examiners may be more cautious and mark in favour of the candidate (Tweed and Ingham, 2010). The instinctive shift in scores awarded can lead to wide variation in candidate scores versus their ability (high, intermediate and lower) to a point that some researchers have recommended that adjustments should be made when an examiner is known to be at the extremes of stringency or leniency. This is added support in favour of multiple examiners to augment examination reliability and validity that should resolve the issue of any shift in scores caused by extremes in examiner marking characteristics.

Scrutinising the assessment metrics by calculating inter-rater reliability is also valuable, for individual examiner ratings may deviate from their normal marking habits, for example a lenient examiner giving an extremely competent candidate a low score. It follows therefore that multi-faceted analysis as an indicator of examiner consistency goes some way to resolving these differences, together with gaining free text data to substantiate their judgement of the performance. These data can aid an understanding of the reasoning behind examiner ratings, and be a valuable resource for individualised feedback for candidates.

**Just what level of consideration is given when assessing a candidate’s work?**

We have all heard of the legendary student hoax to include a spurious sentence in the middle of an assignment, to test whether the examiner *really* read their work thoroughly. But working on the premise that examiners are diligent in their role, verbal protocol analysis has been applied to investigate the basis of examiner behaviours, to clarify and compare rater leniency/ severity. This is achieved by the examiner articulating verbally their cognitive processes to derive the score or decision on a performance which is recorded and subsequently analysed. Although much of the decision-making may parallel the assessment criteria (i.e. content, its synthesis and delivery), there are still individual behaviours where judgement might be affected. For instance, personal preferences for material included within an answer can lead to a disproportionate examiner decisions. A typical example might be: ‘*Great this candidate is the first one to include ****, therefore I will give them*”
some extra marks!’ For students who know their examiner’s ‘pet subject’ and think strategically, this is a potential tactic for success.

The legibility of handwriting is no longer such an issue when work submitted has been word processed. But the narrative, the word count, general presentation, the choice of grammar and vocabulary (i.e. orthography) can alter perceptions and, accordingly, influence rater judgements. Therefore anonymised marking might help to preserve impartiality.

Although it may not be deliberate, an inconsistent application of the assessment criteria results in poor examiner correlation, as do differences in its interpretation, particularly if the criteria are ambiguous. For this reason, the examiner briefing session with discussion to agree the expectations of candidates on precise assessment objectives is invaluable for resolving any discrepancies in the mark sheet and may limit the level of errors – a level of detail which is especially significant for ‘high stakes’ practical assessments. There has been some debate as to whether the novice marker is more accurate than the experienced marker, who relies on more intuition (Ecclestone, 2001). Then again, personal characteristics for marking by novice examiners are sometimes stricter than their experienced colleagues, another cause of inter-rater bias (Ruth and Murphy, 1988). Myford and Wolfe, have made a serious attempt at looking for workable strategies to reduce errors in marking and increase the quality of rater reliability (see Box 9.2).

Why do some students fail?

All students arrive at university with the potential to flourish, especially when taught by academics with a palpable enthusiasm for their discipline. It is recognised that pre-university grades are not good predictors of degree classification; neither is undergraduate performance an indicator for postgraduate success, for students mature at different stages of the academic journey. However, with the complexity in the different learning methods and forms of assessment used in medical and healthcare degree programmes, weaker students may flounder. Therefore a reliable learning infrastructure should be designed to support students in the recognition that in higher education they need to:

• Develop study skills around understanding concepts rather than simply memorising information, and;
• Seek out practice opportunities for skills development and look for immediate and constructive feedback.

There are a minority of students who give an academic performance that is significantly below their potential. Faculty colleagues may have identified
Box 9.2 Strategies to reduce errors in marking

(Taken from Myford and Wolfe, 2003, 2004)

1. Clearly defined descriptors/anchor statements for assessment criteria.
2. Peer observation of rater:
   (a) evaluation to provide self-awareness of undesirable marking traits
   (b) monitor behaviours and correct errors immediately.
3. Raters to grade candidates based on a pre-specified number of examinees per category.
4. Use multiple raters and average ratings for each candidate.
5. Minimise number of ‘check list’ items in a station to reduce risk of ‘rater-fatigue’.
6. Frequent changes of stations for raters to decrease risk of boredom.
7. Use raters from an appropriate background/medical speciality to foster accurate and fine discriminations of attributes.
8. Regular examiner training to improve/maintain quality of ratings.
9. Use statistical/psychometrics to fine-tune candidate scores for rater variation.

them (sometimes subconsciously), but without reliable documentation these concerns remain merely individual observations, and can delay access to academic support if they are not shared with colleagues. Even with the most robust personal mentorship or pastoral care systems, students may be reluctant to seek help.

Student failure is multi-factorial; extrinsic causes are associated with the learning environment, and intrinsic factors may be one or more of the following reasons: inadequate self-motivation, learning approaches that they find uninspiring, low levels of natural aptitude or a social agenda taking priority over studying. The impact of examination stress, anxiety and lack of sleep and mental health issues for young adults away from home for the first time should not be underestimated. But to sort out the problem it needs to be recognised, and then academic support can be initiated.

Ethnicity and gender

Whilst there appears to be no clear reason, evidence from a study of three London medical schools has shown that male students and those from ethnic
minorities frequently underperformed in written and clinical assessments compared with their Caucasian and female peers, (Woolf et al., 2008). These findings have provided tangible data on an intrinsic problem which was known anecdotally, and is not unique to medicine per se but rather a concern for all in higher education. One explanation might be that learning styles or culturally-related barriers manifest themselves as a reticence to participate fully in some clinical workplace learning activities. The consequence would almost certainly be relatively less experience in the application of clinical knowledge. In our multi-cultural community of learning, Schools need to be sensitive to these ethnic and gender differences when planning clinical training and assessment.

From another standpoint, where the candidate and simulated patient (SP) share the ethnicity, there has been subjective evidence to suggest a ‘halo effect’ where higher scores are awarded. But since SP contributions are generally minimal, this is not likely to distort the scores substantially and perhaps could be dismissed as an unfounded concern (Colliver et al., 2001).

**Learning needs and learning disabilities**
An insight into personal learning needs is generally a formula for success. However, some common reasons for failure may not be immediately noticeable owing to: (i) a lack of awareness of their own knowledge base; (ii) a revision plan that neglects unknown information and simply concentrates on information they know; and (iii) ‘question spotting’ as a strategy for revision. Significantly, central to these behaviours may be inadequate metacognitive skills, and some key warning signs that personal tutors and assessors should be alerted to are an inability to use technical vocabulary, or where study notes are learned and repeated verbatim, which would indicate a limited understanding of the material. There is a strong case for all Schools to have an academic support strategy that provides a mechanism for identifying students with a learning disability and workshops to encourage metacognitive skills development – these offerings may then lessen the risk of a poor academic outcome (Garrett et al., 2007).

**Specific learning disabilities (SpLD)**
Undiagnosed learning disabilities, such as dyslexia and dyspraxia, are possible causes of candidates not reaching their full potential. A formal diagnosis of a SpLD such as dyslexia is costly, and frequently it is during the first year of a higher education programme that they are finally assessed. Some schools offer SpLD testing as part of the academic support process for students who are unsuccessful in assessments. One reason might be the examination format
in universities being different from those used in secondary education. It is encouraging that, given the recommended study skills tuition and extra time allotted (normally an extra 25% for a written examination) students with a SpLD can perform as well as their peers. It has been implied, rather disparagingly that, with the alarming number of students presenting with a SpLD, some may ‘perform’ in the diagnostic tests in order to be afforded the benefits of additional educational support and extra time in examinations. But be assured, these scurrilous claims are without solid evidence to support them. What is known is that a candidate’s performance may differ depending on the assessment modality (Gibson and Leinster, 2011).

