Assessment in Medical Education and Training

A practical guide

Edited by

Neil Jackson

Postgraduate Dean of General Practice, London Deanery
Honorary Professor of Medical Education, Queen Mary School of Medicine and Dentistry, University of London

Alex Jamieson

Associate Director, London Deanery GP Department
Course Director (Queen Mary), Joint MSc in Primary Care
Queen Mary School of Medicine and Dentistry, and City University, London

and

Anwar Khan

Associate Director, London Deanery GP Department
University of London

Foreword by

Dame Lesley Southgate

Radcliffe Publishing
Oxford • New York

Radcliffe Publishing Ltd
18 Marcham Road
Abingdon
Oxon OX14 1AA
United Kingdom

www.radcliffe-oxford.com
Electronic catalogue and worldwide online ordering facility.

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library.

ISBN-13: 978 1 84619 046 9

Typeset by Lapiz Digital Services, Chennai
Printed and bound by TJI Digital Ltd, Padstow, Cornwall

Contents

Foreword

This thoughtful, provocative and eclectic book is published at a time of enormous change in the content, structure and quality assurance of postgraduate medical education in the United Kingdom.

The reforms known as Modernising Medical Careers are being introduced at the same time as a new regulator, the Postgraduate Medical Education & Training Board (PMETB), has been established. And the General Medical Council has further developed *Good Medical Practice*, its guidance to the medical profession, setting out the extent and boundaries for contemporary medical practice.

The dynamic system formed by these three forces has fundamental implications for the assessment of postgraduate medical training. The MMC careers framework determines the timing and importance of assessments, the PMETB Principles for Assessment Programmes sets out quality standards for assessment methods and *Good Medical Practice* sets challenges for the design of new or improved methods to assess 'hard to measure' domains such as professionalism. The interplay between the reformed structure of UK postgraduate training, the qualities and attributes that must be exhibited by all registered doctors (*Good Medical Practice*) and the quality assurance requirements for all specialty curricula and their integrated assessment programmes, creates a unique environment for training.

The first chapter emphasises the principles for the design of assessment programmes, and those that follow eleborate those principles by considering assessments for different purposes, or for different contexts. But principally grounded in the world of general practice. Some of the topics are a snapshot of approaches which, while current, will soon be overtaken by events. Others look to the future. The reader, like this writer, will be challenged and stimulated by the variety of views and emphases, but will find the concluding chapter a place for reflection and integration, so essential for all of those who dare to become involved in assessing others, where justice and fairness for doctors and patients must be the prime consideration.

Dame Lesley Southgate
March 2007

About the editors

Neil Jackson first entered general practice in 1974 and worked for 25 years as a full-time Principal in a semi-rural practice in Epping Forest until retiring in 1999. During this time he developed an interest in postgraduate medical education and training and is a former GP Trainer, Course Organiser and Associate Regional Advisor in General Practice. In 1995 he was appointed Postgraduate Dean for General Practice in the North Thames Deanery and has continued in this role in the London Deanery as from 2001. He is also Honorary Professor of Medical Education at Queen Mary School of Medicine and Dentistry, University of London. He is a former MRCGP Examiner and the editor and author of various books, book chapters, peer referenced papers and articles on general practice/primary care education and training issues. For the past several years he has worked as Visiting Family Medicine and Primary Care Consultant in countries of the former Russian Federation, including Georgia and Uzbekistan and more recently Japan and Poland and is currently a Leader Visitor for PMETB.

Alex Jamieson was a GP principal in Hackney from 1981 to 1996, and is now a sessional GP in Tower Hamlets. He is an Associate Director in the London Deanery GP Department, and Senior Lecturer and Course Director (Queen Mary) of the joint MSc in Primary Care, Queen Mary's School of Medicine and Dentistry, and City University, London. As part of his Deanery role, he shares the case management of GP Performance cases in London. Through this casework experience he has developed an interest in assessments in a performance context.

Anwar Khan entered general practice in 1991 after a period of work in clinical genetics and was a Course Organiser for 12 years before joining the London Deanery as Associate Director. He became an Examiner for the MRCGP in 1996 and member of the simulated surgery development group. He is a portfolio GP in East London and has developed an active interest in education and assessment both in the UK and abroad. He has worked as Visiting Family Medicine and Primary Care Consultant in countries of the former Russian Federation, including Georgia and Uzbekistan and more recently has been working in Oman and South Asia in the development of MRCGP(INT). He was a past Chairman of the North-East London RCGP Faculty. He has also worked with Peter Burrows to develop the Freshstart Simulated Surgery for the London Deanery. He continues to be actively involved in developing the Clinical Skills Assessment for the new MRCGP examination.

List of contributors

Dr Reed Bowden is now retired having, for many years, worked as a principal and Trainer in General Practice. He was also a Course Organiser for a conventional VTS and later ran special intensive courses, as well as serving for seven years as an examiner for the MRCGP. He became an Associate Director in the London Deanery GP Department with responsibility for assessment and for helping persistently under-performing doctors with their educational rehabilitation.

Professor Val Wass is a GP who has increasingly developed an academic career in medical education. She studied for a Masters in International Health Professions Education at the University of Maastricht and over the years has had extensive interest in postgraduate and undergraduate assessment. Her experience stemmed from work as convenor of the oral component of the Royal College of General Practitioners Membership Examination (MRCGP) and her subsequent role as Chair of the Examination Board. She led the change from traditional undergraduate clinicals to Objective Structured Clinical Examinations at Kings College medical school. Her PhD compared traditional and new tests of clinical competency and provided evidence to support the new approaches to assessment outlined in this book.

Dr Kamila Hawthorne is a GP and Clinical Senior Lecturer in the Department of General Practice, Wales College of Medicine, Cardiff University. She has also been an examiner for the MRCGP since 1997. She is currently involved in the development of the Clinical Skills Assessment for the new MRCGP. Her undergraduate remit is as Convenor of the Year 3 undergraduate teaching in general practice in Cardiff, and she has published research in the fields of medical education, ethnic minority health issues and diabetes.

Dr Gareth Holsgrove is a graduate of Sussex University and one of the first people in the UK to obtain a BEd, he is also a qualified Teacher of the Deaf, has an MSc in Audiology (Salford University), and a PhD in Education from UEA. Thus, from a background in education and paediatric audiology, he has specialised exclusively in medical and dental education since 1990. During this time, he has held several senior posts including heading departments of medical education in both the UK, as a joint appointment of St Bartholomew's and The Royal London Hospital, and in the national medical school of the United Arab Emirates. He has served as consultant to several medical schools, medical royal colleges and deaneries in the UK and overseas. He is currently Medical Education Adviser to the Royal College of Psychiatrists. His principal interests in medical education are the development of curricula and assessment programmes including workplace based assessment and formal examinations. He is involved with PMETB (the Postgraduate Medical Education and Training Board) and also involved with MMC (Modernising Medical Careers). He has an extensive list of publications and is also a reviewer for OUP and *Medical Education*, and on the editorial board of two journals including *Understanding Medical Education*.

Dr Helena Davies is a Senior Lecturer in Medical Education/Late Effects, University of Sheffield (Honorary Consultant Sheffield Children's Hospital). Helena has a particular research interest in the development, implementation and evaluation of work based assessment. Her team has developed and validated a range of work based assessment tools including a multisource feedback tool (SPRAT) and a patient assessment tool (SHEFFPAT). She has led the research and quality assurance of the National Foundation Assessment Programme and has also undertaken work in relation to assessment with a range of organisations including the RCPCH, RCPATH, National Clinical Assessment Service, PMETB and RCGP.

Dr Anand Mehta is Associate Dean at the London Deanery since 2004. From 1999–2004 he was Director of Medical Education at Mayday University Hospital in Croydon where he still works as a Consultant Physician in Elderly Care Medicine. During his tenure as Clinical Tutor he expanded and formalised the undergraduate and postgraduate training departments at Mayday Hospital and forged closer links with the local medical schools. At the London Deanery he has responsibility for secondary care in the South West London patch as well as some specialty training committees and has recently taken on the Lead role for Managing Poor Performance in Secondary Care which involves working with NCAS and GMC.

Dr Kevin Kelleher is an Associate Dean for PGME at London Deanery and an Honorary Senior Lecturer in Medical Education. He has been a Clinical Tutor and Specialty Training Committee/Specialist Advisory Committee member for many years. He has been accredited by the RCP Physicians as Educators Programme and holds a Diploma in Teacher Education from the University of London. He has been engaged in recent years with all the changes in PGME brought about by Modernising Medical Careers especially in South East London. He is a trained facilitator for the NHS Leadership 360 Appraisal programme and has a particular interest in the area of 'Doctors with Disability'.

Dr Colin Stern qualified from Cambridge University & St Thomas' Hospital, 1966 and then trained in Paediatrics. He carried out research into perinatal immunology for PhD dissertation and also into connective tissue disease in children. He established the Materno-Fetal Immunobiology Research Group (part of the British Society for Immunology). He has developed interest in patients with Chronic Fatigue/ME. He was appointed Consultant Paediatrician to St Thomas' Hospital in 1980 and was Visiting Professor to King Saud University from 1986 to 1987. He was Postgraduate Sub-Dean between 1989–1997 when he was responsible for delivering postgraduate education to the Postgraduate Dean and for running the Postgraduate Centre, and established the Regional Training Scheme for Paediatric SHOs in 1990. In April 2001, he was appointed as Associate Postgraduate Dean to the London Deanery and until September 2004 was responsible for South-East London. At present, he is responsible for ten medical specialities and for career counselling for London trainees. He is the President (Paediatric Section), and Chairman, of the Academic Board of the Royal Society of Medicine.

Professor Pat Lane is the Director of Postgraduate General Practice Education for the South Yorkshire and South Humber Deanery and a visiting professor to the School of Health and Related Research (ScHARR) at Sheffield University. Previously in General Practice for 30 years he has been a GP trainer, VTS Course Organiser and Associate Adviser and the Chairman of COGPED (Committee of GP Postgraduate Education Directors – UK) from 2002–2005.

He has published on the development of recruitment methodology, informatics training and education in general practice.

Professor Fiona Patterson is leading expert in the field of selection, development and innovation. Fiona has over 15 years experience of working at a strategic level with a variety of FTSE 100 organisations. She is currently the Director of the Organisational Psychology research team at City University, London. Fiona is a founding Partner of The Work Psychology Partnership LLP, established 2004, providing advice to many organisations internationally. She was appointed academic advisor to the DTI in 2000 and she currently advises several Royal Colleges on selection issues. Previously, she was Director for postgraduate programmes in the Institute of Work Psychology, University of Sheffield, and held the same position at the University of Nottingham. Prior to her posts in academe, she was Head of Organisational Psychology at The Boots Company and an internal consultant at Ford Motor Company Limited (Global). She has published widely in assessment, especially in relation to selection issues and in validating assessment methods.

Dr Tim Swanwick has a broad range of experience in general practice education and is currently a Director of Postgraduate General Practice Education in the London Deanery, an Honorary Senior Lecturer at Imperial College and a visiting lecturer at the University of Westminster. He is the editor of the Association for the Study of Medical Education's *Understanding Medical Education* series, sits on the editorial board of the journal *Education for Primary Care*, and writes widely on all aspects of GP education and training. Tim is co-author of *The Study Guide for GP Training* (2003), and editor of *The General Practice Journey* (2003) and *The Management Handbook for Primary Care* (2004) and his areas of special interest include assessment, workplace learning and educational leadership.

Dr Nav Chana is a GP in Mitcham, Surrey and an Associate Director in the London Deanery, as well as a current MRCGP examiner. He has been a member of PMETB's workplace assessment sub-committee, and is a current member of PMETB's assessment working group. He is a member of the RCGP's workplace based assessment steering group for the new MRCGP. He is also on the executive board of the National Association of Primary Care.

Dr Robert Clarke is an Associate Director of Postgraduate General Practice in the London Deanery. His educational interests are in the role of reflection in learning, inter-professional learning and communication skills training. He is co-author of a book called *Critical Reading for the Reflective Practitioner* (Butterworth Heinemann, 1998) and is responsible for a clinical skills website (www.askdoctorclarke.com). He is a Fellow of the Royal College of General Practitioners and of the Royal College of Physicians, and is a member of the Higher Education Academy.

Dr Peter Burrows retired in 2006 after 34 years in general practice at Romsey in Hampshire. He has been Chairman and Provost of the Wessex faculty RCGP and an MRCGP examiner since 1980. In 1994 he studied the use of simulated patients in assessment during a sabbatical at McMaster University in Canada. He took part in the development of the MRCGP Simulated Surgery and was convenor of this module of the examination from 1996 to 2000. He has also developed the Freshstart Simulated Surgery, which is used for the assessment of consulting skills of EU doctors, poor performers and others in the London Deanery. He continues to be actively involved in developing the Clinical Skills Assessment for the new MRCGP examination.

Dr Adrian Freeman works as a practising GP in Exeter and as the lead for knowledge assessment in the Peninsula Medical School. He has been an MRCGP examiner since 1996 and is now Convenor of the consulting skills Module (video) of the current MRCGP exam. He is on the board for the MRCGP International and has personal experience of developing Family Medicine assessments in Oman, Dubai and Malta. He assists in developing assessment material for the RCGP, the GMC and Peninsula Medical School.

Dr Moya Kelly is assistant director in postgraduate medical education in West of Scotland with responsibility for vocational training. She has been a GP in Glasgow for 22 years. She has responsibility for the video component of summative assessment and the MCQ.

Dr Murray Lough has been a GP for 25 years in Airdrie in central Scotland and involved in education and educational research for the past 15 years. He was involved in the development and implementation of the audit project for summative assessment and is now responsible for the development of educational research in the West of Scotland deanery.

Dr Jonathan Burton has been a GP for 28 years in Essex and Suffolk. He started as GP Tutor in Colchester in 1986, and is now Associate Director for Educational Research & Development at the London Deanery. He is the editor of the journal, *Work Based Learning in Primary Care*.

Dr Penny Trafford has been a GP in North London for 25 years and has been a trainer and course organiser for Barnet GPVTS. Since 2001, she has been an Associate Director in the London Deanery; her remit within the London Deanery is for GP training in North Central London, the lead for overseas doctors including projects for refugee doctors and resident international medical graduates as well as the EU GP Induction programme.

Dr Sanjiv Ahluwalia is a full-time GP trainer and course organiser in Barnet North London. His affiliations include MRCGP and MILT.

Dr Debbie Cohen holds a joint appointment at Cardiff University across the Schools of Medicine and Psychology. She is the Deputy Director of the Communication Skills Unit (CSU), and Director of the Individual Support Programme (ISP) in the School of Medicine and a Senior Medical Research Fellow in the Unum Provident Centre for Psychosocial and Disability Research, School of Psychology. The CSU provides communication skills training to medical students and doctors in the Welsh Deaneries. It is a research based unit with its main area of expertise in health behaviour change and the development of educational learning programmes. The ISP provides remediation to doctors and undergraduates who are struggling with their performance. Debbie developed this programme five years ago and it now has national recognition. She is member of the National

Clinical Assessment Service steering group reviewing remediation for doctors. Debbie is currently working on her MD, which relates to fitness for work and the GP consultation. This work has been commissioned by the Department for Work and Pensions and aims to produce an e-learning programme for GPs that is informed by motivational interviewing.

Melody Rhydderch is a chartered occupational psychologist based at the Communication Skills Unit at Cardiff University, having worked in the health service since 1990. She has recently completed a PhD at the Centre for Quality of Care at Nijmegen University in the Netherlands. Her research interests and published work span organisational and individual development. At present, she is working on research aimed at identifying bio-psychosocial factors associated with underperformance amongst doctors in collaboration with Dr Debbie Cohen.

Dr Anthea Lints is a Director of Postgraduate Education in the London Deanery as well as a part-time principal in a practice in Chelmsford, Essex. Her Deanery responsibilities include recruitment and career advice.

Dr John Launer is both a GP and family therapist. He currently works as Associate Director and Clinical Supervision Lead at the London GP Deanery, and as Senior Clinical Lecturer in General Practice and Primary Care at the Tavistock and Portman NHS Trust. He has a particular interest in the cross-fertilisation of ideas between primary care and the mental health professions and is the author or co-editor of *Narrative-based Primary Care: a practical guide* (Radcliffe Publishing; 2002), *Supervision and Support in Primary Care* (Radcliffe Publishing; 2003) and *Reflecting on Reality: psychotherapists at work in primary care* (Karnak Books; 2005).

Introduction

> Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted.
>
> Albert Einstein, physicist, 1879–1955

> One test does not improve learning any more than a thermometer cures a fever. We should be using these tests to get schools to teach more of what we want students to learn, not as a way to punish them.
>
> Heubert, Professor of Education, Columbia University

Assessment is arguably the most important method of driving up standards and yet 'there is probably more bad practice and ignorance of significant issues in the area of assessment than in any other aspect of higher education'.[1] This is in spite of the fact that the last two decades have seen major developments in the assessment of clinical competence. This has been evident in the amount of research and written output and in the number of major international conferences that have focused on the topic ('Ottawa' and 'Cambridge' Conferences, etc.). Also, we have much broader notions of what assessment should be doing than in the past. There is potential for information overload to a novice in the growing field of assessment.

The core aim of this book is to be 'a collation of original sources' on assessment throughout medical education. It is set out as an 'assessment journey' highlighting the broad range of references to other current literature in the area of assessment and is intended to act as a portal to the relevant literature on assessment.

Medical science is growing and changing with new drugs, new technology, etc., with the prospect of limitless spending on health. Along with these changes, society is better informed with growing expectations and a growing tendency to litigate when things go wrong. The search for better ways of assessing competence is the order of the day, and legislators need to feel doctors can do a good job; confidence in doctors also affects allocation of resources to healthcare.

Assessment has tended to be a private and intimate affair, but given the importance of assessments in that they determine the student's diplomas and future career, it probably should be scrutinised most carefully. In the past, one accepted the pass marks given out by examining bodies, but we are no longer prepared to be mystified by psychometrics. Society now expects the basis of assessments to be explained and to be able to understand the explanation. Also, assessments have become more closely integrated into the curriculum. The priority is to clarify exactly what we wish to assess and ensure the procedures used provide a valid reflection of the relevant performance.

The concept of assessment has broadened and is seen as having multiple purposes. There has been a move away from comparison/competition among students to what has or has not been learned and a greater emphasis on criterion-referencing than peer-referencing. There is a move to a descriptive approach rather than just raw marks, grades or statistical manipulation. This has led to a greater emphasis on formative assessment to improve practice rather than summative assessment as in the past. Harden uses the bicycle as a useful model when considering the relationship between teaching and assessment.[2] The front wheel represents teaching and learning whilst assessment is represented by the rear wheel; problems occur when the wheels go in different directions or are missing. Assessment can make learning more effective. Norman summarises the relationship between assessment and learning when he says, 'The curriculum tells you what the faculty is doing; the examination systems tell you what the students are doing'. In effect, we are moving from a testing and examination culture to an assessment culture.

Recently, assessment has taken a higher profile and is required to deliver on a wide range of outcomes as follows.

- Support teaching and learning by fostering learning. Feedback to learners and teachers as to what has been learnt and what has not been achieved. It may help to explain why some things have not been learnt and improve teaching through evaluation, consequently, driving the curriculum and teaching.
- Provide information about students, teachers and schools. By providing valid information, it aims to help make sensible and rational decisions about courses, careers, etc.
- Act as selection and certificating device as well as prediction of future success, but it has to be noted that assessment is always a snapshot, and that competence is not always synonymous with performance. It can inform systems for certification and licensure.
- Accountability procedures, in particular, ensuring the safety of the public, but also to set standards and monitor the quality of education. Formulation of policies and directing resources including personnel and money may also be a consequence.

Functions of assessment

We need to define the functions of the assessment in question: is it to diagnose problem students, evaluate teaching/curricula, lead to qualifications, perform selection for jobs, or for sorting people into their roles in society? These different purposes will need different methods, and subsequent chapters will deal with these issues in more detail.

- Professional – supports teaching and learning. This is enabling rather than limiting.
- Accountability – e.g. government and tax-payers need reassurance.
- Certification purposes.

Educational assessment

There is a long-held central argument amongst educationalists that assessment should play a crucial part in any educational process. Whenever and wherever learning occurs then it follows that the learner, the teacher and other interested parties or society at large will express a desire to understand what has occurred in terms of the learning process and its outcomes. It may therefore be reasonably advanced that sound education has a direct link to sound assessment although its purpose will vary according to different situations. Macintosh and Hale classified six possible purposes of assessment:[3]

- diagnosis
- evaluation

- guidance
- grading
- selection
- prediction.

This classification is helpful in understanding the broad spectrum of functions that educational assessment can be used for. Additional purposes may be added to this core list, i.e. motivation and development, which underline a learner-centred approach and a means of supporting learning rather than just to indicate current or past achievement.

Heron[4] called 'the redistribution of educational power' the process whereby assessment becomes not just something which is 'done to' learners but also 'done with' and 'done by' learners.[5] Gipps suggested that educational measurement should be more dynamic where the focus is the learning potential of the student.[6] Assessment may be defined as 'any method that is used to better understand the current knowledge that a student possesses.' This can be of varying degrees of subjectivity and can be formative or summative. The purposes of assessment need to be clearly explained. Staff, students and the outside world need to be able to see why assessment is being used, and the rationale for choosing each individual form of assessment in its particular context. Assessment should be a developmental activity in that it should provide learners with an opportunity to reflect on their practice and their learning; thereby promoting deep learning. The assessment instruments and processes need to be reliable and consistent. Assessment can take many forms, and it can be argued that the greater the diversity in the methods of assessment, the fairer assessment it is to students. These instruments and processes need to be valid in that an assessment method should be chosen which directly measures what it is designed to measure.

In 1984, Frederiksen stated that the time and effort put into learning was determined by what the tests measured, and hence there is a bias against teaching important skills that are not measured.[7] A number of reviews have shown the powerful effects of testing on teaching and the curriculum.[8, 9] Madaus relates this to the Heisenberg 'uncertainty principle': one cannot measure things without affecting them. It is often said that assessment drives learning and there are many examples, but there is the potential that assessment skews learning into a rote-learning and superficial model.[10]

Characteristics of good assessment

1 Validity. The concept of test validity is to what extent an assessment actually measures what it is intended to measure, and permits appropriate generalisations about students' skills and abilities. If the test is valid, we can safely generalise that the student will be likely to do as well on similar items not included in the test. A test might be valid for one purpose but inappropriate for other purposes, and evidence of validity needs to be gathered for each purpose for which an assessment is used. There is the problem of 'assumption of universality' – that a test score has the same meaning for all individuals. However, this often depends on the test's construct validity; if a test assesses the attribute it is supposed to then it is 'valid'. The second problem is uni-dimensionality since many attributes are multi-dimensional.

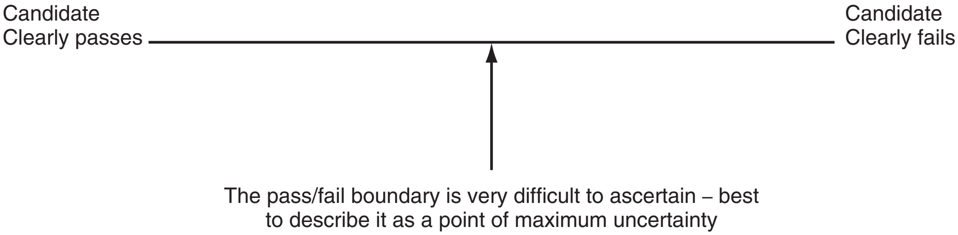
2 Reliability. This attempts to answer if the assessment results for this person or class are similar if they are gathered at some other time or under different circumstances, or if they are scored by different raters. Inter-rater reliability requires that independent raters give the same scores to a given student response.
3 The content of the tests (the knowledge and skills assessed) should match the competencies being learnt and teaching formats used.
4 The test items should represent the full range of knowledge, skills and attitudes that are the primary targets of instruction.
5 Expectations for student performance should be clear.
6 The assessment should be free of extraneous factors that unnecessarily confuse or inadvertently cue student responses.

Ethical issues

Assessment has the potential to improve learning for all students but in the past has acted as a barrier in that assessments have been used to label students. There is a potential to be unfair to minority students, particularly in the assessment of language. Questions of fairness arise not only in the selection of performance tasks but in the scoring of responses, and the training and calibrating of examiners is critical in this regard.

Standard setting

As professional certification grows in importance, the method used to determine the passing scores will come under increasing scrutiny. John Norcini call the decision of who must pass or who must fail as the 'mind of God' and it may be that the perfect standard setting will forever elude us as does the Holy Grail.


Candidate
Clearly passes
Candidate
Clearly fails
The pass/fail boundary is very difficult to ascertain – best to describe it as a point of maximum uncertainty

This is a balancing act based on the following expectations.

1 Expectations based on test – what would constitute as capable performance.
2 Expectations based on candidates – highly qualified, properly selected, well trained.
3 Expectations of institution/society – how many passes we need or can tolerate.

Test expectations

Test-centred methods rely on informed expert judgement but need to demonstrate due diligence and be supported by a body of research. These vary according to test type, and there is a mixture of measuring and judging. In the former, we count points, report a score and aim for precision. However, when measuring, only the measurable is important and there is a problem as to how we give meaning to measurements. With regard to judging, the standard resides within the judge and judgements are encoded as a number/grade. There are many important things that cannot be measured but can be judged. However, there are numerous sources of error (e.g. examiners).

The various methods are discussed more fully in Chapter 1. However, some general principles are that standards setters must understand the purpose of the test, know the contents, and be familiar with candidates' teaching. Also, a significant number of standard setters needs to be involved, and they need to represent all stakeholders.

Institutional expectations

Self-view of institution is an important variable ('can't be a good test if so many pass') but also the law of supply and demand applies. The value of a qualification as perceived by existing holders, employers, society, is an important variable, as are concerns of finances or influence.

Wass discusses whether we are justified in abandoning old methods in favour of the new and the driving forces behind this.[10] There is an increasing focus on professional accountability driven by political pressure and a need to ensure doctors are 'fit for purpose'. There is a current move away from testing written theoretical knowledge to assessments in the workplace and there is potential for assessments to become more integrated.

Assessment drives learning, and studies have shown that the assessments used do determine what students learn[11] and the way in which they learn, as well as determining the way in which trainers teach and what they teach. Ideally assessments should always play a positive role in the learning experiences of students. With an increasing focus on competency-based outcomes, Wass highlights a concern that 'professionalism of practice, i.e. the higher order meta-cognitive levels of performance and understanding' may not be assessed.[12] One needs to consider a combination of methods and we need to be cautious in discarding the traditional methods which may still have a role in testing a more integrated process.

Assessment journey

The assessment journey is discussed further in Chapter 1 which covers the principles of assessment design. Chapter 2 looks at assessment in the undergraduate curriculum, past, present and future. Chapter 3 focuses on assessment in the Foundation Programme, which has an emphasis on the assessment of performance in the workplace. In Chapter 4 the use of records of in-training assessments (RITAs) is critiqued, noting wide variation in what has been developed.

In the chapter on assessment for recruitment (Chapter 5) the use of assessment centres (ACs) for this purpose is described. The authors in Chapter 6 argue for the importance of workplace-based assessment, especially its strength as a developmental tool, with the 'reconnection' of teaching and assessment. Chapter 7 covers the training of supervisors in the context of Foundation Programme assessments, and the 'importance of considering the teacher's aims and objectives ... and the way in which this will be linked with the experience and needs of the learners.' The development and use of simulated surgeries as an assessment tool, and the challenges of standard setting, are covered in Chapter 8. In Chapter 10 the authors argue for the importance of summative assessment for GP training as a driver for an increased understanding of criterion audit method in general practice, albeit that current trends determine this as predominantly historical.

We move with Chapter 10 to the assessment of autonomous professionals, firstly the strengths and weaknesses of self-assessment. Chapter 11 looks at practice-based assessment, the accreditation of practices, and links to quality control and service improvement. In Chapter 12 the authors demonstrate the value of multiple perspectives in the assessment of doctors whose performance has given rise to concern, and argue for the importance of flexibility in remediation. The value of the rigorous selection of assessors, and the mechanism of appeals and complaints processes, is covered in Chapter 13. Finally in Chapter 14 the authors reflect on '... the position of medicine in contemporary culture and society, and what this means for the future of assessment in the medical profession.'

Summary

Methods of assessment are determined by our beliefs about learning and how these are influenced from today's cognitive perspective such that meaningful learning is reflective, constructive, and self-regulated. Acquisition of knowledge and skills is not sufficient alone; we need to be able to apply the knowledge, skills and strategies learnt, and in turn these can be the appropriate targets of assessment. There is a movement away from traditional, multiple-choice tests to assessments that include a wide variety of methods and so provide for more meaningful assessments which can better capture significant outcomes in order to assure their future success.

The scenario below highlights some of the issues discussed above.

Scenario: what sort of assessment and why?

Dr Cool (*AC*) has a background of being an underperforming medical student who needed remedial help, and developed a 'phobia' about examinations, and assessment of all kinds. He became a non-vocationally trained GP and initially had very ambivalent feelings about continuing in a medical career, but became, on the basis of diverse clinical experience, an enthusiast for interpersonal communication and the understanding of context, and the privileged position of the GP in relation to both. He had also been deeply involved over a 10-year period in transforming an underperforming inner-city practice into a flagship teaching and training practice, and realised that 'building up that practice from … nothing to what it was, that this had been a very important learning experience for me … the power of learning from experience … particularly if you understand your own ability to change your life … to take the initiative.' He worked as a GP tutor and felt a 'complete affinity' with GPs who came to the GP lunchtime meetings from smaller and struggling practices, because, ' … I could completely identify with these people, and they felt comfortable with me because I wasn't a threat to them … [and] … because I had [a body of professional] experience quite often that I could share with them.'

Dr Swot (*AS*) has a background of being a high achiever, who, when training as a GP, motivated by 'trying to get some external validation that I'd reached the acquired level of competence,' passed all the exams it was possible to take. 'So when I did O&G I did the DRCOG, when I did Paeds, I did the DCH, when I did Community Paeds, I did DCCH, when I did geriatric medicine, I did DGM, when I did general practice I did MRCGP'.

They now work together with doctors whose performance has given rise to concern, and are in conversation about their very different routes towards an appreciation of the importance of well designed assessments:

AS: Well I suppose throughout my own career I was wary of the fact that there is so much to cover in general practice and doing the vocational training scheme, it was in some ways, I wasn't sure at the end of the six months of every post whether I actually had acquired all of the knowledge. And that probably showed a bit of insecurity in myself. Personally I think it was about trying to get some external validation so I'd reached the acquired level of competence.

AC: Everything I learned about general practice I learned by being a GP through continuing professional development. And because of my rather difficult experiences as a student, I was kind of avoidant in relation to being assessed. We developed the premises, we computerised the practice, we got an excellent practice manager and although I left in the mid-nineties it became a training practice not long after that. So that in a sense is kind of where my world of experience has met with your world of experience and in relation to assessment procedures, their design and their use.

AS: Yes, certainly you've got a much more reflective life. I just wonder for example you mentioned about your knowledge not being good and you avoiding exams but what made you come to that judgement that your knowledge wasn't good? Was it about confidence or was it something objective?

AC: Well I suppose I'd been kind of told that when they passed me in medical school that day, and said to me, 'your knowledge is not good, but its time you had some responsibility'. I'd been kind of told that. What I like to do is to think about things, you know, kind of absorb stuff and think about things and then you know make a decision on the basis of a certain amount of reflection.

AS: There is a degree of honesty of knowing your limits.

AC: Absolutely, and being completely open with people about that, you know, so if I don't know then I'm either going to find out in front of them or I'll speak to somebody who knows, or I'll make sure.

AS: So that's interesting that we've got a very similar thing. The other thing I was quite interested in was what you mentioned about well-designed assessment process and one of the things you talked about a reason for it was defensibility. What are the characteristics of a well-designed assessment process?

AC: Well you know there is this term, 'fit for purpose'. I suppose the questions that you would have to ask, who are your target group, who are the people who are being assessed and what is to be achieved at the end of this assessment process. So in all of those different ways, it's got to be fit for a purpose, hasn't it?

AS: Yes, it's a bit like Wilson's screening criteria. The assessment has to be acceptable to the assessee and the assessor and it needs to have reasonable reliability and validity as well.

AS: And I suppose that you've got a list of competencies that you expect to achieve.

AC: Yes. It might be possible to design an assessment which tests those kinds of competencies from mainstream general practice life and still not get the answer to the question that needs to be asked. For instance, in a performance context, in an ideal world, you would need to go beyond clinical competence per se to get a broader, better picture.

AS: I suppose the assessment process itself does drive the learning doesn't it?

AC: It can skew it, yes.

AS: Skews it – they'll act on it and they'll play on it and they'll be coached by examiners to a level that they can perform and pass. To be cynical about the assessment thing, how much of the assessment can be coached. Another problem we've got is sampling and cost; you may have to limit the areas we test or pick a cheaper test.

AC: These are all the problems in relation to designing and running good assessments aren't they?

AS: Quality assurance is the key.

AC: I suppose in an ideal world I would hope to be able to solve this kind of problem with evidence from assessment processes. In a performance situation, we have limited resources as well as having to defend against legal challenge.

AS: Yes, and again I think it's an important point because most assessments are at risk of legal challenge. A standard setting exercise allows some defensibility. The other thing that's been missing is knowing when to say to doctors that you are not good enough. I think it is unfair to allow doctors the hope that they are able to come back into practice when they are so far below the required standard and have little hope of achieving it.

AC: Well it has so many implications doesn't it, to somebody who's set their heart on doing it …

AS: As for assessment in the real world, what I think we need to be thinking about is making the right diagnosis. To make this diagnosis we need to take a history, examination and then do various tests. It is amazing how colleagues are happy to 'do an assessment such as MCQ' without the preliminary history etc on the basis of a PCT having concerns regarding knowledge. We don't do an MRI scan on children whose mother's are worried about their headaches without taking a history, etc. There seems to be a prevalent view that we can diagnose on the basis of one test! I would suggest that once a diagnosis is made, we then share the management plan in order to improve compliance on uptake of the 'educational prescription'.

AC: That's right. And you know when the CMO makes public any decision about the relationship between appraisal and revalidation for instance, that will give us a clearer idea of the place of appraisal in that. Because if appraisal is going to have some role, there is lots of potential there for looking at the skills of the appraisers.

AS: There is … but the other question is how to assess the assessors. I'm doubtful of people becoming assessors without them being assessed themselves. Similarly appraisers, if they're going to have a role in revalidation, there are skills they need and they need to be assessed to that.

AS: And the other question is, if you are going to start probing you need some guidance on how much probing you do, and how much of an assessor you become and not just an appraiser … .

AC: And they are going to have to defend their actions, you know.

References

1. Boud D. Assessment and learning: contradictory or complementary. In: Knight P (ed.). *Assessment for Learning in Higher Education.* London: Kogan Page; 1995. p. 35–48.
2. Harden RM. Assessment, feedback and learning. In: *Approaches to the Assessment of Clinical Competence.* International Conference Proceedings. Norwich: Page Brothers; 1992. p. 9–15.
3. Macintosh HG, Hale DE. *Assessment and the Secondary School Teacher.* London: Routledge and Kegan Paul. 1976.
4. Heron J. Assessment revisited. In: Boud D (ed.) *Developing Student Autonomy in Learning*. London: Kogan Page. 1981.
5. Harris D, Bell C. *Evaluating and Assessing for Learning.* London: Kogan Page. 1990.
6. Gipps CV. *Beyond Testing – towards a theory of educational assessment.* London: Routledge Falmer. 1994.
7. Frederiksen N. The real test bias: influences of testing on teaching and learning. *Am Psychol.* 1984; **39**: 193–202.
8. Airasian P. Measurement driven instruction: a closer look. *Educational Measurement: issues and practice.* Oxford: Blackwell Publications. 1988. p. 6–11.
9. Corbett D, Wilson B. *Raising the Stakes in State-wide Minimum Competency Testing*. Politics of Education Yearbook. New York: Sage Publications. 1988. p. 27–39.
10. Wass V. *The Assessment of Clinical Competence in High Stakes Examinations – are we justified in abandoning old methods in favour of the new?* Maastricht: Universitaire Pers Maastricht. 2006.
11. Campion P, Foulkes J, Neighbour R, Tate P. Patient-centredness in the MRCGP video examination: analysis of a large cohort. *BMJ*. 2002; **325**: 691–2.
12. Wass V. Is competency based training and assessment the way forward? *BMJ*. Career Focus 2004; **329**: 220–1.

Chapter 1

The principles of assessment design

Val Wass, Reed Bowden and Neil Jackson

Introduction

Education is inceasingly regarded as a life-long continuum. Changes introduced by Modernising Medical Careers aim, through the introduction of the Foundation programme, to provide more support for a doctor's transition from undergraduate to postgraduate training.[1] This is bridged by a more formative approach to assessment focused on performance in the workplace and is radically different from summative methods traditionally used in medical schools. As new structures for training emerge, royal colleges are revising their vocational training curricula and examinations guided by the principles set down by the Postgraduate Medical Education Training Board (PMETB).[2] They aim to support this educational continuum to ensure doctors emerge from training with clear frameworks for keeping up to date and continuing their professional development.

Assessment is intrinsic to these educational changes. New postgraduate curricula are now more focused on achieving competence.[1] There is concern that assessment is becoming too focused on the demonstration of competence and subsequently trivialised.[3] Professionalism in the 21st century requires a higher standard than mere competence. The Royal College of Physicians report on Medical Professionalism highlights the need for professional excellence, not just 'capacity to do something'.[4] The need to develop newer packages of assessment to accommodate this range of needs is becoming clear.[5] Huge demands are being made on assessment methods to address these changes: from testing the 'ability to do' versus 'excellence'; competence of the 'novice' versus the 'expert'; and resolving the tensions between 'revalidation' and 'appraisal'.

This chapter aims to set out the basic principles which underpin the choice and design of assessments, taking a broad view of available methods and processes for standard setting to validate and ensure the processes used are 'fit for purpose'. The basic structure offered supports the subsequent chapters that outline in more detail how assessment is keeping abreast of the challenges presented by changes in education in the 21st century.

Designing assessments

Whether assessment occurs in the workplace or in the examination hall, it must be carefully planned and delivered. Decisions need to be made on key issues (*see* Box 1.1 for summary).

