Chapter 1 Principles of assessment

Professor Olwyn M. R. Westwood[1] *and Dr Ann Griffin*[2]
[1]Queen Mary, University of London
[2]UCL 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, First Edition. Edited by
Olwyn M. R. Westwood, Ann Griffin, and Frank C. Hay.
© 2013 John Wiley & Sons, Ltd. Published 2013 by John Wiley & Sons, Ltd.

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. Educationalists 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 inappropriate 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.

> 'Tests of clinical competence . . . must be designed with respect to key issues including blueprinting, validity, reliability, and standard setting, as well as clarity about their formative and summative function.'
>
> Wass *et al.* 2001

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.

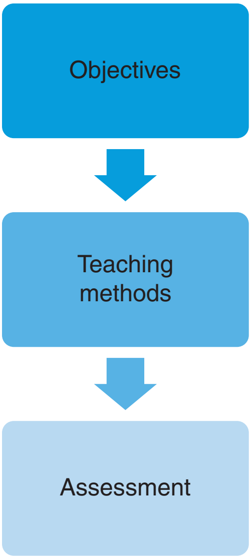

Figure 1.1 The education paradigm.

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

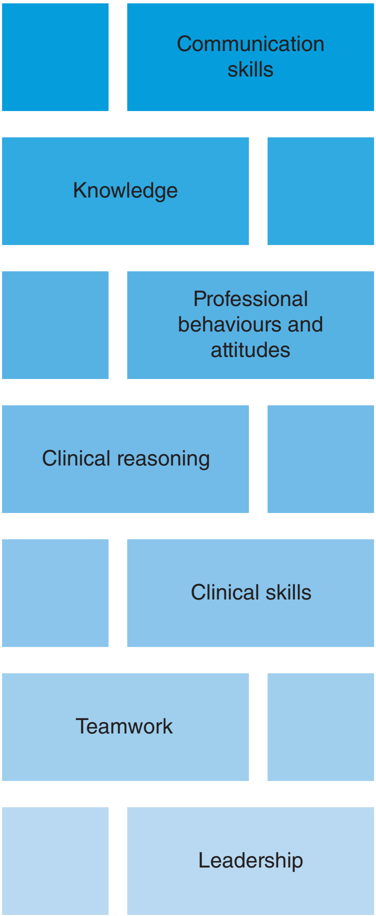

Figure 1.2 The range of assessable attributes.

it his 'taxonomy'. It provides a good way of thinking about the pitch of your questioning or test.