**Language competence**

Universities are keen to attract overseas students because they enrich the experience of the student population as well as the financial incentive. While students might be literate in the *written* language of their studies, their oral communication is not necessarily equivalent. Without minimum competence in oral communication, it is questionable whether they can cope with the rigours of a vocational programme within workplace learning and practical assessments. A language barrier might also explain lower assessment scores which are not a true representation of knowledge and other skills.

Language competence is a genuine concern at many levels; for the patients, carers and colleagues there are real issues around the potential risks to patient safety. Moreover an additional challenge can be when demographics and intolerance have formed part of a country’s politics that then thwart progress. Yet with greater acceptance of cultural and linguistic diversity policies, comes the demand for assessments to be offered in more than one language. With this compromise comes another hurdle of having multi-linguistic and discipline-specific local and external examiners to assure the validity and a standardised approach for assessing sub-populations of students. It is noteworthy that there are far-reaching implications when the validity of judgements made on performance, for example in skills-based or workplace-based assessments (see also Chapter 4) results in a lack of confidence in the assessment outcome of an identifiable sub-population of candidates.

**Students may be reluctant to seek help**

Students in higher education are often the high achievers, so to be struggling with their studies is uncharted territory and some may be loath to seek help, but without it any problems are perpetuated. With the new-found freedom of higher education some students flounder initially, with the absence of
parental encouragement, attendance not monitored as rigorously as at school, and an expectation for self-directed learning. The reluctance to admit their need for academic support may be due to one of a number of reasons.

**Inappropriate study skills**, where students may trust in their own ability to self-remediate without academic support; but this is dependent on them identifying the causes and finding a solution. For those lacking the intrinsic ability or insight into their personal learning needs, the going-it-alone approach is likely to be unsuccessful. Some students find higher education much more difficult because it demands a deeper learning and understanding of concepts, rather than passive learning to ‘recite information like a poetry’ at a later date, or surface learning as seen in students who simply cram for examination. Therefore study skills training in how to develop knowledge and skills application and synthesis should diminish the risk of failure.

**Insecurities** can present of being perceived by faculty or their peers as unable to cope with their studies or deal with personal problems. With some students the deterrent for seeking guidance is their fear of the negative impact that a temporary ‘glitch’ might have on their personal status and long term career prospects (Chew-Graham et al., 2003).

**Competitive behaviours** in a cohort are seen frequently, particularly with courses that recruit the high achievers who have been previously in the top quartile at school. Some can feel discouraged if they do not reach the level they aspire to, so withdraw rather than face up to their problems. Competitiveness can also be particularly hard on those who were recruited via widening access schemes, and who may feel overwhelmed by peers who give the impression of being confident and affluent.

**High levels of anxiety and stress** are common, especially in vocational programmes which have high practice-based workloads that make inroads on students’ time and then impacts on revision schedules. While a little anxiety can motivate, excessive anxiety and the fear of failure can paralyse the ability to study. Occasionally there are students who are conscientious in attendance and in submitting in-course assessments then, for no apparent reason, sabotage their chances of success by a sudden change in behaviour. Others can become dependent on specific faculty members, relying on them to ‘tell them what they need to learn for the examination’. In extreme cases, some forms of procrastination may have a medical basis that requires counselling, or treatment by mental health professionals.

**The nature of workplace learning** can be hard to navigate for relatively young or inexperienced adults. They need guidance in how to make the most
of their time in an environment where the service demands and patients have to take priority over student learning. Disturbing events witnessed may also leave them unable to manage their emotions as effectively. This is where a named tutor for pastoral support is essential to avert a crisis, be it academic or a depressive episode.

*Lack of support* can be frightening for younger students who are getting used to living away from home and without the immediate availability of family encouragement. Any general feelings of isolation can lead to psychological health problems, and where family expectations for a career are at variance with their own, the regular support mechanisms are not available owing to fear of reprisal and this can compound their feelings of loneliness. There has been anecdotal evidence that students in these circumstances have been relieved to be deregistered from a programme, feeling that the academics have made their decision and so taken the pressure off them.

*Lack of exposure or interest in their studies* may result in poor attendance at classes or placement opportunities owing to:

- A genuine lack of interest in their studies, for example demotivation because the programme does not meet their personal expectations;
- Their outside commitments to family or dependents, or the need to work to finance their studies;
- Imbalance between studies and their social activities which take precedence, for example the first year student who is an over-enthusiastic contributor to RAG week activities;
- Personal embarrassment resulting in an inability to operate effectively in the clinical environment. Sometimes this may be on account of their immaturity, or it could be a discord between their cultural norms, that is personal beliefs and ethnic principles, and the expectations of the programme;
- Placement tutors who make students feel unwelcome, often this is not deliberate but owing to pressures of service and targets that take precedence over teaching – so students retreat to the library.

**Steps to support students**

For some students who are struggling, failure comes as a complete surprise to them, albeit this reaction is not shared by the faculty who teach them! Steadfast systems in place to identify students in difficulty are necessary. However any unenthusiastic behaviours of faculty towards the students can be counterproductive. While a less able student may prompt a reaction of ‘more pity – less irritation’, it is frequently the opposite sentiment for those
with the ability but who demonstrate insignificant endeavour. Students need to be aware of their responsibility to be demonstrably active learners, particularly in a workplace setting, so that they gain the attention of busy professionals. Equally, clinical and academic supervisors need to be aware of the impact of their behaviours on student motivation, and faculty mentorship training to provide directives on how to facilitate active engagement. For example:

- To provide unequivocal and interesting learning experiences in practice;
- To encourage students to perform tasks within their personal scope of practice under direct supervision;
- To be empathetic and emotionally supportive, particularly if a student has had a upsetting experience, such as patient death or accident;
- To diarise meetings with students for general discussion regularly throughout the placement, even if there are no problems (Williamson et al., 2010).

Clinical educators are aware of their ethical, legal and professional responsibility to anticipate potential risks in order to avert harm to colleagues, the patients as well as the students in the clinical arena. For those on a placement that is remote from their main campus, access to support services in the face of gruelling timetables and workloads can prove challenging. Therefore a named mentor given at the start of an academic year and on clinical placement provides both structure and personal support. It is prudent for programme leaders to seek out and recruit mentors who actually like students, for they will tend to display an aptitude for identifying the warning signs of a student in difficulty. But again, academics and clinical tutors need adequate training in how to spot the tell-tale signs of a student who is floundering or suspected of having a SpLD.

Three systematic and rather obvious steps to support students who are not coping well with their studies are to:
1. Identify the problem – is it personal, social, academic?
2. Design an action plan that is personalised for the student.
3. Meet up with them regularly to monitor their progress, for example scrutinise their progress in formative assessments, and give immediate and focused feedback.

Monitoring student progress does not have to be arduous, it can happen without much effort through timetabled regular contact, for example in small group sessions. When they include compulsory assignments, it affords a natural opportunity to scrutinise student work. The use of multiple different learning methods and regular testing should support a self-awareness of learning needs in students which in turn can avert a suboptimal academic performance.
References


Further reading

Chapter 10 **Future developments in assessment and feedback**

*Professor Jane Dacre*
UCL Medical School

**OVERVIEW**
This chapter attempts to predict the future of developments in assessments and feedback. It provides insights into current developments and uses these to attempt to scope out what is likely to happen in the future. The term ‘assessment’ has been used to cover both formative and summative assessment activity. The approach has been deliberately high level, as much of the more practical aspects, and the specific detail, has been covered in other chapters.