Box 1.1: Summary of key questions to address when designing and evaluating an assessment

Educational purpose?	Align the assessment with the educational goals and do not create too many assessment hurdles.
Summative or formative?	Be clear on the purpose of the test. Low or high stakes.
Competence or performance?	Check against Miller's triangle. At what level of competency will your assessment measure?
What is the blueprint?	Plan the test against the learning objectives of the course or competencies essential to the specialty.
What is the standard?	Define end point of assessment. Set the appropriate standard, e.g. minimum competence in advance.
Are the methods valid?	Select appropriate test formats for the competencies to be tested. This invariably results in a composite assessment.
What level of reliability?	Sample adequately. Clinical competencies are inconsistent across different tasks. Test length is crucial if high stakes decisions are required. Use as many examiners as possible.
Is it feasible and acceptable?	Practicalities of delivery, e.g. cost, appropriately trained examiners.

What is the educational purpose of the assessment?

Assessment drives learning. Ideally this should not be the case. The curriculum should motivate learning in any clinical course and assessment be planned at a later date to ascertain that the required learning has occurred. In actuality at all levels of education, whether undergraduate[6] or postgraduate[7], students feel overloaded by work and prioritise those aspects of the course that are tested. To overcome this, the assessment package must be designed to mirror and drive the educational intent. The balance is a fine one. Pragmatically, it is the most appropriate engine to which to harness the curriculum. Yet one can be too enthusiastic. Creating too many burdensome time consuming assessment 'hurdles' can detract from the educational opportunities of the curriculum itself.[8] The assessment must have clarity of purpose and be designed to maximise learning. It is important to be clear on both the goal and the direction of travel. Careful planning is essential. In reality the first decision lies in agreeing how to maximise educational achievement. This cannot be an afterthought.

What is the intent of the assessment: formative or summative?

To promote deeper learning, assessment should be *formative*. Students must learn from tests and receive feedback to build on their knowledge and skills. If they do not meet the standard, there should be further opportunities to try again until the competency is ultimately achieved. Feedback should encourage students to identify their strengths and weaknesses and map their progress. Weak students should be identified and given remedial help. This is the focus of assessment in the Foundation Programme.[1] Feedback requires support through trained mentoring; an issue which will be addressed in subsequent chapters on the Foundation Programme and RITAs.

At the same time, with an increasing focus on the performance of doctors and public demand for assurance that doctors are competent to practise, assessment must, at times, have a *summative* function. Tests of clinical competence are necessary to make an end point decision on whether a doctor is fit to practise or not. Such tests generally take a 'snapshot' of ability at a defined moment. The candidate has a fixed time frame and number of attempts in which to succeed. The two forms of assessment are stark in contrast (*see* Box 1.2). Both are necessary.

Box 1.2: Formative versus summative assessment

Formative assessment:
Breaks learning into manageable modules
Allows repeated attempts to master the content of each module
Is not perceived as threatening (low stakes)
Summative assessment:
Is an end-point examination
Can block intended career progression (high stakes)
Is perceived as threatening

This raises a challenge for all involved in medical education. It is difficult for a test to be simultaneously formative and summative. Yet if assessment focuses only on certification and exclusion, the all-important influence on the learning process will be lost. Superficial learning, aimed purely at passing the test, can result. The PMETB principles emphasise the importance of giving students feedback on all assessments to encourage reflection and deeper learning. All those designing and delivering high stakes tests should explore ways of enabling this and make their intentions transparent to candidates.

What aptitudes are you aiming to assess?

Knowledge, competence or performance?

Miller's pyramid (*see* Figure 1.1) provides an important framework for establishing the aim of an assessment.[9] It conceptualises the essential facets of clinical competence. The base represents the knowledge components of competence: '*knows*' (basic facts) followed by '*knows how*' (applied knowledge). The progression to '*knows how*' highlights that there is more to clinical competency than knowledge alone. '*Shows how*' represents a behavioural rather than a cognitive function, i.e. it is 'hands on' and not 'in the head'. Assessment at this level requires an ability to demonstrate a clinical competency.

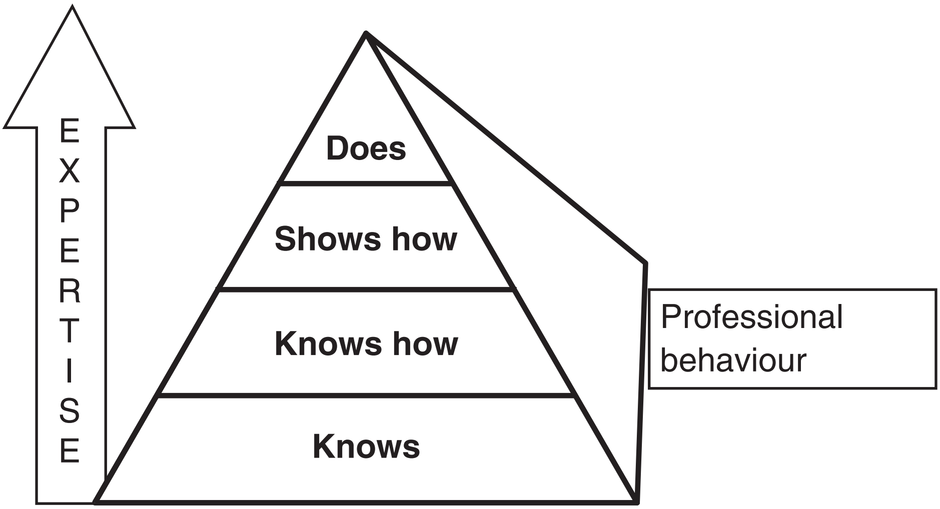


Figure 1.1: Miller's pyramid of clinical competence.[9]

The ultimate goal for a valid assessment of clinical aptitude is to test *performance,* i.e. what the doctor actually *does* in the workplace. Over the last four decades assessment research has focused on developing valid ways of assessing the summit of the pyramid, i.e. a doctor's actual performance.[10,11] Subsequent chapters will explore in more detail the extent to which this has been achieved. We have modified the triangle (Figure 1.1) to include *'professional behaviour'* as a third dimension. Assessment design must develop to address the values and behaviours intrinsic to modern medical professionalism.[2] Methodology for achieving this remains challenging.[12]

At what level of expertise?

Any assessment design must accommodate the progression from novice through competency to expertise. It must be clear against what level the student is being assessed. Developmental progressions have been described for knowledge as in Bloom's taxonomy summarised in Figure 1.2.[13] Frameworks are also being developed for the clinical competency model.[14,15] Work remains to be done in incorporating models of professional development in expertise into the assessment methods (*see* Chapter 6). When designing an assessment package, conceptual clarity is essential to identify the level of expertise anticipated at that point in training. The question, 'is the test appropriate for this level of training?' must always be asked. It is not uncommon to find tasks set in postgraduate examinations which assess basic factual knowledge at undergraduate level rather than applied knowledge appropriate to the candidate's postgraduate experience.

Deciding the content of the assessment: blueprinting

Once the purpose of the assessment is agreed, test content must be carefully planned against the intended learning outcomes, a process known as 'blueprinting'.[16] Medical schools follow the General Medical Council (GMC) guidelines for Undergraduate Education.[17] In the past blueprinting has been difficult for postgraduate collegiate examinations, where curriculum content remained more broadly defined.[18] To address these difficulties and the requirements of PMETB, colleges are now revising their curricula developing clear learning outcomes.

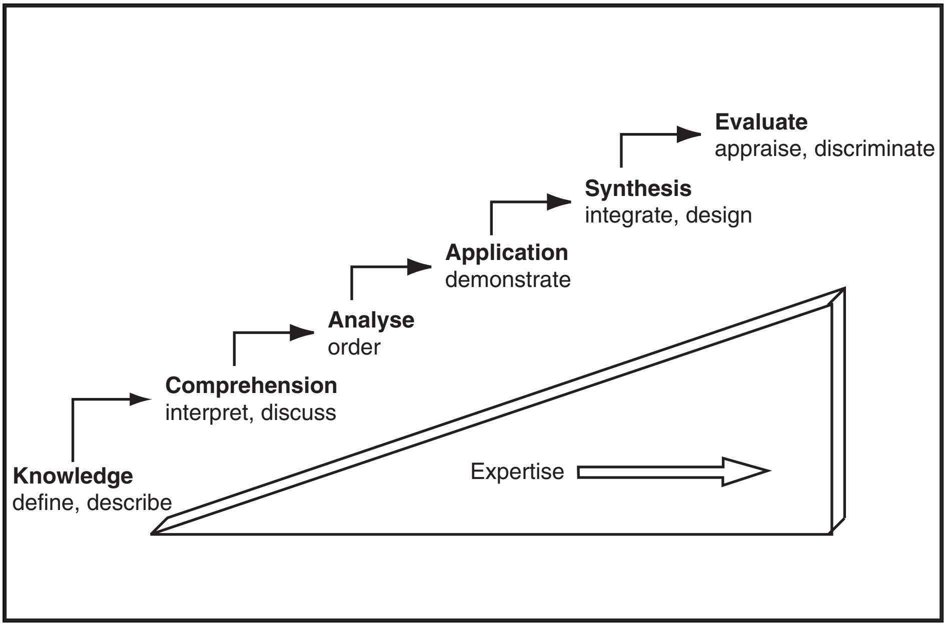


Figure 1.2: Hierarchy of knowledge: Bloom's taxonomy.[13]

Blueprinting requires the following.

- A conceptual framework. A framework against which to map assessments is essential. PMETB is recommending Good Medical Practice[19] is used for UK postgraduate assessments.[2] Alternatives such as the behavioural framework 'knowledge, skills and attitudes' can be employed.
- Context specificity. Blueprinting must also ensure that the contextual content of the curriculum is covered. Content needs careful planning to ensure students are comprehensively and fairly assessed. Professionals do not perform consistently from task to task.[20] Wide sampling of content is essential.[16] Context of learning impacts on clinical competence in a most profound way. This has been the main catalyst to the development of Objective Structured Clinical Examinations[21] and the demise of testing on a single long case.[22] Sampling broadly to cover the full range of the curriculum is of paramount importance if fair and reliable assessments are to be guaranteed (*see* Table 1.1 for an example of a blueprint used to identify stations for a 20-station undergraduate OSCE). Blueprinting written examinations is of equal importance.
- The assessment programme must also match the competencies being learnt and the teaching formats being used. Many medical curricula define objectives in terms of knowledge, skills and attitudes. These cannot be validly assessed using a single-test format. All assessments must ensure the test being used is appropriate to the objective being tested. To assess clinical competence validly, we are moving from a battery of different examinations to an assessment package where performance in the workplace can be included alongside high-stakes examinations such as multiple-choice tests.[11] No single one can be valid, given the complexity of clinical competency itself.

Table 1.1: Example of a blueprint for a 20-station undergraduate OSCE

OSCE case selection blueprint	*Conceptual framework*					
Context: primary system or area of disease	*Diagnosis*	*Examination*	*Management*	*Communication*	*Practical skills*	*Ethics*
Cardiovascular	X				X	
Respiratory		X			X	
Neurological psychiatric	X			X		
Musculo skeletal		X	X			
Endocrine and oncological	X			X		
Eye/ENT/skin		X				
Men's/Women's and sexual health	X					X
Renal/urological			X		X	
Gastro intestinal		X	X			
Infectious diseases				X	X	
Other			X			

- Triangulation. As assessment design develops, the need to combine assessments of performance in the workplace alongside high stakes competency has been increasingly recognised. The complexity of measuring professional performance is becoming better understood.[5] It is important to develop an assessment programme to build up evidence of performance in the workplace and avoid reliance on examinations alone. Triangulation of observed contextualised performance tasks of 'does' can be assessed alongside high-stakes competency based tests of 'shows how'.[23] The GMC's performance procedures, where workplace assessments are triangulated with a knowledge test and an objective structured clinical examination provide such a model.[24]

Deciding who should pass or fail: standard setting

Inferences about examinee performance are critical to any test of competence. When assessment is used for summative purposes, the pass/fail level of a test has also to be defined. Well-defined and transparent procedures need to be set in place to do this.[2]

Norm versus criterion referencing

Comparison of performance to peers, i.e. norm referencing, can be used in examination procedures where a specified number of candidates is required to pass. Performance is described relative to the positions of other candidates and a fixed percentage fail, e.g. all candidates one standard deviation below the mean. Thus the variation in difficulty of the test is compensated for. However, variations in ability of the cohort sitting the test are not taken into account. If the group is above average in ability, those who might have passed in a poorer cohort will fail. This is clearly unacceptable for clinical competency licensing tests, which aim to ensure that candidates are safe to practise.

A clear standard needs to be defined, below which the doctor would not be considered fit to practise. Such standards are set by criterion referencing, where the minimum standard acceptable has to be decided. The reverse problem now faces the assessor. Although differences in candidate ability are accounted for, variation in test difficulty becomes the key issue. Standards should be set for each test, item by item. Various methods have been developed to do this: 'Angoff', 'Ebel', 'Hofstee'.[25,26,27] These can be time consuming but essential and enable a group of stakeholders (not just examiners) in the assessment to participate. PMETB (see Box 1.3) encourages the involvement of lay judges in the standard setting process.[2]

Box 1.3: Summary of PMETB principles for assessment

1 Methods must reflect the assessment's intended purpose/content
2 Reference assessment content to Good Medical Practice
3 Ensure methods used to set standards are in the public domain
4 Involve lay members in the assessment process
5 Have mechanisms for giving students feedback on performance
6 Use appropriate criteria for examiner training
7 Use standardised documentation which is available nationally
8 Be sufficiently resourced

More recently methodology has been introduced using the examiner cohort itself to set the standard. Examiners, after assessing the candidate, indicate which students they judge to be borderline. The mean mark across all examiners (and there is invariably a range) is taken as the pass / fail cut off.[28] The robustness of this method across different cohort of examiners remains to be seen.[29] The choice of method will depend on available resources and the consequences of misclassifying passing and failing examinees.

Evaluating the assessment: validity and reliability

Two key concepts, validity and reliability, are essential when evaluating and interpreting assessments.

- *Validity:* Was the assessment valid? Did it measure what it was intended to measure?
- *Reliability:* What is the quality of the results? Are they consistent and reproducible?

Validity is a conceptual term which should be approached as a hypothesis and cannot be expressed as a simple coefficient.[30,31] It is evaluated against the various facets of clinical competency. In the past these facets have been defined separately acknowledging that appraising the validity of a test requires multiple sources of evidence (*see* Table 1.2).[32]

Table 1.2: Traditional facets of validity

Type of validity	*Test facet being measured*	*Questions being asked*
Face validity	Compatibility with the curriculum's educational philosophy.	What is the test's face value? Does it match up with the educational intentions?
Content validity	The content of the curriculum.	Does the test include a representative sample of the subject matter?
Construct validity	The ability to differentiate between groups with known difference in ability (beginners versus experts).	Does the test differentiate at the level of ability expected of candidates at that stage in training?
Predictive validity	The ability to predict an outcome in the future, e.g. professional success after graduation.	Does the test predict future performance and level of competency?
Consequential validity	The educational consequence of the test.	Does the test produce the desired educational outcome?

It is now argued that validity is a unitary concept which requires these multiple sources of evidence to evaluate and interpret the outcomes of an assessment.[30] Intrinsic to the validity of any assessment is analysis of the scores to quantify their reproducibility. An assessment cannot be viewed as valid unless it is reliable. Two aspects of reliability must be considered.

1 *Inter-rater reliability:* which correlates the consistency of rating of performance across different examiners.
2 *Inter-case reliability:* which quantifies the consistency of performance of the candidate across the cases.

The latter gives a measure of the extent context specificity has been addressed by the assessment blueprint to ensure candidate performance is accurately rank ordered. It is a quantifiable measure which can be expressed as a coefficient either using Classical Test theory[33] or Generalisability analysis.[34,35] A perfectly reproducible test would have a coefficient of 1.0, i.e 100% of the candidates would achieve the same rank order on re-testing. In reality, tests are affected by many sources of potential error such as examiner judgements, cases used, candidate nervousness and test conditions. High-stakes tests generally aim for a reliability coefficient of greater than 0.8, whereas for more formative assessments lower reliability scores are acceptable.

Sufficient testing time is essential to achieve adequate inter-case reliability. It is becoming increasingly clear that, whatever the test format, test length is critical to the reliability of any clinical competence test to ensure breadth of content sampling.[5,6] Increasing the number of judges over different cases improves reliability but to a lesser extent. In an oral examination a sampling framework where a candidate is marked by a series of ten examiners each asking just one question produces a much more reliable test than one examiner asking a series of ten questions.[36,37] Examiners make judgements rapidly.[38] The challenge now is to introduce sample frameworks into workplace-based assessments of performance which sample sufficiently to address issues of content specificity.

What are the practical issues of assessment design?

The practicalities of delivering assessments cannot be ignored. The 'utility equation' defined by Cees van der Vleuten provides an excellent framework for assessment design.[39] It acknowledges that the choice of tool and aspirations for high validity and reliability are constrained by the restraints of feasibility, e.g. resources to deliver the tests and acceptability to the candidates, e.g. level of examination fee. No test can score uniformly high on all five factors. Some trade off is inevitable to ensure the purpose of the assessment is achieved.

The utility equation summarises the position.

$$\text{utility} = \text{reliability} \times \text{validity} \times \text{feasibility} \times \text{acceptability} \times \text{educational impact}$$

Assessor selection and training

In subsequent chapters the contrasting roles of assessors involved in formative and summative processes across the spectrum of assessment will be explored. These range from educational supervision to summative judgements of fitness to progress in high-stakes examinations. Work from the Royal College of General Practitioners emphasises the importance of selecting and training assessors.[40] Just as it cannot be assumed that any professional competent in their work can necessarily teach, the same applies to assessment. Not all teachers can make clear judgements or rank order performance consistently. Selection and training of assessors is essential to ensure they:

- have the skills
- understand the process of the assessment
- can address issues of equal opportunity.[41,42]

For those designing assessments the principles laid down by PMETB emphasise the importance of all these steps in assessment design (*see* Box 1.3). Current revision of assessments by colleges and universities is in place to address these recommendations.

Selecting the most appropriate assessment methods

Assessing the apex of Miller's pyramid, 'the does' is the international challenge of this century for all involved in clinical competency testing. The ensuing chapters will describe in detail progress across undergraduate and postgraduate assessments in both primary and secondary care as we move to do this. Here we aim to provide a brief overview appraising currently available assessment tools in the light of the above principles of assessment design.

The assessment of 'knows' and 'knows how'

Many examinations (undergraduate and postgraduate) focus on the pyramid base: '*knows*' (the straight factual recall of knowledge) and to a lesser extent on the '*knows how*' (the application of knowledge to problem solving and decision making).

Tests of factual recall can take a variety of formats. Multiple-choice question (MCQ) formats are universally the most widely used. Although time consuming to set, these tests have high reliability, because they can easily address issues of context specificity, i.e. a large number of items can be tested and marked within a relatively short time frame. A variety of question formats exist. Increasingly true/false MCQ formats are being replaced by single best answer and extended matching questions using short and long menus of options.[43,44] Some argue that only 'trivial' knowledge can be tested. By giving options, candidates are cued to respond and the active generation of knowledge is avoided. Although reliable, criticism of the validity of the MCQ has stimulated much research into alternative options.

Essays and orals as tests of knowledge have lost popularity over the years. This relates partly to reliability and partly to feasibility. It is difficult to produce highly reliable assessments using either tool because of problems in standardising questions,[37] inconsistency in marking[45] and lack of sufficient testing time to address context specificity. Undue pressure is placed on the examiner resource. Reliability can be achieved using short answer written formats[46] and also through more standardised orals[37] but both are resource intensive. Despite this, orals have remained popular in the UK, and other European countries on the grounds of validity. Many argue that the ability to recall and synthesise information can best be judged in the face-to-face encounter. Unfortunately, validity arguments in this case cannot easily be reconciled with reliability issues. Increased structuring of orals may be a way forward but, even then, attention to validity as well as reliability remains essential.[47]

The 'key feature' test developed in Canada avoids cueing by allowing short written 'uncued' answers to clinical scenarios and limiting the assessment of each scenario only to key issues.[48,49] This enables a large number of scenarios to be covered within a feasible time limit. Using the MCQ format attempts at focusing the content within the question formats using clinical scenarios or scientific extracts for critical appraisal are proving successful. Computer simulations can replace the written or verbal scenarios and, hopefully, with the development of multi-media, can be used to raise the level of clinical testing.[50,51,52] In the past the simulations have been complicated. Dynamic and complex situations have been created which require enormous resources rarely available at university or deanery level. A focus on short simulations to produce the required breadth for tests, which stimulate rather than cue responses, remains a challenge for those developing this test format.

The assessment of 'shows how' and 'does'

The current trend in curriculum development towards competency-based curricula[1] has stimulated increased focus on methods for assessing performance in the workplace at the 'does' rather than the 'shows how' level. Views on assessment methodology are changing.[5]

Originally when the need to address content specificity became apparent there was an international divergence in trends. North America was quick to abandon long cases and orals favouring the knowledge tests described above which covered high content, were reliable and legally defensible. Elsewhere the move away from traditional methods has been more gradual. Objective Structured Clinical Examinations (OSCEs) are now globally well established and orals are used less frequently.[53]

Traditional assessments: long and short cases and orals

These traditional methods stood to be challenged on the grounds of both authenticity and unreliability. Long cases were often unobserved. Thus this method, relying on the candidate's presentation, represented an assessment of '*knows how*' rather than '*shows how*'. Generally, only one long case and three or four short cases were used and context specificity not was not adequately addressed. Attempts have been made to improve the long case format; the Objective Structured Long Examination Record (OSLER)[54] and the Leicester Assessment Package.[55] Observation improves the validity of the long case.[56] Decreasing the length of time available to assess a case and allowing more cases to be assessed within a given testing time may also be an option.

Although unlikely to ever reach feasibility for high stakes testing, a better understanding of the psychometrics of these methods has reopened them to modification for use in the workplace. The 'mini-CEX' format,[57] introduced in the USA, is essentially a modification of an observed long case in the clinical setting. The method takes 'snapshots' of the integrated assessment by focusing on one of a range of predetermined areas, e.g. observation of history taking, the physical examination or the management of the case but not the entire process. Furthermore it is emerging that less than ten cases may be enough for a reliable judgement of clinical competency to be made.[58]

The Objective Structured Clinical Examination (OSCE)

As a potential solution to the problems of adequate sampling and standardisation of cases, the OSCE has gained increasing popularity on both sides of the Atlantic.[21] Candidates rotate through a series of stations based on clinical skills applied in a range of contexts. The structured assessment which provides wide sampling of cases, each with an independent examiner, improves reliability but this examination format is expensive, labour intensive and a challenge to feasibility. Validity may be lost at the expense of reliability as complex skills, requiring an integrated professional judgement, become fragmented by the relatively short station length (generally 5–10 minutes).[3,59] Assessment of communication skills and attitudinal behaviours can be included. Interestingly these skills are also proving to be context specific and to have low generalisability across clinical contexts.[60,61] OSCEs are also proving less objective than originally supposed. Scoring against a checklist of items is not ideal.[62] The global performance may reflect more than the sum of the parts.[3] Global ratings are increasingly used but neither offer a true 'gold standard' of judging performance.[63,64] Rater training is required to ensure consistency and care has to be taken not to discriminate.[42]

The use of standardised patients versus real patients remains an area of interest. Simulations are becoming the norm as it proves increasingly difficult to use real patients.[65] Extensive training to ensure reproducibility and consistency of scenarios is carried out.[66] Given the high reliabilities required of the North American licensing tests, the high costs of training can be justified but, perhaps, at the cost of validity. Performance in an OSCE is arguably not the same as performance in real life.[67]

The assessment of 'does'

The real challenge lies in the assessment of actual performance in practice, i.e. the tip of the pyramid. Increasing attention is being placed on this in the postgraduate assessment arena.[8,24] Revalidation of a clinician's fitness to practise and the identification of poorly performing doctors are increasingly areas of public concern.

Any attempt at assessment of performance has to balance the issues of validity and reliability. Interestingly modifications of the more traditional methods are now coming to the fore. Assessments of clinical competencies in the Foundation Programme are workplace based. They incorporate adaptation of the observed long case (mini-CEX), direct observation of procedures in the workplace (DOPs) rather than in the OSCE[2] and an 'oral' type case based discussion. There is a swing away from the OSCE back to more traditional methods modified to address the issue which led to their demise, i.e. context specificity.

Similarly most knowledge tests can be improved to test at the '*knows how*' rather than '*knows*' level but fail to assess higher up Bloom's taxonomy at the synthesis and evaluation level (*see* Figure 1.2 on page 15). Workplace assessments, e.g. audit projects and portfolios may well prove the answer to assessing a student's ability to evaluate and synthesise knowledge in the workplace. The use of the portfolio will form the subject of later chapters. Broadly defined as a tool for gathering evidence and a vehicle for reflective practice, a wider understanding is developing of the potential of portfolio use in assessment.

What it adds in validity to formative assessment weighs against its reliability for use in summative purposes.[67,68] The 'Learning Portfolio' for the Foundation programme provides an interesting example.[2]

Whether these methods can ever achieve more than medium stakes reliability given the difficulties of standardising content and training assessors remains to be seen. The ensuing chapters will cover these issues in more detail.

Summary

Further research into the format and reliability of workplace-based assessment and the use of portfolio assessment is essential.[69] In the past assessment formats tended to focus too heavily on knowledge-based competencies. Assessment at the apex of Miller's pyramid, '*the does*', is the international challenge of the 21st century for all involved in clinical competence testing. In addition research is needed on the assessment of attitudinal behaviours and how these inform the development of medical professionalism. We need to understand much more about the outcomes of assessment. Important tensions remain to be resolved between educational aspirations to support students formatively and the public's aspirations to ensure doctors exiting from specialty training are reliably judged as 'fit for purpose'. Many challenges face us. The ensuing chapters will extend and highlight the debates surrounding the issues raised in this preliminary chapter.

References

1. Modernising Medical Careers. www.mmc.nhs.uk
2. Southgate L, Grant J. Principles for an assessment system for postgraduate training. Postgraduate Medical Training Board. www.pmetb.org.uk
3. Talbot M. Monkey see, monkey do: a critique of the competency model in graduate medical education. *Med Educ.* 2004; **38**: 587–92.
4. Cumberlege J *et al.* Doctors in society: medical professionalism in a changing world. *Clinical Med.* 2005; **5**: Supp. or www.rcplondon.ac.uk/pubs/books/docinsoc/
5. Vleuten CPM van der, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ.* 2005; **39**: 309–17.
6. Wass V, Vleuten CPM van der, Shatzer J, Jones R. Assessment of clinical competence. *Lancet.* 2001; **357**: 945–9.
7. Dixon H. Candidates' views of the MRCGP examination and its effects upon approaches to learning: a questionnaire study in the Northern Deanery. *Educ for Primary Care.* 2003; **14**: 146–57.
8. Swanwick T, Chana N. Workplace assessment for licensing in general practice. *BJGP.* 2005; **55**: 461–7.
9. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med.* 1990; **65**: S63–7.
10. Rethans J, Norcini J, Baron-Maldonado M, Blackmore D *et al.* The relationship between competence and performance: implications for assessing practice performance. *Med Educ.* 2002; **36**: 901–9.
11. Schurwith L, Southgate L, Page G, Paget N *et al.* When enough is enough: a conceptual basis for fair and defensible practice performance assessment. *Med Educ.* 2002; **36**: 925–30.
12. Cushing A. Assessment of non-cognitive factors. In: Norman GR, Vleuten CPM van der, Newble DL (eds). *International Handbook of Research in Medical Education Part 2.* Dordrecht: Kluwer; 2002. p. 711–56.

13. Bloom BS. *Taxonomy of Educational Objectives*. Longman: London; 1965.
14. Eraut M. *Developing Professional Knowledge and Competence*. London: Falmer Press; 1994.
15. Dreyfus HL, Dreyfus SE. *The Power of Human Intuition and Expertise in the Era of the Computer*. Free Press, New York; 1986.
16. Dauphinee D, Fabb W, Jolly B, Langsley D *et al.* Determining the content of certifying examinations. In: Newble D, Jolly B, Wakeford R (eds). *The Certification and Recertification of Doctors: issues in the assessment of clinical competence.* Cambridge: Cambridge University Press; 1994; p. 92–104.
17. The General Medical Council Education Committee. *Tomorrow's Doctors: recommendations on undergraduate medical education.* London: General Medical Council; 1993. www.gmc.org.uk
18. Hays RB, Vleuten CPM van der, Fabb WE, Spike NA. Longitudinal reliability of the Royal Australian College of General Practitioners Certification Examination. *Med Educ.* 1995; **29**: 317–21.
19. Good Medical Practice. www.gmc.org.uk
20. Swanson DB, Norman GR, Linn RL. Performance-based assessment: lessons learnt from the health professions. *Educ Res.* 1995; **24**(5): 5–11.
21. Newble D. Techniques for measuring clinical competence; objective structured clinical examinations. *Med Educ.* 2004; **38**: 199–203.
22. Wass V, Vleuten CPM van der. The long case. *Med Educ.* 2004; **38**: 1176–80.
23. Messick S. *The Interplay of Evidence and Consequences in the Validation of Performance Assessments.* Research Report 92. Princeton, NJ: Educational Testing Service; 1992.
24. Southgate L, Cox J, David T, Hatch D. The General Medical Council's Performance Procedures: peer review of performance in the workplace. *Med Educ.* 2001; **35**: 9–19.
25. Cusimano MD. Standard setting in Medical Education. *Acad Med.* 1996; **71**: S112–20.
26. Norcini J. Setting standards on educational tests. *Med Educ.* 2003; **37**: 464–9.
27. Champlain de A. Ensuring the competent are truly competent: an overview of common methods and procedures used to set standards on High Stakes Examinations. *J Vet Med Educ.* 2004; **31**: 62–6.
28. Wilkinson TJ, Newble DI, Frampton CM. Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score. *Med Educ.* 2001; **35**: 1043–9.
29. Downing SM, Lieska GN, Raible MD. Establishing passing standards for classroom achievement tests in medical education: a comparative study of four methods. *Acad Med.* 2003; **78**: S85–7.
30. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ.* 2003; **37**: 830–7.
31. Messick S. Validity. In: Linn RL (ed.) *Educational Measurement (3e).* New York: American Council on Education Macmillan; 1989. p. 13–104.
32. Crossley J, Humphris G, Jolly B. Assessing health professionals. *Med Educ.* 2002; **36**: 800–4.
33. Cronbach LJ, Shavelson RJ. Measurement of error of examination results must be analysed. *Educ Psych Measurement.* 2004; **64**: 391–418.
34. Brennan, RL. *Elements of Generalisability Theory.* Iowa: American College Testing Program; 1983.
35. Shavelson RJ, Webb NM. *Generalisability theory: a primer*. Newbury Park, CA: Sage Publications; 1991.
36. Swanson DB. A measurement framework for performance based tests. In: Hart IR, Harden RM (eds). *Further Developments in Assessing Clinical Competence.* Montreal: Can-Heal; 1987. p. 13–45.
37. Wass V, Wakeford R, Neighbour R, Vleuten CPM van der. Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioner's Membership Examination's oral component. *Med Educ.* 2003; **37**: 126–31.

38. Williams RG, Klamen DK, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Med.* 2003; **15**: 270–92.
39. Vleuten CPM van der. The assessment of professional competence: developments, research and practical implications. *Adv in Health Sci Educ.* 1996; **1**: 41–67.
40. Wakeford R, Southgate L, Wass V. Improving oral examinations: selection, training and monitoring of examiners for the MRCGP. *BMJ.* 1995; **311**: 931–5.
41. Roberts C, Sarangi S, Southgate L, Wakeford R *et al.* Education and debate: oral examinations – equal opportunities, ethnicity, and fairness in the MRCGP. *BMJ.* 2000; **320**: 370–4.
42. Wass V, Roberts C, Hoogenboom R, Jones R *et al.* Effect of ethnicity on performance in a final objective structured clinical examination: qualitative and quantitative study. *BMJ.* 2003; **326**: 800–3.
43. Case SM, Swanson DB. Extended matching items: a practical alternative to free response questions. *Teach Learn Med.* 1993; **5**: 107–15.
44. Case SM, Swanson DB. *Constructing Written Test Questions for the Basic and Clinical Sciences*. National Board of Examiners: Philadelphia; 1996.
45. Frijns PHAM, Vleuten CPM van der, Verwijnen GM, Van Leeuwen YD. The effect of structure in scoring methods on the reproducibility of tests using open ended questions. In: Bender W, Hiemstra RJ, Scherbier AJJA, Zwierstra RP (eds). *Teaching and Assessing Clinical Competence.* Groningen: Boekwerk; 1990. p. 466–71.
46. Munro N, Denney ML, Rughani A, Foulkes J *et al.* Ensuring reliability in UK written tests of general practice: the MRCGP Examination 1998–2003. *Med Teacher.* 2005; **27**: 37–45.
47. Simpson RG, Ballard KD. What is being assessed in the MRCGP oral examination? A qualitative study. *BJGP.* 2005; **515**: 430–6.
48. Page G, Bordage G, Allen T. Developing key-feature problems and examinations to assess clinical decision-making skills. *Acad Med.* 1995; **70**: 194–201.
49. Farmer EA, Page G. A practical guide to assessing decision-making skills using the key features approach. *Med Educ.* 2005; **39**: 1188–94.
50. Cantillon P, Irish B, Sales D. Using computers for assessment in medicine. *BMJ.* 2004; **329**: 606–9.
51. Schuwirth LWT, Vleuten CPM van der. The use of clinical simulations in assessment. *Med Educ.* 2003; **37**(Suppl 1): 65–71.
52. Guagnano MT, Merlitti D, Manigrasso MR, Pace-Palitti V *et al.* New medical licensing examination using computer-based case simulations and standardized patients. *Acad Med.* 2002; **77**: 87–90.
53. Harden RM, Gleeson FA. ASME Medical Educational Booklet no. 8 Assessment of medical competence using an objective structured clinical examination (OSCE). *Med Educ.* 1979; **13**: 41–54.
54. Gleeson F. The effect of immediate feedback on clinical skills using the OSLER. In: Rothman AI, Cohen R (eds.). *Proceedings of the sixth Ottawa conference of medical education.* Toronto: University of Toronto Bookstore Custom Publishing; 1994. p. 412–15.
55. Fraser R, Mckinley R, Mulholland H. Consultation competence in general practice: establishing the face validity of prioritised criteria in the Leicester assessment package. *BJGP.* 1994; **44**:109–13.
56. Wass V, Jolly B. Does observation add to the validity of the long case? *Med Educ.* 2001; **35**: 729–34.
57. Norcini JJ, Blank LL, Duffy FD, Fortuna GS. The mini-CEX a method for assessing clinical skills. *Ann Intern Med.* 2003; **138**: 476–81.
58. Durning SJ, Cation LJ, Markert RJ, Pangaro LN. Assessing the reliability and validity of the Mini-Clinical Evaluation Exercise for Internal Medicine residency training. *Acad Med.* 2002; **77**: 900–4.

59. Shatzer JH, Wardrop, JL, Williams, RC, Hatch TF. The generalizability of performance of different station length standardised patient cases. *Teach Learn Med.* 1994; **6**: 54–8.
60. Colliver JA, Willis MS, Robbs RS, Cohen DS *et al.* Assessment of empathy in a standardized-patient examination. *Teach Learn Med.* 1998; **10**: 8–11.
61. Colliver JA, Verhulst SJ, Williams RG, Norcini JJ. Reliability of performance on standardised patient cases: a comparison of consistency measures based on generalizability theory. *Teach Learn Med.* 1989; **1**: 31–7.
62. Reznick RK, Regehr G, Yee G, Rothman A *et al*. Process-rating forms versus task-specific checklists in an OSCE for medical licensure. *Acad Med.* 1998; **73**: S97–S99.
63. Regehr G, MacRae H, Reznick R, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med.* 1998; **73**: 993–7.
64. Swartz MH, Colliver JA, Bardes CL, Charon R *et al.* Global ratings of videotaped performance versus global ratings of actions recorded on checklists: a criterion for performance assessment with standardized patients. *Acad Med.* 1999; **74**: 1028–32.
65. Sayer M, Bowman D, Evans D, Wessier A *et al.* Use of patients in professional medical examinations: current UK practice and the ethico-legal implications for medical education. *BMJ.* 2002; **324**: 404–7.
66. Vleuten CPM van der, Swanson DB. Assessment of clinical skills with standardised patients: state of the art. *Teach Learn Med.* 1990; **2**: 58–76.
67. Ram P, Grol R, Rethans JJ, Schouten B *et al.* Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of validity, reliability and feasibility. *Med Educ.* 1999; **33**: 447–54.
68. Driessen EW, Tartwijk van J, Overeem K, Vermunt JD *et al.* Conditions for successful reflective use of portfolios in undergraduate. *Med Educ.* 2005; **39**: 1221–9.
69. Royal College of Physicians Working Party. Doctors in society: medical professionalism in a changing world. *Clin Med.* 2005; **5**: Suppl 1 or www.rcplondon.ac.uk/pubs/books/docinsoc/

Chapter 2

Assessment in the undergraduate curriculum

Kamila Hawthorne

> The control of the licensing power is the most important function of the medical boards. Acting on behalf of the State, it is their duty to see that all candidates for license are properly qualified. They stand as guardians of the public and the profession, and here their responsibilities are indeed great.
>
> (Osler, 1885)

Introduction

When the general public are asked what they expect from their doctors, the usual response is that they want a doctor who is competent and who listens to them.[1] The concept of assessment of medical students in order to pronounce them safe and fit to practise has come a long way since the time of Hippocrates and his dictum to 'be of benefit and do no harm'. Undergraduate assessment is the prerequisite to licensure, the basic qualification to practise a health profession. It is seen as a means of checking that students have learnt the basics in terms of knowledge, skills and attitudes in a broad range of general and specialty medicine to a pre-determined level of proficiency, and that the public can feel protected by the transparency of the process. The last few decades have seen a number of high profile cases in the UK of practitioners who have not practised to the expected standard,[2] who have behaved unprofessionally, or without due regard to the expected ethical and compassionate standards of medical practice,[3] or who have been downright fraudulent or criminal in their activities.[4] These cases have resulted in a number of wide ranging reviews of the regulation of the medical profession, with increasing scrutiny of the processes currently used to measure professional competence and behaviour and its implications for public safety.