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

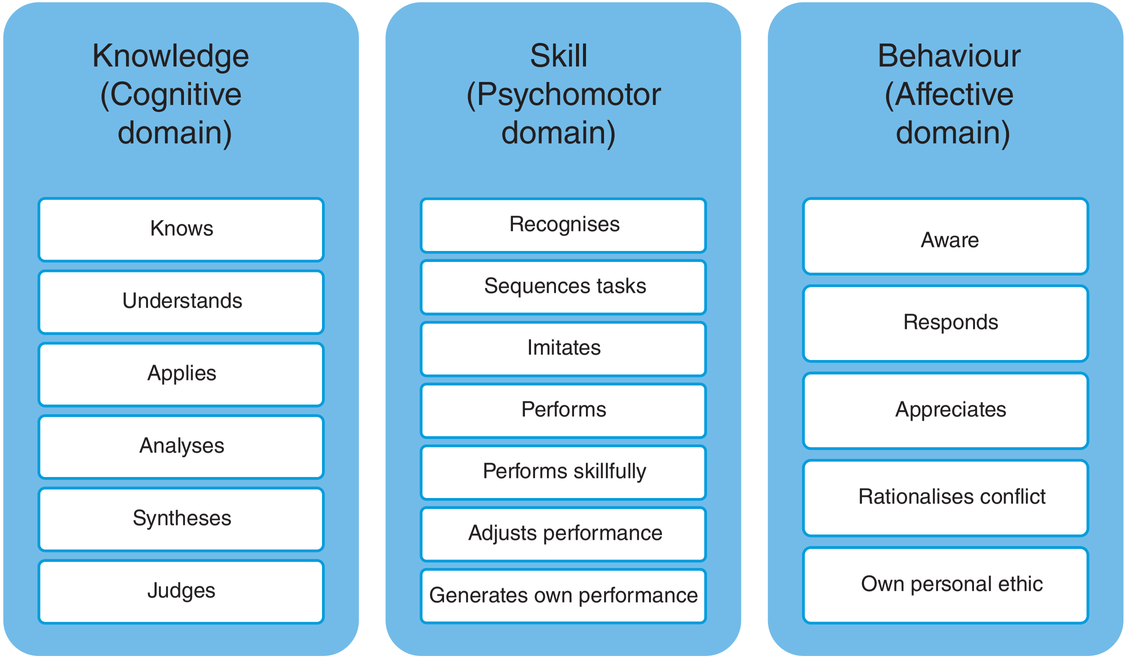

Figure 1.3 Hierarchy of knowledge, skills and performance based on Bloom's taxonomy.

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 unco-operative 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

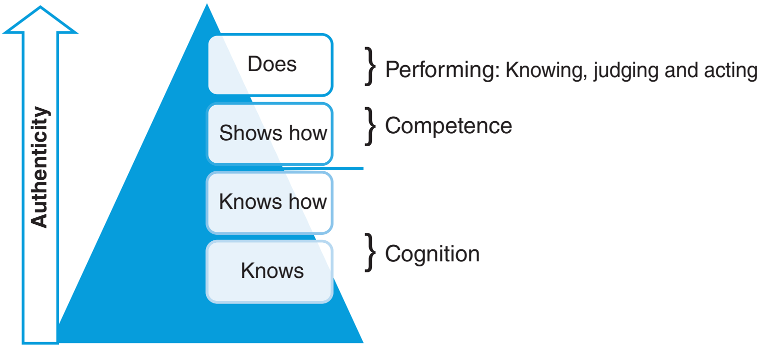

Figure 1.4 Assessment of competence and performance, based on Miller.

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 practise (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 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).

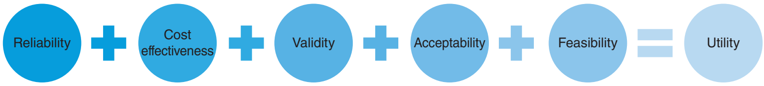

Figure 1.5 The van der Vleuten equation. Adapted from Van der Vleuten (1996).

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

Bloom, B.S., Englehart, M.D., Furst, E.J., *et al.* (1956). *Taxonomy of Educational Objectives: Cognitive Domain*. McKay.

Boursicot, K., Etheridge, L., Setna, Z., *et al.* (2011). Performance in assessment: Consensus statement and recommendations from the Ottawa conference 2011. *Medical Teacher*, 33: 370–383.

Miller, G.E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, 65: 563–567.

Van der Vleuten, C.P.M. (1996). The assessment of professional competence: developments, research and practical implications. *Advances in Health and Scientific Education,* 1: 41–67.

Wass, V., Van der Vleuten, C., Shatzer, J. and Jones, R. (2001). Assessment of clinical competence. *The Lancet*, 357: 945–949.

Further reading

Crossley, J., Humphris, G. and Jolly, B. (2002). Assessing health professionals. *Medical Education*, 36: 800–804.

Dent, J., and Harden, R., (eds) (2001). *A Practical Guide for Medical Teachers*. Churchill Livingstone.

Friedman Ben-David, M. (2000). AMEE Guide No. I8: Standard setting in student assessment. *Medical Teaching*, 22: 120–130.

Harden, R.M. and Gleeson, F.A. (1979). ASME medical education booklet no 8: assessment of medical competence using an objective structured clinical examination (OSCE). *Journal of Medical Education*, 13: 41–54.

Gordon, M.J. (1991). A review of the validity and accuracy of self-assessments in health professions training. *Academic Medicine*, 66: 762–769.

McManus, I.C., Thompson, M. and Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education*, 6: 42.

The National Student Survey: http://www.thestudentsurvey.com/ [accessed Nov 2012].

Schon, D. (1983). *The Reflective Practitioner: How Professionals Think in Action*. Basic Books.