**Introduction**
Even very clever people are unable to predict the future – as illustrated by Lord Kelvin of the Royal Society of London, who predicted in 1883 that x-rays would prove to be a hoax. Who would have thought that the mainframe computers, which occupied whole buildings in the 1960s, would evolve into smart phones which are almost universal and fit into a pocket? So predicting future developments in assessment and feedback is an impossible task. We can, however, take examples of past endeavours and use those to provide educated guesses about what may be in store. Other chapters of this book have provided insights into what is new in assessment, in relation to specific aspects related to the testing of knowledge (Chapter 2), competence (Chapter 3), performance (Chapter 4) and professionalism (Chapter...
5). Chapters have also covered the frameworks for assessment with the development of new practices and processes (Chapters 6 and 8) and evaluation and feedback (Chapter 7). In this chapter I will attempt to draw on current examples of innovation in assessment and feedback policy, practice and research, and use these examples as a basis for what is likely to become mainstream assessment practice over the next few years, and further into the future. It is a high level overview, drawing on the current challenges in the healthcare education environment and looking towards future changes to educational context and practice.

Healthcare reforms will inevitably influence the process of education of healthcare professionals. In the UK, the passing of the Health and Social Care Bill in 2012 has prompted the greatest potential change in healthcare delivery since the beginning of the NHS in 1948. President Obama’s emerging health policy in the USA will significantly alter access to care, which will influence healthcare provision. The recognition of the importance of the social determinants of health, the correlation of social gradients with wellness and life expectancy, the aging population, and the increased average age at death for men and women will shift the focus to prevention of long term conditions and the delivery of care in the community. The move to greater professional accountability and a better educated world population will encourage our patients to become more demanding of their clinicians, and more vocal about their healthcare choices. Our assessments will need to prepare students and trainees for this.

The delivery of healthcare is becoming increasingly expensive. High-tech treatments which prolong life expectancy and treatments for more and more trivial and non-life-threatening conditions are becoming available, at a time when the economic climate means that there is less money in the healthcare system. Recognition that resources are finite has led to new concepts in healthcare, such as rationing expensive therapies. It will also increase the pressure on the delivery of healthcare education. It is up to the educators to ensure that we maintain the quality of the assessments in order to maintain the high standards of the healthcare workforce.

The organisations in which healthcare education training and assessment is delivered are also responsible for leading research in the biological, clinical and social sciences. Research funding to support this endeavour is increasingly competitive, and funded projects are expected to be delivered more economically. Collaboration between departments and between higher education institutions is increasing, and there is an emphasis on applied research, with outputs that can be rapidly translated from bench to bedside and to the patient’s home. It is the responsibility of the educators to ensure that our training and assessments are able to keep students and trainees up to date.
As assessment is known to drive learning, this means that educators will need to be prepared to engage with the latest developments and be flexible enough to test any new translational concept, or innovation in genetics, nanotechnology, proteomics and other cutting edge research areas.

These changes provide the context in which the education training and assessment of clinicians will sit in the future, and will provide the environment in which we will be delivering our assessments and our feedback.

The remainder of this chapter will explore future developments in assessment and feedback in the context of this evolving healthcare environment. The focus will be on regulatory, academic, economic, sociological, technical and education research changes which are currently underway, and will influence the developments in assessment and feedback for the future.

**Regulatory change and its influence on assessment and feedback**

Healthcare professions are regulated in order to protect the public. Most healthcare practitioners are expected to be registered with a professional body in order to be allowed to practice. The overarching duties of a healthcare regulator are: to keep an up to date register of practitioners, to provide clear standards of practice, to oversee education and to deal ‘firmly and fairly’ with those who fail to demonstrate adequate standards of behaviour or practice. There is some variation internationally in the way these generic duties are implemented, but recent concern involving some high profile cases of healthcare professionals who have not demonstrated adherence to the established code of practice has highlighted the importance of what has become known as ‘right touch’ regulation. Future assessment and feedback methodology for medical schools and doctors is required to be scrutinised by the General Medical Council. This will introduce more uniformity in future assessment systems, and will encourage the sharing of good practice and innovations between assessing organisations.

**Standardisation of assessment practice**

The work of the UK’s General Medical Council provides an example of the regulatory changes which are likely to impact on the future of assessment and feedback of health professionals. Their document, Good Medical Practice, has outlined the generic standards expected of a doctor working in the UK. This document has provided a framework for the development of a comprehensive and consistent blueprint of the duties of a doctor. Several UK medical school and postgraduate assessments have adapted this blueprint in the development of their own assessments. The latest version of this document
will update the overall framework, which will be reflected in assessments being designed by medical education establishments. The guidance is updated regularly in a rolling timetable of review, which will ensure curricula and assessment systems are up to date with current clinical practice.

In the UK, the regulation of training curricula for undergraduate and postgraduate curricula and assessment systems has encouraged change, and has provided the opportunity for greater consistency in assessment standards and approaches. This is a welcome change, and is likely to continue to support effective change in assessment processes.

**National testing**
A few years ago, the then GMC Education Committee led discussions on the appropriateness of a National Examination for medical students in the UK. The consensus view at the time was that this was a step too far. Medical schools value their academic independence, and see the diversity in the culture in schools as a strength. Since that time, however, the external environment has changed. The quality of care received and patient safety have become paramount in NHS strategy. The old style diversity in the medical schools results in a gradient in the performance of doctors from different schools. Indeed research has shown this to be so, by taking graduates from different schools and comparing performance in postgraduate assessments (the MRCP and the MRCGP), and in preparedness for practice (McManus et al., 2008; Wakeford et al., 1993; see also Goldacre et al., 2010). The schools at the lower end of the gradient have questioned the validity of this approach while those at the top welcomed it. The GMC has reopened this debate and, as a result, the Medical Schools Council is piloting a sample of shared knowledge test questions to look at feasibility and to explore any differences. Future medical school qualifying examinations are likely to consist of at least a shared core of written test questions, and this could eventually lead to some kind of national qualifying examination.

**Stimulating innovation**
In the UK, the merger of the Postgraduate Medical Education and Training Board with the GMC has brought the regulation of all stages of medical education together under a single regulator: the GMC. This has prompted a fresh look at the regulation of medical education across all stages of a medical career. A review of the stages of medical education was carried out by Lord Patel in 2010. This highlighted some areas for innovation which are being taken forward.

In Europe, to achieve specialist status a doctor must achieve a Certificate of Completion of Training or CCT. This is a transferable qualification within
Future developments in assessment and feedback

the European Economic Area. It provides evidence of successful performance in local European assessment systems. For doctors coming to the UK from other countries, they must either pass the Professional and Linguistics Assessment Board (PLAB) test, or undergo an assessment process to be awarded a Certificate of Eligibility for Specialist Registration (CESR) or Certificate of Eligibility for GP registration (CEGPR). Both these assessment systems are being reviewed and will be updated in line with current best practice in assessment and clinical practice. These changes are likely to have a significant impact on international medical graduates in the future. Potential candidates will be expected to be successful in a more contemporary and robust assessment process.

There is currently disquiet about the appropriateness, legality and feasibility of English language testing of doctors coming to the UK who have qualified overseas. These two future assessment methods will need to develop innovative assessment instruments for assessing both language and cultural competence, in a valid and reliable way, which is in alignment with UK and European law.

The current successful documentation and publication of healthcare curricula has also highlighted some areas which will need to be addressed in the future development of assessments. There are some gaps highlighted by effective blueprinting of examinations, the so called ‘too difficult pile’, of areas which appear in curricula, but are rarely adequately assessed because it is difficult to write good quality assessment instruments. These areas include: the assessment of a team; the observed assessment of ethics; the assessment of complex activities which involve integration of several skills; and the assessment of leadership. The regulator expects the full curriculum to be assessed, and this requirement will stimulate the Schools to innovate in their approaches to both formative and summative assessments. The current healthcare curricula contain reference to these areas but, as yet, there are few published data or evidence based instruments to rely on. The future of assessment and feedback will benefit from developments in this area.