Previously acceptable cultures of self-regulation of the profession are slowly becoming less and less palatable to the public, as the potential for misuse is illustrated by some of the cases above. Although much of this affects postgraduate training and assessment, its effect also alters the way we see the beginnings of medical training – in the medical schools themselves, and Osler's statement to the 18th Annual Meeting of the Canadian Medical Association in 1885 is still relevant today.[5]

History of undergraduate assessment

Medicine has been first a guild activity and then a university subject since the Middle Ages, when the concept of 'competence' became an explicit virtue. The teaching of medicine at this stage was dominated by Arab and Greek medical lore, and the rigour of the university setting provided scholastic standards, centred in learning and reading, rather than in research. Around this time legal regulation of the profession started, although licensure of physicians and surgeons still differed greatly. The Royal College of Physicians began the process of self-regulation of the profession (1518), but it was only in 1858 that the Medical Act required that a person proclaiming to be a physician had to demonstrate evidence of appropriate qualifications from recognised educational institutions in order to have their name entered on a national registry.[6] The advent around this time of the General Medical Council (GMC) in the UK allowed the formal binding of self-regulation and registration, with the GMC being given the statutory responsibility of monitoring, advising and regulating medical education.

The role of the GMC in developing undergraduate assessment in Britain

While the GMC has a statutory responsibility to set and maintain the standards of basic medical education (undergraduate education and the first year of registration), the day-to-day organisation of the curriculum is still left to individual medical schools (Medical Act 1983). The Education Committee of the GMC issues regular recommendations as guidance,[7] with the principles of 'Good Medical Practice' as the basis for medical education.[8] It coordinates a quality assurance process of regular visits to ensure standards are maintained. A new format for monitoring the running and outcomes of medical schools began in 2004, with a GMC visiting team of academic teachers and lay members. These visits consist of detailed questionnaires on selection procedures for entry, the content of the medical curriculum, monitoring and mentoring of students, and the undergraduate assessment strategy, and require the provision of evidence of these activities to a specified standard. Visits also include attendance and reporting on a selection of assessments. The evidence collected is collated and reported in order to decide whether or not to recommend recognition or renewed recognition of individual UK Primary Medical Qualifications (PMQs) to the Privy Council (PMQ is the first medical degree awarded by a UK medical school).

Since 1980, the GMC has been calling for a reduction in the factual overload in medical teaching and an increase in the promotion of self-directed learning, critical thought, communication skills with patients and other team members, and development of professional behaviour and attitudes. The aim of these measures is to raise the standards of professional competence and enable medical graduates to build successful relationships with patients and work effectively with colleagues.[9] It has also clearly stated the principles of assessment which should be followed by medical schools (*see* Box 2.1). As a result, British medical schools have taken up the challenge to change or adapt their courses and assessment processes. As assessment drives learning,[10] medical undergraduates in their turn have been stimulated to adopt more adult-style learning behaviours leading to preparation for the life-long learning ethos the GMC has been championing. The next section of this chapter will describe these changes.

Box 2.1: GMC principles of assessment and student progress[7]

Assessment:

1 Schemes of assessment must support the curriculum and allow students to prove they have achieved the curricular outcomes. Professional attitudes and behaviour must also be assessed.
2 Student performance in both the core and student selected components of the curriculum must be assessed and contribute to the overall result.
3 A range of assessment techniques should be used, as appropriate for the curricular outcomes. Medical schools must be able to provide evidence that the schemes are valid and reliable, and that they have processes for setting standards and making decisions about student performance.
4 When students get close to graduating, their knowledge, skills, attitudes and behaviour must be thoroughly assessed to determine their fitness to practise.
5 Schemes of assessment must be open, fair and meet appropriate standards.

Student progress:

1 If students feel they have made a wrong career choice they should be able to gain an alternative degree, or to transfer to another degree course.
2 Only those who are fit to practise as doctors should be allowed to complete the curriculum and gain provisional registration. Others should be advised of alternative careers to follow.
3 There must be robust and fair procedures, including an appeals process, to deal with students who are causing concern.
4 Students should be informed of these procedures so they understand their rights and obligations.

Early forms of assessment in undergraduate medical schools

Up until the last 20 years, medical undergraduate assessments consisted of a selection of multiple-choice papers, written essay-style or short-answer papers, long and short clinical cases and oral (viva) examinations. They concentrated on medical knowledge, clinical examination and diagnostic skills. Viva assessments were used to explore understanding and decision making. This selection of assessments tended to be unstructured and case specific (i.e. as candidates show a variance in performance with different cases or problems, if only a few cases are presented during an assessment, it gives a biased picture of the candidate's ability). Clinical assessments in particular depended on 'luck' – for example, an arbitrary selection of cases that might not reflect what was commonly seen in the real world, getting a garrulous patient during an unsupervised 'long case', or an assessor who asked obscure or awkward questions. There were no pre-determined marking schedules so marking could be quite unregulated. The reliability of these assessments (i.e. their ability to give a true picture of students' abilities) and their validity in terms of assessing the complexities of a real encounter with a patient was increasingly in question. In addition, vivas were used to help decide if candidates with borderline results in their written papers should pass. Clearly, using such an unreliable tool for a critical decision is untenable, and these days this type of assessment is less important, with more efforts instead on making the original test as reliable as possible so that there is better confidence in the pass/fail mark.[11] As with many types of examination, it was not clear if these assessments were good measures of a student's proficiency in medicine, or an indication that he/she was good at passing examinations. This, together with pressure from the GMC for more reliable and holistic assessments that included measures of students' communication skills and attitudes to a diversity of patients and medical team working, has led to medical educationalists looking for ways to standardise tests and to present them in more valid ways. Much of the research on technical aspects of test development has occurred in the North American context. This is possibly due to the pressure of litigation against licensing institutions in these countries, resulting in the need to be able to defend assessment decisions.[12] However, there has been a global interest in assessment, with sharing of ideas and information over the last couple of decades.[13]

Current assessment practices

The Miller pyramid model (also discussed earlier on page 13) is a simple and clear method of looking at where a planned assessment might sit in the validity stakes relating to real life performance (*see* Figure 2.1).[14] The base of the triangle is the basic '*knowledge*' foundation, above it is the '*knows how*', or applied knowledge segment, and above this is the '*shows how*' segment (hands-on demonstration of the competency *in vitro*). The apex of the triangle is the '*does*' area; the activity the candidate will demonstrate in normal everyday practice. It is the part of the pyramid the assessors are most interested in, but also the part most difficult to test. The lower sections often serve as a proxy for the apex, to greater or lesser degrees of success.

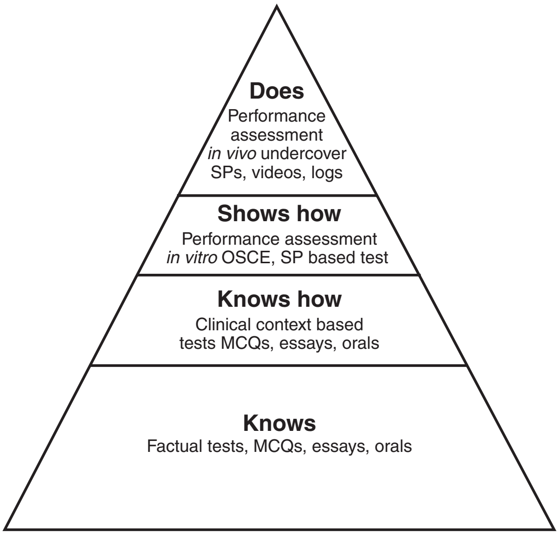


Figure 2.1: Miller's pyramid model.

Certain basic principles now underpin the planning and design of medical assessments.

1 Assessment should be tailored to the learning outcomes and closely match the curriculum.[11]

2 The breadth of test information required about a student's fitness to practise requires multidimensional assessment to measure it. Assessments need to be designed to test specific areas of competence – for example, computer marked multiple-choice questions (MCQs) are considered to be a good test of factual knowledge, objective structured clinical examinations (OSCEs) test specific practical clinical skills, and portfolio based assessments can test performance in the workplace. The range of testing modalities can be selected (with their strengths in mind) to be both complementary and to triangulate results.

3 Within each testing modality, the following aspects need to be considered:
   - reliability is a measure of the consistency and accuracy with which a test measures what it is supposed to. The test method is designed and standardised as far as possible, to ensure reliability (*see* Box 2.2). Essay questions have fairly low reliability (because it can be difficult standardising marking) and low validity. Structured short-answer papers have higher validity because more precise instructions can be given to candidates, and higher reliability because more structured marking schedules can be designed.
   - case specificity is avoided by making sure a range of contexts and testing domains is included.
   - the test carries face, content (i.e. a representative sample of the items it is designed to test) and criterion validity (correlates with other accepted assessment methods). The most basic evidence of validity of an assessment comes from documenting the links between the content of the assessment and the objectives of the curriculum it is designed to test. For example, are the patient problems set relevant and important to the curriculum? Will the stations assess skills that have been taught within the curriculum?
   - the test is feasible, depending on the setting up preparations, cost, duration of the test and numbers taking it. For example, an OSCE may be very good at assessing essential clinical skills, but it can become a logistic nightmare to find sufficient patients if a large number of students are being tested at one time.
   - the test is acceptable to those taking it – they consider it a fair test of their abilities, in line with its design and purpose.

4 There should be a significant and properly managed formative element to all assessments, so that strengths and weaknesses can be identified and fed back to candidates. Formative feedback is a powerful tool in focusing student learning, and should be designed to lead to the correction of weaknesses.[15]

5 Summative assessment should be criterion referenced (i.e. marked against an externally set standard), not norm referenced (scores ranked on a normal distribution and the pass mark adjusted to achieve a pre-determined pass rate). The latter method, commonly used in the past, can result in a variation in standards year on year (as year groups vary in their average ability), and does not ensure that minimum standards are attained.

6 Checklists only result in scores, it is the judges who set the standards.[16] A clear and fair method of standard setting should be chosen for the assessment at the

planning stage, and time needed for this activity. The modified Angoff approach (*see* Box 2.3) is one of the methods suitable for clinical assessments and takes place at the time of the assesment; while the Ebel method is often used for knowledge based assessments. In this latter method, questions are categorised according to their difficulty and appropriateness. The assessors then estimate the percentage of items in each category that a borderline candidate is likely to answer correctly.

7 Assessment drives learning, and this should also be borne in mind when designing the learning outcomes.[10]

Box 2.2: Factors in clinical skills assessment that lead to lower reliability

- Too few stations or too little testing time (case specificity)
- Checklists or items that do not discriminate (too easy or too hard)
- Unreliable patients or inconsistent role players
- Idiosyncratic examiners
- Administration problems

(Most improvement in reliability comes from increasing the number of patients.)

Box 2.3: An example of standard setting in a clinical skills assessment, using the modified Angoff approach

A reference group of markers for the assessment (the 'judges') are asked to imagine the performance of a borderline candidate in that assessment. Going through their marking checklist, each 'judge' decides what he/she feels a borderline (just passing) candidate is likely to score for each item. They discuss their 'virtual' score with each other, and resulting from this discussion, may revise their predictions for each item score. This score determines the pass mark for that station/question. The overall pass mark is based on the combination of the station/question pass marks.

Assessment methods currently in use

Written assessments

Choosing the most appropriate type of written assessment is often difficult, and issues of cost, time available to mark and feasibility also enter into the decision to choose a particular method. Essay questions and short answer questions are amongst the most widely used methods, but are time consuming and expensive to mark properly, and of limited reliability. Reliability can be improved by setting questions carefully, indicating how detailed an answer is required, and training assessors to use a systematic marking schedule. However, over-emphasis can lead to a fragmentation and trivialisation of the question. Their advantages lie in their flexibility of response – the candidate can show off their creativity and lateral thinking abilities.[17] The issue of whether or not to double-mark papers (to increase the reliability of the mark, to protect the department from complaints about unfair marking, or for assessor training) must also be considered.[18]

Adaptations of short answer questions are 'key feature' questions and 'extended matching questions'[19,20]. The purpose of the former is to measure candidates' problem-solving abilities, by describing a realistic scenario followed by a series of questions requiring essential decision making. Although these have reasonable validity and reliability, they are time consuming to produce and large numbers need to be written as their case specificity lends itself to easy memorisation by students. Extended matching questions are easier to devise, and also test applications of knowledge in problem solving situations. They consist of a list of options in related areas and a series of questions whose answers are selected from the options (*see* Box 2.4). Scoring can be relatively easily adapted to an opscan method so that papers do not have to be marked by hand. However, it takes practice to devise these types of questions, and some themes are difficult to fit into this format.

True/false tests are quick to answer and can cover broad areas of knowledge. The answers must be defensible (there is no room for 'maybe'), and take time to research and construct. Single, best option multiple-choice questions also have the advantage of high reliability. Multiple true or false questions take these tests a step further – now questions can be asked for which there is more than one correct answer. Construction and scoring can be complicated, answers must be defensible, and there needs to be a balance of correct options and reasonable distractors.

Box 2.4: Examples of a key feature question and an extended matching question for undergraduate medical students

Key feature question

Case:
You are a general practitioner, who saw a little girl this morning with a temperature of 38.5 °C, a blotchy rash that blanches on pressure, and flu-like symptoms. Her mother rings you later the same day during your evening surgery to say that she appears listless, has not eaten or drunk and the rash you observed earlier that day is still there. She is not sure if it blanches on pressure.

Which of the following is the best next step?

i) Ask the mother to give the girl another dose of paracetamol and call back later if it is no better
ii) Prescribe amoxicillin for a presumed otitis media
iii) Suggest she takes the child to the out-of-hours centre later that evening if she is no better
iv) Arrange to see the child immediately

Extended matching question

Management of diabetes in the community:

a) Trial of diet control for 3 months
b) Start on twice daily insulin injections

c) Start on oral hypoglycaemics
d) Refer to hospital diabetes specialist
e) Do diabetes annual review
f) 3-monthly blood tests and review
g) 6-monthly blood tests and review
h) Refer to podiatrist
i) Refer to obstetrician
j) Refer to hospital diabetes specialist nurse
k) Refer to practice nurse
l) Do fasting blood sugar
m) Do haemoglobin A1c blood test

Select the most appropriate course of action for the following patients:

i) Woman, mid-20s, 14 weeks pregnant, found to have glycosuria on routine testing in GP ante-natal clinic
ii) Woman, mid-50s, BMI 30, just found to have fasting blood sugar 15 mmol/L, HbA1c 8.5%
iii) South Asian woman, mid-30s, complaining of fatigue, thirst and dry mouth

Skills based assessments

As discussed previously, traditional long and short cases lack the basic requirements for validity and reliability, due to their variation in content from day to day, and to case specificity. A wider sampling frame is provided by Objective Structured Clinical Examinations (OSCEs), and variations on the theme. OSCEs have been in practice for several decades, and are a flexible test format based on a circuit of stations, each designed to get the candidate to demonstrate a clinical or consulting skill.[21,22] Standardised patients (lay people trained to present clinical problems in a realistic and repeatable way) set the scene for testing communication skills and manikins and models allow candidates to demonstrate their proficiency in carrying out a practical task such as urinary catheterisation or taking a blood sample. This makes it possible to test a wide range of situations and skills that could not otherwise be tested in an objective, standardised and repeatable way with real patients. Both clinicians and lay markers can assess stations via a task specific checklist, or checklist and rating scale.[23] Marks can be awarded for demonstrating the correct process, assessment of clinical findings and their interpretation, differential diagnoses, and the attitude and behaviour of the student towards the patient. To ensure calibration and standardisation of marking, markers assess one station repeatedly for the duration of the assessment. Stations are easy to mark, and there is the potential to include a formative element as immediate feedback on performance could be included in the test format. The method allows comparisons of performance both within and between groups of students taking the same test. Over a period of time OSCEs can be designed to increase in difficulty in a stepwise progression, to assess the increasing ability of the student to deal with complex situations and integrate consulting with clinical skills. Variations on the OSCE theme include PACES (Practical Assessment of Clinical Examination Skills), used by the Membership of the Royal College of Physicians, and OSLERs (the Objective Structured Long Examination Record), a structured presentation of an unobserved long case.

OSCE methodology has been criticised for its tendency to reduce clinical scenarios into fragmented checklists, making it less likely to assess the student's ability to integrate clinical, communication and consulting skills.[24,25] In addition, all these assessments are expensive, administratively complex and time consuming to design and set up (*see* Box 2.5), and feasibility can be compromised by the numbers of students in year groups that need to be assessed. Their running can be disruptive to hospital outpatient clinics – co-operation of NHS administration, NHS and academic consultants and support from patients are vital. For example, a 20-station OSCE for 240 Final Year medical students in Cardiff in 2005 closed the outpatient department of a major hospital for three days, required 50 examiners, 30 real patients and 20 simulated patients per session, and generated 4800 Opscan marksheets.

Box 2.5: Summary of PMETB principles for assessment

1. Method must reflect the assessment's intended purpose/content.
2. Reference assessment content to Good Medical Practice.
3. Ensure methods used to set standards are in the public domain.
4. Involve lay members in the assessment process.
5. Have mechanisms for giving students feedback on performance.
6. Use appropriate criteria for examiner training.
7. Use appropriate standardised documentation which is available nationally
8. Be sufficiently resourced

Work-based assessments

Methods of work base assessments have developed along the assumption that measuring what the candidate does in practice is a better reflection of their day-to-day performance than an artificial test situation (the apex of Miller's pyramid). Initially this type of assessment was based on process measures and clinical outcomes, such as numbers of patients seen and procedures undertaken, and morbidity and mortality rates for clinical practitioners. The process is becoming more sophisticated now, and includes personal development, attitudes, depth of understanding on a topic, team-working and communication skills. There are limitations to this method for medical students, as they do not practise clinically independently, but there is still scope to measure their performance by means of portfolio course work, working diaries detailing cases, or projects undertaken and reflective accounts of teaching and clinical situations witnessed.

Portfolios can include a variety of different types of data, such as process information, clinical outcomes, significant event analyses, diaries, case studies and patient survey results. All of these are good training for students who are setting out on a working lifetime of regular appraisal and ongoing personal development. Again, there are issues of the time needed to read and mark written work, or to watch videos of consultations and comment upon them, and a judgement in these cases can only be made on what has been recorded. The reliability of marking written work can be increased by benchmarking material to be assessed and by standardising the criteria by which they will be marked and compared. To comply with Miller's definition, observation of people at work needs to be routine or covert to exclude artificiality, but proxies such as ratings and peer reviews by supervisors, peers and patients can be used. Whether such methods are robust and comprehensive enough to cover problems of content specificity are not known.

Applying and assessing adult learning techniques to undergraduate medical education

Children learn in a largely passive and teacher-led manner,[26] which does not engage the learner's motivation to learn and apply what has been learnt to real situations. Neither does it encourage students to develop the habits and techniques of ongoing learning for themselves throughout their professional lives. The GMC has strongly advocated 'active' or 'deep' learning, a process that involves students in their own education, harnessing their interest and motivation, and allows them to integrate new knowledge and skills in a way that will result in longer retention and better understanding.[27]

This has led to medical schools redefining their curriculum, reducing the quantity of pedagogical teaching and increasing the proportion of student-directed learning. It takes the form of self-selected written projects and assignments throughout the medical course. Assessing these can be a challenge, because all the caveats regarding reliability, validity, case specificity and feasibility discussed earlier apply to these exercises. McMaster's 'triple jump' test is devised to assess students' competence at self-directed learning, in which the method of learning is as important as the learning itself.[28] The first step requires the student to think about a chosen 'problem' and after discussion to select further tasks to help learn more about the problem. The following time of private study is the second step, during which the student may make use of a variety of different sources or types of information. The final step is the report back to the tutor, based on the learning gained by steps 1 and 2. The tutor provides feedback on the way the student has approached and dealt with the 'problem'. Team-working skills are encouraged by getting students to work on some pieces of work together, which are then marked. Again, there are difficulties in assigning a mark to a group of students without really knowing how much responsibility for that work has been taken by the individual members.

What thinking is shaping the future of undergraduate assessment?

Are assessments a good indicator of a student's proficiency? Do they reliably indicate that a medical student on graduation is safe to practise? How do you measure professional attitudes and behaviours, as promoted by the GMC? Unfortunately the GMC does not give guidance on how to do this. Simulated patients to test communication skills in OSCEs go part way to assessing behaviours in clinical settings but are limited by their artificiality. 360 degree peer review techniques and attitude and conduct rating forms completed by workplace based tutors give an idea of the student's day-to-day conduct with patients, peers and other medical staff.[29,30] There are few well-documented studies of instruments that can be used to measure professionalism in formative or summative assessment. Often published reports include no independent objective observation and consist of self-reported behaviour and attitudes.[31]

The increasing specialisation of medicine and dramatic increases in student numbers mean that many students cannot cover all specialties during their undergraduate courses. This can lead to them 'missing out' on clinical teaching in important areas such as breast disease, varicose veins, hernias and testicular lumps. It is a common cause of complaint by students when they get such cases in an assessment. The numbers of students now passing through teaching rotations mean that they are less likely to be personally known to their teachers. Students that may be a cause for concern are more difficult to identify in the early stages, as they may only be present on a rotation for a short time, and may be hidden by the numbers in the group. Teachers need clear guidance on what to do about poorly performing students if they are identified – what indicators might pick them up, how to handle the situation sensitively, and who the problem should be reported to. The curriculum needs to take these issues into account in its design of teaching and assessment.

Postgraduate medical training is currently undergoing profound changes as a result of the Modernising Medical Career project, with a push to standardise and develop reliable well designed assessments mapped against the requirements of 'Good Medical Practice'.[8] Undergraduate schools meanwhile have no common curriculum and contrasting educational approaches, with little evidence to show that graduates of equivalent competency are produced regardless of the medical school they attended.[32,33] This has resulted in calls for a national competency based curriculum for medical schools, and a national licensing process.[34] Judging by the emotive response to these ideas in the medical press, it seems unlikely that such cornerstone changes will come from the medical schools themselves.[35]

What is undergraduate assessment likely to look like in the future?

Undergraduate medical assessment is only the beginning of a student's career. Lifelong habits of learning to do one's own needs assessment, knowing where to find and critically appraise reliable information, and the motivation to keep up to date, all need to be learnt and assessed at medical school. The increase in numbers of medical students is causing some pressure on medical schools to find sufficient clinical teachers to teach and assess them, and sufficient clinical material to teach them. This is driving the development of distance learning methods, using computer-assisted learning packages, which allow students to work interactively with self-learning programmes. Many of these programmes have an assessment component that can be completed online, marked electronically and provide instant feedback.[36] In addition, they can deliver high quality images, drawings and multi-media presentations in large numbers, increasing the testing capacity beyond that of traditional paper-based tests.

High fidelity electronic simulations are being developed, and best evidence suggests they facilitate learning under the right conditions. They give instant feedback, for example on resuscitation exercises, and allow for repetitive practice. The difficulty of the clinical presentation can be varied to suit the level of the student, all within a controlled 'safe' environment.[37]

Undergraduate assessments should help focus learning during the course, in identification of individual's strengths and weaknesses (and provide opportunities for the latter to be improved), and ultimately give the public confidence in the quality and performance of its doctors. In order to do this, there must be a significant and managed formative element that delivers timely constructive information to students, while the summative assessment must be criterion referenced. It is likely that continuous assessment will become more prominent in medical courses, with more workplace based assessment. In addition to testing knowledge, clinical and consulting skills should be tested, and attitudinal behaviour towards the diversity of patients likely to be encountered as well as to their other colleagues in the health professions. Ethics, clinical governance and other aspects of professionalism also need to be part of the learning and testing environment. As demands increase for undergraduate assessments to be reliable measurements of students' abilities, with some predictive validity of their future performance as doctors, they are likely to be subjected to even closer scrutiny by outside licensing bodies, with greater roles for external clinical examiners and lay representation on assessment panels.[38]

References

1. Meryn S. Improving doctor–patient communication. *BMJ*. 1998; **316:** 1922–30.
2. Dyer C. Bristol doctors found guilty of serious professional misconduct. *BMJ*. 1998; **316**: 1924.
3. Bauchner H, Vinci R. What have we learnt from the Alder Hey affair? *BMJ*. 2001; **322**: 309–10.
4. Baker R. Implications of Harold Shipman for general practice. *Postgrad Med*. 2004; **80**: 303–6.
5. Osler W. On the Growth of a Profession. *Can Med Surg*. 1998; **14**: 1239–55.
6. Dale Dauphinee W. Licensure and certification. In: Norman GR, Vleuten CPM van der, Newble D (eds). *International Handbook of Research in Medical Education*. Dordrecht: Kluwer Academic Publishers; 2002. p. 835–82.
7. Tomorrow's Doctors, 2003. www.gmc-uk.org/education/undergraduate/tomorrows_doctors.asp
8. *Good Medical Practice*. London: GMC. www.gmc-uk.org/guidance/library/GMP.pdf
9. Burge S. Undergraduate medical curricula: are students being trained to meet future needs? *Clin Med*. 2003; **3**(3): 243–7.
10. Newble DI, Jaeger K. The effects of assessments and examinations on the learning of medical students. *Med Educ.*1983; **17**: 165–71.
11. Wass V, Vleuten C van der, Shatzer J, Jones R. Assessment of clinical competence. *Lancet*. 2001; **357**: 945–9.
12. Jolly B, Wakeford R, Newble D. Implications for action and research. In: Newble D, Jolly B, Wakeford R (eds). *Certification and Recertification in Medicine*. Cambridge: Cambridge University Press.
13. Newble D. Introduction to Part 2. In: Norman GR, Vleuten CPM van der, Newble D (eds). *International Handbook of Research in Medical Education Vol. 7. Part 2*. Dordrecht: Kluwer Academic Publishers; 2002.

14. Miller GE. The assessment of clinical skills/competence/performance. *Academic Med.* 1990; **65**: 563–7.
15. Lowry S. *Medical Education.* London: BMJ Publishing Group; 1993.
16. Smee S. Skill based assessment. In: ABC of learning and teaching in medicine. *BMJ.* 2003; **326**: 703–6.
17. Schwirth L, Vleuten C van der. Written assessment. In: Cantillon P, Hutchinson L, Wood DF (eds). *ABC of Learning and Teaching in Medicine*. London: BMJ Books; 2004. Chapter 9.
18. Cannings R, Hawthorne K, Hood K, Houston H. Putting double marking to the test: a framework for assessing if it is worth the trouble. *Med Educ.* 2005; **39**: 299–308.
19. Case SM, Swanson DB. Extended-matching items: a practical alternative to free response questions. *Teach Learn Med.* 1993; **5**: 107–15.
20. Bordage G. An alternative approach to PMPs: the 'key-features' concept. In: Hart IR, Harden R (eds). *Further developments in assessing clinical competence: proceedings of the Second Ottawa Conference.* Montreal: Can-Heal Publications; 1987. p. 59–75.
21. Harden RMcG, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured clinical examination. *BMJ.* 1975; **i**: 447–51.
22. Newble DI, Dawson B, Dauphinee WD, Page G *et al.* Guidelines for assessing clinical competence. *Teach Learn Med.* 1994; **6**: 213–20.
23. Gorter S, Rethans JJ, Scherpbier A, van der Heijde D *et al.* Developing case-specific checklists for standardised-patient-based assessments in internal medicine: a review of the literature. *Academic Med.* 2000; **75**: 1130–7.
24. Hodges B, Regehr G, McNaughton N, Tiberius RG *et al.* OSCE checklists do not capture increasing levels of expertise. *Academic Med.* 1999; **74**: 1129–34.
25. Wilkinson TJ, Frampton CM, Thompson-Fawcett M, Egan T. Objectivity in Objective Structured Clinical Examinations: checklists are no substitute for examiner commitment. *Academic Med.* 2003; **78**(2): 219–23.
26. Knowles MS (ed.). *Andragogy in Action: applying modern principles of adult learning.* San Francisco: Jossey-Bass; 1984.
27. Marton F, Saljo R. On qualitative differences in learning. 1. Outcome and process. *Brit Educ Psychol.* 1976; **46**: 4–11.
28. Painvin C, Neufeld V, Norman G, Walker I *et al.* The triple jump exercise – a structured measure of problem solving and self-directed learning. *Annual Conference on Research in Medical Education.* 1979; **18**: 73–83.
29. Martin J, Lloyd M, Singh S. Professional attitudes: can they be taught and assessed in medical education? *Clin Med.* 2002; **2**(3): 217–23.
30. Howe A, Campion P, Searle J, Smith H. New perspectives – approaches to medical education at four new UK medical schools. *BMJ.* 2004; **329**: 327–31.
31. Veloski JS, Fields SK, Boex JR, Blank LL. Measuring professionalism: a review of studies with instruments reported in the literature between 1982 and 2002. *Acad Med.* 2005; **80**(4): 366–70.
32. Fowell SL, Maudsley G, Maguire P, Leinster SJ *et al.* Student assessment in undergraduate medical education in the United Kingdom 1998. *Med Educ.* 2000; **34**(Suppl): S1–49.
33. Goldacre MJ, Lambert T, Evans J, Turner G. Preregistration house officers' views on whether their experience at medical school prepared them well for their jobs: national questionnaire survey. *BMJ.* 2003; **326**: 1011–2.
34. Wass V. Ensuring medical students are 'fit for purpose'. *BMJ.* 2005; **331**: 791–2.
35. http://bmj.bmjjournals.com/cgi/eletters/331/7520/791
36. Cantillon P, Irish B, Sales D. Using computers for assessment in medicine. *BMJ.* 2004; **329**: 606–9.

37. Issenberg B, McGaghie WC, Petrusa ER, Lee Gordon *et al.* Features and uses of high fidelity medical simulations that lead to effective learning: a BEME systematic review. *Medical Teacher.* 2005; **27**(1): 10–28.
38. Vleuten C van der. Validity of final examinations in undergraduate medical training. *BMJ.* 2000; **321**: 1217–19.

Chapter 3

Assessment in the Foundation Programme

Gareth Holsgrove and Helena Davies

Introduction

The Foundation Programme is a major reform of postgraduate medical education, arising from the recommendations of *Modernising Medical Careers (MMC) – the next steps*.[1] The first cohort of Foundation students entered the Programme in August 2005. The Programme will bring significant changes to the first two years of postgraduate medical education and encourage the development of professionalism in a variety of specialties and settings. The curriculum for the Foundation Programme, and for subsequent specialist training, must map to the domains of *Good Medical Practice*.[2] These are:

- good clinical care
- maintaining good medical practice
- teaching and training
- appraising and assessing
- communicating with patients
- working with colleagues
- probity
- health.

Therefore it is important that assessment also blueprints to these same domains as well as to the Foundation curriculum.

In addition to covering the domains of *Good Medical Practice* (GMP), assessment in the Foundation Programme focuses on performance (i.e. what the doctor actually does, day-in, day-out, in the workplace) in order to ensure that the minimum standards for career progression have been met. Since a doctor's performance is demonstrated in the workplace, this is also the most obvious place to assess it. Therefore, all the assessment in the Foundation Programme is workplace based.

The Foundation Programme, and, hence, assessment within the Programme, will not re-visit the undergraduate course but will build upon it. Assessment will be grounded in relevant clinical content, particularly generic clinical competencies and the ability to identify and respond appropriately to acutely ill patients. In order to ensure that most doctors successfully complete the Programme, they will receive focused, relevant and timely feedback and there will also be an 'early warning' system. This will identify failure to participate, or to make appropriate progress, and will trigger remedial action such as transfer to a different programme, or intensive teaching and closer supervision.

The design and delivery of the Foundation Programme is clearly important in enabling these objectives to be achieved, and within that design the methods and conduct of assessment are of fundamental importance. This is both to ensure that doctors successfully completing the Programme (which is what the great majority are expected to do) really have attained the requisite standards, and to ensure that problems are identified early. However, we cannot be sure that standards have been reached without having appropriate evidence, and problems cannot be corrected unless they are first identified. The assessment strategy in the Foundation Programme has been developed to ensure that both these purposes are fulfilled. Moreover, since these criteria will remain important throughout training, it is likely that the Foundation assessment methods will continue to play a significant role in assessment throughout specialist training and continuing professional development.

The purpose of assessment

As noted in the introduction, the principal purposes of assessment in the Foundation Programme are to ensure that the required standards are being achieved and to identify any problems at an early stage. There are four essential characteristics of an assessment system that are needed in order to achieve both these objectives.

1 The focus must be on performance in the workplace.
2 The assessment must provide evidence of performance.
3 Evidence must be triangulated whenever possible.
4 Complete records must be kept.

A further stated purpose of the Foundation assessment programme is the provision of feedback in order to optimise the educational impact of the assessment programme.

Of course, the assessment methods themselves need to be appropriate for the task, and assessment must be properly carried out. We shall return to these points below. First, though, we should give a little more detail about the four characteristics of performance, evidence, triangulation and record-keeping.

Performance

Traditionally, assessments in medical education have focused to a major extent on high-stakes formal examinations, and these, in turn, have been predominantly concerned with the recall of factual knowledge. Unsurprisingly, candidates preparing for such examinations have concentrated on learning lots of facts. This is an example of consequential validity, which is the effect that assessment has on learning and associated behaviour. Learning lots of facts, particularly in relative isolation, is difficult and boring. The situation can be made even more difficult and boring if the facts to be learned are obscure, trivial, or cannot be perceived to be of any lasting importance to the learner. Unfortunately, it is not unknown for many medical exams, even contemporary ones, to be open to criticism along these lines.

It is far more relevant to assess the application of knowledge, rather than simply its recall – after all, knowing a lot of facts is not much use if you do not know what to do with them. It is also more appropriate to make assessments in real-life situations, or realistic simulations of such situations, rather than in the artificial setting of the examination hall. The famous illustration of the hierarchy of knowledge, application, competence and performance is Miller's pyramid (*see* Figure 3.1) which shows how progressively higher levels of activity are developed from a foundation of lower ones, with performance at the top of the pyramid. Satisfactory performance is dependent on acquisition of the lower levels of competence. If problems are detected in assessments of performance, then diagnostic assessment of lower-level components including knowledge might be necessary to pinpoint them.

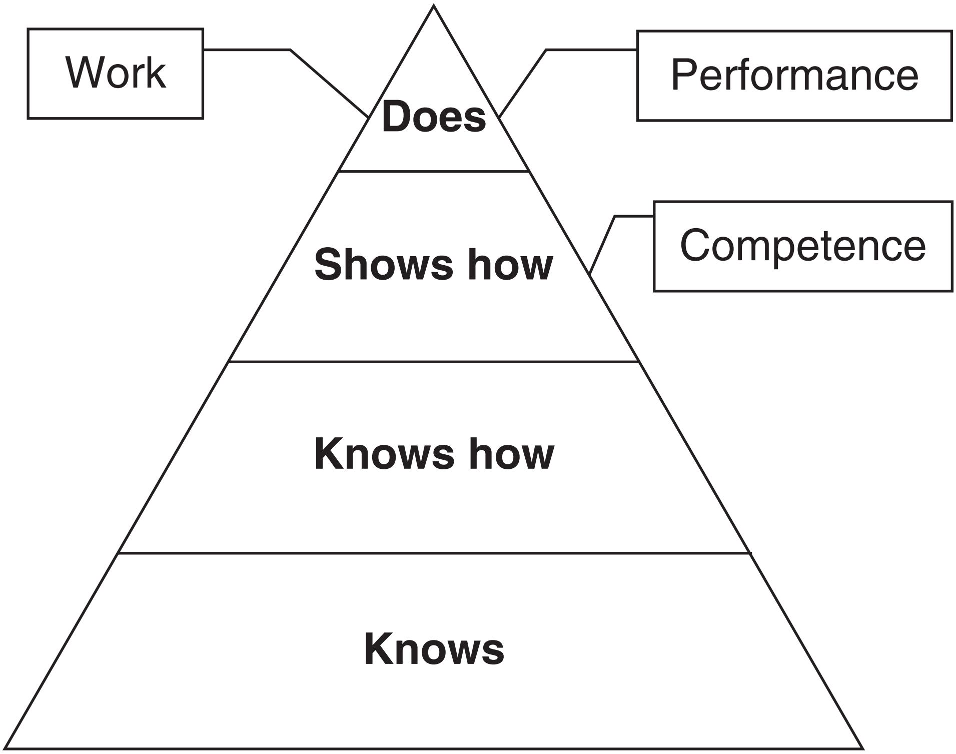


Figure 3.1: Miller's pyramid model.[3]

Evidence

In medicine it is extremely important to base decisions on evidence. This applies not only to clinical decisions concerning the safe and appropriate management of patients, but also to decisions regarding the performance and progress of students. In fact, it goes beyond even this, to include the quality assurance of all aspects of training and assessment.

The Postgraduate Medical Education and Training Board (PMETB), the UK statutory authority for standards in training and assessment, requires that assessments are reliable, valid, feasible, fair and defensible.[4] In order to meet these requirements, assessments must produce evidence of each doctor's performance and progress, whether satisfactory or not. Furthermore, those responsible for carrying out the assessments must be accountable for both the process and the findings. Design of the Foundation Programme was informed by the PMETB principles.

However, it is not enough simply to have good assessment methods – it is also necessary to provide training for those who will be responsible for using them. Training for assessment in the Foundation Programme was initiated with a series of national *Modernising Medical Careers* workshops during 2004 and 2005, in which both authors of this chapter were involved. These workshops have been supplemented with local training events organised by deaneries and medical royal colleges, and there are also written training materials and training CDs for assessors and guidance notes for students.*

Triangulation

An important principle when gathering evidence on which important decisions are based is that, if possible, pieces of evidence should not exist in isolation. Evidence of satisfactory (or, especially, unsatisfactory) performance should be supported by additional evidence. If at all possible, this evidence should be produced at different times, under different circumstances, witnessed by different people and gathered using different methods. This process, triangulation, is supported by the four assessment methods for Foundation because most aspects of the curriculum are assessed by more than one method and by more than one assessor, and all assessments are carried out several times. This means that a student would not be put at a major disadvantage simply because they were having a 'bad day' (which can happen in traditional formal exams). It also enables small but persistent problems, which might be missed on a single assessment or on which the trainee might be given the benefit of the doubt on a one-off assessment, to be identified over time as issues that need to be addressed.