In order to be legally defensible, a written assessment is required to demonstrate validity and reliability (see Chapters 3 and 8). These concepts have been explored in other chapters. It is possible to do this by adherence to good practice in blueprinting, question design and standard setting. This is currently possible for large cohorts of students or trainees, using established statistical methods. The increased number of GMC approved small specialties in medicine has made it increasingly difficult to provide a robust measure of statistical reliability. Work is underway to produce valid and reliable assessment methods for cohorts of assessment candidates which have been traditionally too small to assess reliably. The need to demonstrate compli-
Re-assessment with regulatory standards of assessments has encouraged research in this area, and it is likely that new metrics will emerge. Recent research has been published supporting a new method of written assessment for very small cohorts. To achieve this, individual questions are piloted in volunteers to establish their psychometric characteristics. Mathematical modelling then allows a virtual cohort to be developed for comparison purposes. This method is now in use for the GMC tests of competence in the Fitness to Practise procedures, where single doctors are being evaluated.

Re-validation

Patients and the public expect their healthcare professionals to be up to date and fit to practise. Future providers of patient care will be expected to have undergone an assessment process which demonstrates this. The high level guidance for the local implementation of re-validation processes will be provided by the healthcare regulatory bodies. The GMC is committed to introduce re-validation for all doctors with a licence to practice who are on the UK Medical Register. Other regulators will likely follow suit. For doctors, this process begins at the end of 2012, and will roll out from then. Doctors will be required to provide evidence of their fitness to practise in a rolling five year programme. Assessment instruments have been developed which can provide performance metrics as an indicator of a doctor’s fitness to practise. These will include as a minimum data set: documentation of annual engagement in an enhanced appraisal process, with the completion of a personal development plan, multi-source feedback from colleagues and patients and evidence of continuing professional development (CPD). Novel assessment methods will need to be created and evaluated to enhance the current suite of workplace-based assessment instruments, to include, for example, the assessment of management and leadership and the assessment of teamwork.

The implementation of re-validation for all healthcare professionals over the next decade is likely to enhance standards of practise and will require profound changes in healthcare workers’ attitudes to lifelong learning and continuous assessment. Employers will need to embrace the necessary re-organisation of clinical service delivery to allow time and resources for the personal and professional development of staff.

Workplace-based assessments (WPBA) were discussed in Chapter 4. Although they have been implemented widely, there is a disappointing lack of published evidence for the effectiveness of these methods as a reliable and effective summative assessment tool. There is also scant evidence of the effects on improvements in patient care. This has led to the view that WPBA methods should in future be used as formative instruments. They are an
Future developments in assessment and feedback

extremely useful method of ensuring that a teacher or trainer takes the time to provide feedback to trainees about live observations of their performance in clinical practice. They are not however robust as stand-alone summative assessment methods. The GMC has now made recommendations for their use as supervised learning events (SLEs). Successful completion of a series of these will form the basis of discussion in a progress review. It is likely that the evidence base for the usefulness of WPBA will be developed further, as more educators pilot and publish new modifications to the instruments and their use.

Academic changes influencing assessment and feedback

The education of healthcare professionals has, in the past, been provided by colleagues who were less interested or less motivated to engage in traditional clinical research. This has resulted in a two tier system, which is particularly evident in some research intensive universities, where educators have not focused on the scholarship related to innovation in education and assessment practice. As a result, they have often felt under-valued, and have been overlooked for academic promotion. Progress in curriculum design, and in the development and evaluation of assessment and feedback methods, together with the publication of results of the evaluation of methods and the information that this provides to educational scholars, will provide a significant opportunity for educators to demonstrate their scholarship ability and to engage in ground breaking educational research. Robust assessment research is at the forefront of this development, and has been the focus of a recent increase in output in peer reviewed journals. The application of mixed method approaches to research is becoming more acceptable, as is the use of qualitative research methods. The inclusion of healthcare education into the Research Excellence Framework (REF) has meant that guidelines have been written to clarify what can be submitted as an educational research activity. This panel has developed an approach which is inclusive, and will encourage emerging departments to submit their research activity. The act of a REF submission will focus the activities of emerging research centres in healthcare education, and will support the improvement in the quality of submissions in healthcare education. In the future, I anticipate that education research will blossom, and research into assessment methods and outputs will lead the way. There are, however potential risks. Colleagues in research intensive institutions, who are not in a unit likely to make a REF submission, may feel that they are in a vulnerable position.

The design and delivery of assessment and feedback methods is becoming professionalised, and this trend will increase. Future healthcare educators
have a responsibility to ensure that this new knowledge is shared with colleagues through staff development activity. Indeed, the development of such courses on assessment and feedback should be the subject of robust and publishable evaluation in the future.

Future developments and innovations in assessment methods, and the outcome data they will provide, will need to be widely publicised to the academic medical education community. I envisage an increase in the number of conferences focusing on assessment and innovations in feedback delivery and an increasing emphasis on medical education and assessment in the traditional as well as new academic journals, both paper based and electronically.

**Economic changes influencing assessment and feedback**

The next few years are likely to be difficult for all public sector organisations, as the economic downturn takes its toll. This is likely to have a negative influence on the development of assessments, the developments in feedback and the ability to implement innovations in healthcare education in general. There is already some disquiet about the increase in fees for enrolment in healthcare courses at undergraduate level. The financial burden of assessment on postgraduate trainees will be added to the debt already accrued, and will be likely to influence the progress that can be made on the development and research side. Currently, postgraduate clinical training programmes are expected as part of clinical career development, but are not tax-deductable, and are expensive. The design and delivery of assessments is labour intensive, especially if they are to be produced to the high standards that we have come to expect. High fee paying trainees and graduates are likely to demand more for their money. This, combined with the change in culture around assessment in general and a more legalistic society, means that we are likely to have more appeals about assessment judgements and processes. This will result in the providers of assessments becoming more aware of the possibility of legal challenge and developing ever more comprehensive and defensive regulations.

Assessments for healthcare workers are funded as part of the monies given to universities via the Higher Education Funding Council for England (HEFCE; or equivalent for the other countries within the United Kingdom)-teaching budget allocation, and to the NHS via the Multi-Professional Education and Training budget (MPET). The component of this reserved for undergraduate medical education is known as the Service Increment For Teaching (SIFT). These assessment budgets are not ring fenced and, although they are very large, there is a lack of transparency around their use. As health-
Future developments in assessment and feedback

Care budgets come under strain, it is likely that in future there will be more pressure on the education budgets for undergraduate education and assessments. Trusts and universities will need to guard against the raiding of education budgets to support clinical service and research. The design of future assessment and feedback methodology would be wise to take into account the cost pressures in the system and to future proof assessment budgets by making them as transparent and lean as possible. The Department of Health in England has made recommendations to review health education funding after Health Education England has become embedded. This should result in better transparency of funds, but is also likely to result in an overall reduction in the money which goes to NHS Trusts, especially those in big medical schools and in London. It remains to be seen how the affected Trusts deal with this dilemma, and whether the funds currently available are reduced, resulting in less protected time for formative assessment and feedback, and less money to support clinical and practical assessments in clinical skills.

Sociological change influencing assessment and feedback

The medical profession is changing. From 2013, doctors in primary care will be predominantly female, and this will have extended to the whole profession by 2017. Although it has recently been shown that women perform better in formative and summative assessments than men, there are unknowns about how this shift in female to male ratios will impact on assessment and feedback (Elston, 2009). Women may choose different specialties to enter than men, focusing on those where there can be greater autonomy and fewer technological activities, for example primary care and psychiatry. This will skew the results of assessments, and may result in problems with equity and with the distribution of men and women in the career pathways available to them. The brightest people in the profession may not want to work in the specialities which need them. This will add to the complexity of the difficult task of workforce planning.

Medicine is an increasingly diverse profession, with doctors becoming increasingly geographically mobile. Difference in performance in clinical assessments related to cultural issues will need to be addressed in the future to avoid inadvertent discrimination against some minority groups. Induction programmes will need to be put in place for future new and international healthcare professionals in order to address the difficulties that they experience on coming into the NHS. These will need to be accompanied by robust and constructive feedback and support to the participants, to help them understand the cultural issues related to working in the UK.
Demographic change in the population we serve will result in an increase in the numbers of frail elderly patients, and an increase in long term clinical conditions which are better managed closer to home. There will also be a huge increase in patients with diseases of the elderly, including dementia. A recent informal review of the content of the MRCP question banks to establish whether these areas were adequately covered in the examination, suggested that the conditions would need to have greater emphasis in our current assessment processes to reflect the disease burden that our future healthcare professionals will see.

The increased geographical mobility of the healthcare professions reflects globalisation in general. Patients as well as doctors now travel more widely and live overseas. They bring a different pattern of illness to the healthcare system. Curricula, blueprints and assessments will need to reflect this in future.

**Technical developments and innovations in assessment practice**

The Internet has revolutionised the way we live and work. It has also changed the way we deliver education and training, and is beginning to influence methods of assessment and feedback. Most healthcare students now have access to a smart phone, and some healthcare training organisations are using these to provide learning resources to students. Documentation of workplace-based assessments and their feedback is increasingly being captured on an e-portfolio, and access to the e-portfolio through a smart phone or other mobile device is becoming possible. Future developments in mobile technology will facilitate the use of such devices on the wards and in clinics. At Leeds and Brighton and Sussex medical schools, pilot projects are underway to explore the roll out of these systems to large numbers of students. The access to patient information and to educational resources at all times is changing the way we practise medicine. In relation to assessments, it will mean that test developers will be planning assessments of candidates’ decision making and critical thinking skills, which will need to take into account their ability to access and evaluate available learning resources, possibly during the assessment. There are an increasing number of online learning modules – whether for undergraduates or postgraduates. There are a lot of organisations offering short (10 minute or so) online modules on specific subjects which doctors can use as part of their CPD. In order for these to count as CPD, or as certificated modules, robust assessments will be required in the future. It is still problematic to attribute the assessment to the appropriate learner with absolute certainty, but these technical problems are being rapidly overcome.
At the moment, e-portfolios are generally still under development, but are likely to become the mainstay of the recording of workplace-based assessments and feedback. The current system which is most used is the NHS Education Scotland software. Currently, there are still a number of technical problems with these systems, but these issues are likely to be resolved in the near future. It is likely that competitors will appear, and the standards of the products will be raised by this.

Simulation training has increased significantly over the past five years. Realistic computer based simulation is not only a potential source of training, but also a way of gathering data on the performance of doctors. The virtual reality surgical simulators developed at Imperial College London have led the way to the automatic gathering of evidence about the performance of practical skills (Wilson et al., 1997). The assessment of surgical knot tying is an interesting example. An experienced surgeon, who is more proficient at this activity, will make fewer minor additional movements. These can be recorded by a sensor on the back of the surgical trainee’s hand, and his dexterity measured and compared to that of his colleagues.

Video recording of team-based simulation events are likely to increase, and recordings can be viewed to provide direct feedback to trainees, which can then be used as a basis for a formative assessment. Summative assessments of a team-based activity continue to be a challenge for test developers. The difficulties arise from the evaluation of individual members of a team in competition with each other.

With the high pressure to make best use of our healthcare facilities in the future, the use of clinical space for examinations is likely to be increasingly limited, due to the cost of any reduction in the numbers of patients seen. In addition, the professionalisation of the delivery of clinical assessments, both to provide robust results and to be legally defensible, has led to the development of assessment centres. These are venues which are bespoke to ensure as standardised a clinical assessment as possible. Automatic timing devices, and a cadre of specially trained surrogate patients, provide a much more consistent clinical examination. In the USA, such centres are run by non-clinical staff and use only standardised patients. In the UK, where there is more emphasis on the quality of clinical method and diagnostic acumen, real patients with physical signs are invited to attend, but may need to be trained to provide a consistent history. Future clinical examinations are likely to need to be delivered in these centres.

Computer Based Testing (CBT) is likely to become mainstream. It will allow candidates to take a test in a computer centre. The advantages of this are that the candidate will be able to take the examination near to his home and during an assessment window that suits his needs. This will reduce the
need for all candidates for a particular test to take the same day off. It will also reduce the opportunity for copying from another candidate, as different tests can be made available on different days, and careful blueprinting and mathematical equating can be used to ensure that the different test forms are equivalent. For international written assessments, where currently candidates sit at the same time all over the world, this will be a significant advance. Challenges include the costs, the potential technical problems – which could be disastrous for a high stakes assessment – and the burden for the question setters in providing enough questions to facilitate the production of several test forms. Finally, although it will reduce copying, there are other more creative methods to access even the most secure question banks.

An additional advantage of CBT is the opportunity to have more flexible methods of assessment. It allows increasing use of still and video images. It will also allow adaptive testing, where the test can be tailored to the candidate’s ability, or speciality. This is the future direction, but it is not without its challenges.

Currently, images used in assessments need to be of very good quality, and to come with the consent of the patient. The free use of images on the Internet will raise again the problems associated with putting patient material on an electronic platform. Any developments in this area will need to be aware of the potential legal challenges over copyright and confidentiality.

The other potential benefit of the implementation of CBT to large scale national assessments is a potential reduction in costs. Modelling to date does not suggest that this is likely to be the case, as the costs of the hardware and the software and increasing the size of the bank to allow for window testing make up for the potential reduction of costs associated with the current computer marked but paper-based methods.

**Developments in assessment research**

Assessment research will need to focus on the development of new and robust methods for developing test instruments in areas which are traditionally difficult to assess. These are:

- Summative assessment of performance in the workplace;
- Assessment in teams;
- Assessment of complex actions;
- Assessment of leadership skills.

**Mining the assessment data sets**

The increasing role of the healthcare regulator has provided new sources of data about health professionals, and an expectation of increasing transpar-
ency in the performance data collected about assessments and candidates. Developments in computer software and record keeping systems has allowed these systems to collate information about healthcare professionals and their performance in assessments (Dacre et al, 2009). A culture of increasing transparency is emerging, and it is becoming possible to publish anonymised data showing trends in the performance of students, trainees and qualified healthcare professionals. With the increasingly robust collection of examination performance data and its linkage with other data sets, such as data on age, ethnicity and gender, the regulators – the higher education institutions, postgraduate clinical training organisations and the Medical Royal Colleges – will begin to be able to provide much more granular data on the students in the healthcare professions, and their characteristics in relation to a successful career outcomes or otherwise. The State of Medical Education and Practice, published in 2010 by the GMC, is the first step in the collation and exploration of an enormous data set that provides huge potential insights to the medical profession. It is likely that other healthcare professional regulators and assessment organising bodies will follow this lead.

**Recent education research that will influence policy change**

In the future, the increase in data sets and in the quality of research evidence will allow research findings to influence developments in medical education and assessment. The following two examples highlight this. Recent concern has been expressed about the variability in the way assessments at postgraduate level are standardised. This is an interesting area of future research, which should help those of us in assessment develop an evidence based and practical way forward. The number of attempts allowed at healthcare examinations is widely variable. In some assessments, such as the GMC’s PLAB test, there is no limit on the number of attempts. In others, there is an arbitrary limit placed on the number of times a candidate is allowed to sit. There is no clear consensus, and no clear evidence base. The arguments for a limit to the number of attempts focus on perceived patient expectations, those against focus on the human rights of the candidate. Until there is data on how long a candidate continues to improve, and his likely chances of passing over time, there will be difficulties in making robust rules on this issue. Recent research by McManus and Ludka (2012) have suggested that many candidates do improve their scores incrementally, but some, even with improvement, will never reach the pass mark.