The principle of triangulation is now recognised as such an important issue in the assessment of professional competence that in contemporary best-practice we are moving away from a focus on individual assessment methods and towards designing programmes of assessment. Nowhere is this better described than in the excellent article by van der Vleuten and Schuwirth in which the authors explain why assessment is no longer seen as a measurement problem, but as an instructional design problem.[5] 'Assessment in medical education addresses complex competencies and thus requires qualitative and quantitative information from different sources as well as professional judgement. Adequate sampling across judges, instruments and contexts can ensure both validity and reliability.'[5] (p. 309) Although there is further development to be done and new methods yet to be added to the assessment toolbox, the four assessment methods for the Foundation Programme represent an important step towards achieving this. The magnitude of this step can be judged by comparing assessment in the Foundation Programme with that traditionally used in the undergraduate curriculum and in typical Royal College Membership and Fellowship exams.

* Available on the MMC website www.mmc.nhs.uk

Record-keeping

In the Foundation Programme, and, indeed, beyond, responsibility for arranging assessments will fall mainly upon the students. They should be appropriately prepared for this by the undergraduate curriculum because, since the inception in 1990 of the first undergraduate medical curriculum to follow the principles set out in *Tomorrow's Doctors* medical students have become increasingly active participants in, rather than passive recipients of, learning and assessment.[6]

Throughout their training, doctors will not only be responsible for their own learning and workplace based assessments, but they will also need to maintain records of progress and attainment. At the end of specialist training, doctors will apply to PMETB for a Certificate of Completion of Training (CCT) so that they can be entered on the Specialist Register. PMETB will require evidence in support of the CCT application, and it will be the student's responsibility to provide this. The best way to do this is by maintaining a record of assessments throughout training. This, in its turn, means that records should be complete and properly organised. Therefore, it is very much in the student's interest to maintain good records of assessment and attainment right from the start of the Foundation Programme.

Accurate and complete records are also very useful to mentors, educational supervisors and the postgraduate deanery. Decisions about progress through the curriculum, and any problems requiring particular attention, will be based on proper records. Students must be clear that they have a responsibility to keep all of their assessments and that failure to do so is a probity issue. An occasional unfavourable assessment will not disadvantage a student. Indeed, if they can demonstrate reflection and learning from this, perhaps by written comments on the form itself this may be taken as evidence of critical self appraisal. A small minority of students may wish to remove unfavourable assessments but these are likely to be the 'problem' students. A common characteristic of these students is that they often have complete lack of insight into their own poor performance.

Assessment in the Foundation Programme

Having considered the purpose and principles of assessment, we can now turn our attention to the assessment methods used in the Foundation Programme. Prior to the implementation of the Foundation Programme in 2005 it was recognised that effective work-based assessment was essential to its success. A working party was established in early 2004 led by the London Deanery to plan and then evaluate an assessment programme for the Foundation years. Design of the programme was informed by available evidence and the stated purposes of the programme. In addition to ensuring that the required standards are achieved and identifying those students in difficulty, the programme was designed to utilise the consequential validity of assessment positively. Since the majority of students will be satisfactory, the assessments should generate feedback that will facilitate their personal development in line with a quality improvement model. Furthermore assessments need to be feasible in the range of specialties and clinical contexts they cover.

The tools

Four assessment tools are being utilised currently for Foundation assessment.

1 Multi-source feedback(MSF); mini-PAT in most UK Deaneries – TAB in some.
2 Case-based discussion (CbD).
3 Mini-clinical evaluation exercise (mini-CEX).
4 Direct observation of procedural skills (DOPS).

Multi-source feedback (MSF)

Multi-source feedback is a means of assessment based on the collection and collation of views from a range of sources, which may include colleagues and/or patients.[7–11] MSF has been used in a management setting for many years as part of performance management and appraisal and is being increasingly utilised in a healthcare setting. Alternative terminology for MSF includes peer assessment, 360° feedback and peer review. MSF has the advantage of describing the method in a way that is universally understood and does not depend on contextual definitions – for example, of who constitutes a 'peer'. There is evidence to support its use within medicine from the USA, Canada and the UK. The first published work supporting its use in this context came from Paul Ramsey,[12] who concluded that 11 raters were needed to achieve acceptable reliability (using Generalisability theory[13,14,15]). Factor analyses suggested two broad categories: humanistic/psychosocial and clinical management/cognitive skills. Importantly, scores were not significantly biased by who was responsible for selection of raters or the relationship of the raters to the doctor undergoing assessment. Subsequently a considerable body of work on MSF has been undertaken in Canada and more recently in the UK.[7,16–21] For the Foundation programme the majority of deaneries is currently utilising mini-PAT. This is an MSF tool mapped to *Good Medical Practice* and derived from the Sheffield Peer Review Assessment Tool (SPRAT).[7,19] It consists of 15 questions and a global rating scale across the five key domains of GMP with an additional yes/no question in relation to concerns about health and probity (www.hCAT.nhs.uk, www.mmc.nhs.uk). There are separate cohorts for F1 and F2, and trainees nominate eight raters for each round of mini-PAT. They are also required to complete a self-rating. Initially mini-PAT was managed as a paper based, centrally coordinated system with feedback provided showing a self-rating for each question, compared with the mean for their assessors and a cohort of peers. Analysis can be done electronically and the report is fed back face-to-face by their supervisor. Web-based completion is being piloted in early 2006. Other MSF tools in use include Team Assessment of Behaviours (TAB).[20] TAB is a 360° tool developed by the West Midlands Deanery and consists of four questions aimed to identify concerns/problems in relation to a student's attitude and/or behaviour in relation to the four areas covered by the questions (maintaining trust/professional relationships with patients, verbal communication skills, team working and accessibility).

Case-based discussion (CbD)

Case-based discussion is a structured discussion of an actual case seen by a student.[22–6] The discussion arises from an entry in the notes made by the student and aims to explore the thinking that underpinned that entry. For example, prompts might include: what did they think was wrong; how did they think the investigations they did might help them sort the patient's problem out; did they consider the ethical issues surrounding the case? The particular strength of CbD lies in the ability to explore clinical reasoning. However, it should not be a viva-type interaction exploring the student's knowledge of the clinical problem. Students are asked to bring along two or three sets of notes to discuss with an assessor who then selects one for discussion and assessment. Work using CbD in the USA (where it is known as chart stimulated recall, CSR) as part of recertification of practising doctors showed the scores were highly correlated with performance in an exam using standardised patients, and CbD scores distinguished between doctors who had been 'referred' because of concerns and those about whom there were no concerns.[27] CbD has also been extensively used by the GMC as part of its performance procedures[28] and is utilised by the National Clinical Assessment Service (NCAS). The CbD tool for the Foundation programme is specific for this purpose but builds on previous methods.

Mini-clinical evaluation exercise (mini-CEX)

Mini-CEX was developed initially by the American Board of Internal Medicine (ABIM) and has been used in a range of settings both in a postgraduate and undergraduate context.[22–26,29,30] Available evidence suggests that 4–6 interactions are required for satisfactory reproducibility. Support for validity has been provided by demonstration of an increase in ratings over a year of training,[24] correlation with SP scores,[23] and correlations with other methods of assessment.[26] Holmboe demonstrated that using scripted tapes, faculty were able to distinguish between unsatisfactory, satisfactory and superior performance.[29] The Foundation mini-CEX is an 'anglicised' (with permission) version of the ABIM tool.

Direct observation of procedural skills (DOPS)

Direct observation of procedural skills is a tool developed by the Royal College of Physicians (RCP) for the assessment of procedural skills. It is similar to mini-CEX except that the focus is on technical rather than communication/clinical judgement skills. The theory underpinning DOPS is derived from other observational methods of assessing technical skills such as the objective structured assessment of technical skills (OSATS).[31–5] The RCP has developed both a generic version of DOPS and a procedure-specific DOPS. For Foundation assessment a generic version is used.

Practical considerations

All four tools utilise a six-point rating scale where '4' is satisfactory ('meets expectations for completion of F1 (or F2 as appropriate)'. '1' and '2' represent 'below expectations for completion of F1 (or F2)', '5' and '6' are 'above expectations' and '3' is 'borderline'. Based on available evidence in relation to reliability students are asked to submit six each of mini-CEX, CbD and DOPS during the year. Since the two biggest threats to reliability for all clinical assessments are clinical content (content specificity) and variation between assessors, students are asked to sample as widely as possible across assessors and clinical problems. Ideally they would have a different assessor and a different clinical problem category for each of their six assessments with each tool. Using different assessors also spreads the assessment load, which improves feasibility. Utilising a range of assessors including, for example, specialist registrars as well as consultants, and nursing staff and other healthcare professionals where appropriate, will also spread the assessor load.

For most deaneries, mini-PAT is being centrally managed and CbD, mini-CEX and DOPS undertaken using triplicate pads with a copy being submitted centrally for scanning, a copy kept by the student and a copy provided for their educational supervisor. Students are provided with an assessment profile at the end of the year that will be used to inform their end of year 'sign-off'.

Feedback for students

One of the key design considerations for the Foundation Programme was the importance of maximising feedback for all students. All four tools are designed to help in identifying areas of strength and also developmental needs. This does not of course guarantee that individual assessors will do so – or that if they do, a student's development needs will be fed back constructively and linked in with personal development planning – though this, of course, is the intention.

Training in feedback skills and objective setting is an important and ongoing need for educational supervisors. It is of particular importance because there is evidence that while focused specific feedback can be a powerful stimulus for development, feedback that identifies problems but is not specific or credible may do more harm than good.[21,36–8] There is evidence, particularly from the organisational psychology literature, of the benefit of face-to-face feedback with the opportunity for discussion. This is especially important where the feedback contains elements which the individual may find difficult.

Feedback for mini-CEX, CbD and DOPS is provided at the time of the encounter, and this is an important positive characteristic of each of these tools.

For mini-PAT, the collated feedback is made available electronically to local administrators for distribution to educational supervisors. The recommendation is that the supervisor should feed this back to the trainee in a face-to-face meeting, with the opportunity for further follow-up meetings if necessary.

Quality assurance

The Foundation Programme has been planned with careful attention to the purposes of the programme and the available evidence in relation to work-based assessment, including evidence to support reliability and validity. Factors such as feasibility, educational impact and feedback were also important considerations. However, all assessment processes require quality assurance (QA). This is not a one-off process, but one that should be ongoing and iterative. All postgraduate assessment programmes in medical education should have regular QA undertaken against the principles set out by the PMETB.[4] Centralised implementation and data collection will greatly facilitate this. As well as evaluation of reliability and validity, identification of possible systematic sources of bias is important to ensure the fairness of the programme. Evaluation should include qualitative as well as quantitative aspects. It is likely that the Foundation assessment programme will evolve over the years with refinement of the tools and improvements in their associated training, as well as the introduction of appropriate additional methods (for example, the current programme does not include any assessments undertaken by patients). It is also clear that some of the assessment tools may need modification for specialties such as pathology, psychiatry and public health.

Prior to national implementation of foundation programmes in August 2005, a number of deaneries ran foundation pilots. Evaluation of the assessment methods for several hundred students in foundation pilots between January and July 2005 was funded by the Department of Health and will be reported during 2006. Detailed QA for approximately 5000 students in foundation posts which began in August 2005 is currently being undertaken at the time of writing (early 2006).

How do we know if students are in difficulty?

A student may be identified as being in difficulty through the assessment programme by a number of means. MSF may highlight problems that supervisors were previously aware of, or identify new problems. These concerns may relate to globally poor performance or to concerns in a limited area of practice such as communication with colleagues. 'Flags' which are aimed at picking up concerns at an early stage, either because the total aggregate score is low or because particularly worrying words have been included in the free text (e.g. dangerous, serious) are being implemented as part of the management of mini-PAT. For the other tools, assessors will know if they have particular concerns about an individual interaction and should feed this back both to the student and their educational supervisor. The educational supervisor should review the student's portfolio and assessments on a regular basis. A student's failure to engage in the process will be an important means of identifying that they might not be progressing satisfactorily. Central data management allows regular reports on participation on a trust or deanery basis to be provided.

Where a student is identified as potentially being in difficulty it is likely that further diagnostic work will be needed to explore the concerns and clarify how best to address them. Where any concerns are raised the usual deanery mechanisms for students in difficulty would operate.

Robust evaluation of the effectiveness of the assessment tools should include longitudinal follow-up of the students in order to examine predictive validity, an area in need of research.

The future

Selection into specialty training is currently under review. In order to be eligible to participate in selection, students will need to have satisfactorily completed their required assessments. However, it is important to remember that the assessment process was not designed to rank students and should not be used for this purpose. Whether or not it may be possible to band students in the future on the basis of their performance in the assessment programme is under discussion.

Given the huge number of assessments being undertaken, continual review of the processes and improvements in efficiency wherever possible are essential. An important future development is likely to be the use of electronic records of assessment. Online MSF would reduce the data administration load and streamline the process. Online pilots of mini-PAT are being undertaken for the second round of mini-PAT in 2006. CbD also lends itself well to web-based completion as it is undertaken in a planned way and a computer could be made available.

Careful planning informed the current Foundation Programme and both the initial pilot and the first year of its implementation are being evaluated. However, ongoing evaluation and QA is essential and it is likely that the foundation assessment process will evolve over time in response to this. This may include the use of additional tools such as patient assessment but could also include reducing the assessment load if it is demonstrated that reproducible, valid judgements about a student can be made with fewer assessments. This is most likely to be the case for students who are making satisfactory progress – those in difficulty will probably continue to require a more thorough assessment programme.

References

1. *Modernising Medical Careers (MMC) – the next steps.* London: Department of Health; 2004.
2. GMC. *Good Medical Practice*. 2001 [cited 2002 October]. Available from: www.gmc-uk.org/standards/
3. Miller G. The assessment of clinical skills/competence/performance. *Academic Med.* 1990; **65**(Suppl): S63–7.
4. PMETB. *Principles for an Assessment System for Postgraduate Medical Training*. 2005 [cited 2005 January]. Available from: www.pmetb.org.uk/pmetb/publications/
5. Vleuten CP van der, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ*. 2005; **39**(3): 309–17.
6. GMC. *Tomorrow's Doctors*. 2003 [cited 2005 January]. Available from: www.gmc-uk.org/education/undergraduate/tomorrows_doctors.asp
7. Davies H, Archer J. Multi-source feedback: development and practical aspects. *Clin Teacher.* 2005; **2**(2): 77–81.
8. Evans R, Elwyn G, Edwards R. Review of instruments for peer assessment of physicians. *BMJ*. 2004; **328**(7450): 1240.
9. Lockyer J. Multisource feedback in the assessment of physician competencies. *J Contin Educ Health Prof.* 2003. **23**(1): 4–12.
10. Norcini JJ. Peer assessment of competence. *Med Educ.* 2003. **37**(6): 539–43.
11. Fletcher C. Performance appraisal and performance management: the developing research agenda. *Occup Organiz Psychol*. 2001; **74**: 473–87.
12. Ramsey PG, Wenrich MD, Carline JD, Inui TS *et al.* Use of peer ratings to evaluate physician performance. *JAMA*. 1993; **269**(13): 1655–60.
13. Cronbach L. My current thoughts on coefficient alpha and successor procedures. *Educ Psychol Meas.* 2004; **64**(3): 391–418.
14. Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. *Med Educ.* 2002; **36**(10): 972–8.
15. Streiner D, Norman G. *Generalizability. Health Measurement Scales: a practical guide to their development and use.* Oxford: Oxford University Press; 1995.
16. Hall WC, Violato R, Lewkonia J, Lockyer H *et al.* Assessment of physician performance in Alberta: the physician achievement review. *CMAJ.* 1999; **161**(1): 52–7.
17. Violato C, Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. *BMJ.* 2003; **326**(7388): 546–8.
18. Violato CA *et al.* Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. *Acad Med.* 1997; **72**(10 Suppl 1): S82–4.

19. Archer JC, Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training. *BMJ.* 2005; **330**(7502): 1251–3.
20. Whitehouse AA, Hassell L, Wood D, Wall M *et al.* Development and reliability testing of TAB a form for 360 degrees assessment of Senior House Officers' professional behaviour, as specified by the General Medical Council. *Med Teach.* 2005; **27**(3): 252–8.
21. Sargeant JK, Mann K, Ferrier S. Exploring family physicians' reactions to multi-source feedback: perceptions of credibility and usefulness. *Med Educ.* 2005; **39**(5): 497–504.
22. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med.* 1995; **123**(10): 795–9.
23. Boulet JR, McKinley DW, Norcini JJ, Whelan GP. Assessing the comparability of standardized patient and physician evaluations of clinical skills. *Adv Health Sci Educ Theory Pract.* 2002; **7**(2): 85–97.
24. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. *Ann Intern Med.* 2003; **138**(6): 476–81.
25. Norcini JJ, Blank LL, Arnold GK, Kimball HR. Examiner differences in the mini-Cex. *Adv Health Sci Educ Theory Pract.* 1996; **2**(1): 27–33.
26. Durning SJ, Cation LJ, Markert RJ, Pangaro LN. Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. *Acad Med.* 2002; **77**(9): 900–4.
27. Norman G, Davis D, Painvin A, Lindsay E *et al.* Comprehensive assessment of clinical competence of family/general physicians using multiple measures. In: *Proceedings of the Research in Medical Education Conference*. 1989.
28. Southgate L, Campbell M, Cox J, Foulkes J *et al.* The General Medical Council's Performance Procedures: the development and implementation of tests of competence with examples from general practice. *Med Educ.* 2001; **35**(Suppl 1): 20–8.
29. Holmboe ES, Huot S, Chung J, Norcini JJ *et al.* Construct validity of the miniclinical evaluation exercise (mini-CEX). *Acad Med.* 2003; **78**(8): 826–30.
30. Holmboe ES, Yepes M, Williams F, Huot SJ. Feedback and the mini clinical evaluation exercise. *Gen Intern Med.* 2004; **19**(5, Pt 2): 558–61.
31. Faulkner H, Regehr G, Martin J, Reznick R. Validation of an objective structured assessment of technical skill for surgical residents. *Acad Med.* 1996; **71**(12): 1363–5.
32. Martin JA, Regehr G, Reznick R, MacRae H *et al.* Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg,* 1997; **84**(2): 273–8.
33. Goff B, Mandel L, Lentz G, Vanblaricom A *et al.* Assessment of resident surgical skills: is testing feasible? *Am J Obstet Gynecol.* 2005; **192**(4): 1331–8; discussion 1338–40.
34. Grober ED, Hamstra SJ, Wanzel KR, Reznick RK *et al.* The educational impact of bench model fidelity on the acquisition of technical skill: the use of clinically relevant outcome measures. *Ann Surg.* 2004; **240**(2): 374–81.
35. Beard JD, Jolly BC, Newble DI, Thomas WE *et al.* Assessing the technical skills of surgical trainees. *Br J Surg.* 2005; **92**(6): 778–82.
36. Brett JF, Atwater LE. 360 degree feedback: accuracy, reactions, and perceptions of usefulness. *J Appl Psychol.* 2001; **86**(5): 930–42.
37. Fidler H, Lockyer JM, Toews J, Violato C. Changing physicians' practices: the effect of individual feedback. *Acad Med.* 1999; **74**(6): 702–14.
38. Walker A, Smither J. A five year study of upward feedback: what managers do with their results matters. *Personnel Psychol.* 1999; **52**: 393–423.

Chapter 4

Record of in-training assessments (RITAs)

Anand Mehta, Kevin Kelleher and Colin Stern

Reforms to the Specialist Registrar Grade

The changes to the Specialist Registrar grade introduced in 1995/6, generally know as the CALMAN (Curriculum Appraisal Length of training Management of training Assessment National standards) reforms, were the UK's response to the requirements laid down by the European Specialist Order of 1978. The Order placed a minimum time limit of 4 years of Higher Specialist Training in order to achieve a Certificate of Completion of Specialist Training (CCST) and was intended to harmonise the framework of such training across Europe.[1]

In Britain this was eventually seen as an opportunity to improve the quality of the training that was being delivered and, at the same time, to collect better evidence that not only was training being delivered, but that it was being delivered effectively.

Collaboration was established between the royal colleges and faculties on the one hand and the postgraduate deaneries on the other, working together in the newly formed specialist training committees (STCs). The colleges and faculties set the curricula and the standards that should be met by trainers and by students. The royal colleges established various committees such as the Joint Committee for Higher Medical Training (JCHMT), Joint Committee for Higher Surgical Training (JCHST) and Joint Committee for Post-Graduate General Practice Training (JCPGPT) to enable implementation of such curricula.[2] Their role has now been subsumed into Post-Graduate Medical Education and Training Board.[3]

The RITA process

All students who enter into Higher Specialist Training (HST) undergo a series of work-based assessments with their educational supervisors, who also provide an annual report on the student's performance. All students in the SpR grade, including LATs, FTTAs and flexible students should expect to attend a RITA panel on an annual basis. The RITA process is governed by the guidelines laid out in sections 11, 12 and 13 of the *Guide to Specialist Registrar Training*.[4] Assessments are documented and reviewed, usually annually, by RITA panels commissioned by the postgraduate dean (PGD). Each panel includes representatives of the PGD, the Regional Specialty Training Committees, the Specialty Advisory Committee (SAC) and/or the colleges. The panel will review in detail the Training Record, will explore with the student the depth of experience and understanding on which it has been based and consider individual trainer reports. The main functions of the RITA panel are:

- to ensure the students have had appraisal and feedback from their educational supervisor
- to review the accuracy of the proposed CCST date
- to check the completeness of the student's record/log book
- to plan the next year of training
- to provide careers advice
- to review the quality of the training posts.

The penultimate year assessment (PYA) is somewhat different from RITA panels in other years to ensure evenness of standards and impartiality and in order to meet the rigorous demands of the PMETB. The RITA assessments held in the penultimate year are conducted by panels that include an external assessor, and/or specialty representatives selected by the SAC, who have had no involvement in the training of those being assessed. The PYA focuses mainly on the Training Record and on the individual reports and assessment forms that it contains or which may be developed.

On the basis of the documentation, augmented by direct enquiry, the panel must satisfy itself that the requirements of training as set out in the specialty curriculum have been met so far. They will provide the opportunity to review the training programmes of the specialty or specialties in the region in which the assessments are being carried out. Such importance is attached to the assessment role; all those who assume it undertake a course of training in assessment techniques.

There is a tendency to use the terms annual assessment, annual review and RITA rather loosely and sometimes synonymously, which has the potential for confusion. The RITA is not an assessment; it is merely a record of assessment results.[5] The postgraduate deaneries manage the education and act as the auditors of the quality of education that is delivered locally.[6]

Figure 4.1 shows the processes used to manage higher specialist trainee (HST) in medicine and Box 4.1 shows the types of forms used to record the outcome of the RITA process.

It is a guiding principle that training, appraisal and assessment should be supported and carried out locally, in the workplace, by the named educational supervisor and their teams. Colleges and their affiliates have developed programmes

Box 4.1: Forms used to record the outcome of the RITA process

A – core information on the student
B – changes to core information
C – record of satisfactory progress within the specialist registrar grade
D – recommendation for targeted training (stage 1 of 'required additional training')
E – recommendation for intensified supervision or repeated experience (stage 2 of 'required additional training')
F – record of experience outside training programme
G – final record of satisfactory progress

A copy of the RITA form is forwarded to the college with another copy retained at the deanery.

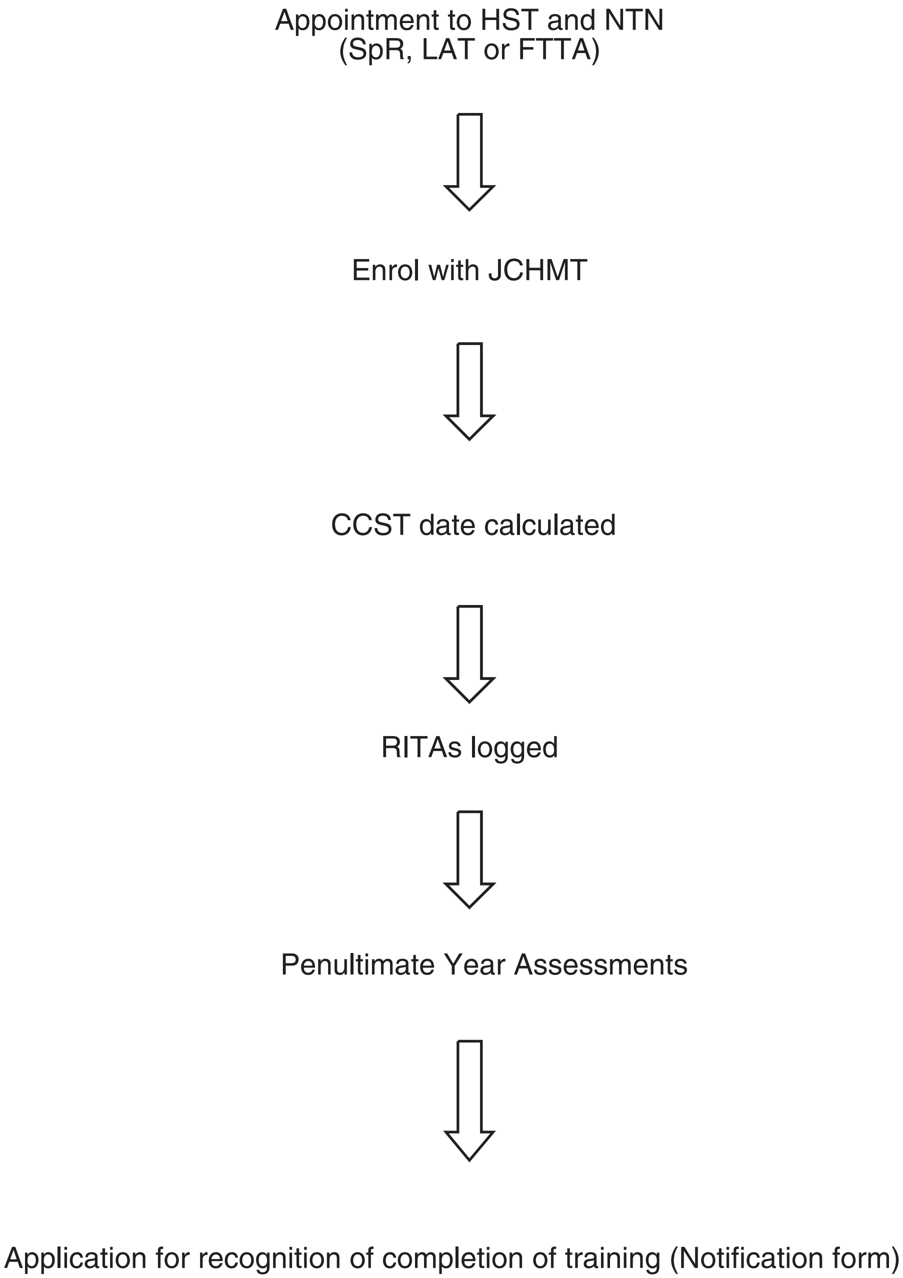

Appointment to HST and NTN
(SpR, LAT or FTTA)
Enrol with JCHMT
CCST date calculated
RITAs logged
Penultimate Year Assessments
Application for recognition of completion of training (Notification form)

Figure 4.1: Processes used to manage higher specialist trainee (HST) in medicine.

over recent years to enable educational supervisors to train and educate themselves for the role.[7–10] Examples include the Royal College of Physicians 'Physicians as Educators' series[11] and the Royal College of Paediatrics Appraisal and Assessment package.[12] Each college has recommended the qualities and skills that should be developed through appraisal and most have designed documentation that should be completed at assessment.[13,14] These documents are sent to the programme directors and, for some specialties, chairs of a panel specifically to review the most recent period of training.

The application of the RITA process has undergone a series of metamorphoses that vary according to the specialty. Where there is a strong and experienced team of local educational supervisors, the process has followed the recommended model. Where there is a greater variety of local educational standards and experience and especially where the development of craft skills are key drivers, RITA panels have developed methods of working in which a degree of trainee assessment is made.[15,16] The general view is that this is not appropriate, because consultants whose contact with the student is occasional and with whom they do not normally work are not in a position to assess them. However, skill-station assessments, in which specialist tasks are undertaken and rated against a validated scale, are a regular responsibility of some RITA panels.[18] Some specialty groups have considered including assessments in a patient or scenario simulator because of proven validity.[19] This is an area which will expand to inform the RITA process in future and is being piloted in Foundation Programmes in London Deanery.

Current reliability of workplace assessment

The degree of reliability of workplace assessments is a continuing weakness of the present system.[20,21] There are two main reasons for this. First, in general, few educational supervisors have been trained in the standardised evaluation of the clinical skills of their specialist registrars. Second, the ways in which records of skills evaluations have been recorded have not been designed to lead to the development of criterion referencing.

There is a significant difference between craft specialties and those that have a primarily medical and counselling style. There is greater scope for criterion referencing for craft specialties, whereas the establishment of an agreed standard for the other specialties, and for those generic skills of a similar type, is more difficult, because of the subjective nature of such assessments.[22] In non-craft specialties educational supervisors 'sign-up' a student as competent in a range of tasks after an arbitrary period of time during which they may have undertaken an arbitrary number of procedures. Until now there has been no robust mechanism for formal assessment and poor performance was variably recognised and unreliably addressed.[17] When one examines various specialty curricula, scoring systems for assessing attributes can vary.

On the other hand, it has been accepted that the opinion of an experienced consultant on the performance of an individual student has authority, particularly when the performance is satisfactory. However, although an experienced consultant is also a reliable source of evidence on the poor performance of a doctor, there is a risk that, for a minority, such an adverse evaluation may be based upon a brief period of under-achievement secondary to a temporary outside influence, such as illness or personal stress.[23] Doctors challenge such adverse subjective opinions often successfully. Data available from the JCHMT, for example, shows that issuing of RITA D or E grades is low indicating the vast majority of doctors make satisfactory progress.[41] This is hardly surprising given the paucity of validated evidence.[24]

Range of current methods

Most evaluation forms list a number of clinical, management, communication and academic skills based on relevant curricula and aligned to the General Medical Council's *Good Medical Practice*[25] and ask the educational supervisor to grade the student as satisfactory or less than satisfactory.[26] Some systems include grades above and below these two, but the lowest grade is rarely employed and the higher ones only occasionally.[27] Each grade may be given a typical descriptor, to assist the assessor in selecting one. A paradigm of this system is used for much of nurse training, but here the descriptors are criterion referenced and this allows for much greater inter-observer consistency.

In an increasing number of specialties, objective structured skills assessment systems have been developed, notably in general surgery and cardio-thoracic surgery. These assessments employ the technical developments in clinical modelling to set a standardised practical skills task, such as repairing a piece of simulated small bowel. Repeated evaluations made of students at different stages of their training has permitted the developers of these systems to describe levels of performance. First, that have medians and standardised variances and, second, that map accurately to the standard expected from a student at a particular stage of training. However, it is a challenge to maintain standards as there are constant additions made to curricula. It is rarely clear how the attainment of such defined skills should be measured.

The recent introduction of the Foundation Programme for doctors in their first two years of postgraduate training includes formal evaluations of their skills and of their performance.[26,28] The tools that have been chosen with which to make these assessments are the mini-Case Examination (mini-CEX); Case-Based Discussions (CbDs); Directly Observed Procedural Skills (DOPS); and the mini-Peer Assessment Tool (mini-PAT). While, at present, there are no data available from these assessments to permit one to claim that they are valid, reliable and reproducible, the accumulation of reports of these assessments may be expected to allow such claims to be made in the future.[10,29] Some of these assessment tools have been piloted by the JCHMT for Medical Specialty SpRs in 2003–04. This study showed the methods were feasible, reliable and valid for use in the workplace for the assessment of SpRs and by extrapolation could be used to assess all grades of doctors in training.[30] A number of these assessments sampling over a wide range of a doctor's practice should help build up a representative picture of the quality of a doctor's overall practice. These methods are to be rolled out for use in 2005. A Knowledge Based Assessment (KBA) is also being developed in the 'Best of Five' MCQ format for introduction in 2006. The fundamental importance of the introduction of such methods cannot be overestimated.[17] It is the evidential basis of performance assessments that is lacking at present from British postgraduate medical education and the introduction of these methods offers the possibility of providing the evidence of performance that is needed.[31,32]

The introduction of continuum training

Foundation programmes mark the beginning of the second and final phase of the reform of postgraduate medical training in the UK. August 2007 will mark the introduction of continuum training and, during the succeeding years, the Specialist Registrar grade will disappear as a discrete post. Instead, postgraduate education from the second Senior House Officer year to the completion of training will be continuous. Within this continuum, evaluations of each student's performance will build upon those achievements recorded in their portfolios, so that, when they are judged to have reached the point at which they should receive a Certificate of the Completion of Specialist Training (CCST), the portfolio will contain unambiguous evidence that justifies the issuing of a CCST.

Ipso facto, it is imperative that appropriate adaptations of the mini-CEX, mini-CbD, DOPS and mini-PAT are made that permit appropriate benchmarking of a student at each stage of their training, so that their progress may be charted.

Further, students are often uncertain about the choice of an appropriate subspecialty and evaluations of performance at specific clinical tasks could provide useful information about aptitude and facilitate subspecialty selection. Assessments at later stages of training should require increasing amounts of independent performance and of clinical responsibility culminating in a level that equates within those of a consultant.

It is an issue of current debate as to whether national assessment centres will be developed to deliver a strategy that will produce reliability and economies of scale into the assessment arena. The experiences of the health authority, the National Clinical Assessment Service (NCAS) formerly known as the National Clinical Assessment Authority (NCAA), will inform this decision.

It is a mistake to imagine that the award of a CCT marks the end of medical education. The impact of the European Working Time Directive has been to reduce the opportunities for clinical contact and practice.[33,34,35] Consequently, consultants at appointment have very much less experience than their predecessors at the same stage of their careers and, while they are competent, they have had less chance to become skilful. The strength of appointment to a Senior Registrar post was to perform as a quasi-consultant, while remaining under the support and supervision of an established consultant. In the future, newly appointed consultants will have the support and supervision of a senior colleague in a comparable manner. It seems sensible that the same assessments methods should become part of the portfolio that is submitted for the Continuous Professional Development (CPD) of consultants, so that skills development can be mapped against outcome and used as an integral part of job planning.

Future of the RITA process

There are likely to be three improvements to the appraisal and assessment procedures for doctors achieving their CCT via continuum training.[36] They may be listed broadly under the headings of *training*, *evidence* and *validation*. The changes will require some modification of the monitoring of training that occurs in the RITA process.

Training

General practice has led the way in the care that has been taken in training doctors as primary care physicians. The assessment and certification of primary care physicians as 'trainers' is a model for the way in which we should ensure that those consultants charged with the responsibility of training the consultants of the future have the skills and abilities to do the job.[37] Another example from primary care is the way in which training is seen as work that requires both time and recompense. The introduction of the new consultant contract provides the opportunity to designate a specific number of programmed activities (PAs) as an integral part of the job planning process.

In an attempt to assure the quality of educational supervision for the Foundation Programme, the London Deanery rolled out a training programme in 'Educational Supervision for the Foundation Programme'. This included some practical experience in the use of assessment tools, such as the mini-CEX and CbDs. Development of a series of training programmes, designed to assure that the standards of educational supervision are as similar as possible with respect to all trainers, would lay the basis for training consistency and reliability.[38]

Evidence

The adoption of the assessment tools already described, in parallel with a refinement of the assessment records used by most specialties, will provide evidence of two kinds. First, evidence of competence as shown by the mini-CEX, CbD and DOPS. Second, evidence of performance as demonstrated by mini-PAT and the summary assessment forms presently in use, but with some refinement. The latter is important, because it reflects a considered evaluation of the student's accomplishments in workplace clinical situations.

The portfolio, complete for each student on the award of the CCT, should continue to grow as additional skills are acquired.

Validation

Validation is a regular process employed by the General Medical Council to assure that doctors remain fit to practise. In this context it refers to the need to validate that the methods employed to assure that students are competent are valid.[39]

Assessments of the competence and performance of students are carried out across a wide variation of training environments and by many assessors. Ultimately, one would wish to be able to demonstrate that the standards of training and assessment are the same for every doctor in training. This is not possible at present. It is apparent that using the methods and comparison already described and by collecting these data, eventually one will be able to define the tolerances of the system and to validate the methods used.[40]

Once this point has been reached, the performance criteria that need to be achieved by a student in a given specialty at a particular stage of training will be more precisely defined and the achievement of a CCT could be said to be criterion referenced. Undoubtedly the accumulated evidence presented and quality assured by the RITA process will be a method for the revalidation for doctors in training.

Specialty training committees (STCs)

Currently, the specialty training committees (STCs) that review the performance of those students in a local programme do so annually. RITAs on individual students take place more frequently only when there have been problems.

Some specialties have slimmed down the process so that, for the majority of students, it is a paper exercise, when the RITA panel examines and accepts standard documentation that confirms that the student has achieved the training objectives that were set at the start of the training year. Separate face-to-face meetings with students experiencing difficulty in achieving their objectives are arranged. As there are always only a few such learners, more time can be spent with them, in order to consider their difficulties and plan appropriate further placements.

STCs will take responsibility for managing doctors in training for the whole of continuum training from foundation schools. It is likely that the relationship between the royal colleges and the STCs will become closer. It is vital that the partnership between the royal colleges, who set the curriculum and approve training posts, and the postgraduate deaneries, who act as the guarantors of quality control for training is strengthened.

The ideal RITA model is one that requires face-to-face meetings with only those students who are making less than satisfactory progress. However, the significant increase in evidence of competence will lead to ever-larger portfolios. It would be easier for students to bring them for review than to mail them. This makes the development of on line documentation to support training very important. Not only would access to the training record by the RITA panel be facilitated, but standardisation and early warning of training problems also.

Conclusion

The public need to be confident that the award of a CCT indicates that a doctor has completed training to a satisfactory and agreed standard. The profession has a duty to ensure that each doctor's portfolio contains unequivocal evidence of this and to develop a training structure that supports the doctor in completing his or her training. By building upon the methods for the assessment of competence introduced for the Foundation Programme and improving the consistency of the RITA process, continuum training can meet these demands and assure the skills and ability of the specialists of the future.