Cheating in assessment is another area where use of research data has helped to develop policy. The future of assessment regulations should as far as possible aim to be evidence-based. An example of this is with the use of anomalous pair detection in written examinations. Software has been
developed which detects sequences of pairs of answers to computer marked questions which are more similar than chance would predict. If the pairs relate to the answers that two candidates who are sitting in an examination hall in close proximity have given, there is a significant chance that one candidate has copied from the other (an anomalous pair). Should either of the same candidates be found in an anomalous pair in a subsequent examination sitting, there is a significant chance that the candidate has been copying. This research evidence has now resulted in referral of candidates to the GMC, and sanctions being placed on the cheating doctor’s registration.

References


Useful websites

GMC: The Patel Review: Future Regulation of Medical Education and Training 2010 http://www.gmc-uk.org/Recommendations_and_Options_for_the_Future_
Future developments in assessment and feedback

GMC: Good Medical Practice Framework for Appraisal – Ready for Validation 2012
absolute (criterion referenced) methods
  p-values, 152
  standard setting, 3, 52, 100–102, 103
academic changes, 191–192, 196–197
Academy of Medical Royal Colleges, 71
acute care assessment tool (ACAT), 69
ageing population, 193–194
Angoff method, standard setting, 54, 101–102
anomalous pair detection, 198
anonymised marking, 169, 176
answer cueing, 20, 22, 23–25, 161
anxiety, 180
test (exam stress), 160, 162, 163
appeals, student, 171, 192
appraisal, 132–133
assessment, 1–11
  academic change, 191–192, 196–197
  assessable attributes, 5–8
  blueprinting, 3, 45, 105–106, 107, 110, 189
  competence, 2–3, 8–9, 38–56, 59
  computer-based, 106–111, 135–136, 137, 194–196
definition, 4
economic change, 192–193
evaluation contrasted, 4
evaluation of tests see psychometrics
examination boards, 168–173
examiners see examiners
feedback see feedback
future developments, 185–198
knowledge, 6–8, 9, 16–36
need for, 4
performance, 2–3, 8–9, 38, 59–74
professionalism, 9–10, 76–92
psychometrics, 141–165
regulatory change, 187–191
reliability see reliability of tests
research into, 191–192, 196–198
sociological change, 193–194
standard setting, 3, 52–54, 94–105, 152
student failure, 169–171, 176–182
teaching–learning–assessment
  alignment, 5, 45, 105–106, 107, 110, 189
technical developments, 194–196
trends, 2–3
  see also above future developments
validity see validity of tests
assessment centres, 195
assessors see examiners
attendance problems, students, 181
attitudes, professionalism, 90
behaviour (affective domain), 7, 8
behavioural assessments, 90, 91
bias
  examiner, 173
  in tests, 158–164
blind double marking, 169
Bloom’s taxonomy of assessable attributes, 5–8
blueprinting, 3, 45, 105–106, 107, 110, 189
Bologna Declaration, 169–170
borderline groups, sequential OSCE, 54–55
borderline methods, standard setting, 52–54, 101–102
borderline students, examiners’ options, 170
budget allocations, 192–193
candidate failure see student failure
career pathways, gender factors, 193
case-based discussions (CbDs), 68–69, 82–83
case specificity problem, 17, 18–19, 106, 154
Certificate of Completion of Training (CCT), 188–189
Certificate of Eligibility for GP Registration (CEGPR), 189
Certificate of Eligibility for Specialist Registration (CESR), 189
cheating, 163–164, 195–196, 197–198
checklist rating, OSCE scoring, 51, 174
classical test theory (CTT), 147–149, 154–155
clients see service user assessment;
  service user images; service user simulators
clinical practice assessments see competence assessment;
  performance assessment
clinical reasoning tests, 28–32
case-based discussion, 68
Cohen method, standard setting, 103–105
compensated marks, 170
competence assessment, 38–56
  dimensions, 39
  evaluation criteria, 39–40
  long case examination, 39, 40
  OSATS, 39, 44
OSCE see objective structured clinical examination
OSLER, 39, 40–42
performance and, 8–9, 38, 59
reliability, 153, 154, 175
test formats, 39
  see also specific formats
trends, 2–3
competitiveness, 180
complaints, student, 171
computer-based assessments, 106–111, 135–136, 137, 194–196
computer-based cheating detection, 164, 197–198
condoned fails, 170
certainty-based marking, 27–28, 161
consensus statements on performance assessment, 73
constructed answer questions, 16–17
  see also essay-type questions
construct validity, 146
  see also validity of tests
contexts, professionalism, 91
convergent validity, 158
core curriculum, 3
criteria-guided workplace-based assessment, 71, 72
criterion referenced (absolute) methods
  p-values, 152
  standard setting, 3, 52, 100–102, 103
critical incidents, 84
Cronbach’s alpha, 151
cueing, 20, 22, 23–25, 161
cultural issues, 177–178, 193
curricula
regulatory change, 188, 189
teaching–learning–assessment
alignment, 45, 105–106, 107, 110, 189
cut score setting see standard setting
decision-making skills assessment
case-based discussion, 68
key feature problems, 28
script concordance items, 30
demographic change, 193–194
diagnosis script questionnaires, 30
script concordance items, 30–32
directly observed procedural skill
(DOPS), 67–68
discriminant validity, 158
dissertations, 32–35
divergent validity, 158
doctors’ interpersonal skills
questionnaire (DISQ), 132
double marking, 169
doves see examiners, leniency/severity
economic changes, 192–193
educational institutions
examination boards, 168–173
professionalism, 91–92
educational paradigm, 5
education budgets, 192–193
education research, 191–192, 197–198
emotions, feedback, 122
English language testing, 189
environmental factors, 91, 171
e-portfolios, 194–195
essay-type questions, 17–18
marking, 17–18, 19, 111
modified (MEQs), 17, 18–19
professionalism, 85
short answer (SAQs), 20, 85, 111
ethnicity, 177–178, 193
European Credit Transfer and
Accumulation System (ECTS), 169–170
European Economic Area,
qualifications, 188–189
evaluation
assessment contrasted, 4
of tests see psychometrics
exam failure see student failure
examination boards, 168–173
examination performance data,
196–197
examiners, 167–176
bias, 173
borderline student options, 170
designing assessment criteria, 168
error-reduction strategies, 177
external, 171–173
inter-rater reliability, 153–154, 162,
175, 176
judgement process, 173–175
leniency/severity, 10, 50, 173,
174–176, 177
marking see marking
moderation of assessments, 169, 172
OSCE circuits
feedback, 50
judgement process, 173–174
marking, 43, 47, 51–54, 82, 110,
174, 175
professionalism, 82
standard setting, 52–54
training, 48–50
of professionalism, 82, 83
verbal protocol analysis, 175–176
workplace-based assessments, 70–71, 174
exam marking see marking
exam questions see written questions
exam stress, 160, 162, 163
extended matching questions (EMQs),
17, 20, 22–23, 24, 25–26
external examiners (EEs), 171–173
failure at exams see student failure
feedback, 114–138
appraisal, 132–133
on assessment, 126, 134–136, 137
benefits, 116
case of the month scenarios, 135–136, 137
definitions, 115–116, 117
definitions, 115–116, 117
emotions, 122
facilitation challenges, 117–122, 124–126
facilitation models, 128–129
facilitators defined, 117
facilitator–learner relationship, 121–122, 125, 127–128
formative–summative balance, 119–120
future developments, 190–195
importance, 116–117
language factors, 120, 122
meta-feedback skills, 126, 134–136
multi-source, 9, 10, 87, 130–132
OSCE, 50
power relations, 122
practical suggestions, 126–129
principles, 124–126
range of sources, 117, 118
recipient factors, 120–122, 127–128
roles in facilitating, 117–122, 124–126
service user, 86–87, 130–132
theories, 123–124
typology, 118, 119
workplace-based assessment, 70–71, 190–191
formative assessments, 10, 95
appraisal, 133
balance with summative, 95, 119–120
feedback, 116, 119–120, 133
feedback on, 134–135
Kolb’s learning cycle, 61
professionalism, 80, 83, 84
simulation, 60, 61–66, 108, 195
true–false questions, 21
workplace-based, 60, 70, 73–74, 190–191
formula scoring (negative marking), 20, 161

Galton, Sir Francis, 142–143
gender differences, 177–178, 193
generalisability theory (GT), 154, 156
General Medical Council (GMC)
appraisal, 133
assessment practice standardisation, 187–188
international graduates, 188–189
national testing, 188
performance assessments, 70, 190–191
performance data, 197
re-validation, 190, 191
geographical mobility, 193, 194
global rating scales, 51
GROW(N) model, 128, 129
guessing, 161
hawks see examiners, leniency/severity
healthcare changes, 186
academic, 191–192, 196–197
economic, 192–193
regulatory, 187–191
sociological, 193–194
healthcare institutions, professionalism, 91–92
healthcare teams, simulations, 64, 195
high fidelity simulations, 62–63
OSCE, 55–56
reason for, 60, 61–62
high stakes assessments see summative assessments
history taking
long case examination, 40
OSLER, 40–42
Hofstee method, standard setting, 103, 104
illness, student, 170
images, use in assessments, 196
institutional phenomenon,

professionalism as, 91–92
interdisciplinary simulations, 64
international healthcare professionals, 188–189, 193
Index 205

Internet, on-line assessments, 109–111, 135–136, 137, 194–195, 196
interpersonal phenomenon, professionalism as, 91
interpersonal skills questionnaire, 132
inter-rater reliability, 153–154, 162, 175, 176
item response theory (IRT), 109, 155, 156

Johari’s window, 121

key feature problems (KFPs), 28–30, 31–32
knowledge assessment, 16–36
assessable attributes hierarchy, 6–8
case-based discussion, 68
constructed answer questions, 16–17
see also essay-type questions
dissertations, 32–35
essay-type questions, 17–20, 85, 111
key feature problems, 28–30, 31–32
Miller’s pyramid, 8, 9
multiple-choice questions see multiple-choice questions
portfolios, 35–36
professionalism, 84–85
projects, 32–35
script concordance items, 30–32
selected answer questions, 16, 17
see also multiple-choice questions
Kolb’s learning cycle, 61
Kuder–Richardson 20 (KR 20), 151

language
feedback, 120, 122
students’ competence, 179
testing, 189
learning
assessment for, 4, 60, 61, 95, 119
assessment–teaching alignment, 5, 45, 105–106, 107, 110, 189
factors in student failure, 178–181
feedback, 116, 119–126, 134–135
formative–summative assessment
balance, 95, 119–120
Kolb’s learning cycle, 61
portfolios for, 35
professionalism, 78–79, 80, 83, 84
simulations for, 61–66, 108, 195
workplace-based assessments, 60, 70, 73–74, 190–191
learning environments, assessment, 91
learning resources, on-line, 194
Likert scales, WPBAs, 67, 68
Log book, 35, 71–72
long case examination, 39, 40
see also objective structured long examination record
mannequins, 61, 62, 63
marking, 168–169
anonymised, 169, 176
blind double, 169
compensated marks, 170
computer-assisted, 110–111
condoned failure, 170
certainty-based, 27–28, 161
double, 169
erroreduction strategies, 176, 177
essay-type questions, 17–18, 19, 111
leniency/severity, 10, 50, 173, 174–176
long case examination record, 40
mark capping, 170
moderation, 169, 172
negative, 20, 161
OSCE, 43, 47, 51–54, 82, 110, 174, 175
OSLER, 40–42
projects and dissertations, 34–35
threats to test validity, 161, 162
see also scoring systems; standard setting
Marlow–Crowne social desirability scale, 161
Membership of the Royal College of Physicians (MRCP) 145, 149n
mentors, 182
meta-cognitive skills, 178
meta-feedback skills, 126, 134–136
Miller’s pyramid of competence, 8, 9, 59, 60, 79, 80
mini clinical evaluation exercises (mini-CExs), 66–67, 174
mobile technology, 194
moderation of assessments, 169, 172
modified Angoff methods, standard setting, 54, 102
modified essay questions (MEQs), 17, 18–19
multiple-choice questions (MCQs), 17, 20–28
calendar-based cases of the month, 135–136, 137
calendar-based cheating detection, 164
calendar-based item generator, 109
certainty assessment, 27–28
Nedelsky method standard setting, 101
optical marking, 110
professionalism, 84–85
scoring, 20, 27–28, 101, 161
types, 20, 21–23
writing, 24–27
multi-source feedback (MSF), 9, 130–132
professionalism, 10, 87

national testing, 188
near misses, 84
Nedelsky method, standard setting, 100–101
negative marking, 20, 161
NHS budget allocations, 192–193
norm referenced (relative) methods, standard setting, 3, 52, 99–100, 103
objective structured assessment of technical skills (OSATS), 39, 44

objective structured clinical examination (OSCE), 39, 42–56
assessor severity, 175
assessors’ judgement process, 173–174
assessor training, 48–50
blueprinting, 45
candidate instructions, 45, 47
calendar-based assisted marking, 110–111
developing OSCE stations, 45–48, 82
feedback, 50
guidance for assessors, 45, 47, 48
marking, 43, 47, 51–54, 82, 110, 174, 175
planning, 44–50
professionalism, 82
purpose, 45
reliability, 153, 154, 175
scoring, 51–52, 174
sequential, 54–55
‘serious concerns’ documentation, 50
simulated patient information, 45, 47
simulated patient scoring, 52
simulated patient training, 48–49
simulation fidelity, 55–56
standard setting, 52–54
objective structured long examination record (OSLER), 39, 40–42
optical markers, 110
Ottawa Conference 2010, professionalism, 88–92