References

1. Calman KC, Temple JG, Naysmith R *et al*. Reforming higher specialist training in the United Kingdom – a step along the continuum of medical education. *Med. Educ.* 1999; **33**: 28–33.
2. The Green Guide.
   www.nimdta.gov.uk/downloads/home/01.01_the_green_guide_pdf.pdff
3. www.pmetb.org.uk/
4. The National Health Service Executive. *A Guide to Specialist Registrar Training.* London: The National Health Service Executive; 1998.
5. Schuwirth LWT, Van Leuten CPM. Changing education, changing assessment, changing research? *Med. Educ.* 2004; **38**: 805–12.
6. Bache J, Broon J, Graham D. In-training assessment for specialist registrars; views of trainees and trainers in the Mersey Deanery. *Royal Soc Med*. 2002; **95**: 612–13.
7. Howley LD. Performance assessment in medical education: where we've been and where we're going. *Evaluation & the Health Professions*. 2004; **27**: 285–303.
8. Lynch DC, Surdyk PM, Eiser AR. Assessing professionalism: a review of the literature. *Med Teacher.* 2004; **26**: 366–73.
9. Duffy FD, Gordon GH, Whelan G *et al*. Participants in the American Academy on Physician and Patient's Conference on Education and Evaluation of Competence in Communication and Interpersonal Skills. Assessing competence in communication and interpersonal skills: the Kalamazoo II report. *Academic Med.* 2004; **79**: 495–507.

10. Frohna JG, Kalet A, Kachur E *et al*. Assessing residents' competency in care management: report of a consensus conference. *Teach Learn Med.* 2004; **16**: 77–84.
11. Perkins GD, Rayner HC. Physicians as educators. *J Royal College of Physicians of London*. 2000; **34**: 112.
12. Royal College of Paediatrics and Child Health website: www.rcpch.ac.uk/education/appraisal.html
13. JCHMT curriculum guidance pages: www.jchmt.org.uk/specialtyHome.asp
14. JCHST information and links to curriculum information: www.jchst.org/
15. Norcini JJ. Work based assessment. *BMJ*. 2003; **326**: 753–5.
16. Flynn KJ, Payne S. Getting the best out of RITA. *BMJ.* 2002; **325**: S52.
17. Arnold L. Assessing professional behaviour: yesterday, today, and tomorrow. *Academic Med.* 2002; **77**: 502–15.
18. Smee S. Skill based assessment. *BMJ.* 2003; **326**: 703–6.
19. Paisley AM, Baldwin PJ, Paterson-Brown S. Validity of surgical simulation for the assessment of operative skill. *Brit J Surgery.* 2001; **88**: 1525–32.
20. Williams RG. Have standardized patient examinations stood the test of time and experience? *Teach Learn Med.* 2004; **16**: 215–22.
21. Downing S M. Validity, on the meaningful interpretation of assessment data. *Med Educ.* 2003; **37**: 830–7.
22. Carraccio C, Englander R. Evaluating competence using a portfolio, a literature review and web-based application to the ACGME competencies. *Teach Learn Med.* 2004; **16**: 381–7.
23. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med.* 2003; **15**: 270–92.
24. Ringsted C, Henriksen AH, Skaarup AM *et al*. Educational impact of in-training assessment (ITA) in postgraduate medical education: a qualitative study of an ITA programme in actual practice. *Med Educ.* 2004; **38**: 767–77.
25. The General Medical Council. *Good Medical Practice*. London: The General Medical Council; 1998.
26. Bann SD, Datta VK, Khan MS *et al*. Attitudes towards skills examinations for basic surgical trainees. *Int J Clin Pract.* 2005; **59**: 107–13.
27. Wass V, Vleuten C van der, Shatzer J, Jones R. Assessment of clinical competence. *Lancet.* 2001; **357**: 945–9.
28. The Foundation Programme Committee of the Academy of the Royal Colleges. *Curriculum for the foundation years in postgraduate education and training*. The Foundation Programme Committee of the Academy of the Royal Colleges; 2005.
29. Sturmberg JP, Atkinson K, Farmer EA. Research and Development Subcommittee, Board of Examiners, The Royal Australian College of General Practitioners. Standards and performance – attainment and maintenance of professional capabilities. *Austr Fam Phys*. 2005; **34**: 371–3.
30. JCHMT.org.uk/assessment/studysynopsis.asp Report of RCP (UK) evaluation of performance assessment methods for specialist registrars in medicine.
31. Schuwirth LWT, van der Vleuten CPM. Merging views on assessment. *Med Educ.* 2004; **38**: 1208–10.
32. PMETB Workplace based assessment. A paper from the PMETB workplace based assessment subcommittee:
www.pmetb.org.uk/media/pdf/g/c/Workplace_Based_Ass_paper_FINAL_20051.pdf
33. Gossage JA, Modarai B, McGuinness C *et al*. The modernisation of the surgical house officer. *Ann R Col Surg Eng.* 2005; **87**: 369–72.
34. Marron CD, Byrnes CK, Kirk SJ. An EWTD-compliant shift rota decreases training opportunities. *Ann Roy Coll Surg Engl.* 2005; **87**: 246–8.
35. Paice E, Reid W. Can training and service survive the European Working Time Directive? *Med Educ.* 2004; **38**: 336–8.

36. Schuwirth LWT, van der Vleuten CPM. Changing education, changing assessment, changing research? *Med Educ.* 2004; **38**: 805–12.
37. Jackson N. Assessment and work based learning in primary care. *Work Based Learning in Primary Care*. 2003; **1**: 89–92.
38. Norcini JJ. Current perspectives in assessment of performance at work. *Med Educ*. 2005; **39**: 880–89.
39. Catto G. Improving professional competence – the way ahead? *Int J Qual Hlth Care.* 2003; **15**: 375–6.
40. Vleuten CPM van der, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ*. 2005; **39**: 309–17.
41. Tunbridge M, Dickinson D, Swan P. Outcomes of specialist registrar assessment. *Hosp Med.* 2002; **63**: 684–7.

Chapter 5

Assessment for recruitment

Fiona Patterson and Pat Lane

Introduction

Selecting the wrong doctor for the job could have serious consequences; the potential costs (both human and financial) are substantial. In the UK, the Modernising Medical Careers (MMC) change programme has placed more emphasis on delivering robust methods for assessing and developing doctors throughout the career life-cycle. Similarly, the recently published PMETB Principles of Assessment have accelerated the need for fair and transparent selection processes.[1] In this chapter, we describe the key concepts associated with competency-based selection and the academic literature. The application of best practice is demonstrated through a case study showing how a selection system was developed and validated for doctors applying for training in general practice.

The selection process

Figure 5.1 summarises the key stages in the selection process. A thorough job analysis is the foundation to an effective selection process and is used to guide choice of selection methods. The outputs from a job analysis should detail the tasks and responsibilities in the target job and also provide information about the particular behavioural characteristics required of the job holder.[1] The analysis will provide an inclusive description of the job-relevant knowledge, skills, abilities and attitudes that are associated with competent job performance.

The next stage in the process is to identify selection tools (e.g. work simulation exercises, interviews, application forms) that can be used to examine whether candidates display the required characteristics or not. These tools are then used to assess candidates and selection decisions are taken. Conducting empirical validation studies allows the quality of the selection process to be monitored. Best practice selection is an iterative process, where the selection system evolves over the course of time. Specifically, evaluation information can help inform improvements such as updating the selection criteria targeted.

Key concepts

When choosing the assessment method(s) it is important to make sure that the assessment method is accurate (*reliable*), relevant (*valid*), *objective*, *standardised*, administered by trained professional(s), and monitored. Evaluation of the system is essential to ensure that selection tools are also *fair*, *defensible*, *cost-effective* and *feasible*. Feedback is used to continually improve the selection system to enhance accuracy and fairness. Further, there are legal reasons for ensuring accurate selection procedures are used and it is essential for compliance with current employment law.

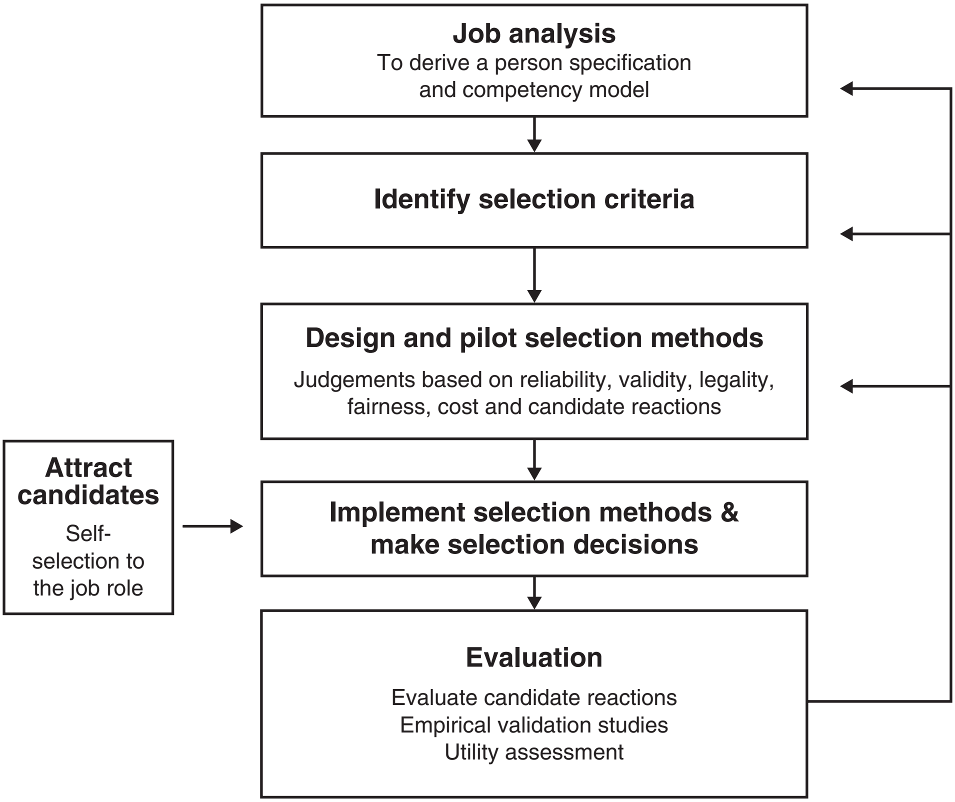

Job analysis
To derive a person specification and competency model
Identify selection criteria
Design and pilot selection methods
Judgements based on reliability, validity, legality, fairness, cost and candidate reactions
Attract candidates
Self-selection to the job role
Implement selection methods & make selection decisions
Evaluation
Evaluate candidate reactions
Empirical validation studies
Utility assessment

Figure 5.1: Selection system design and validation process.

Conducting validation studies can become very complex in practical terms since researchers would rarely use one single predictor to make selection decisions and applicants will be judged on multiple assessment criteria. Given the multifaceted nature of job analysis information, recruiters are likely to design multiple selection tools to assess these criteria. Therefore, recruiters must decide whether a job applicant must score highly on all assessment criteria (non-compensatory) or whether high scores on some criteria can make up for low scores on another (compensatory). In practice, recruiters might assign different weightings to various assessment criteria, depending on the nature of the job role. For example, if clinical knowledge is the most important criterion and applicants do not achieve a certain score at shortlisting, they may not be considered further.

In summary, paying attention to best practice criteria for designing selection methods is crucial. We suggest 12 key issues that should be reviewed when designing and implementing a selection system, as follows.

1 Establishing reliability and validity of the tool.
2 Positive employee/candidate reactions.
3 Ensuring ease of interpretation.
4 Ensuring generality of use.
5 Minimising costs and maximising value.

6 Practicality.
7 Expertise required for analysis and interpretation of information generated by the tool.
8 Utility.
9 Fairness perceptions.
10 Educational impact/value.
11 Generates appropriate feedback.
12 Procedures are in place for ongoing validation, evaluation and renewal of assessment tools.

Selection methods – the research evidence

There are many different selection methods available and their relative accuracy is well-documented in the research literature.[2] A full review is beyond the scope of this chapter, so we focus on two methods, interviews and 'assessment centres'.

Interviews

Selection interviews are undoubtedly the most widely-used selection method for short-listed candidates. However, research has consistently shown that interview effectiveness varies greatly depending on the degree of structure employed, and the extent to which it targets evidence on key competencies. A structured, competency-based interview, where candidate responses are assessed against standardised rating scales, can be a very effective selection method. Conversely, open-ended, unstructured interviews, are a relatively unreliable selection method. Because different selection methods measure different competencies more effectively than others it is desirable to use several methods in a single selection process.

Assessment centres

An assessment centre (AC) is a selection method, not a place. ACs make use of a combination of different selection tools and allow candidates to be assessed by multiple assessors. ACs were first used during World War II to select military personnel. However, it was not until the American company AT&T applied ACs to identify industrial managerial potential in the 1950s that the idea developed as a selection method. Since this time, ACs have become widely used as a tool for recruitment and selection. Recent survey data indicates that the use of ACs is increasing more rapidly than any other selection method in the UK, with 65% of organisations (employing more than 1000 people) reporting using the AC method.[3] It is only recently that this approach been used in medicine.[4,5]

ACs are often the core of competency-based selection systems, as they combine a range of assessment techniques to achieve the fullest and clearest indication of competence.[6] ACs typically involve a one-day assessment of applicants, using different methods, including various work sample exercises such as group discussions, in-tray exercises, simulations and so on. Gains are made in reliability and validity because ACs make use of a combination of different exercises (using a multi-trait, multi-method approach) and use standardised scoring systems to measure key competencies. Best practice suggests that ACs use work-related exercises, allowing behavioural observation by independent, trained assessors.

As far as predictive validity is concerned ACs perform well.[7,8] A review by Schmitt *et al.* found ACs were the best method of predicting job performance, with a mean correlation of 0.43 between overall AC rating and measures of job performance.[9] In summary, the research evidence clearly shows that ACs are better predictors of performance than interviews alone.[10–13]

Candidate reactions

An increased emphasis has been placed recently on the importance of candidates' reactions to different recruitment methods.[14] Considerable research has attempted to determine applicants' view on selection methods. Research has tended to explain the different factors that affect applicant reactions using theories of organisational justice.

- Distributive justice focuses on perceived fairness regarding equity (where the selection outcome is consistent with the applicant's expectation) and equality (the extent to which applicants have the same opportunities in the selection process).
- Procedural justice refers to the formal characteristics of the selection process such as information and feedback offered, job-relatedness of the procedures and methods, and recruiter effectiveness.

Anderson *et al.* suggest that four main factors seem to account for positive or negative applicants' reactions where selection methods are:[15]

1 based on a thorough job analysis and appear more job relevant
2 less personally intrusive
3 do not contravene procedural of distributive justice expectations
4 allow applicants to meet in person with the recruiters.

Other literature suggests that applicants prefer multiple opportunities to demonstrate their skills and that the selection system is administered consistently for all applicants.

Fairness issues

Fair selection and recruitment is based on:

- having objective and valid criteria (developed through a job analysis)
- accurate and standardised assessment by trained personnel
- monitored outcomes.

There has been a great deal of research exploring the extent to which selection procedures are fair to different subgroups (such as ethnic minorities or women) of the population. First, it needs to be made clear that a test is not unfair or biased simply because members of different subgroups obtain different scores on the tests. Men and women have different mean scores for height; this does not mean that rulers are unfair measuring instruments. However, it would be unfair to use height as a selection criterion for a job, if the job could be done by people of any height, since it is important for selection criteria to be job-related. Normally, of course, the extent to which a selection method is related to job performance can be estimated by validation research, and it is clear therefore that fairness and validity are closely related.

To demonstrate how a competency-based selection system is designed and implemented, we present a case study for doctors applying for training in general practice. The work started in 1996 and we have summarised the developments over the course of several years.

Case study: A competency-based selection system for general practice

Background and context

The NHS had been in existence for many years before doctors were given the opportunity to work as trainee assistants under supervision. Organised training schemes for general practice were created, principally between 1969 and 1973. This training was voluntary (hence vocational) but was made mandatory in 1981 by the Joint Committee on Postgraduate Training for General Practice (JCPTGP). On the completion of training, doctors would submit statements of satisfactory completion of training to the JCPTGP and in return received a certificate enabling them to practise, unsupervised, as a GP. The subsequent performance of a number of vocationally trained doctors was less than satisfactory and some had their names erased from the general medical register by the GMC. In response to a growing concern that more robust evidence of competence was required the process of summative assessment was introduced and this became mandatory in 1996.

For over 25 years GP training was largely focused upon developing the trilogy of knowledge, skills and attitudes. In the mid-1990s recruitment to general practice slipped to its lowest level for 15 years. The applicant/placement ratio had dropped from its peak of 15/1 in 1981 to around 2/1 or less. Many training schemes were resorting to re-advertising vacancies up to four times for each intake. It was not uncommon to have unfilled vacancies. Significant contributory factors to this situation were:

- the introduction of the Calman reforms of specialist training in the NHS, which had streamlined the pathway to becoming a consultant
- a generally negative medical press about general practice (fundholding had run out of steam and impending new NHS organisational changes in primary care were not welcomed by GPs).

Development of a new selection process

During debates in the late 1990s about the implications of the transfer of funding for GP training the GP directors were concerned that:

- those doctors who could 'operate' the system known as patronage made greater progress than others
- a number of doctors were being referred to the GMC within months of graduation
- it was noted that a significant number of GP trainees were quitting training courses because they had chosen the wrong career
- around 5% of doctors were failing summative assessment.

Whilst it was recognised that improving and changing the existing selection system would not be a panacea to addressing these concerns, research and practice in other occupations suggested that it was a valid starting place. The development of a new GP selection process followed best practice stages, first in identifying the key criteria for selection and second in developing, piloting and validating the selection methods.

To identify appropriate selection criteria (between 1996 to 1999), Patterson, Lane and Ferguson conducted three independent job analysis studies to define a behavioural competency model for general practice.[1] The knowledge, skills, abilities and other attributes (KSAOs) that accurately and consistently appeared to define competent GP performance, were systematically elicited and 11 key competencies emerged (*see* Box 5.1).

Box 5.1: Key competencies in GP performance

1.	*Empathy and sensitivity*	(recognising patient's thoughts and feelings)
2.	*Communication skills*	(active listening, clarity of explanation)
3.	*Problem-solving*	(identifying root cause and making diagnosis)
4.	*Professional integrity*	(respect, vocational enthusiasm)
5.	*Coping with pressure*	(calm under pressure, recognising limitations)
6.	*Clinical expertise*	(clinical process awareness, identifying options)
7.	*Managing others and team involvement*	(collaborative style supports others)
8.	*Legal, ethical and political awareness*	(aware of responsibilities)
9.	*Learning and personal development*	(reflects and learns from others)
10.	*Organisation and administration skills*	(prioritises conflicting demands, efficient)
11.	*Personal attributes*	(flexible, sense of humour, shows initiative, decisive)

Importantly, each of these is embodied by a cluster of indicators that define, in behavioural terms, the knowledge, skills and abilities for each competency. The model was validated by general practitioners and patients. After extensive consultation, it was decided that certain domains are best targeted at selection (e.g. communications skills, professional integrity), and others are best addressed during training (e.g. legal, ethical and political awareness).

The three postgraduate deaneries in the old Trent Region (Sheffield, Nottingham and Leicester) collaborated to develop the methods of selection, including competency based application forms, competency based referees' reports and an assessment centre. Figure 5.2 summarises this multi-stage process to develop a competency-based selection system. Of particular importance are the feedback loops, highlighting the operation of an iterative, systemic approach to selection.

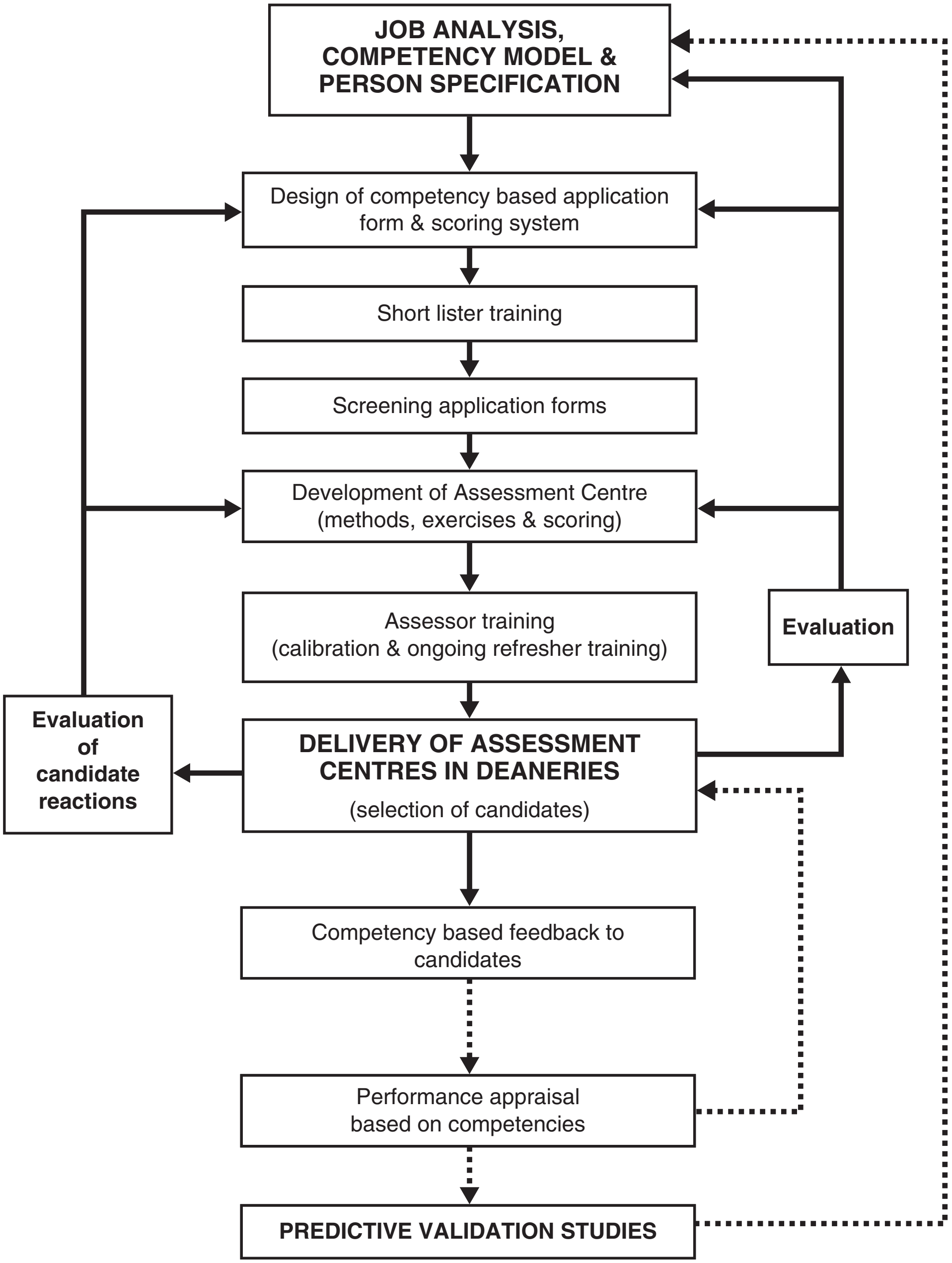

JOB ANALYSIS, COMPETENCY MODEL & PERSON SPECIFICATION
Design of competency based application form & scoring system
Short lister training
Screening application forms
Development of Assessment Centre (methods, exercises & scoring)
Assessor training (calibration & ongoing refresher training)
Evaluation
Evaluation of candidate reactions
DELIVERY OF ASSESSMENT CENTRES IN DEANERIES
(selection of candidates)
Competency based feedback to candidates
Performance appraisal based on competencies
PREDICTIVE VALIDATION STUDIES

Figure 5.2: Design and validation of the GP competency-based selection process.

Application forms

Application forms have traditionally been designed to collect information regarding educational qualifications and work experience. This information is essential to demonstrate capability to do the job, however, ranking decisions here are difficult because work history and qualifications are difficult to reliably differentiate, and sifting is usually based on fairly limited information. A new application form was developed to include competency-based questions where applicants are required to supply more focused work experience information, relating to the demonstration of the target competencies. An example question could be 'Describe a situation when you have demonstrated empathy and sensitivity when dealing with a patient. What did you do and what was the outcome?' Candidates are asked to briefly outline a situation they have encountered and how they dealt with it. There is not a 'correct' answer, as responses are scored according to agreed competency-based criteria. Several of these questions are posited and shortlisters are trained to assess responses using standardised rating scales, so that reliability is addressed. Also, two shortlisters examine the same application forms so that inter-rater reliability can be assessed. Clearly, it is suggested that responses to such questions are easily 'fakeable' in that applicants could just make up their responses. In practice, however, some applicants fail to demonstrate that they understand the difference say between empathy and sympathy, and that awareness of behaviours associated with empathy is actually part of the question! These competency questions provide additional 'non-academic' information to make decisions about sifting. The evaluation studies showed that scores on these questions were significantly positively related to how applicants performed at the AC.

References

References are traditionally open-ended and research has consistently shown that reference information has limited use in selection (due to leniency effects, etc.). To improve reliability, competency-based, standardised rating scales were provided to referees. Evaluation showed that reliability of reference forms and of referees was significantly improved.

During the initial development of the assessment centre, five exercises were developed based on the agreed competencies. An assessment centre day (run simultaneously in each deanery), designed for selecting doctors to train for general practice, was constructed. The exercises used were specifically designed so that several competencies could be assessed during each exercise. Standardised rating scales and checklists were used throughout the process to optimise objectivity and assessors were rigorously trained.

The five exercises piloted and used in 2000 were as follows.

1 *Simulation exercise* (20 minutes; candidate acts as doctor and a role player acts as patient, in a given scenario).
2 *Group exercise* (30 minutes; a group of four candidates is asked to resolve work-related issues).
3 *Written exercise* (30 minutes; candidates work independently to prioritise six on-call issues, and justify chosen sequence).
4 *Competency-based structured interview* (20 minutes; candidates provide evidence based on specific previous experience).
5 *Technical interview* (20 minutes; candidates respond to questions relating to clinical practice).

Practical issues

Each exercise is observed by a different assessor with an aim to reduce potential for bias. The assessments are used to construct a matrix specifying which competencies are to be assessed in each activity. Assessors are specially trained in behavioural observation and recording, ensuring that individual ratings can be explained via the competency model. As a result, when the final 'wash-up' discussion takes place, a rounded picture of each individual emerges, grounded in specific demonstrated behaviours.

A 'wash-up' meeting is held at the end of the AC process where assessors can be audited by a facilitator. In this way, any inconsistencies in scoring can be discussed using evidence based on observed behaviour. Hence, several trained assessors determine the final selection decision and a trained facilitator guides the process.

Evaluation and validation

After reviewing the outcomes of the first two years, it was decided to make two alterations in 2003 to streamline the process. Exercise 4 was removed because the assessors found it was duplicating the shortlist competency questions (and therefore had limited added value). The technical interview was replaced with a comprehensive MCQ paper to assess general medical knowledge (which enabled a wider assessment to be made and it was machine marked). At this time, a comprehensive validation study was conducted during 2003/04 to assess the reliability, fairness and validity of the new selection system and the system is currently undergoing further extensive evaluation. Some of the results have recently been published demonstrating that the new selection system demonstrates good internal and predictive validity.[5]

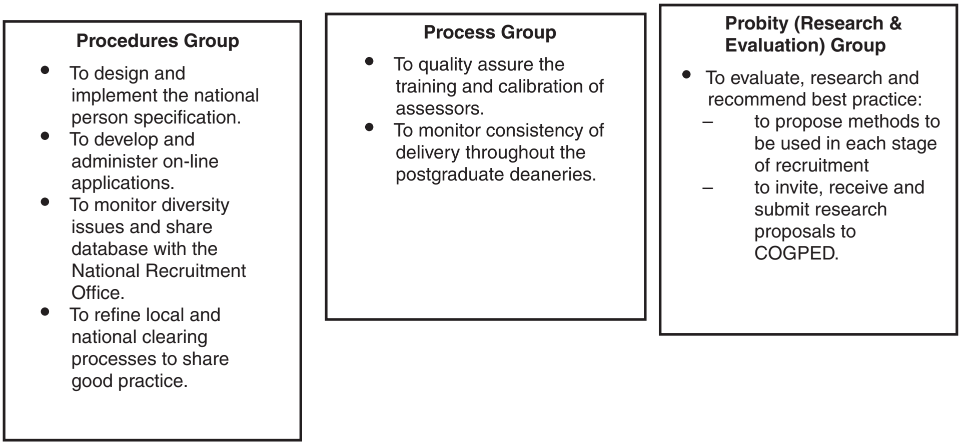
Utility and cost-effectiveness

Adopting a competency-based approach takes time and resources to develop in the first place. The benefits, however, soon outweigh the costs of recruiting the wrong person for the job – both for that person, the profession and, crucially, the patient. In 2005, the average cost for the selection process per appointed GP trainee is around £400–£450. Currently, over 2,600 new doctors are recruited into training in general practice per annum and the failure rate of summative assessment is just under 5%. Training costs including salaries, allowances, supplements and education costs average £83,700 (2004/05) per annum and for a 3-year programme this amounts to over £250,000 per doctor. Thus, prevention of one failure recoups the cost of the selection of over 500 doctors into GP training.

Unlike most other selection procedures, an AC generates a range of in-depth information about candidate knowledge, skills and attitudes. The information collected can be used to generate individual development plans for doctors, so that potential performance deficiencies can be targeted more accurately. More accurate training plans are likely to lead to improved performance. Equally for those candidates not successful in gaining a place on a training scheme, detailed competency based feedback on their performance may be constructive in their career planning or personal development.

Current and future directions

From 1 April 2000 all costs related to training in general practice, in England, were transferred out of the General Medical Services (GMS) budget into the Medical and Dental Education Levy (MADEL) under the management of the postgraduate deans and directors of Postgraduate General Practice Education (DsPGPE). GP directors predicted this change would generate many opportunities to improve GP training, as well as transform the recruitment process. The direction of travel had to be towards a national, equal opportunities based recruitment of doctors into GP training and there is now a deanery-centralised National Recruitment Office.[16] The National Recruitment Steering Group (NRSG) of COGPED has now developed a nationally agreed recruitment system guided by three working groups (procedures, process and probity) (*see* Figure 5.3).



Procedures Group

- To design and implement the national person specification.
- To develop and administer on-line applications.
- To monitor diversity issues and share database with the National Recruitment Office.
- To refine local and national clearing processes to share good practice.

Process Group

- To quality assure the training and calibration of assessors.
- To monitor consistency of delivery throughout the postgraduate deaneries.

Probity (Research & Evaluation) Group

- To evaluate, research and recommend best practice:
  – to propose methods to be used in each stage of recruitment
  – to invite, receive and submit research proposals to COGPED.

Figure 5.3: 'Procedures', 'process' and 'probity' – the three working groups that guide the national recruitment system.

COGPED is currently working towards a standardised national approach to recruitment across all deaneries. Figure 5.4 shows a stage model summary of this new competency-based system. The new system will need to be ratified by the Postgraduate Training Committee of the Royal College of General Practitioners and has to meet the standards set by PMETB. New systems must be in place for all doctors entering GP training from August 2007. While such an approach will bring benefits in terms of reducing duplication of effort and improving standardisation of criteria, fairness and defensibility, there are also challenges in balancing the needs of local ownership of selection practices. Implementation of large-scale selection processes is dependent on the engagement of local trainers and course organisers and any future process will need to ensure that these key personnel are involved in future developments. In fact, the system would not have developed or evolved without the consultation, goodwill and collaboration of these key stakeholders in the process.

Having thoroughly evaluated the system in general practice, this competency-based selection approach is now being developed and used in many secondary care specialties (e.g. obstetrics & gynaecology, paediatrics, anaesthetics). The primary research has spring-boarded developments where common criteria across many medical specialties have been identified for selection (e.g. communications skills, professional integrity, problem-solving). Further, there are now pilot projects underway to investigate best practice selection into medical school via the Council for Heads of Medical Schools (CHMS).[17]

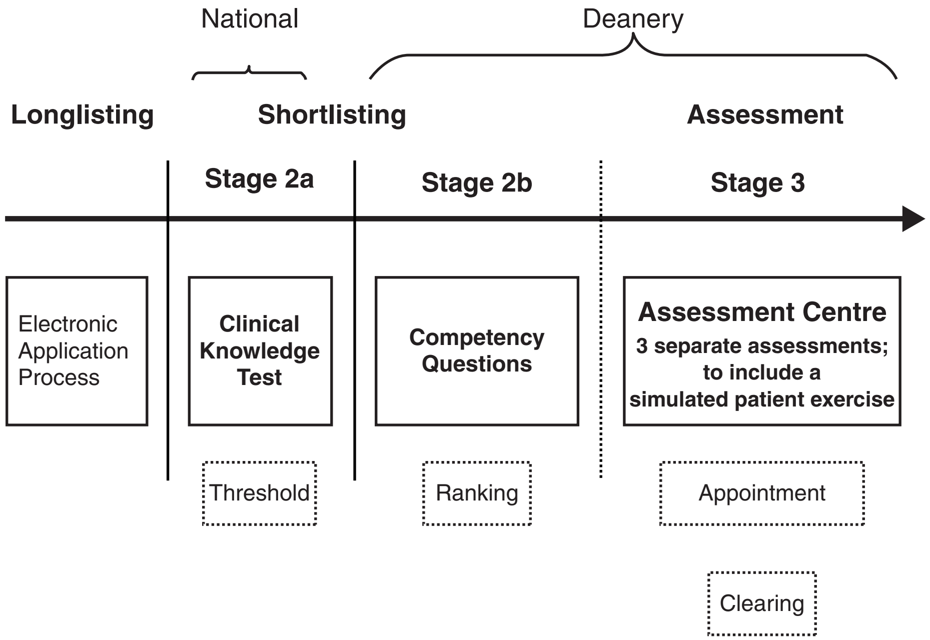


Figure 5.4: Overview of the Proposed GP Specialty Selection Process for 2007.

Summary

Research over the past three decades has provided a much clearer picture of the criterion-related validity of different selection procedures. No selection system is infallible, but a constructive framework for minimising the risks can be provided. The quality of a selection system is heavily dependent upon its design. Poor initial research will inevitably compromise the process: a poor job analysis, for instance, will potentially lead to an inaccurate selection criteria; sub-optimal assessment tools will weaken the validity of the ratings; insufficient training of assessors may undermine the objectivity of the process. ACs are increasingly popular and widely used, especially in organisations with larger numbers of employees or where the costs of mistakes are high. Contrary to popular belief, once designed, ACs are more cost-effective to run than panel interviews.[5] Further, initial costs are usually recouped, for instance, if the quality of individuals selected either raises the efficiency of the workforce and/or reduces attrition rates. Looking forward, selection systems evolve over time as organisational needs and job roles change and develop. This evolution process will be driven by the commitment and feedback of many stakeholders – the selection system designers, the assessors (lay and clinical), the administrators and, importantly, the candidates.

Acknowledgements

The authors would particularly like to thank many contributors to the work described in the case study to design a selection system for general practice. Notably, the three Associate Directors for Trent, LNR and SYSH deaneries: Dr Sathya Naidoo, Dr Roger Price and Dr Steve Ball.

References

1. Patterson F, Ferguson E, Lane PW, Farrell K *et al.* Competency model for general practice: implications for selection, training and development. *Br J Gen Pract.* 2000; **50**: 188–93.
2. Robertson IT, Smith M. Personnel selection. *J Occup Organ Psych.* 2001; **4**: 441–72.
3. Industrial Relations Service. The state of selection: An IRS survey. *Employee Devel Bull.* 1997; **85**: 8–18.
4. Patterson F, Lane P, Ferguson E, Norfolk T. A competency based selection system for GP trainees. *BMJ.* 2001; **323**: 2.
5. Patterson F, Ferguson E, Norfolk T, Lane P. A new selection system to recruit general practice registrars: preliminary findings from a validation study. *BMJ.* 2005; **330**: 711–4.
6. Woodruffe C. *Development and Assessment Centres: identifying and assessing competence.* London: Chartered Institute of Personnel and Development; 2000.
7. Damitz M, Manzey D, Kleinmann M, Severin K. Assessment center for pilot selection: construct and criterion validity and the impact of assessor type. *Appl Psychol: Int Rev.* 2003; **52**: 193–212.
8. Lievens F, Harris MM, Van Keer E, Bisqueret C. Predicting cross-cultural training performance: the validity of personality, cognitive ability, and dimensions measured by an assessment center and a behavior description interview. *J Appl Psychol.* 2003; **88**: 476–89.
9. Schmitt N, Gooding RZ, Noe RA, Kirsch M. Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Pers Psychol.* 1984; **37**: 407–22.
10. Harel GH, Arditi-Vogel A, Janz T. Comparing the validity and utility of behavior description interview versus assessment center ratings. *J Manag Psychol.* 2003; **18**: 94–104.
11. Bobrow W, Leonards JS. Development and validation of an assessment center during organizational change. *J Soc Behaviour Personality.* 1997; **12**: 217–36.
12. Hough L, Oswald FL. Personnel selection: looking toward the future – remembering the past. *Ann Rev Psychol.* 2000; **51**: 631–64.
13. Salgado, JF. Personnel selection methods. In: Cooper CL, Robertson IT (eds). *International Review of Industrial and Organizational Psychology.* 1999; **14**: 1–54.
14. Anderson N, Lievens F, van Dam K, Ryan AM. Future perspectives on employee selection: key directions for future research and practice. *Appl Psychol: Int Rev.* 2004; **53**: 487–501.
15. Anderson N, Born M, Cunningham-Snell N. Recruitment and selection: applicant perspectives and outcomes. In: Anderson N, Ones D, Sinangil H, Viswesvaran C (eds). *Handbook of Industrial, Work and Organizational Psychology.* 2001; **1**: 200–18.
16. Lane P, Sackin P. GP Registrar Recruitment. *Educ Prim Care.* 2003; **14**: 11–14.
17. Patterson F, Kerrin M, Carr V. Piloting the use of assessment centres for graduate entry to medical school. *Technical Report to Department of Health and CHMS.* 2005; 1–6.

Chapter 6

Workplace-based assessment for general practice training

Tim Swanwick and Nav Chana

Introduction

A number of themes emerge from the theoretical base that informs the contemporary practice of medical education. Amongst these are the acceptance that the learner is an active contributor in the learning process, that the context in which learning takes place is important, that learning is integrally related to the understanding and solution of real-life problems, that the past experiences and knowledge are critical in learning, that the learners' values, attitudes and beliefs strongly influence their learning and that the ability to reflect on one's practice is crucial for lifelong learning.[1]

In parallel with these developments, assessment is also changing. Traditionally, medical assessment focused on ritualistic end-point summative judgements conducted far away from the place of work. Knowledge-based tests, often uncoupled from examination of understanding or application, were twinned with simulated clinical encounters using surrogate or volunteer patients. To round it off, and to add a little spice, a viva or oral examination of dubious validity and reliability had to be endured as a final rite of passage. Things have moved on. Increasingly, assessments of medical competence now take us to the higher echelons of Miller's pyramid (*see* Figure 6.1) from the lower cognitive levels of '*what do you know*', traditionally tested in MCQs, to the '*what do you do*' of consultation observation and performance indicators.[1] The realisation that medical expertise is not a simple summation of stable and generic constructs has also led a movement away from the 'one trait-one instrument approach to assessment'[3] and coupled with an evolution in thinking about validity and reliability, medical education and training are gradually seeing a replacement of reductionistic tests by assessment approaches designed to assess medical competence in a more holistic and integrated way.[3]

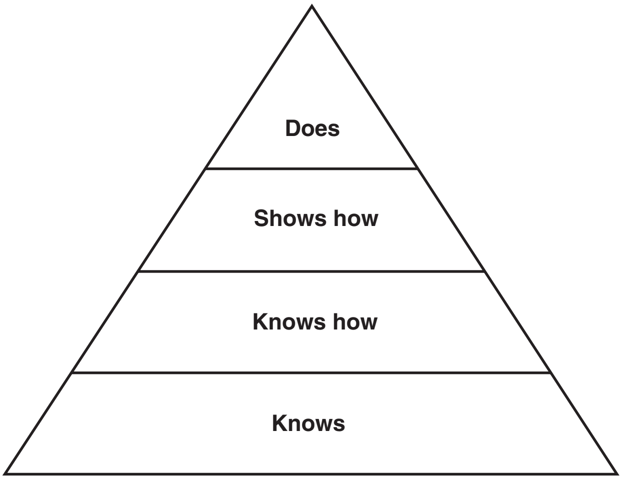


Figure 6.1: Miller's pyramid.[2]

But the assessment of the competence of doctors is inherently problematic. There are real differences between what doctors do in controlled assessment situations and their actual performance in professional practice[4] and degrees of correlation between the two have been shown to be extremely variable.[5] Whilst competence indicates what people can do in a contextual vacuum, under perfect conditions, performance indicates how people behave, in real life, on a day-to-day basis.[5] And it is performance, how well the doctor carries out their work 'in the wild', that we are ultimately interested in.

In recognition of the complexity of professional practice there is a need to consider assessment as a programme of activity requiring the assimilation of quantitative and qualitative information from different sources.[6] Some of this may come from standardised one-off testing '*in vitro*' but assessing doctors in their actual working environment offers enormous opportunities to gather data about an individual across multiple contexts and at different points in time. In this way, a 'rich picture' of that doctor may be constructed, reflecting not just what they can do in a controlled examination situation, but what they actually do, at work, with real patients.

Workplace based assessment

Workplace based assessment has been defined as the assessment of working practices based on what doctors actually do in the workplace, and predominantly carried out in the workplace itself.[7] This might include direct observations of performance as well as assessments specifically undertaken in the working environment. Exactly what form workplace based assessment takes though is often contingent on its purpose. All assessment can be positioned along a continuum between assessment for learning and assessment for the purposes of accountability, with assessment for certification somewhere up towards the right-hand end (*see* Figure 6.2).[8] North American approaches to 'work-based' assessment have tended to emphasis hard data, collected for accountability purposes and related to patient outcomes, process and volume.[9,10]

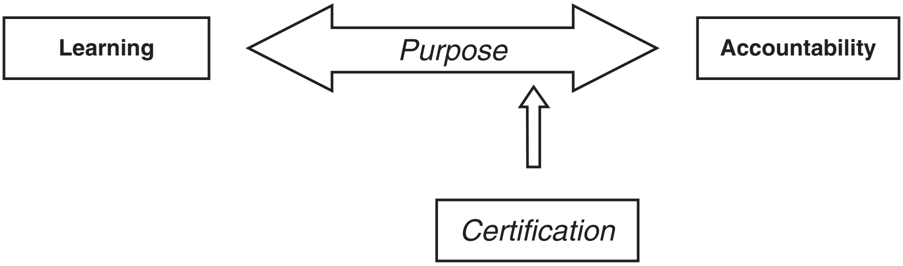


Figure 6.2: The purpose of assessment.

In the remainder of this chapter we shall discuss workplace based assessment in the context of medical education and training and in training programmes, where such patient-orientated data is difficult to attribute to an individual student. Furthermore, in the design of workplace based assessments for training purposes, educational impact becomes an increasingly important component of test usefulness and, as such, in-training assessments need to be more than mere audits of clinical activity.

Why workplace-based assessment?

It is worth rehearsing the main arguments for including a workplace based assessment within a programme of educational assessment. The first of these concerns the re-coupling of teaching learning and assessment. Assessment should be an integral part of educational planning, not a 'bolt on' extra at the end. Doctors should know what is expected of them and have an opportunity to demonstrate attainment over time and in a variety of contexts.

The second relates to the notion that assessment is more valid the closer it gets to the activity one wishes to assess.[11] If you want to know how a doctor consults, watch him do it. In the lexicon of test design, this is known as authenticity. Authenticity is particularly important when dealing with the assessment of medical expertise, as expertise appears to be domain specific and contextual.[12] As assessment is a potent driver for learning,[13] it is imperative that assessment focuses on what is considered important, rather than what appears to be easiest to assess.

The final argument is that some competency areas e.g. professional development, probity and team-working simply cannot be assessed effectively in any other way as they are impossible to disentangle from system (e.g. practice facilities) or personal influences (e.g. health). Assessment of performance in the workplace provides us with the only route into many aspects of professionalism.

Medicine has had a long preoccupation with objective standardised testing believing it better for patient safety. This view is changing as illustrated by the wholesale introduction of workplace based assessment in the English Foundation Programme. The Postgraduate Medical Education and Training Body which has now assumed responsibility for overseeing specialist training for all doctors, recognises the importance of reassuring the public about the safety and competence of its doctors, but also now recommends that this is based, at least in part, on what those doctors actually do within the workplace itself.[7]

Issues to consider in the design of a programme of workplace based assessment

Any assessment system for postgraduate training must now meet a number of principles laid down by the Postgraduate Medical Education and Training Board.[14] In addition, it is has been argued[15] that a programme of workplace based assessment for general practice training should be underpinned by a number of fundamental principles, namely that it should be:

- competency based
- developmental
- based on the collection of evidence using an appropriate variety of methods
- triangulated
- quality assured.

Clearly, the design of any assessment system will also need to take into account its utility. The utility, or usefulness, of an assessment has been defined as a product of its reliability, validity, feasibility, acceptability and educational impact.[13]

Competency-based

There is growing criticism of competency based education and assessment, largely because of the notion of competencies being overly simplistic, atomistic and reductionist.[16,17] There is also widespread confusion of the terms 'competence' and 'competency' which are often considered (erroneously) to be interchangeable. Despite these criticisms, it is possible for a competency-based approach to education and training to be made to work provided that 'competencies' are defined holistically as general attributes within a context rather than as discrete bite-sized pieces of behaviour.

Competence then becomes 'a complex structuring of attributes needed for intelligent performance in specific situations'.[12] It is important to note that the concept of 'intelligent performance' comes from an integrated approach to constructing competencies and a move towards the more holistic construct of competence, based on outcomes defined by an overarching curriculum.

To illustrate these points, a list of holistic competency areas, developed for the workplace assessment of trainee general practitioners in the new membership examination of the Royal College of General Practitioners (nMRCGP), are shown in Box 6.1. In each competency area, the scope of the competency is defined in a succinct statement which is then explicated through a series of graded word pictures. Success in this particular assessment requires attainment of a specific standard (that deemed to reflect competence) in each of the 12 areas.

Box 6.1: Competency areas within the draft nMRCGP Enhanced Trainer's Report

1 *Communication and consultation skills* – about communication with patients and the use of recognised consultation techniques.
2 *Practising holistically* – about the ability of the doctor to operate in physical, psychological, socioeconomic and cultural dimensions.

3 *Data gathering and interpretation* – about the gathering and use of data for clinical judgement, the choice of physical examination and investigations, and their interpretation.
4 *Making a diagnosis/making decisions* – about a conscious, structured approach to decision-making.
5 *Clinical management* – about the recognition and management of common medical conditions in primary care.
6 *Managing medical complexity and promoting health* – about aspects of care beyond managing straightforward problems, including the management of co-morbidity, uncertainty, risk and the approach to health rather than just illness.
7 *Primary care administration and IMT* – about the appropriate use of primary care administration systems, effective record-keeping and information technology for the benefit of patient care.
8 *Working with colleagues and in teams* – about working effectively with other professionals to ensure patient care, including the sharing of information with colleagues.
9 *Community orientation* – about the management of the health and social care of the practice population and local community.
10 *Maintaining performance, learning and teaching* – about maintaining the performance and effective continuing professional development of oneself and others.
11 *Maintaining an ethical approach to practice* – about practising ethically with integrity and a respect for diversity.
12 *Fitness to practise* – about the doctor's awareness of when his/her own performance, conduct or health, or that of others, might put patients at risk and the action taken to protect patients.

Developmental

Workplace based assessment offers the opportunity to link training, learning and assessment more effectively and the potentially developmental nature of this form of assessment is a key feature. Developmental assessment is defined as the 'process of monitoring student's progress through an area of learning so that decisions can be made about the best ways to facilitate future learning'.[18]

Educational impact is enhanced then when competencies are both made explicit and satisfactory progression is defined within them. Eraut brings these two strands, the descriptive and the developmental, together in arguing that 'a professional person's competence has at least two dimensions, scope and quality'.[19] Scope is defined as being what a person is competent in; that is the range of roles, task and situations. The quality dimension concerns judgements about the quality of that work on a continuum from novice to expert. Glaser outlines the characteristics that differentiate the performance of experts from novices, 'as proficiency develops, knowledge becomes increasingly integrated, new forms of cognitive skills emerge, access to knowledge is swift, and the efficiency of the performance is heightened'.[20] The expert's knowledge base becomes increasingly 'coherent, principled, useful and goal orientated'. Developmental progressions in the literature, such as that described by Dreyfus and Dreyfus[21] may be helpful in constructing developmental continua. Such continua have the advantage of explicitly illustrating the direction of travel for students rather than merely pointing out the level below which they should not fall.

Table 6.1 gives an example of progression statements in relation to the competency area of 'Communication and consultation skills' from the proposed Training Record component of workplace based assessment module for the nMRCGP.[22]

Table 6.1: Example of a competency progression statement from the nMRCGP Enhanced Training Record

1	*Communication and consultation skills*		
This competency is about communication with patients, and the use of recognised consultation techniques.			
Insufficient Evidence	**Needs Further Development**	**Competent**	**Excellent**
From the available evidence, the doctor's performance cannot be placed on a higher of this developmental scale.	Develops a working relationship with the patient, but one in which the problem rather than the person is the focus. Produces management plans that that are appropriate to the patient's problem. Provides explanations that are relevant and understandable to the patient, using appropriate language. Achieves the tasks of the consultation but uses a rigid approach.	Explores the patient's agenda, health beliefs and preferences. Elicits psychological and social information to place the patient's problem in context. Works in partnership with the patient,negotiating a mutually acceptable plan that respects the patient's agenda and preference for involvement. Explores the patient's understanding of what has taken place. Flexibly and efficiently achieves consultation tasks, responding to the consultation preferences of the patient.	Incorporates the patient's perspective and context when negotiating the management plan. Whenever possible, adopts plans that respect the patient's autonomy. Uses a variety of communication techniques and materials to adapt explanations to the needs of the patient. Appropriately uses advanced consultation skills such as confrontation or catharsis to achieve better patient outcomes.

Implicit in this is the notion that the assessment must provide detailed formative and developmental feedback to the learner. This raises the tension of potentially mixing formative and summative assessment but it is possible to address this through the careful design of the assessment system. 'Separating the interpretation of evidence from its elicitation, and the consequent actions from the interpretations' is a way around the problem.[23] Such an approach supports the process of ongoing evidence collection throughout the training period, but with regular, well circumscribed, staging reviews at which the developmental framework is reviewed and the learner's progress through it, judged.

Based on the collection of evidence using an appropriate variety of methods

As discussed above, collecting 'sufficient' evidence is essential in making a judgement about the competence of the learner. Workplace based assessment in common with all other medical assessments suffers with the problem of content specificity, as the assessment of competence appears to be domain specific and contextual. Therefore, a large number of samples of performance is required to achieve adequate reliability.[24]

In the assessment of 'work' in contrast to traditional assessments there is no single 'controlled' method that can be developed. It is more helpful to think in terms of identifying the basis of judgements, deciding how the information, or evidence, will be gathered and threats to validity and reliability avoided.[10]

The importance of gathering evidence using a variety of methods gives rise to the notion of a 'tool-box' of approved methods. In considering the individual tools it is worth recognising that, even unstandardised, they can be made sufficiently reliable provided enough sampling occurs, and the tools are used sensibly and expertly.[6] However, it is important to remember that the tools form part of the overall assessment programme. Attention should focus on the reliability of the entire programme of assessment, not just the individual tools themselves. A selection of currently available workplace based assessment tools appropriate for use in general practice is shown in Box 6.2.

Box 6.2: Workplace based assessment: examples of tools

Tool	*Comment*
Mini-CEX	In the mini-CEX,[25] an observer assesses one candidate completing a focused interview or examination, the assessment is recorded in a standard format, and takes 15–20 minutes to complete. Mini-CEX is based on assessment of multiple encounters within a hospital setting. Its applicability to a general practice setting is not known.
Longitudinal evaluation of performance	Similar to the mini-CEX, the longitudinal evaluation of performance,[26] piloted as a formative tool for the assessment of dental trainees, uses direct observation of the trainee in clinical practice and relies on judgements of an evaluator across eight broad categories. For each category, a

	judgement is made on a nine-point scale, which allows for developmental progression.
Video assessment	Various video assessment methods are already in use in general practice, including the existing summative assessment methodology[27] and the video module of the membership examination of the RCGP.[28] Other models have been developed elsewhere in the world, most notably in Holland.[29]
Direct observation of procedural skills (DOPS)	DOPS[30] requires the educational supervisor to directly observe the student undertaking a certain procedure and to make judgements about specific components of the procedure. DOPS has now been incorporated into the workplace based assessment of foundation medical students and is the subject of evaluation.
Direct observation of consultation skills	The Leicester Assessment Package, originally described by Fraser,[31] has been used to provide systematic formative feedback to postgraduate and undergraduate students after the direct observation of six consecutive and largely unselected patients.
Case-based discussion	Case-based discussion,[32] or chart-stimulated recall,[33] involves a structured oral interview involving a trained assessor reviewing selected cases provided by an assessee. The presentation of each patient case may take 5–10 minutes, and typically a case-based discussion may last 30–60 minutes.
Multi-source feedable	Multi-source feedback has been demonstrated to be a reliable and valid tool for the assessment of professional behaviours.[34]
Patient satisfaction surveys	The Physician achievement review[35] is rolled out to all Alberta physicians every five years. Raters include systematically selected patients, a self-questionnaire and questionnaires distributed by the participating physician to medical colleagues and non-physician co-workers. Summated responses are sent to the participating physician in text and graphic format. The consultation and relational empathy measure (CARE)[36] has been successfully evaluated as a means of measuring patients' perceptions of relational empathy in the consultation.
Written assignments	Written assignments in a variety of formats are already currently used as part of the assessment process for Summative Assessment of GP Registrars in the UK.[37,38] Significant event analysis reporting has also been piloted for use with established practitioners.[39]

The framework currently proposed for the workplace-based assessment module of nMRCGP will consist of an evidenced training record that will apply across the entire training programme informed by some externally assessed work-based tools such as a patient satisfaction questionnaire and a web-based multi-source feedback tool. These will provide core information which will provide external validity as well as feeding into the record itself. Evidence of the trainee's progress will be also be obtained using locally administrated tools such as: case based discussion, mini-CEX, a consultation observation tool and DOPS. Because of the complexity and nature of the competency areas under test, there will also remain the flexibility to record naturally-occurring events in the workplace.

Workplace based assessment becomes more feasible if the process of collecting evidence is learner-led with the educational supervisor responsible for overall coordination. The educational supervisor is also be involved in making summative judgements, a process legitimised where 'the synthesis of the evidence and the process of its judging is made explicit'.[16]

Triangulated

The confidence in the veracity and reproducibility of judgements in a workplace based assessment can be improved through triangulation both within the workplace based assessment as well as triangulation with other assessments. An overarching assessment strategy will be essential in which workplace based assessment is supported by rigorous tests, e.g. those of 'knowledge' and 'skills for clinical method'. It is recommended that there should be an overarching assessment strategy for the whole training period, and that this is mapped to a blueprint. Workplace based assessment forms part of this including other appropriate methods.

Quality assured

In viewing the quality assurance of a programme of workplace based assessment it is helpful to review the utility equation described earlier. Workplace based assessment has strengths in the areas of validity (by virtue of its authenticity), educational impact and acceptability (because it reconnects teaching and learning) and feasibility (through local assessment).

There are, however, problems with demonstrating its reliability using traditional psychometric approaches. As Southgate points out, 'establishing the reliability of assessments of performance in the workplace is difficult because they rely on expert judgements of unstandardised material'.[32] In workplace based assessment like any other form of assessment there are several potential threats to reliability:[40]

- inter-observer variation (the tendency for one observer to mark consistently higher or lower than another)
- intra-observer variation (the variation in an observer's performance for no apparent reason – the 'good/bad day' phenomenon)
- case specificity (the variation in the candidate's performance from one challenge to another, even when they seem to test the same attribute).

Despite these challenges reliability in workplace based assessments can be maximised through a series of measures outlined by Baker, O'Neil and Linn.[41]

- *Specification* of standards, criteria, scoring guides.
- *Calibration* of assessors and moderators.
- *Moderation* of results, particularly those on the borderline.
- *Training* of assessors with retraining where necessary.
- *Verification and audit* through quality assurance measures and collection of reliability data.

It is clear therefore that the implementation of workplace based assessment will require a complementary training programme, arrangements for calibration, a procedure for the moderation of results and a raft of quality control and reliability checks. But it will be worth the effort. The more that teachers can be engaged in assessment, in selecting methodologies, generating standards, discussing criteria etc., the more they will be empowered in the educative process.

Conclusion

Assessments conducted in the workplace are of high validity and serve to reconnect teaching and assessment. A competency-based model accords with the overall contemporary emphasis of medical assessment but caution is advised lest defined competencies become over-atomised. In order to enhance educational impact, the use of holistic competencies within a developmental continuum is recommended. Such a continuum has the advantage of explicitly illustrating the direction of travel for trainees rather than merely pointing out the level below which they should not fall.

To further strengthen the link between teaching and assessment, and to deal with the practical expediencies of wide scale implementation, a workplace based assessment should be locally assessed and based on the collection of evidence. The determination of 'sufficient' evidence should be pre-defined and triangulation built in as an essential feature in order to enhance the reliability of judgements made.

Clearly there is much to be done in the development of a workplace based assessment to create a vehicle for assessment that is robust, fair, comparable and consistent and further research in this area is urgently required. To get it right will not only reduce the current assessment burden on students, but also harness the involvement of medical teachers. In doing so, we have the opportunity to create a powerful tool for professional development.

Acknowledgement

This chapter has been adapted from an original article by the authors that first appeared in the *British Journal of General Practice* in July 2005.

References

1. Kaufman D, Mann K, Jennett P. Occasional publication: *Teaching and Learning in Medical Education: how theory can inform practice*. Edinburgh: The Association for the Study of Medical Education; 2000.
2. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med.* 1990; **65**(9): 63–7.
3. Schuwirth L, Vleuten C van der. Changing education, changing assessment, changing research. *Med Educ.* 2004; **38**: 805–12.

4. Rethans J, Norcini J, Baron-Maldonado M, Blackmore D *et al.* The relationship between competence and performance: implications for assessing practice performance. *Med Educ.* 2002; **36**: 901–9.
5. Schuwirth L, Southgate L, Page G, Paget N *et al.* When enough is enough: a conceptual basis for fair and defensible practice performance assessment. *Med Educ.* 2002; **36**: 925–30.
6. Vleuten C van der, Schuwirth L. Assessing professional competence: from methods to programmes. *Med Educ.* 2005; **39**(3): 309.
7. Postgraduate Medical Education and Training Board Workplace Based Assessment Subcommittee. *Workplace Based Assessment.* London: Postgraduate Medical Education and Training Board; 2005. www.pmetb.org.uk/index.php?id=664
8. Swanwick T. Work based assessment in general practice: three dimensions and three challenges. *Work Based Learn Prim Care.* 2003; **1**(1): 99–108.
9. Norcini J. Recertification in the United States. *BMJ.* 1999; **319**: 1183–5.
10. Norcini J. Current perspectives in assessment: the assessment of performance at work. *Med Educ.* 2004; **39**: 880–9.
11. Wiggins G. A true test: toward more authentic and equitable assessment. *Phi Delta Kappan*. 1989; **70**(9): 703–13.
12. Gonczi A. Competency based assessment in the professions in Australia. *Assess Educ.* 1994; **1**(1): 27–44.
13. Vleuten C van der. The assessment of professional competence: developments, research and practical implications. *Adv Hlth Sci Educ.* 1996; **1**: 41–67.
14. Postgraduate Medical Education and Training Board. *Principles and Standards for Assessment.* London: Postgraduate Medical Education and Training Board; 2003.
15. Swanwick T, Chana N. Workplace assessment for licensing in general practice. *BJGP.* 2005; **55**: 461–7.
16. Leung W-C, Diwakar V. Competency based medical training: review. Commentary: The baby is thrown out with the bathwater. *BMJ*. 2002; **325**(7366): 693–6.
17. Talbot M. Monkey see, monkey do: a critique of the competency model in graduate medical education. *Med Educ.* 2004; **38**(6): 587–92.
18. Masters G. Developmental assessment: what, why, how? In: *Conference on Advances of Student Learning*. China: Chinese University of Hong Kong; 1997.
19. Eraut M. *Developing Professional Knowledge and Competence*. London: Falmer Press; 1994.
20. Glaser R. Toward new models for assessment. *Int J Educ Res*. 1990; **14**(5): 477.
21. Dreyfus H, Dreyfus S. *Mind Over Machine: the power of human intuition and expertise in the era of the computer*. Oxford: Basil Blackwell; 1986.
22. Royal College of General Practitioners. *nMRCGP Enhanced Trainer's Report* (draft). London: RCGP; 2005.
23. William D, Black P. Meanings and consequences: a basis for distinguishing formative and summative functions of assessment. *Brit Educ Res J.* 1996; **22**(5): 537–8.
24. Hays R, Davies H, Beard J, Caldon L *et al.* Selecting performance assessment methods for experienced physicians. *Med Educ.* 2002; **36**: 910–17.
25. Norcini J, Blank L, Duffy F, Fortna G. The mini-CEX: A method for assessing clinical skills. *Ann Intern Med.* 2003; **138**: 476–81.
26. Prescott L, Norcini J, McKinlay P, Rennie J. Facing the challenges of competency-based assessment of postgraduate dental training: Longitudinal Evaluation of Performance (LEP). *Med Educ.* 2002; **36**: 92–7.
27. Campbell L, Howie J, Murray T. Use of videotaped consultations in summative assessment of trainees in general practice. *BJGP*. 1995; **45**(392): 137–41.
28. Royal College of General Practitioners. *Video Assessment of Consulting Skills in 2004: Workbook and Instructions.* London: Royal College of General Practitioners; 2004.
29. Ram P, Grol J, Rethans B, Schouten C *et al.* Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of reliability and feasibility. *Med Educ.* 1999; **33**(6): 447–54.

30. Department of Health. DOPS (Direct Observation of Procedural Skills). In: *Modernising Medical Careers*. London: DoH; 2005. www.mmc.nhs.uk/pages/assessment/dops (accessed 21 March 2007).
31. Fraser R, McKinley R, Mulholland H. Consultation competence in general practice: establishing the face validity of prioritised criteria in the Leicester assessment package. *BJGP*. 1994; **44**: 109–13.
32. Southgate L, Cox J, David T, Hatch D *et al.* The General Medical Council's Performance Procedures: peer review of performance in the workplace. *Med Educ*. 2001; **35**: 9–19.
33. Cunnington J, Hanna E, Turnbull J, Kaigas T *et al.* Defensible assessment of the competency of the practicing physician. *Academic Med.* 1997; **72**: 9–12.
34. Norcini J. Peer assessment of competence. *Med Educ.* 2003; **37**: 539–53.
35. Violato C, Hall W. Alberta Physician Achievement Review. *CMAJ.* 2000; **162**(13): 1803.
36. Mercer S, Maxwell M, Heaney D, Watt G. The consultation and relational empathy (CARE) measure: development and preliminary validation and reliability of an empathy-based consultation process measure. *Fam Pract*. 2004; **21**(6): 699–705.
37. Evans A, Singleton C, Nolan P, Hall W. Summative assessment of general practice registrars' projects. *Edu Gen Pract*. 1996; **7**: 229–36.
38. Lough J, McKay J, Murray T. Audit and summative assessment: 2 years' pilot experience. *Med Edu*. 1995; **29**: 101–3.
39. Bowie P, McKay J. An educational approach to significant event analysis and risk management in primary care. In: Swanwick T, Jackson N (eds). *The General Practice Journey: the future of educational management in primary care*. Oxford: Radcliffe Medical Press; 2003.
40. Crossley J, Davies H, Humphries G, Jolly B. Generalisability: a key to unlock professional assessment. *Med Educ.* 2002; **36**: 972–8.
41. Baker E, O'Neil H, Linn R. Policy and validity prospects for performance-based assessment. *Am Psychol.* 1992; **48**(12): 1210–18.

Chapter 7

Preparing teachers for work-based teaching and assessing

Robert Clarke

Introduction

This chapter describes the process of setting up a programme to prepare educational supervisors for the Foundation Programme (FP) in general practice, with a particular focus on its work-based assessment framework, the theoretical basis and evidence for which have been described in Chapters 3 and 6. A learner-centred approach will be proposed both to the formative assessment of foundation programme doctors and to the preparation of teachers who are themselves 'the learners' during the preparatory period. Methods of teaching the assessment framework and undertaking peer-calibration exercises will be discussed, which build on the themes of an introductory course for teachers in primary care.

Key themes for potential educational supervisors are that:

- learning needs analysis – the first stage in formative assessment
- Socratic methods are suited to a learner-centred approach
- an important role for educational supervisors is to help the learner make their own judgements
- asking awareness-raising questions and giving descriptive feedback helps this process
- a key task of the educational supervisor is to help in the interpretation of the results of work-based assessment from different people, using different methods at different times (triangulation)
- teaching and assessing are an important part of continuing professional development as a practitioner.

A framework for the elements of educational planning, based on the acronym 'AILMENTS' (*see* Box 7.1), will be applied to a workshop on the FP assessments.

Box 7.1: 'AILMENTS': preparing for teaching

Preparing for teaching
('lesson planning')

A	Aims
I	Intended learning outcomes
L	Learning needs analysis
M	Methods
E	Evaluation

N	Next steps
T	Testing (assessment)
S	Summary

'AILMENTS'

This acronym represents a process of preparation that is significantly different from other approaches in that it emphasises equally the importance of considering the teacher's aims and objectives for the session and the way in which this will be linked with the experience and needs of the learners. It is particularly this process of connecting with the learner by undertaking a needs analysis that seems so often to be missed out.[1] This part of the process is essential, can be planned, but may often be perceived as dangerous as it may force the teacher to adapt or even abandon the planned teaching programme, sometimes at very short notice, in order to meet the needs of learners. However, good teachers usually espouse such flexibility and, by emphasising learning needs analysis at an early stage of preparation, thought will be given to possible alternative directions that the teaching may need to take in order to be effective. Some of the most successful teaching sessions have resulted from the teacher having very clear aims, coming with a plan for what might happen and then abandoning that plan in the light of a needs analysis. The general term 'preparing for teaching' is preferred to 'lesson planning' as the latter has pedagogical overtones with an instructional rather than participatory approach to teaching.

Context

Modernising Medical Careers (MMC) sets out a framework for the educational support and assessment of junior doctors in the first two years post-qualification.[2] The programme is supported by work-based assessments, which are mainly intended to be supportive, but which contain an early warning system to highlight doctors who are either not engaging with the assessment programme or who are performing below the standard expected.[3] The assessments (*see* Table 7.1 and Chapter 3) sample from a range of clinical, interpersonal and practical skills and professional behaviours.

Table 7.1: The four assessment methods

Type of assessment	*Acronym*	*Tested by*	*Main focus*	*Also tests*
Clinical evaluation exercise	mini-CEX	Sitting in	Clinical skills	Professionalism Communication
Peer assessment tool	mini-PAT	Assessment in all aspects of work	Professionalism	Clinical care Communication
Case-based discussion	CbD	Case review	Clinical reasoning	Professionalism
Direct observation of procedural skills	DOPS	Observing practical procedures	Practical skills	Communication Professionalism

The onus is on the individual doctor to collect the evidence for these at a time when he or she feels ready. Each of the four assessment methods have both formative and summative elements in that feedback should always be given to help the learner's development (formative) and the assessments need to be passed (summative) in order to demonstrate achievement of the FP competencies, even if this requires several attempts.

Preparing for teaching and assessing

Training in formative assessment is an essential component of any preparatory programme for teachers since both the assessment of educational need and the giving of feedback require specific skills.[4] The importance of appropriate training of supervisors has been particularly emphasised with respect to the assessments used in the Foundation Programme.[5,6,7] In London, a deanery team was established to help prepare new teachers to take on the large numbers of general practice placements required by the Foundation Programme. Over the course of three years, new cohorts of educational supervisors participated in a course which started with a two-day introduction to educational theory and practice, followed by locality-based, facilitated learning sets which met monthly over the subsequent six months and culminated in a final day where the learning sets came together for review and further development.

Learning sets

One of the challenges for the educational supervisor programme was that F2 placements in general practice had not yet started, so that some new teachers had limited teaching opportunities. The learning sets were very highly evaluated (*see* Box 7.2) and so the format of a two-day introduction followed by learning sets was retained, but future learning set support would be provided after the educational supervisor had started with the first F2 doctor.

Box 7.2: Evaluation of learning sets: examples of participants' views

'I learned a lot of facilitation skills and consolidated many of the ideas from the introductory course and found out how to apply them to my teaching.'
'Practical aspects of the course were particularly useful including analysis of videos and tutorials.'
'The group worked well, was valuable and enlarged my world.'
'In a partnership you can be stuck within 4 walls and it's good to see what other doctors do.'
'Learning needs analysis was a particularly useful skill and I also learned the importance of not having to be an expert on everything.'
'I learned about feedback by asking questions to take people on a journey.'
'Peer support with colleagues was very valuable. For example, a colleague was using the Internet during consultations and I tried this with a recent patient who had an alcoholic father. The patient came back to me having looked at the website and saying "This is me, this is all about me".'
'It was brilliant. I learned a lot about different styles of teaching.'

'It broke my isolation. I learned the importance of checking learning needs early on in teaching.'
'We found it particularly helpful to use "time out" techniques when teaching on the consultation and experimented with different methods of role play.'
'I learned about different styles of listening and about using concepts based on the humanities to help with consultation analysis.'
'We were encouraged to keep a reflective teaching log and I have found this particularly helpful.'

The introductory programme

The educational supervisor programme was intended as an entry-level programme, an introduction to teaching in primary care, with an explicitly skills-based focus and no end-point assessment or academic validation. The programme was later developed further as a collaborative venture between the five undergraduate departments of general practice in London and the London Deanery, enabling those who attend to become either undergraduate teachers or educational supervisors of F2 doctors. The majority of participants expressed a desire to teach both medical students and F2 doctors.

Box 7.3: Introduction to teaching in primary care: outline programme

Day one
Session 1
Introduction to needs assessment
Principles of assessment and giving feedback
Session 2
Teaching methods and learning styles
Preparing for teaching
Day two
Session 1
Teaching practice on clinical topics
Session 2
Teaching practice on the consultation

Assessment is a major theme of the introductory programme with clear links made between assessment of learning need and the tailoring of teaching methods to those needs as well as to the learning style of the individual. Many practitioners come to the course with implicit assumptions about how the course will be taught, based on traditional didactic teaching and about their future roles as passers-on of information.[8] Alternative roles are explored and specific training in Socratic (questioning) techniques provided.[9,10] These are developed through teaching practice based on topics and on the consultation where learner-centred approaches are emphasised.[11] The use of awareness raising questions[10] is rehearsed in the safety of a small group and the giving of specific descriptive feedback on consultations is encouraged, allowing the learner to make their own judgements.[11]

By the end of the two-day course, practitioners have broadened their view of their educational role and have had several opportunities to practise learning needs assessment and giving feedback within teaching sessions. Those attending are given information about the Postgraduate Certificate for Teachers in Primary Care, the one-year Masters-level programme required of potential GP trainers, and several have already been inspired to undertake this further training. The final assessment and selection of educational supervisors is separate from the preparatory course and is based on a peer-review process similar to trainer selection, with a practice visit by the patch Associate Director, including a review of the educational supervisor's teaching.

Foundation Programme assessments

In addition to this generic introduction to teaching, potential supervisors are required to attend a one-day course introducing the FP and its assessments. This workshop is informed by the preparatory work of the educational supervisor programme, which establishes the skills required for learner-centred teaching, learning needs analysis, effective teaching methods, giving feedback, assessment and evaluation. The workshop is followed by application and selection to become an educational supervisor and ongoing support in locality-based learning sets (*see* Box 7.4).

Box 7.4: Outline programme for F2 workshop

Session 1
Modernising Medical Careers: the purpose of the programme
Group work – what would the curriculum look like in my practice?
Principles of work-based assessment
Session 2
Participants rotate around four 45-minute stations trying out the four different assessment methods
Planning next steps

With this context in mind, the planning for and implementation of a series of workshops will be explored, with the aim of sharing the lessons learned in the process. The outline programme of the workshop is shown in Figure 7.1.

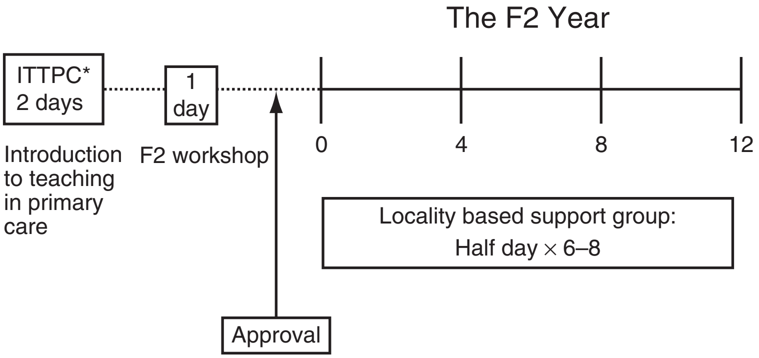

The F2 Year
ITTPC* 2 days
1 day
0
4
8
12
Introduction to teaching in primary care
F2 workshop
Locality based support group: Half day × 6–8
Approval

Figure 7.1: Preparing and supporting educational supervisors.

Aims

The first point in planning any educational activity, whether a brief 'lesson', or a complete curriculum, should be a consideration of its aims.[12,13] The Foundation Programme is a new development in medical education and will be formally introduced into general practice for the first time in August 2006.[14] This led the deanery to recruit new teachers and to consider how best to introduce both established and new teachers to the curriculum and assessment framework of the Foundation Programme, and to encourage them to become educational supervisors.

There was also a need to check educational supervisors' views on what ongoing support they would require, particularly given the results of evaluations which had shown how well received and effective the learning sets had been. The planning team also wanted to explore ways of enhancing communications between primary and secondary care, and saw the Foundation Programme as an ideal opportunity for such collaboration.

Most importantly, the assessment framework of the Foundation Programme was recognised as a new approach to work-based assessment, predicated on making multiple assessments by different individuals, using the idea of triangulation, derived from qualitative research[15] and based on real performance,[16] which trainees find highly acceptable.[17,18] Educational supervisors needed both an introduction to the assessment methods and an opportunity for calibration in defining standards. The most appropriate format for such learning would be in facilitated small groups using a peer-review process.

Our core aims for the workshops were to:

- give participants a sense of ownership of the curriculum
- provide an opportunity to understand the administration of placements
- provide an introduction to the assessment methods
- provide an opportunity to calibrate assessments against peers
- consult educational supervisors about ongoing support.

Intended learning outcomes (ILO)

It is helpful to consider ILO so long as one recognises that lots of unintended learning may occur and that this may be just as or more important than the intended learning.[19] For this reason, it is essential to be flexible about the delivery of an educational programme as new needs, previously unrecognised by those involved in planning, may emerge. ILOs are a useful planning device, which may well have to be abandoned or refined as the teaching unfolds, but which should not be used as an instrument of control by the teacher. There should be considerable flexibility in the degree to which ILOs are specific and measurable, as experience suggests that the more specific one tries to be, the less useful is the ILO as a statement of direction of travel. The intended outcomes for each of the above aims were for participants to:

- develop a plan for how the Foundation Programme placement would be implemented in the educational supervisor's practice
- have a good understanding of the practicalities of GP placements and to know where to direct further enquiries
- understand the rationale for the assessments and to have participated in mock assessments, including giving feedback to learners for each of the four assessments

- have understood the concept of inter-observer reliability through comparing marking with peers and to have gained an understanding of what is an acceptable level of performance for F2
- have considered their own needs for support as educators.

Learning needs analysis

The first session of the programme (*see* Figure 7.1) was designed to establish the participants' needs in relation to the Foundation Programme: in other words to find out what they already knew and what they need to know. Pre-course reading included an outline of the Foundation Programme and a summary of the assessment methods. At the start of the one-day workshop, a resource folder was issued containing further information and detailed notes on the assessment methods from the MMC website.