paper-based tests see written assessments; written questions
parallel forms reliability, 150
pass score setting see standard setting
patients see service user assessment; service user images; service user simulators
patient satisfaction surveys, 86–87, 131–132
peer review, 10, 87, 130–131, 134
Pendleton’s rules, feedback, 128, 129
performance assessment, 59–74
appraisal, 132–133
competence and, 8–9, 38, 59
consensus statements on, 73
data transparency, 196–197
log-books, 71–72
portfolios, 72–73
reason for, 60
reliability, 153–154
simulation, 59, 60, 61–66
trends, 2–3, 190–191
workplace-based, 8–9, 38, 59–60, 66–71, 73–74, 174, 190–191
personal development portfolios see portfolios
personality traits, 90
physical examination
long case examination, 40
OSCE, 42–56
OSLER, 40–42
portfolios, 10, 35–36, 72–73, 111, 194–195
postgraduate theses, word length, 34
power relations, feedback, 122
predictive validity, 157–158
procedural skills, OSATS, 44
professionalism, 9–10, 76–92
assessment approaches, 79–81
definitions, 77–78
as individual phenomenon, 90–91
as institutional/societal phenomenon, 91–92
as interpersonal phenomenon, 91
learning, 78–79
observation in vitro, 81–83
observation in vivo, 83–84
Ottawa 2010 expert group, 88–92
paper-based tests, 84–86
peer review, 87
proxy measures, 88
self-assessment, 87–88
service user assessment, 84, 86–87
teaching, 78–79
workplace-based assessments, 10, 82–83
Professional and Linguistics Assessment Board (PLAB), 189
projects, 32–35
psychometrics, 141–165
definitions, 142–143, 144
educational assessments as, 144
regulatory change, 189–190
role, 142
test reliability, 144, 145, 147–156
test validity, 144, 145–146, 156–164
Quality Assurance Agency (QAA), 168–170, 172–173
quality of tests see psychometrics
questionnaires
 doctors’ interpersonal skills, 132
patient satisfaction, 86–87, 131–132
questions, exam see written questions
reflective practice
documenting, 35, 36, 73, 85
feedback, 124
professionalism, 85
simulation and, 63
regulatory change, 187–191
relative (norm referenced) methods,
standard setting, 3, 52, 99–100, 103
reliability of tests, 11, 39–40, 144
blueprinting, 45
classical test theory, 147–149, 154–155
competence assessments, 153, 154
computer analysis, 109–110
definition, 145
establishing and evaluating, 147–156
generalisability theory, 154, 156
importance, 145
internal consistency, 150–151
inter-rater reliability, 153–154, 175, 176
item response theory, 155, 156
item statistics, 147, 151–153
parallel forms method of assessing, 150
performance assessments, 153–154
p-value (item difficulty), 152
regulatory change, 189–190
reliability coefficients, 147, 148–149, 155
small cohorts, 189–190
split-half method of assessing, 150
standard error of measurement, 151
test re-test method of assessing, 149–150
validity and, 145
written tests, 148–153
research, into assessment, 191–192, 196–198
Research Excellence Framework (REF), 191
research projects, 32–35
resits
  future rules, 197
  mark capping, 170
re-validation, 190–191
right-scoring, MCQs, 20
Ronseal test, 11
  see also validity of tests
Royal College of Physicians 145, 149n
sampling, 106
  classical test theory, 154–155
scenarios, assessing professionalism, 85–86
scoring systems
  general requirements, 168
  multiple-choice questions, 20, 27–28, 101, 161
  OSCE, 51–52, 174
  threats to test validity, 161, 162
  see also standard setting
script concordance items (SCIs), 30–32
selected answer questions, 16, 17
  see also multiple-choice questions
self-assessment, 10, 87–88, 133, 134
self-awareness, 63
sequential OSCE, 54–55
‘serious concerns’ documentation, 50
service user assessment, 130–132
  professionalism, 84, 86–87
service user images, 196
service user simulators, 45, 47, 48–49, 52, 62
  assessment centres, 195
  halo effect, 178
  undercover, 84
short answer questions (SAQs), 20, 85, 111
significant events, 84
simulated client observation, 84
simulation(s), 2–3, 59, 60, 61–66
  assessment centres, 195
  computer-based, 108, 195
  OSCE, 43, 45, 47, 48–49, 52, 55–56, 82
  professionalism, 81–83, 84
single best answer questions (SBAs), 17, 20, 21–22
situational judgement tests (SJT), 85
situation-specific professionalism, 91
skill (psychomotor domain), 7, 8
  simulated practice, 61–66
  see also competence assessment;
  performance assessment
skills-based scoring, OSCE, 51–52
smart phones, 194
socially desirable answers, 161
societal phenomenon, professionalism as, 91–92
sociological changes, 193–194
specific learning disabilities (SpLD), 178–179
split-half reliability, 150
standard error of measurement (SEm), 151
  sequential OSCE, 55
Index 209

standard setting, 3, 94–105
absolute methods, 3, 52, 100–102
see also below compromise methods
compromise methods, 102–105
definition, 98
European Credit Transfer and Accumulation System, 169–170
judging panels, 97
need for, 94, 96
OSCE, 52–54
p-values (item difficulty), 152
relative (norm referenced) methods, 3, 52, 99–100
see also above compromise methods
test modality–method relation, 96–97, 99
station-based scoring, OSCE, 51–52
still images, 196
stress, 180
exam, 160, 162, 163
structured (supervised) learning events (SLEs), 38, 191
student failure, 169–171, 176–182
anxiety, 180
see also stress, exam
appeals, 171, 192
attendance problems, 181
compensated marks, 170
competitiveness, 180
complaints, 171
condoned, 170
ethnicity, 177–178
gender differences, 177–178
insecurities, 180
language competence, 179
learning disabilities, 178–179
learning needs, 178
mark capping, 170
meta-cognitive skills, 178
mitigating circumstances, 170
reluctance to seek help, 179–181, 182
resits, 170, 197
stress, 180
see also stress, exam
study skills, 180
support systems, 178–179, 180–182
study skills training, 180
summative assessments, 10, 95
appraisal, 133
balance with formative, 95, 119–120
competence, 38
evaluation see psychometrics
feedback, 116, 119–120, 133
feedback on, 134, 135
OSCE standard setting, 52–54
pass score setters, 97
professionalism, 80
ture–false questions, 21
workplace-based assessment, 70
supervised (structured) learning events (SLEs), 38, 191
support for students, 178–179, 180–182
surgical skills
computer-based assessment, 195
objective structured assessment, 44
surveys, patient satisfaction, 86–87, 131–132
systematic bias in tests, 158–164
teaching, professionalism, 78–79
teaching–learning–assessment alignment, 5
blueprinting, 45, 105–106, 107, 110, 189
team working simulations, 64, 195
technical developments, 194–196
test anxiety (exam stress), 160, 162, 163
test quality see psychometrics
test re-test reliability, 149–150
transparency, performance data, 196–197
ture–false questions, 17, 20, 21
true scores theory, 147–148, 155
undercover observers, 84
users see service user assessment; service user images; service user simulators

validity of tests, 11, 39–40, 144, 156–164
badly written questions, 160–161, 162, 163
blueprinting, 45
cheating, 163–164
construct, 146
construct-irrelevant variance, 159–164
construct under-representation, 159
definition, 145–146
divergent (discriminant), 158
exam paper errors, 160
exam stress, 160, 162
guessing, 161
predictive, 157–158
regulatory change, 189–190
reliability and, 145
small cohorts, 189–190
sources of evidence for, 156–157
threats to, 158–164
values, professionalism, 90
van der Vleuten equation, 11

verbal protocol analysis, 175–176
video recordings/images, 195, 196

workforce planning, 193
workplace-based assessments (WPBAs), 8–9, 38, 59–60, 66–71, 73–74
e-portfolios, 194–195
examiners, 70–71, 174
future of, 190–191
professionalism, 10, 82–83
writing skills, 36
written assessments
cheating, 163–164, 195–196, 197–198
computer-assisted free text marking, 111
computer-based, 195–196
reliability, 148–153
small cohorts, 189–190
validity, 157–164
see also written questions
written questions
answer cueing, 20, 22, 23–25, 161
badly written, 160–161, 162, 163
exam paper errors, 160
knowledge assessment, 16–32
professionalism assessment, 84–86
question banks, 108, 109, 162
see also written assessments
Keep up with critical fields

Would you like to receive up-to-date information on our books, journals and databases in the areas that interest you, direct to your mailbox?

Join the Wiley e-mail service - a convenient way to receive updates and exclusive discount offers on products from us.

Simply visit www.wiley.com/email and register online

We won’t bombard you with emails and we’ll only email you with information that’s relevant to you. We will ALWAYS respect your e-mail privacy and NEVER sell, rent, or exchange your e-mail address to any outside company. Full details on our privacy policy can be found online.