After a brief introduction to the day and to the opportunities presented by the Foundation Programme, the first small group exercise was to consider how the programme would work in participants' practices. This exercise was successful in engaging everybody and stimulated questions and discussion. The issues raised in each small group were briefly shared in the large group. This helped achieve the first two intended learning outcomes (developing a plan for the practice and knowing where to go for further information) as well as establishing the context and relevance of the next part of the day. Many practical questions were dealt with in this session (*see* Box 7.5), which helped to prepare the ground for considering the assessments.

Box 7.5: Themes from group work on 'How will the curriculum work in my practice?'

Making contact with the F2 doctor before the attachment
Educational agreement
Planning the induction programme, including IT training
Meeting the primary healthcare team
Planning learner centred, realistic goals
Philosophy/ethos of programme:
– listening, communication, professional skills
– changing role of primary care
– communication between primary and secondary care
– recommended introductory reading
– assessment of learning style
– what to call the F2 doc?
Pastoral care – roles of supervisor and practice manager
Developmental assessment, formative feedback
Problems and role of Foundation Programme Director
GPRs often have a three-month 'dip' with low confidence: dealing with this in an F2 doctor near the end of the attachment
Finding rooms, juggling spaces, sitting-in
Number of clinical sessions, appointment durations and level of supervision
Indemnity

In terms of a satisfying a hierarchy of needs, participants' agendas were addressed and the small groups given a chance to start working before the main focus of the day: learning about the assessment programme.

Methods, part 1: rationale for the assessment programme

The teaching methods used involved a traditional but short didactic presentation, explaining the educational and philosophical basis of the assessment programme (*see* Chapters 3 and 6). Reference was made to the fact that, for many assessments, there is a trade-off between validity and reliability and this idea was illustrated by reference to Miller's pyramid.[16] The importance of assessing real performance in the workplace was emphasised, as was the formative aspect of each method. In addition, the idea of triangulation, from qualitative research, was found to be a helpful way of explaining the overlap in attributes measured by the different techniques. This new view of validity – getting at the truth from multiple different viewpoints, each of which may have a lower reliability than a single, narrow measurement, for example a multiple-choice test of knowledge – seemed to be effective in explaining the rationale of the assessment programme. The other method that enabled participants to gain a quick understanding was making a link between the new framework and existing assessments, already undertaken by general practitioner trainers;[20,21] for example the similarities between case-based discussion (*see* Table 7.1) and problem-case analysis.

Methods, part 2: calibration through assessment stations

The second teaching method was based on small-group work. Each group rotated around four assessment stations, at each of which one of the assessment methods was explored, with an opportunity to try out the assessments 'live', including giving feedback to learners using role play, a component of assessment which is frequently overlooked.[22] This format had been used in the national programme introducing the Foundation Programme, where it was found to be effective.[3] In order to adapt to the needs of general practice, it was necessary to produce some new stimulus material. For example, for direct observation of procedural skills (DOPS), two short videos were produced demonstrating levels of performance which were clearly above and below the standard expected on completion of F2. This was a useful method of calibration because it facilitated a discussion about what was and was not a satisfactory level. For case-based discussion (CbD), participants had been asked to bring a case of their own.

> 'Please bring with you a copy of the actual clinical record (Lloyd George or computer printout) of one of your own recent patient encounters. It may be helpful to choose a patient presenting with an acute illness as this is a particular focus of the foundation programme. The case record will be used as the basis for discussion in one of the workshops, exploring clinical reasoning, differential diagnosis and decision making about whether or how to investigate, involve others practitioners, refer and treat.'

The cases brought by participants were used for discussion and marked according to the F2 CbD assessment tool. With each method, an informal learning needs assessment was performed: the small group was asked what it already knew about the assessment method under discussion, prior to reviewing the documentation provided by MMC and trying out the method using the stimulus materials already discussed.

It was noted that the small groups became quickly engaged in these practical exercises. At first, there was a fairly wide range of marks and opinions about standards within each group, but as they started rotating around the assessment methods, a process of peer-calibration occurred, with increasing consistency of results (norming). This was surprisingly rapid, applied to all four groups and was independent of the order of the assessments. The concept of inter-observer reliability was introduced during one of the stations, which participants found helpful. It was also useful to ask in what ways performance would be expected to be different from that of a medical student and that of a GP registrar. In these ways, the implicit published standards ('meets expectation' etc.) were operationalised by peer review.

This was intended to be the start of an ongoing process of training in the assessment methods, which will be continued in the locality-based learning sets.

Methods, part 3: learner-centred formative assessment

Each workshop presented stimulus material for one of the assessment methods and once calibration was achieved through discussion of standards, participants were invited to role play the giving and receiving of feedback. Techniques were suggested for encouraging the 'teacher' to elicit a self-assessment by the 'learner', based on the preparatory course described above. Methods of giving specific, non-judgemental feedback on performance were also explored, inviting the learner to evaluate the effectiveness of their clinical behaviour against their stated goals.[23] Parallels were drawn between patient-centredness in the consultation and learner-centredness in formative assessment.[24]

Evaluation

The day ended with a question and answer session in which any remaining issues were discussed which had not been dealt with in the assessment stations. This enabled us to elicit what had been learnt as well as being a practical ending for those attending, who were able to focus on 'Where do we go from here?'. In addition, we used a simple written feedback form to evaluate the workshop, the results of which were combined with our own evaluations. A written plan of the day was recorded, together with these reflections, for modification in the planning of subsequent workshops.

Next steps

The views of workshop participants, both from the needs assessment and their written evaluations, were used to make a case for continuing the learning sets. The views expressed during the previous evaluation of the educational supervisor course were valuable and were shared with those attending the assessment workshop. Many participants felt that becoming involved in teaching and assessing had considerably helped their continuing professional development, with plenty of examples of positive and unintended learning outcomes.

Testing (assessment)

There was no formal assessment of participants' skills (either as teachers or as assessors) during the workshop. Such assessments are incorporated into the appointments procedure described above. Attendance at the one-day workshop was, however, a prerequisite for application to become an educational supervisor.

Summarising the workshops

The conclusion of the day involved a collective reflection on the process of getting closer agreement on grading as the afternoon progressed as well as emphasising the importance of continuing work in support groups. Practical issues and curriculum plans were summarised so that participants left with a sense of what had been achieved and what were the next steps.

Preparation of the F2 workshops was learner-centred in that it involved active planning of how the needs of participants would be elicited with respect to the Foundation Programme.

Conclusion

In conclusion, the workshops on assessment were learner-centred, consistent with the model of education explored in the introductory course for teachers in primary care. They were built upon Socratic methods of teaching and assessing which the potential educational supervisors had already experienced. Engagement of participants in discussion of the curriculum enabled their learning needs to be expressed and addressed at an early stage.

Small-group workshops were successful in actively involving those attending through the practical tasks of trying out the different assessment methods. They also provided an opportunity for rapid peer calibration with clarification of standards and for rehearsal of giving feedback to learners through role play.

Locality-based, facilitated learning sets have been highly evaluated as a method of providing support to new teachers on this introductory programme. Further work will be needed to ensure that such support continues to be effective once educational supervision of FP doctors has started. Preparation for such educational activity is likely to be successful if assessments of learning need are planned and implemented, as required by the 'AILMENTS' framework. Once established, it seems likely that some learning sets will move rapidly from facilitated to self-directed groups[25] and these are of particular relevance to teachers in primary care.[26] Participation in work-based teaching and assessing can contribute greatly to the continuing professional development of practitioners.

Acknowledgment

The author is grateful for the support of the Deanery team planning and implementing this programme, particularly Drs Anthea Lints, Tim Swanwick, Mandy Platts, Nav Chana and Budgie Savage and for the administrative support of this project by Sarah Rowe, Christine Madekwe and Nikki Whitelock.

References

1. Pitts J. Pathologies of one-to-one teaching. *Educ Gen Pract.* 1996; **7**: 118–22.
2. Department of Health. *Curriculum for the Foundation Years in Postgraduate Education and Training*. London: The Stationery Office; 2005.
3. Southgate L, MMC conference team. *MMC Foundation Programme Assessment: tools of the trade.* Conference Document. MMC; 2005.
4. Brookfield S. *Understanding and Facilitating Adult Learning: a comprehensive analysis of principles and effective practices.* Milton Keynes: Open University Press; 1986.
5. Norcini J. Work based assessment. *BMJ.* 2003; **326**: 753–5.
6. Norcini J, Blank L, Arnold G, Kimball H. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Int Med.* 1995; **123**: 795–9.
7. Holmboe E. Faculty and the observation of trainees' clinical skills: problems and opportunities. *Academic Med.* 2004; **79**: 16–22.
8. Claxton G. Implicit theories of learning. In: Claxton G, Atkinson T, Osborn M, Wallace M (eds). *Liberating the Learner*. London: Routledge; 1996. p. 46–56.
9. Royal College of General Practitioners. *The Future General Practitioner: learning and teaching.* London: British Medical Journal; 1972.
10. Neighbour R. *The Inner Apprentice*. Newbury: Petroc; 1992.
11. Silverman J, Kurtz S, Draper J. The Calgary-Cambridge approach to communication skills teaching: agenda-led outcome based analysis. *Educ Gen Pract.* 1996; **7**: 288–99.
12. Dent J, Harden R. *A Practical Guide for Medical Teachers.* Edinburgh: Churchill Livingstone; 2001.
13. Quinn F. *The Principles and Practice of Nurse Education (4e).* London: Chapman and Hall; 2000.
14. Department of Health. *Operational Framework for Foundation Training.* London: The Stationery Office; 2005.
15. Mays N, Pope C. *Qualitative Research in Health Care*. London: BMJ Publishing Group; 2001.
16. Miller G. The assessment of clinical skills, competence and performance. *Academic Med.* 1990; **65**: 63–7.
17. MMC, BMA. *The Rough Guide to the Foundation Programme.* London: The Stationery Office; 2005.
18. Alberto AL, Ruth H, Jorge T, Jorge P *et al.* A qualitative study of the impact on learning of the mini clinical evaluation exercise in postgraduate training. *Med Teacher*. 2005; **27**(1): 46–52.
19. Rowntree D. *Assessing Students.* London: Kogan Page; 1987.
20. Hays R. *Practice-based Teaching: a guide for general practitioners.* Melbourne, Australia: Eruditions Publishing; 1999.
21. Middleton P, Field S. *The GP Training Handbook.* Oxford: Radcliffe Publishing; 2000.
22. Holmboe E, Yepes M, Williams F, Huat S. Feedback and the mini clinical evaluation exercise. *J Gen Int Med.* 2004; **19**: 558–61.
23. Kurtz S, Silverman J, Draper J. *Teaching and Learning Communication Skills in Medicine.* Oxford: Radcliffe Publishing; 1998.
24. Neighbour R. *The Inner Consultation (2e).* Newbury: Petroc Press; 1996.
25. Burton J. Learning groups for health professionals: models, benefits and problems. *Work Based LearnPrim Care.* 2003; **1**:93–9.
26. Burton J. Work based learning in action: collaborative learning and personal learning. In: Burton J, Jackson N (eds). *Work Based Learning in Primary Care*. Oxford: Radcliffe Publishing; 2003. p. 25–48.

Chapter 8

Simulated surgery for the assessment of consulting skills

Peter Burrows

This chapter will give an account of the development of MRCGP Simulated Surgery, its current use as a consulting skills component of the examination and its likely evolution to the Clinical Skills Assessment (CSA) of the new MRCGP, which is to be the licensing examination for admission to the specialist register of general practice.

The MRCGP examination

In 1980 the gold standard for assessment of doctors completing their vocational training in general practice was the MRCGP examination. It consisted of two parts: the written papers, comprising the Traditional Essay Question, the Modified Essay Question and the Multiple-Choice Question which were all taken on the same day. Candidates were then called forward for Orals at Princes Gate or in Edinburgh if their combined scores were sufficient. Orals consisted of two half-hour interviews, each with two examiners. The candidates were presented with long scenarios containing impossible dilemmas, to see if they could find a solution that the examiners had missed. The marking was generally subjective and highly unreliable.

However, the examination underwent constant re-evaluation and change throughout the 1980s. The search for validity and reliability led to the Traditional Essay Question being dropped and the introduction of the Critical Reading Question. Pressure was mounting for a clinical component to be introduced. The examiners realised that talking or writing about what you would do in a given situation might not test the same thing as observing what you actually did. However, it was difficult to see how this could be done. The long case and short cases with physical signs, used in the clinical examinations of Finals, MRCP and FRCS, did not seem to address the skills that were considered important in general practice.

Objective Structured Clinical Examination (OSCE)

One promising method of assessment for clinical skills was the OSCE, the 'Objective Structured Clinical Examination',[1,2] invented in the late 1970s by Ronald Harden in Dundee for examining medical students. Students would progress around a circuit of 'stations' at each of which they would undertake a clinical task, such as taking a BP, examining a knee, or interpreting an ECG. Their performance was observed by examiners, who marked them according to predetermined criteria. A bell would sound and the students would progress to the next station until they had completed the whole circuit.

Walker and Walker[3] first described the use of the OSCE in assessment of general practice trainees in 1987. The college carried out pilots of the OSCE[4] in 1989 using 18 five-minute stations and a mixture of written and observed tasks. The advantages were that a large number of competencies could be assessed in conditions that were the same for all candidates. The fact that clinical skills were being addressed, was viewed with enthusiasm, but the isolation of clinical tasks was considered unrepresentative of general practice and inappropriate for fully trained GPs, so the project was shelved.

Simulated patients

Another important development in the 1980s was the use of simulated patients in medical education, which was pioneered by Paula Stillman[5] in Massachusetts. Simulated, or 'standardised patients' (SPs) in the American terminology are lay persons or actors trained to portray a clinical scenario as if they were a real patient interacting with a student or clinician. Cases can be written with specific teaching objectives in mind and SPs can be trained to provide a consistent and repeatable presentation. The great advantage for teachers is the availability of the patient for scheduled sessions. The use of SPs also overcomes the problem of obtaining consent from real patients and the risk of harming them by exposure to poorly performing students.

The potential of SPs in assessment was soon realised and checklists were produced for marking students' performances. Howard Barrows[6] is credited with early development in this field. The use of SPs was combined with OSCE methodology and used by many medical schools in the USA, Canada and Holland. The validity and reliability of such examinations were studied and the state of the art was summarised by Cees van der Vleuten and David Swanson[7] in 1990. One important principle recognised was that problem-solving skills were very context specific, and a large number of cases had to be assessed in order to infer that a candidate possessed adequate clinical competence. Rethans *et al.* [8] introduced the idea of assessing general practitioners' performances with simulated patients to British readers in the *BMJ* in 1991, and this was further explored by Kinnersley and Pill[9] in 1993.

Origins of the MRCGP simulated surgery

During the late 1980s the 'Simulated Office Oral' was developed by the Canadian College of Family Physicians for their certification examination.[10] This was an OSCE involving consultations with simulated patients, observed and marked through a one-way mirror. Here the physician remained in the consulting room and was visited by a series of patients, which had better fidelity for family practice. Philip Tombleson, who was convener of the panel of MRCGP examiners, had observed these examinations and persuaded the Examination Board to explore the use of this methodology. A development group was set up in 1991, which included Liz Bingham, Peter Burrows, Rob Caird and Neil Jackson with Gareth Holsgrove as education adviser.[11]

An early problem was terminology: the consultation in Britain is called a 'visit' in North America, whereas they use the term 'consultation' for referral (the GP consults the specialist) and our visit is a 'house call'. The Simulated Office Oral (SOO) is so-called because the GP's surgery is the 'office'. We opted for the

Simulated Surgery, which is equally confusing to our North American colleagues, who think it concerns surgical training in the operating room! They use 'standardised patients' (SPs), while we use the term 'simulated patients' or role players.

Role players

The role players in the MRCGP simulated surgery were friends and colleagues of members of the organising group; some were patients and others practice staff; none were professional actors, although several had amateur dramatic experience. They were approached informally and agreed to help. They were remunerated with a sessional fee and travel expenses.

They would be given a briefing about their case with details of their age and occupation and what they had come to see the doctor about. They were told about important features of the history, but warned not to deliver these directly without the doctor making appropriate enquiries. They were briefed about underlying worries and concerns, which it was expected that the doctor should elicit. However, they were given scope to mould the role to their own perception of the patient and permission to react naturally to the approach and style of the doctor. Above all, they were to try and achieve consistency between enactments so as to give each candidate a fair and equal opportunity to score marks.

We tried to match the cases to our role players, but avoided using actual sufferers from the condition being portrayed as we felt this might risk harming them if they were badly treated or misinformed by poor candidates. Two other groups were excluded: actors and actresses who we found tended to embellish the role with drama and emotion, and doctors, who, we discovered, could not resist the temptation to teach the candidates!

Case writing

We started writing cases drawn from our own experience in practice. We took simple presentations such as back pain, sore throat, or anxiety and depression. We added some less common scenarios such as new angina and diabetes, which we felt it important that the candidates should be able to manage correctly. We looked at the ability to give health advice in a heavy drinker and a new pregnancy. We wrote a case about the recurrence of metastases in breast cancer to test breaking bad news. And we made a foray into confidentiality with a case where a mother comes asking for her 17-year-old daughter's test results.

We had cases about genital herpes and urethral syndrome to test candidates' comfort with handling sensitive personal issues. We also tried to test the ability to deal with social issues using a lady who was agonising over whether to put her aging parents in a residential home. This was later discarded, due to its cultural sensitivity; Asian candidates had no dilemma – the lady should give up her job and devote herself to nursing them at home. We would often find that a case played quite differently from the way we expected when it was written. For example, where we had expected that most candidates would prescribe a drug, we found that many, both good and poor candidates, did not, so that prescribing was not a discriminating element of the case.

We learnt early on that you could not really test diagnostic skills in the simulated surgery. If candidates were explaining different diagnoses and treating different conditions, you could not mark them in a comparable way, so we made the diagnosis very plain and concentrated on how candidates gathered history and elicited the patient's concerns. What we *could* test were consulting skills, and there was no other part of the examination where this could be done. Clinical skills were a necessary context and could not be separated from the assessment of consulting skills, but it was in the latter that you could recognise important differences in candidate behaviour.

Domains of consulting skills

We categorised consulting skills into five domains, which seemed to be relatively discrete and, moreover, described the tasks of the consultation in chronological order (*see* Box 8.1).

Box 8.1: Domains of consulting skills

Information gathering:
Taking a focused and efficient history
Performing appropriate physical examination
Obtaining information from the records provided

Doctor–patient interaction:
Facilitating expression of the patient's story
Eliciting the patient's concerns and expectations
Using listening skills and non-verbal cues

Communication:
Explaining the problem and options for treatment
Negotiating the patient's agreement
Using appropriate language, checking understanding

Clinical management:
Devising a safe and effective management plan
Rational prescribing, investigation and referral
Rational use of time and resources

Anticipatory care:
Recognising implications for patient and others
Appropriate follow-up and surveillance
Opportunistic health promotion and advice

- 'Information gathering' is about gathering bio-medical information that is essential for making a diagnosis and excluding potentially serious conditions. Although we do not simulate physical signs, we expect candidates to undertake physical examinations when appropriate in such a way that they would find them if they were present.
- 'Doctor–patient interaction' measures patient-centredness and we find it remarkably discriminating. Good listening, open questions and non-verbal

skills will facilitate the telling of the patient's story and allow the candidate to pick up cues that are present.

- 'Communication' looks at verbal skills and rewards the candidate who can explain their understanding of the problem in a way that the patient can understand. Particular skills such as negotiating the acceptance of treatment and breaking bad news are included here.
- 'Clinical management' examines decision-making skills and formulating a plan of treatment. Like the video component we are keen to reward the offering of options and we would penalise poor use of resources, such as ordering MRI scans for simple backache. We also ensure that the candidate can recognise and manage urgent situations safely.
- 'Anticipatory care' is about what happens when the patient leaves the surgery, how follow-up will be arranged and about the implications of the patient's illness for other people. Appropriate health promotion is also included.

These categories are useful in that we can divide up the marks awarded for a case equally between the domains, although sometimes a domain will feature more than once in a case (for example, history taking and physical examination). Recent research has validated the domains, showing that a candidate's performance in one domain may be independent of their performance in another, and that this pattern may hold across a number of different cases in the circuit. Thus we have quantitative ways of identifying the candidate who takes an excellent history, but ignores the patient's concerns, or the good communicator who is nevertheless clinically inept. This gives us an objective basis for offering formative feedback, which is offered to failing candidates who enquire about their performance. At the same time we believe that we have devised and validated a useful model of the consultation.

Marking system

Devising a marking system proved to be the most difficult part of the project. Initially, we had generic marking schedules, which were applied to every case. However, these blunted the discriminating features of our case scenarios, because they covered so many general points that most candidates gained similar marks, even though their performance was plainly dissimilar. We therefore moved to case-specific marking schedules, different for each scenario, and covering the key tasks of the case. This is a major strength of the simulated surgery versus video or direct observation, because the cases are pre-written, and therefore expected good or poor performance can be defined prior to the assessment. Furthermore, the marking can be confined to those key tasks, which are essential for the doctor to address in order to achieve a successful outcome of the consultation. Georges Bordage in Canada enunciated this 'key features' principle.[12]

Unlike an undergraduate OSCE, where there is usually agreement on the correct way to perform a clinical task, in a GP consultation there may be several ways of doing it that are equally valid and successful in outcome. So the traditional checklist of a mark for every element (one for this, one for that), of the history, examination, explanation and treatment is not appropriate. Checklists also encourage candidates to adopt a scattergun or grapeshot approach in the hope of gaining more marks. Instead, we decided to use a grading system based on the observer's judgement, namely how well did the candidate perform within each domain, on a five-point scale between excellent and unacceptable. Research studies have shown that global judgements yield higher reliability, better construct validity and better concurrent validity than checklists in SP based examinations.[13]

The use of examiner judgements versus observers' marks has had far-reaching implications for the simulated surgery. This contrasts with the Leicester model developed by Justin Allen and Ali Rashid,[14–17] in which the role players themselves are trained to mark the case using a schedule based on what the doctor did and how the patient felt. The use of experienced GPs as examiners in our model allows flexibility of judgement at a far more sophisticated level and is likely to be more acceptable to candidates undertaking a high-stakes examination. However, this has contributed substantially to the costs and manpower required for the examination and has retarded the widespread adoption of simulated surgery in the MRCGP.

Validity and reliability

We needed to test our cases to ensure that they measured what we intended them to (construct validity) and revise them if necessary. We therefore held a number of pilot trials using VTS course registrars as candidates in venues around the South of England.[18] We had eight core cases that were used in each pilot, with two or three new ones on each occasion. Eighty-seven candidates were examined and we were able to distinguish between them with a wide range of marks (29–93%, mean 68%, SD 11.4%) (*see* Figure 8.1). Furthermore we seemed to be able to place correctly those whom the course organisers identified as their high flyers and their weaker brethren. This provides an indication of good concurrent validity.

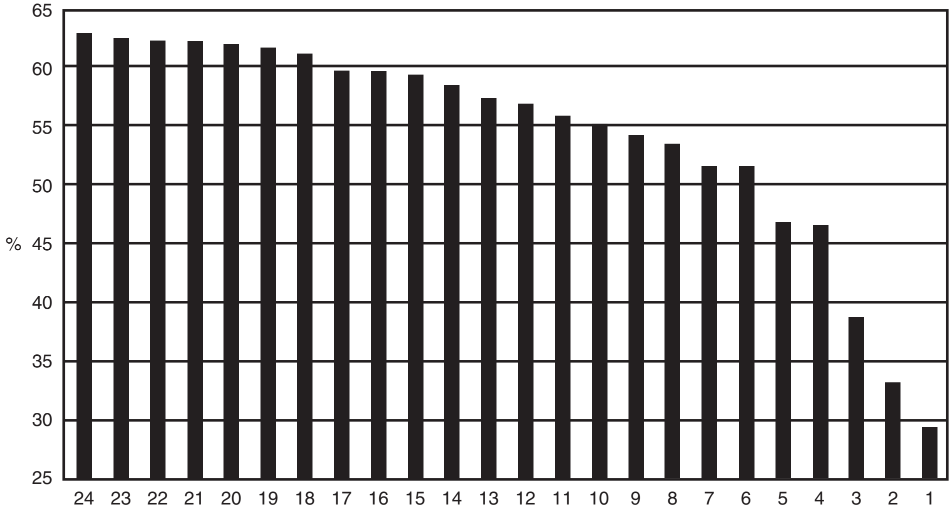


Figure 8.1: Low end of score distribution.

The reliability[19] of an examination indicates the reproducibility of the results and hence the likelihood that candidates would achieve the same scores if you tested them again. In most examinations it is measured by Cronbach's Alpha coefficient, which is a measure of internal consistency and shows the degree to which candidates' final scores are consistent with their case scores. It is expressed on a scale of 0 to 1 and the Alpha of our eight-case pilots was 0.65. An Alpha of 0.8 or above is considered to be the gold standard for a high-stakes examination, and to reach this level more judgements need to be made, implying a longer exam with more cases. Our initial calculations suggested that we needed 16 cases, and indeed we launched the exam with 20 but soon reduced the number to 12.

Additional reliability can be achieved with training of role players and markers to improve the consistency of their performance. Another contribution can be made by monitoring the correlation of candidates' performance in each case with their performance on the other cases in the circuit. Cases with a low or negative correlation are reviewed and some have been discarded from the casebank. The 12-case examination has returned Alpha coefficients of between 0.74 and 0.87 in the past 7 years.[26]

Standard setting

Separating good from poor performance reliably is the essential function of any examination. But how good is good enough and how poor is unacceptable? Figure 8.2 shows the marks distribution in the pilot trials. Where should the pass mark be? Traditionally in medical exams it is set at 60%. This would fail 17 of our 87 registrars, which is 20% and arguably too many. We can be confident that the last three should fail, so why not set the passmark at 45%? However, that represents pretty minimal performance and why go to all this trouble to eliminate just 3.5%? Standard setting is an arbitrary business and dependent on the needs of the authority setting the exam. Using the Hofstee 'compromise' method allows one to offset the desired marks against an acceptable failure rate[20] and we came up with a pass mark of 53.3%, which fails seven candidates (8% of the cohort).

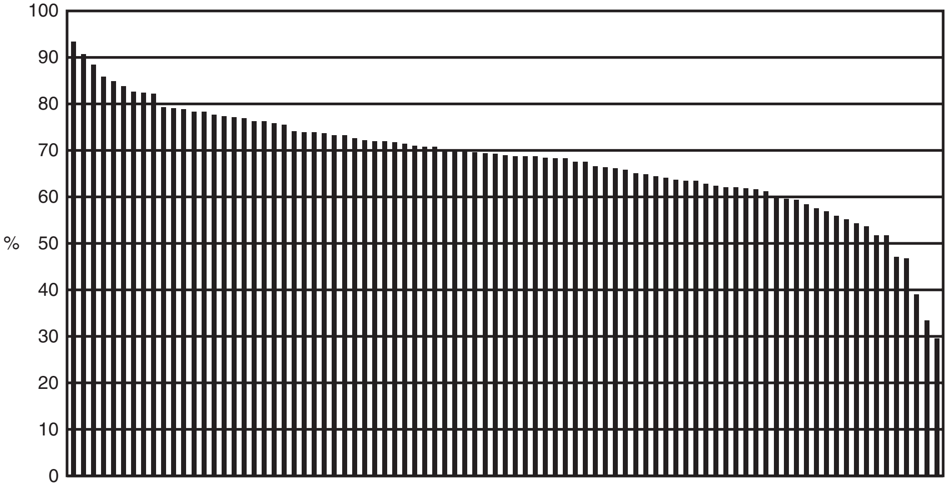


Figure 8.2: Distribution of scores for 87 candidates.

Standard setting should ideally be criterion referenced, but application of the Angoff technique as used in written papers is highly resource-intensive and takes no account of the conditions in an SP examination that may lead to variations in difficulty. However, in SP exams one can make use of the judgements of the examiner group who have just observed the candidates interacting with the SPs. This is the basis of the 'contrasting groups' and 'borderline groups' methodologies[21,22] developed in Canada and North America.

We devised a modified 'contrasting groups' method,[23] which sets the pass mark at a point where 50% of examiners would pass a candidate. Subsequently we adopted a 'borderline candidate' method, asking the examiners to rate a hypothetical 'just passing' candidate on the case, that they had just marked. These marks were then combined and applied as the pass mark for that examination. It would vary a little between examinations, depending on the harshness of the examiners and the difficulty of the cases selected. Eventually, however, we have settled on a fixed pass mark, which is just above the score that would be obtained by a candidate who is awarded a borderline ('b') grade on every element of the examination. The justification for this is that each grading decision is referenced to the examiner's own implicit standard of adequate performance.

Video module and eligibility

While this work was going on, the college was developing a parallel assessment of consulting skills, the video component.[24] Peter Tate led a development group, which created a method for assessing the performance of candidates on a videotape of live consultations made in their own surgery. Analysis of the skills needed in good consulting was refined by consensus of the panel of examiners and a detailed set of criteria was developed that could be applied by trained observers. Candidates know what these criteria are, and are invited to submit a videotape of consultations, which demonstrates their mastery of them. The video component does lack the advantage of providing a standardised challenge to the candidates, but it has significant advantages in terms of manpower and cost. Whereas in the video, three candidates are examined per examiner day, in the simulated surgery the number is only two, not to mention the additional cost of employing role players.

The Exam board therefore decided to restrict the simulated surgery to those candidates who were unable to make and submit a video. This effectively excluded current registrars who could use video in their training practices, but catered for principals, those not in current practice, overseas candidates, those who consult in foreign languages and those with an ethnic patient population who will not consent to video. So we started with an atypical cohort of candidates and a correspondingly high failure rate. More recently, with the modular MRCGP, many candidates, especially women GPs, delay taking the consulting skills component until after the completion of their vocational training, when they are taking a career break or doing locum work. They cannot therefore prepare a video and are thus eligible to enter the simulated surgery. We also see a proportion of those who are resitting their consulting skills module after an earlier failure to pass the video. Simulated surgery is provided as a free choice instead of video assessment for candidates taking Membership by Assessment of Performance (MAP).

Establishing the simulated surgery

In July 1997 the first MRCGP Simulated Surgery examination was held at Princes Gate. Twenty-three candidates were examined in three days using 20 cases. The ten role players each played one case in the morning and a different case in the afternoon. The organisation was brilliant, the results highly discriminating, the reliability magnificent and the cost astronomical! After some frenzied extrapolation, we decided to reduce the examination to 12 cases, a compromise between the number that could be fitted into a single circuit and the likely loss of reliability. This proved to be a successful decision and the simulated surgery has run 12 case circuits with high reliability ever since.[25]

The following year, we moved to Millbank, the RAMC headquarters and the tradition of residential venues for the team of examiners and role players was established. This allows the bedrooms in which we sleep to be used as consulting rooms for the examination the next morning. It overcomes the problem of using screened off sections of a hall, where candidates who have completed their consultation can listen in to the next case being done by their neighbour. It has also encouraged a remarkable esprit de corps to develop between examiners and role players, born of socialising together at meals and evening entertainments.

A year later, we moved to the Beaumont Conference Centre in Windsor. This had the space for us to scale up our operation and run two parallel circuits simultaneously on different floors of the building. We were now able to examine 52 candidates in a day, giving us a capacity of 128 candidates in a 2½-day examination. Careful synchronisation of each role between the two role players and calibration of the marking between the two examiners involved was required, and this necessitated a short training session at the beginning for each 'quartet', who were provided with a training video of the role. Analysis of the marking showed high concordance between the pairs of examiners in terms of mean, standard deviation and range of marks awarded for the same case on different circuits. Consistency of role playing was also monitored by 'floating' observers.

Subsequent refinements to the examination have included increasing the range of grades which the examiners use to parallel those of the oral exam (a nine-point scale ranging from 'outstanding' to 'dreadful'); also a reduction of the number of domains to four by combining 'Management' and 'Anticipatory Care'; the core group is also testing ways of using the role players' views on the interpersonal aspects of the consultation, which are clearly valid, but not always reflected in the examiners' grades. This is an important representation of the 'patient's voice' which will add another dimension of validity to the assessment of doctors entering practice. The results of the last seven years of the simulated surgery have been reviewed and a report has been accepted for publication in *Education for Primary Care*.[26]

Organising the examination

Organising a session of the simulated surgery is a complex business requiring advance planning, booking of personnel, preparation of paperwork, arranging the rotation, timing the consultations and recording the marks. Attention to detail is crucial; a missing record, a muddled rotation or an absent role player can crash the whole examination. A major strength of our examination has been the administration. Frances Cloyne who was Faculty Support Manager in Wessex took on the task of organising it and developed the formidable expertise that we came to depend upon. This part of the examination was organised in Wessex until 2005, when the task was taken on by the examination department at Princes Gate.

So what happens at a simulated surgery examination?[27,28] Candidates are given a time and place to attend, and they get a briefing from one of the organisers. They are told to expect a surgery of 12 patients, which resembles a normal surgery in a practice where they are the new doctor. They will find an appointment list and brief records of each patient in the consulting room. A whistle is blown and the patient enters, followed by an examiner who is there to observe and plays no part in the consultation. The patient will say why they have come, and a normal consultation should ensue. Physical examination may be carried out, if appropriate to the case, but intimate examinations are not expected and will be declined. After 10 minutes a second whistle is blown and if the consultation has not reached a natural conclusion before this, the patient will get up and leave as no further marks can be gained. After a two-minute break while the examiners record their grades, the whistle is blown again, and the next patient comes in. This continues, with a break for tea or coffee after six cases, until each candidate has seen all of the 12 patients.

The grades are entered into an Excel spreadsheet, which converts them into marks that are then summed and expressed as a percentage of the maximum available for the case. A mean of the 12 case scores is taken as the final examination score. The pre-set pass mark is applied to determine whether a candidate passes or fails. This is a fully compensatory system in which good performance on some cases will balance out poor performance on others. This acknowledges that even the best candidates may do less well on some cases and recognises the context specificity of clinical competence. A proposal to require a minimum standard of performance in each domain, using subsidiary pass marks applied to the domain subscores has been considered but not implemented. Merit classification is awarded to those candidates in the top 25% of the rank order for each examination.

The future of simulated surgery

From July 2007 summative assessment and the MRCGP will be combined into a single examination, which will not only admit candidates to the college, but also become the licensing examination for entry into general practice. This may be a different standard from the existing MRCGP, an issue that is causing a lot of controversy among members! The Joint Committee on Postgraduate Training for General Practice (JCPTGP) has been replaced by the Postgraduate Medical Education and Training Board (PMETB). It has devised a tripartite assessment, which, in common with other specialties, will be administered by the Royal College. This will consist of a workplace based assessment, an MCQ knowledge test and an examination of clinical skills. This clinical skills assessment (CSA) will use the methodology of the simulated surgery and is under development by a working group of the panel of examiners. The purpose of the CSA has now been defined as, 'An assessment of the doctor's ability to integrate and apply clinical, professional, communication and practical skills in a general practice setting'.

Unfortunately the decisions about how the CSA will be organised are still under discussion and not yet in the public domain. However, the principles of the new examination are clear and the CSA will undoubtedly take the form of a station-based examination. It may be that not all stations will be patient consultations, and some may be mini-orals in which decision-making skills are examined by a live examiner, perhaps relating to a case, which has just been presented previously on the circuit. A 14-case circuit is probably at the limit of feasibility if two circuits are to be run in one day. This will restrict the scope for extending the content without loss of reliability.

The logistics involved are daunting. The existing simulated surgery will have to be scaled up from 240 to 2500 candidates a year. It is unclear at present whether the examination will be held on a regional basis or in a central examination venue. It is certainly possible to syndicate the same set of cases and run them simultaneously in a number of venues. This is done in Canada, where the LMCC OSCE is run in 17 centres, across five time zones, and in two languages simultaneously.[29] The new examination will require the recruitment and training of a large number of role players. They cannot perform for more than two and a half days without suffering fatigue and loss of consistency.[30] A much larger group of examiners will be required and this ought to increase the sense of ownership of the examination amongst members. It is important not to develop an elite corps of assessors who are perceived as distant from the trainers and working GPs. This is both an opportunity and a challenge for the college.

References

1. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ.* 1979; **13**(1): 41–54.
2. Harden RM. What is an OSCE? *Med Teacher.* 1990; **10**(1): 19–22.
3. Walker R, Walker B. Use of the objective structured clinical examination for assessment of vocational trainees for general practice. *J Roy Coll Gen Practs.* 1987; **37**(296): 123–4.
4. McAleer S, Walker R. Objective structured clinical examination (OSCE). *Roy Coll Gen Practs. Occasional Paper.* **46**: 1990; 39–42.
5. Swanson DB, Stillman PL. Use of standardized patients for teaching and assessing clinical skills. *Evaluation Hlth Prof.* 1990; **13**(1): 79–103.
6. Barrows HS, Williams RG, Moy RH. A comprehensive performance-based assessment of fourth year students' clinical skills. *Med Educ.* 1987; **62**(10): 805–9.
7. Vleuten C van der, Swanson D. Assessment of clinical skills with standardised patients: state of the art. *Teach Learn Med.* 1990; **2**: 58–76.
8. Rethans JJ, Sturmans F, Drop R *et al.* Assessment of the performance of general practitioners by the use of standardized (simulated) patients. *BJGP.* 1991; **41**(344): 97–9.
9. Kinnersley P, Pill R. Potential of using simulated patients to study the performance of general practitioners. *BJGP.* 1993; **43**(372): 297–300.
10. Sawa RJ. Assessing interviewing skills: the simulated office oral examination. *J Fam Pract.* (1986); **23**(6): 567–71.
11. Bingham L, Burrows P, Caird R *et al.* Simulated surgery: a framework for the assessment of clinical competence. *Educ Gen Pract.* 1994; **5**: 143–150.
12. Page G, Bordage G, Allen T. Developing key-feature problems and examinations to assess clinical decision-making skills. *Acad Med.* 1995; **70**(3): 194–201.
13. Regehr G, MacRae H, Reznick RK *et al.* Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med.* 1998; **73**: 993–7.
14. Rashid A, Allen J, Thew R *et al.* Performance based assessment using simulated patients. *Educ Gen Pract.* 1994; **5**: 151–6.
15. Allen J, Rashid A. What determines competence within a general practice consultation? Assessment of consulting skills using simulated surgeries. *BJGP.* 1988; **48**: 1259–62.
16. Allen J, Evans A, Foulkes J *et al.* Simulated surgery in the summative assessment of general practice training: results of a trial in the Trent and Yorkshire Regions. *BJGP.* 1998; **48**: 1219–23.

17. Sutcliffe R, Evans A, Pierce B *et al.* Simulated surgeries – feasibility of transfer from region to region. *Educ Gen Pract.* 1998; **9**: 203–10.
18. Bingham L, Burrows PJ, Caird R *et al.* Simulated surgery – using standardised patients to assess the clinical competence of GP registrars – a potential clinical component for the MRCGP examination. *Educ Gen Pract.* 1996; **7**: 102–11.
19. Streiner DL, Norman GR. *Health Measurement Scales – a practical guide to their development and use*. Oxford: Oxford University Press; 1989. ch. 8.
20. Cusimano MD, Rothman AI. The effect of incorporating normative data into a criterion-referenced standard setting in medical education. *Acad Med.* 2003; **78**(10 Suppl): S88–90.
21. Clauser BE, Clyman SG. A contrasting-groups approach to standard setting for performance assessments of clinical skills. *Acad Med*. 1994; **69**(10 Suppl): S42–4.
22. Dauphinee D, Blackmore D, Smee S *et al.* Using the judgements of physician examiners in setting the standards for a national multi-center high stakes OSCE. *Adv Hlth Sci Educ.* 1997; **2**(3): 197–9.
23. Burrows PJ, Bingham L, Brailovsky CA. A modified contrasting groups method used for setting the passmark in a small scale standardised patient examination. *Adv Hlth Sci Educ.* 1999; **4**: 145–54.
24. Tate P, Foulkes J, Neighbour R *et al.* Assessing physicians' interpersonal skills via videotaped encounters: a new approach for the MRCGP. *J Hlth Comm.* 1999; **4**: 143–52.
25. Burrows PJ and Bingham L. The simulated surgery – an alternative to videotape submission for the consulting skills component of the MRCGP examination: the first year's experience. *BJGP.* 1999; **49**: 269–72.
26. Hawthorne K, Denney ML, Bewick M *et al.* Simulated surgery – an exam for our time? Summary of the current status and development of the MRCGP simulated surgery module. *Educ Prim Care.* 2006; **17**: 138–46.
27. Burrows P. Consulting skills assessment by simulated surgery. In: Moore R (ed.). *The MRCGP Examination (3e).* London: The Royal College of General Practitioners; 1998. p. 101–13.
28. Bewick M. A guide to the simulated surgery. *Practitioner.* 2004; **248**(1658): 372–3.
29. Reznick RK, Blackmore D, Dauphinee D *et al.* Large-scale high-stakes testing with an OSCE: report from the Medical Council of Canada. *Acad Med.* 1996; **71**(1 Suppl): 19–21.
30. Denney ML, Wakeford R, Hawthorne K *et al.* Experiences of simulated patients and candidates in the MRCGP simulated surgery examination. *Educ Prim Care.* 2006; **17**: 354–61.

For an extensive and regularly updated bibliography about the use of Standardised Patients, readers are referred to the online database compiled by University of Texas Medical Branch (UTMB) at http://oed.utmb.edu/SP/bibliography.htm

Chapter 9

Video assessment

Adrian Freeman

Introduction

Using video recording allows unobtrusive assessment of candidate performance. Such assessments are often labelled high fidelity, implying good validity. As with most assessments the utility of video assessment is a function of many components and not just of validity. Assessment of performance moves to the top of Miller's triangle, a measurement of what the candidate actually does.[1] Direct observation of clinician behaviour has a long tradition in medical assessment.[2] However, it can be difficult to make reproducible and standardised assessments. This can create compromises to reliability. On the other hand observations of simulated activity such as OSCEs or simulated patients allow standardisation with improved reliability but simulations can weaken validity. Video creates a permanent valid record of performance that can be assessed with reliability.

The video is merely a means of recording and can therefore record many different activities to be assessed. This may then create an assessment of psychomotor skills such as surgical techniques, communication skills, simulations or even complete practice.[3] Leaving a video camera constantly running will give a 'spy in the cab' observation of what a clinician is actually doing in practice. However, the limitation is that someone has to take the time to watch the video to make the assessment.

Formative to summative assessment

Video observations have been used in many aspects of medical education. They have been valuable as a teaching tool in both undergraduate and postgraduate training. As a formative assessment they can monitor improvement in communication skills. In family medicine video has been used to assess consulting skills for nearly 15 years.[4] What began as a formative teaching tool was seen as a good opportunity for a high fidelity summative assessment. In the UK video assessment has become part of the licensing requirements for general practice.[5,6] Videotapes of consultations recorded in the surgery allow assessment of consulting and communication skills. As with most examinations accuracy of assessment in videos is enhanced by having structured report forms.[8] The details of that structure are ideally informed by blueprinting of the assessment and then consensus agreement on marking criteria. Inevitably subsequent piloting of the process will further refine a marking schedule. For example, the MRCGP video marking schedule was derived from an initial Delphi exercise of over 100 practising GPs who rated what they thought were important parts of a consultation.[6] This large list was then honed down through pilot work to 15 competencies or performance criteria that worked well in an assessment. These criteria were grouped into five areas: discovering the reason for the patient's attendance, exploring the problem, tackling the problem, explaining the problem and making effective use of the consultation. The standards of competence were agreed by the profession and candidates submit videos of consultations that they have selected to demonstrate those competencies. That process of case selection can be a critical part of video assessment. Some of the contexts of the consultations can be specified, such as a consultation involving a child and a consultation involving a mental health case. Candidates can spend a long and probably unnecessary time trying to select the absolute perfect combination of consultations. An alternative method would be to judge sequential consultations, i.e. remove the choice from the candidate. This would inevitably require more judgement time and is unlikely to enhance the reliability.

The issues for standard setting in a video assessment are the same as for other assessments. The purpose of the assessment should be clear as different end points may require different methods of standard setting. Most commonly with video, as an assessment of performance, standards will be set using a criterion reference, i.e. how many of the selected competencies or tasks have to be demonstrated to pass. The established methods of Angoff and borderline regression have been used and described successfully with video assessment of general practice.[7] The methods chosen should take allowance of the fact that the material presented is often variable in context and content.

Validity and reliability

A recent article by Downing and Haldanya[9] addresses some threats to validity relevant to video assessments; in particular how many cases and the effect of judges. The number of cases refers to the threat of construct under-representation. It is well known that performance and knowledge in one domain does not predict others. For example, a candidate scoring highly on a cardiology case may score very poorly on a respiratory case. Multiple short observations are more defensible than in depth analysis of a few.[10] The blueprinting process should indicate the areas of knowledge/performance that the assessment should be sampling in. It is therefore preferable for the assessment to be of observations of different types of clinical encounter/material and for there to be as many observations as possible. The number of observations is limited by the feasibility of the time and resources to judge them. Increasing complexity of the clinical material allows a more realistic assessment than low challenge simple encounters. However, over-complex situations such as multiple patients consulting with multiple problems on the same occasion create difficulties of marking. As with all assessments best practice is to give feedback to failing candidates of how they could improve. One of the common reasons for failing the MRCGP video examination is that candidates have submitted low challenge consultations that do not allow them to fully demonstrate their skills. Feedback on this point allows them to submit more appropriate cases in the future.

The other threat to validity in video assessment has been labelled as construct irrelevant variance and refers to the variance of raters (judges).[9] In particular this refers to the 'halo effect' where a judge is influenced for either the good or bad by candidate performance in the first case. For subsequent cases the rater judges according to that initial belief of candidate performance rather than the actual performance in subsequent cases. However, there are other aspects of rater variance such as interpretation of marking scales, severity of the judgments, difficulty in rank ordering, etc. It is theoretically possible to make statistical adjustments to the final score to take these factors into account. However, 7–11 independent judges have been estimated as being sufficient to compensate for this problem.[11]

A Dutch study has looked specifically at the effect on reliability of the number of judges and cases for a general practice video assessment. This generalisibility study demonstrated good reliability with one judge observing 12 cases or two judges watching a total of eight cases.[12] That same study concluded that video assessment of GPs in daily practice was not only reliable but also valid and feasible. Similar results were found in a study of video analysis of Australian general practitioners.[13] The British Royal College of GPs consultation assessment module uses seven judges independently watching one each of a total of seven cases.[14]

Educational impact

Several recommendations can be made about judges to improve the performance of the assessment.[11]

1 Educate and train the judges.
2 Establish the meaning of the ratings/competencies.
3 Provide time for judgement.
4 Judge specific performances.
5 Give the judges feedback about their performance.

It is said that assessment drives learning and educational impact is one of the stated components of utility. Campion and others looked in detail at the performance of over 2000 candidates in the MRCGP video examination.[14] They particularly looked at what is known as the patient-centred competencies in the consultation. Whilst there are many issues about the precise definition of patient centredness[15] these competencies measure areas which have been shown to improve patient care. The initial cohort study indicated that these training doctors were in fact not good at patient-centred behaviours. Subsequent review three years later showed a real increase in demonstration of these behaviours.[16] The assessment had indicated to candidates that these were important aspects of patient care and they were expected to be demonstrated. The candidates and their teachers had taken that message on board and the assessment had clearly driven learning.

Acceptable method?

Unless the judges have significant linguistic capabilities then one limitation is that the clinical encounter should be in a language that the judge understands. Of course there may be cultural reasons which would not allow video recording of patient care.

An important aspect of video assessment of doctors consulting with real patients is to obtain the consent of the patient. As the videotapes will be seen by independent judges it is vital that explicit consent is sought from the patients. The consent should clearly state the purpose of the recording, who will see it and the arrangements for subsequent destruction of the video. The patient should not feel that they are under pressure and should understand that non-consent would have no detrimental effect on their clinical care. The seeking of consent should be independent of the direct involvement of the doctor and after the video the patient should be given time for reflection before giving final consent. It is understandable that some patients will not consent. An early study of video in UK practice showed that the percentage was small but particularly identified that consent was more likely to be withheld by younger patients and those with mental health problems.[17]

One assumption is that knowledge of being observed will lead the doctors to behave in a different manner to normal. However, an early study of consulting behaviour in the UK showed that doctors are not affected by having the consultations video recorded.[18] Another study showed that general practitioners felt that video assessment was more recognisable to the doctors as normal practice than a multiple-station examination.[19] In this comparison of assessment of practising physicians the same doctors submitted videos of their consultations for assessment and took part in a multiple-station simulated examination. In fact that study suggested that video assessment was favoured to multiple-station examination using simulated patients in validity, reliability and feasibility.

Summary

A video-based assessment is close to assessing what a candidate actually does and provided attention is paid to certain aspects, the assessment has a high utility.

References

1. Miller G. The assessment of clinical skills/competence/performance. *Acad Med.* 1990; S63–67.
2. Wass V, van der Vleuten C. The long case. *Med Educ.* 2004; **38**(11): 1176–80.
3. Weller JM, Bloch M, Young S, Maze M *et al.* Evaluation of high fidelity patient simulator in assessment of performance of anaesthetists. *Br J Anaesth.* 2003; **90**(1): 43–7.
4. Cox J, Mulholland H. An instrument for assessment of videotapes of general practitioners' performances. *BMJ.* 1993; **306**(6884): 1043–6.
5. Campbell LM, Murray TS. Summative assessment of vocational trainees: results of a 3-year study. *Br J Gen Pract.* 1996; **46**: 411–14.
6. Tate P, Foulkes J, Neighbour R, Campion P *et al.* Assessing physicians' interpersonal skills via videotaped encounters: a new approach for the Royal College of General Practitioners Membership examination. *J Hlth Comm.* 1999; **4**(2): 143–52.
7. Hobma SO, Ram PM, Muijtjens AMM, Grol RPTM *et al.* Setting a standard for performance assessment of doctor–patient communication in general practice. *Med Educ.* 2004; **38**(12): 1244–52.
8. Noel GL, Herbers JEJ, Caplow MP, Cooper GS *et al.* How well do internal medicine faculty members evaluate the clinical skills of residents? *Ann Int Med.* 1992; **117**: 757–65.
9. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ.* 2004; **38**(3): 327–33.
10. Norman G. Research in clinical reasoning: past history and current trends. *Med Educ.* 2005; **39**(4): 418–27.
11. Williams RG, Klamen DA, McGaghie WC. Special article. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med.* 2003; **15**(4): 270–92.

12. Ram P, Grol R, Rethans JJ, Schouten B *et al.* Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of validity, reliability and feasibility. *Med Educ.* 1999; **33**(6): 447–54.
13. Hays R, Spike N, Gupta TS, Hollins J *et al.* A performance assessment module for experienced general practitioners. *Med Educ.* 2002; **36**(3): 258–60.
14. Campion P, Foulkes J, Neighbour R, Tate P. Patient centredness in the MRCGP video examination: analysis of a large cohort. *BMJ*. 2002; **325**(7366): 691–2.
15. Epstein RM, Franks P, Fiscella K, Shields CG *et al.* Measuring patient-centered communication in patient-physician consultations: theoretical and practical issues. *Soc Sci Med*. 2005; **61**(7): 1516–28.
16. Campion P, Tate P. Patient centredness improving? *BMJ Rapid Responses*. 2003; **12**: 8.
17. Coleman T, Manku-Scott T. Comparison of video-recorded consultations with those in which patients' consent is withheld. *BJGP*. 2006; **48**: 971–4.
18. Pringle M, Stewart-Evans C. Does awareness of being videorecorded affect doctors' consultation behaviour? *BJGP.* 1990; **40**: 455–8.
19. Ram P, van der Vleuten C, Rethans JJ, Grol R *et al.* Assessment of practising family physicians: comparison of observation in a multiple-station examination using standardized patients with observation of consultations in daily practice. *Acad Med*. 1999; **74**: 62–9.

Chapter 10

Summative assessment

Moya Kelly and Murray Lough

History

In 1975 the Joint Committee for Postgraduate Training for General Practice (JCPTGP) was set up with the responsibility for conferring the rights of independent practice. The JCPTGP contained representatives from General Medical Service Committee (GMSC) and the Royal College of General Practitioners (RCGP), with nominees from other bodies such as universities. In 1979 it became a requirement for doctors to complete a year as a trainee in general practice to achieve certification and by 1982 it had become necessary to complete two years of approved hospital posts and a trainee year. For each of these posts, the responsible trainer had to complete a statement of 'satisfactory completion'. The meaning of this term was rather vague and the interpretation of 'satisfactory' equated to completing the appropriate time in the post. The situation was clarified in 1990 by the Chairman of the JCPTGP, GMSC and the RCGP who stated that the doctors should have reached an acceptable standard of competence.[1] Despite this it was unusual for a registrar not to receive a satisfactory completion certificate as this could only be triggered by the informed signature of the trainer. Between 1989 and 1992 the proportion not receiving certificates was 0.26%.[2]

There was an assumption by the public, shared to some extent by the medical profession, that doctors entering on restrictive practice, either as general practitioners or as consultants were of proven competence. This was not the case and a review group was set up in 1995 to look at the identification of poorly performing doctors. Their report indicated that there was a significant problem and recommended that for general practitioners 'systems for objective assessment against a national framework should be introduced as soon as possible'.[3] A system of summative assessment was therefore developed and put in place for all general practice registrars on 4 September 1996.[4,5]

Development of the system

The two key questions to be addressed were as follow.

1 What should be assessed?
2 How should it be assessed?

Before deciding on specific instruments it was necessary to define the attributes to be tested. It was impossible to decide whether a doctor was competent without first deciding what aspects of knowledge, skills and behaviour make up competence. In general practice this is not an easy task because of the wide-ranging nature of the job.

The Joint Committee sets out the following basic attributes required in a doctor at the end of training:

- adequate knowledge
- adequate problem-solving skills
- adequate clinical competence
- adequate consulting skills
- adequate skills in producing a written report of practical work in general practice
- adequate performance of skills, attitudes and knowledge.

In order to assess such a wide range of attributes a variety of tools was developed[4] containing the following components:

- a multiple true/false paper
- submission of videotaped consultation
- a trainer's report
- submission of an audit project.

In order to pass summative assessment a trainee had to pass all four components.

Multiple true/false paper

Multiple-choice question (MCQ) papers test factual knowledge and problem-solving skills and allow candidates to be ranked. For summative assessment purposes an additional step was required to devise a pass mark that equated with the minimal acceptable knowledge base. The process of defining this pass mark involved using the Angoff and the Hofftee techniques.[6,7] In this a group of experienced general practice principals analysed the paper question by question and produced a figure for the percentage of trainees with minimum acceptable competence whom they would expect to answer the questions correctly. By this means a preliminary pass mark was determined which could then be further modified. Summative assessment MCQ is offered four times per year and is free to candidates. It is a three-hour examination that initially consisted of 260 true/false questions and 40 extended matching questions. The number of extended matching questions was gradually increased and the paper currently consists of 170 true/false questions, 80 extended matching questions and ten single best answer questions.

The true/false section of the paper is composed of 20% of questions that are taken from an earlier paper and are called 'anchor' items. 50% of the questions have been used previously but more than 18 months ago and 30% are new questions. The extended matching questions are usually new each time but those that are not new will have been used more than two years previously. There is no negative marking. The true/false questions cover topic areas that are found in general practice with the proportions being internal medicine (medicine, therapeutics, surgery, psychiatry, geriatrics) 45%, child health, womens' health, external medicine (ENT, ophthalmology, dermatology) 50% and practice management 5%. Extended matching questions use a blueprint for topic areas and cover diagnosis, investigations, management, therapeutics and health promotion.

MCQ pass rates

The pass rates for the MCQ component are high. In the first five years almost 9,000 candidates took the MCQ at least once and of those only 63 had not passed at that time. As explained before there is no fixed pass mark. This is decided at the standard setting meeting which happens once a year.

The paper is marked by optical scanning and the detailed analysis is carried out on each paper to check reliability which has always been above 0.8. All items are analysed statistically for facility and discrimination values.

The former checks how easy the question was for the candidates and the latter relates to the relationship between passing individual questions and performance in the test as a whole. The expectation is that those who score highly on the tests as a whole are more likely to answer individual questions correctly. Questions that are easy, as measured by the facility score often show a weak or negative correlation.

If a candidate fails they can repeat the MCQ paper, up to three attempts, in their registrar year. If the candidate fails on three occasions they would have a meeting with their regional director to look at their educational needs and assess the requirements for further training. Each candidate is given their pass mark, the pass rate, failure rate and the minimum and maximum marks, the mean mark and one standard deviation for their registrar group. No feedback is given on particular areas of weakness as the number of items in any one area is too small to give a meaningful result for the candidate. Since 1996, registrars who satisfy the MRCGP examiners that their performance in the MRCGP MCQ is adequate (pass or passed with merit) are exempt from undertaking the COGPED MCQ. Only a small number of registrars uses this option to pass the component, the majority taking the summative assessment MCQ which is a no-cost option for registrars.

Before the final mark is decided the paper is reviewed and questions that are found to have errors or that have a poor facility and discrimination are suppressed. In the first five years of summative assessment 7,653 candidates undertook the process, 273 (3.6%) were unsuccessful. Of these failures ten had failed the MCQ.

The MCQ is a good way of testing factual knowledge and problem solving. The summative assessment MCQ has identified individuals with significant knowledge gaps. Some may criticise a summative assessment process that allows individuals to retake the component several times during the year. However, the evidence shows that the more often the MCQ is sat and failed the less likely an individual is to pass.

The MCQ could be further developed by expanding the number of extended matching and single best answer questions. It is going to be merged with the MRCGP MCQ in 2007 as part of the new MRCGP that will replace the existing summative assessment.

Consulting component

The challenge of assessing consulting skills was to develop a system that had reliability and validity while being feasible and practical to implement. An ideal assessment performance would be where a doctor was unaware that the assessment was taking place. Work in this area has been carried out particularly in Holland with the use of simulated patients.[8,9] This method would not be without its problems in the National Health Service where such patients would need to be treated as temporary residents and where the length of appointments is not within the control of the GP registrars. At the time of the development of the summative assessment process a number of methods were available that looked at consultation competence.[10,11,12] In each of these methods results were presented in numerical format which required a relatively arbitrary decision as to the cut-off point for minimal acceptability. These scales were not designed specifically for the identification of the non-competent GP but were very useful in giving formative feedback.[10,11,12]

Observation of performance in the workplace undoubtedly has face validity. However, reliability can be much more difficult to achieve, especially if markers are used to make a judgement. The use of simulated patients has an advantage in that each candidate can be presented with the same set of problems to deal with. This is very helpful if it is wished to rank candidates and impressive reliability figures can be achieved with this system. Simulations tend to be used to assess specific skills rather than overall competence in most cases.[13,14] The system developed for summative assessment used real consultations. This allowed the GP registrar to record the consulting session at a time of his/her choosing and to select consultations that they felt demonstrated their competence. There was no evidence that the presence of a video camera in the consulting room affected patient satisfaction.[15]

COGPED video

The COGPED video looked at generic skills such as listening, negotiating and making reasonable decisions at minimum competence level. Registrars submitted a tape of two hours duration with a minimum of eight consultations on the tape. An instrument was developed looking at broad criteria – listening, action and understanding. A judgement was made independently by an assessor for each consultation as to whether this was satisfactory or whether they had doubt about a registrar being competent. Having watched a minimum of six consultations an overall global judgement was made as to whether the individual should pass or be referred. This type of global assessment has been shown to be more reliable than numerical assessment. The reliability of the instrument was tested in real patient consultations and found to be adequate and in the pilot had a failure rate of around 5%.[16] No GP registrar failed the process until their performance had been reviewed by at least six assessors, four of whom are from outside the local region (*see* Figure 10.1).

Initially the assessment of consulting skills was solely by the COGPED video method. In January 2000 other methods were approved for the assessment of consulting skills. These included simulated surgery[17] and in the summer of 2001 a pass in the video component of the MRCGP was accepted as a pass in the summative assessment consulting skills module. A fail in the MRGP automatically entered the summative assessment process and was sent along with other COGPED tapes to first level assessors.

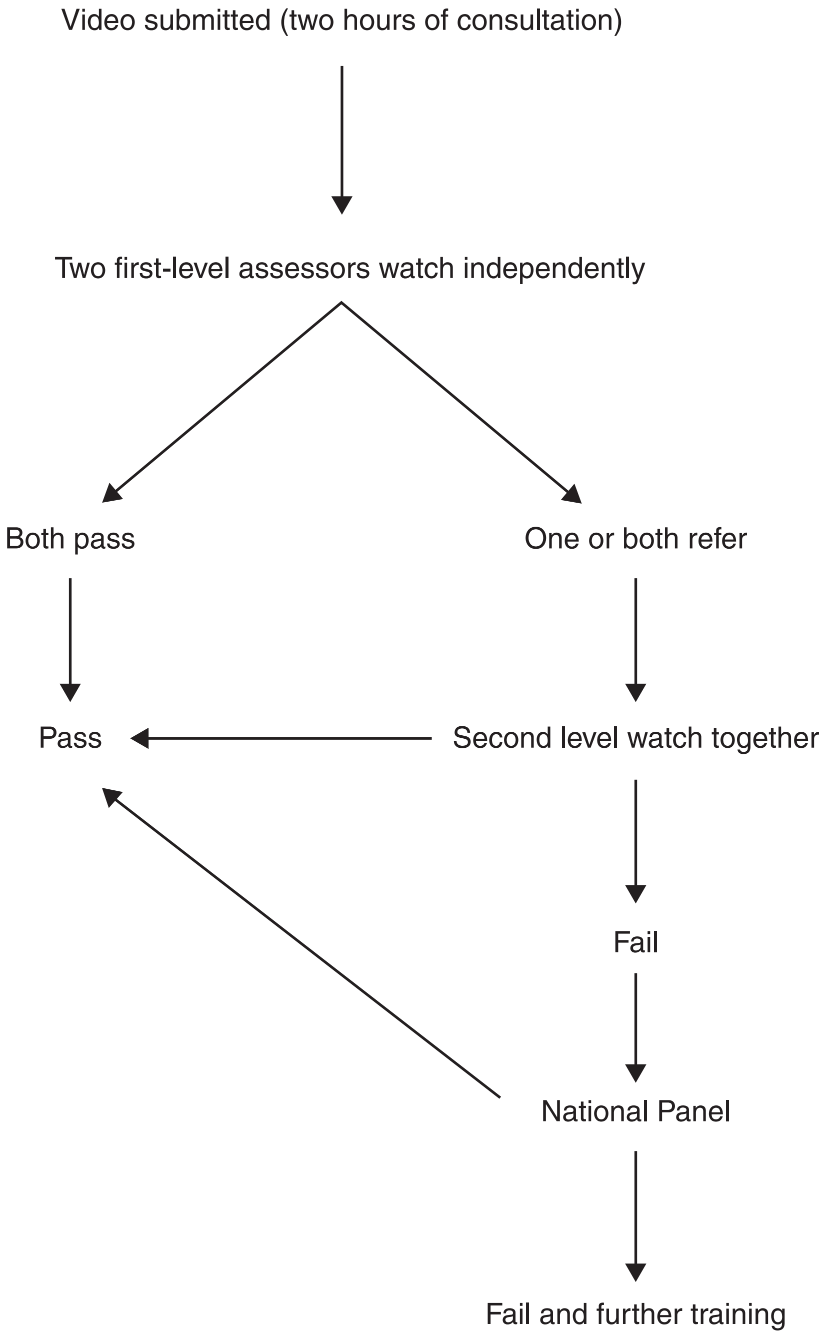

Video submitted (two hours of consultation)
Two first-level assessors watch independently
Both pass
One or both refer
Pass
Second level watch together
Fail
National Panel
Fail and further training

Figure 10.1: Video assessment process.

All assessors are trained and experienced general practitioners. At first level the assessors watch the tape independently having selected a minimum of six consultations from the material submitted. The judgement is as described above and there is a referral rate of 20% of tapes viewed to second level.[16] The role of the second level assessors is to look at material referred by first level and make a judgement as to whether the registrar is above or below minimal competence. They watch the tape together and judge the consultation independently before discussion. Second level must agree the final pass/fail decision. If they fail a tape it is then sent to another pair of second level assessors from a different deanery who are called the National Panel. The process they follow is the same as that described for the second level assessors but their role is to ensure fairness and equity. Their decision is the final one and they can overturn a second level judgement.

Critique of the system

The system developed has shown to be feasible and thousands of registrars have now undertaken the COGPED method of assessment. In terms of validity the video is a way of observing the doctor at work and does look at real performance rather than surrogate measures such as patient satisfaction or trainer opinion and therefore can claim face validity. Outcome validity would require a longer term follow-up with large numbers of registrars who had been deemed competent by different assessment methods.

The original research used videotapes from ten registrars and these were assessed independently by 25 assessors. The conclusions were that using two assessors for registrar tapes produced a 95% probability of identifying an unsatisfactory registrar (sensitivity), while identifying 20% of satisfactory registrars (specificity). In today's assessment climate the reliability of the tool would need to be determined, inter-rater reliability and the relationship between individual consultation judgement and global judgements. The original work did show that assessors' judgements were unlikely to change after watching four consultations.

All first level and second level assessors are calibrated on an annual basis. Protocols for maintaining assessor competence both at first and second level have been developed and are used as part of the quality assurance mechanism for deanery visits.

To monitor the fairness of the system a process of quality control was devised and has evolved over time. First level quality control is carried out by a small number of first level assessors and has shown consistency in their judgements and has been specifically calibrated for this purpose. The same process was carried out to identify second level assessors to participate in quality control. Initially a sample of one in ten first level deanery passes and one in five second level deanery passes were selected. From October 2000 the sampling was increased for first level passes to one in eight and for the first time all National Panel passes were included in the quality control process.

The accuracy of the data, the tracking and monitoring of the summative assessment system was the deaner's responsibility and is dependent on a good administrator and database system. The whole system is monitored nationally by a single summative assessment administrator who liaises with the deanery administrators in the collection of the appropriate data. Bi-annual feedback is sent to the Joint Committee that has now been superceded by the Postgraduate Medical and Education Training Board and to the directors on their deanery performance.

National results

From 1 October 1996 until the 31 March 2005, 16,410 general practice registrars have undergone the summative assessment process. The number of registrars sitting the MRCGP single route has increased year on year as has the pass rate and those who gained merit. The implication for this is that there has been a decrease in the number of tapes coming through the COGPED system and deaneries will therefore need to look at the impact of this on the calibration of their assessors (*see* Table 10.1).

Table 10.1: MRCGP single route

Diet	*Number of GPRs*	*% of GPR eligible*	*% passed MRCGP*	*% of passes who gained merit*
Spring 01	570	50.5	81.6	17.6
Autumn 01	187	36	79.1	19.6
Total 2001	***757***	***43.3***	***81***	***18.1***
Spring 02	686	53.6	73.6	20.4
Autumn 02	313	53.3	75.4	22.9
Total 2002	***999***	***53.5***	***74.2***	***21.2***
Spring 03	864	64.8	81.4	27
Autumn 03	399	46.4	83.2	30.7
Total 2003	***1263***	***55.6***	***81.9***	***28.2***
Spring 04	1010	64.8	84.1	25.4
Autumn 04	520	60.8	81.9	26.3
Total 2004	***1530***	***62.8***	***83.3***	***25.7***
Grand total	***4549***	***53.8***	***80.5***	***23.3***

As trainers and assessors have become more confident in the system the number of failures purely related to the video has increased (*see* Table 10. 2).

Table 10.2: Number of fails related to video

Year	*No. (%)*
1996/97	22 (2)
1997/98	31 (2)
1998/99	40 (3)
1999/2000	46 (2)
2000/01	29 (2)
2001/02	49 (4.9)
2002/03	41 (5.1)
2003/04	50 (5.8)
2004/05	53 (6.4)

COGPED video contributes to 33% of the overall failures in summative assessment.

The number of registrars choosing to take the simulated patient surgery (SPS) is small. In 2003/04 172 (8.7%) undertook this process and in 2004/05 123 (5.6%). The SPS contributed to 5% of summative assessment failures in 2004/05.

Table 10.3: Quality control

	First level			Second level			National Panel		
	Number	*Agree (%)*	*Disagree (%)*	*Number*	*Agree (%)*	*Disagree (%)*	*Number*	*Agree (%)*	*Disagree (%)*
1996/97	101	95 (4)	6 (6)	34	30 (88)	4 (12)	–	–	–
1997/98	67	61 (91)	6 (9)	20	18 (90)	2 (10)	–	–	–
1998/99	107	99 (93)	8 (7)	43	41 (95)	2 (5)	–	–	–
1999/2000	141	130 (92)	11*(8)	40	37 (92.5)	3 (7.5)	8	7 (88)	1 (12)
2000/01	78	68 (87)	10 (13)	31	30 (96.8)	1 (3.2)	5	5 (100)	0 (0)
2001/02	59	53 (89.8)	6 (10.2)	22	22 (100)	0	9	9 (100)	0
2002/03	82	70 (85.4)	12 (4.6)	20	20 (100)	0	20	6 (54.5)	14 (45.5)
2003/04	51	46 (90.2)	5 (9.8)	17	16 (94.1)	1 (5.9)	4	4 (100)	0
2004/05	57	52 (91.2)	5 (8.8)	27	26 (3.3)	1 (8.7)	18	15 (88.2)	2 (11.8)

* Three should have been rejected on quality grounds; and one rejected as low challenge.

Quality control

As described earlier a selection of passes from first level, second level and National Panel were independently reviewed. Table 10.3 shows the results from 1996 until 2005. Where a professional judgement has been made it is not always possible to have 100% agreement but it shows over time there is consistency in judgement at the different levels. Any deaneries where aberrant behaviour is found has this information fed back to the assessors and if necessary receives a visit from the Summative Assessment Administrator.

The results of submissions to National Panel from second level are shown in Table 10.4. Again this shows that as the system has become more established there is increasing agreement between second level and National Panel. The role of National Panel is to moderate behaviour of second level assessors and the results would show that they are carrying out this role effectively. The third aspect of quality control is that of MRCGP passes. The system of assessment of the MRCGP and summative assessment are different in that the MRCGP looks at performance whereas summative assessment looks at competence. Initially when the MRCGP single-route system was put in place, a sample of one in four MRGP passes was put blindly through the COGPED mechanism. The numbers identified that would possibly have failed the summative assessment route were very small and in 2004 it was decided to reduce the sampling to one in eight. When this occurred there was an initial rise in summer 2004 to 5.1% and this will need to continue to be monitored (*see* Table 10.5).

Table 10.4: Results of video submission to National Panel (1 October 1996 to 30 September 2001)

	No. referred (%)	*Agree* (%)*	*Disagree** (%)*
1996/97	84 (6.7)	63 (75)	21 (25)
1997/98	74 (4.9)	54 (74)	19 (28)
1998/99	89 (5.6)	65 (73)	24 (27)
1999/2000	98 (5.8)	89 (91)	9 (9)
2000/01	86 (5.2)	63 (73)	23 (27)
2001/02	56 (6.3)	44 (79)	12 (21)
2002/03	82	69 (84)	13 (16)
2003/04	75	66 (88)	9 (12)
2004/05	103	89 (89)	14 (11)

* i.e. candidate should fail; ** i.e. candidate should pass.

Table 10.5: Quality control of MRCGP single route*

Videos that passed the MRCGP→ COGPED quality control marking as part of the 'blinding' process				
	Total	*1st level pass*	*2nd level pass*	*2nd level fail*
Summer '04	98	68 (69.4%)	25 (25.5%)	5 (5.1%)
Autumn '04	51	36 (70.6%)	14 (27.5%)	1 (2%)
Total	**149**	104 (69.2%)	39 (26.2%)	6 (4%)

* Since starting the quality control system the numbers have been very small and it was decided to decrease the sampling from one in four to one in eight in the last 12 months.

Structured Trainer's Report

The final component of summative assessment process is the Structured Trainer's Report.[18,19] The Trainer's Report is the only instrument that attempts to assess a registrar's practical skills. The Structured Trainer's Report has 35 competencies in three broad areas.

1 Specific clinical skills, e.g. using an auroscope, carrying out a vaginal examination or gaining venous access.

2 Patient Care. Looking at:

- making a diagnosis
- patient management
- clinical judgement.

3 Personal skills. Incorporating:

- organisational skills
- professional values
- personal and professional growth.

The assessment can be performed in a number of ways which can include direct observation either by the trainer or another member of the primary care team, case analysis, tutorials or using simulated patients or mannequins.

Completion of the Trainer's Report should be carried out throughout the year. For each competency there is a description of fail criteria to help guide the trainer. The trainer should document sufficient information for each element to enable him/her to make a judgement as to whether the GP registrar has achieved a standard for independent practice at the final assessment stage. For a registrar to pass the trainer should have satisfied him or herself that the registrar has minimal competence in all elements.

Content validity of the Structured Trainer's Report was carried out using a sample of doctors who recently completed the training and there was agreement that it did contain items considered to be important by recently trained doctors.[20] However, there was concern expressed whether a report was the right medium for assessment for some of the items. This was borne out by further work which demonstrated a discrepancy between the judgements and the Trainer's Report and the doctors' abilities to carry out certain clinical procedures.[21]

There is no doubt that the report has allowed trainers to make a much more informed decision as to whether or not a registrar is ready for independent practice. GP registrars can fail on the Trainer's Report alone but failure is more often linked to fail in one of the other modules. In the first five years the Trainer's Report contributed to 33% of summative assessment fails. This has steadily increased year on year as trainers have become more confident in the system and in 2004/05 15% of registrars failed the Trainer's Report alone but passed other components.

While the content validity of the Structured Trainer's Report has been looked at in detail there is no reliability data and there is no quality control mechanism for the Trainer's Report. A new Trainer's Report is being developed that will be informed by workplace based assessment, video and case-based discussion. Work is currently being carried out to pilot this and look at its reliability.

Inclusion of an audit project

Background

In 1992 the JCPTGP decided that at the end of their training a doctor should have, in addition to other things, 'adequate skills in producing a written report of practical work in general practice'.[22] The word 'adequate' was not defined. When the JCPTGP published its policy document on summative assessment there was considerable debate about the need for a broad range of options to represent practical work.[23] Examples given were: literature reviews, business plans or a piece of research being carried out during the hospital component of vocational training (Toby J, personal communication, 1994).

A written report revealed the ability to communicate an idea or concept which might promote change. Trainees are exposed to many examples of written reports of practice work during the training years. As advocates for their patients many written reports may have crucial implications. Appropriately written referral letters and legal reports are two examples. The urgency with which they are dealt may depend on the manner in which they were written. A badly prepared or poorly written report would therefore be deemed a demonstration of competence below the standard acceptable of a practising general practitioner. The argument for including a broad range of material in the report of practical work was therefore persuasive.

Balanced against this, however, was the need for a fair assessment. Consistency in the material being submitted was therefore seen as an overriding necessity if a fair assessment was to be achieved. An audit was seen to be a method of identifying learning needs[24] and could be useful in problem solving.[25] Data collection, awareness of relevant literature, negotiated teamwork and discussion of change all involved a certain amount of action on the part of the trainee and could therefore be justified as practical work. Committing the audit to a written format helped to focus on the need for change where such change had been clearly identified. The choice of subject for the audit project tested whether the doctor was able to balance the importance of the topic with the feasibility of investigating the quality of care in the time available. In essence, the trainee was demonstrating his or her ability to monitor and, if required, to improve the quality of care being provided, described by the GMC as, 'a basic principle of good practice'.[26] It was strongly argued that failure to demonstrate an example of this principle was accepted as being important enough to require a period of extra training to ensure that audit method was understood as judged by the successful submission of an audit project.

Assessing an audit project

Irvine[27] advised that an audit project should include the following:

- subject of audit
- background
- reason for the audit
- methods
- results
- changes recommended
- repeat audit, if possible.

Crombie and Davies[28] described the need to answer three questions in any data collection exercise.

- Why was it done?
- How was it done?
- What did it find?

In order to introduce trainees to considering the management of change, a fourth question was added.

- What next?

A series of discussions involving trainers and trainees in the west of Scotland established 14 elements considered to be part of an audit project appropriate for training in the mid-1990s. As a result of a content validity exercise carried out by 155 trainers, ten elements were found to receive agreement from at least 80% of them. The outcome was that collecting a second set of data for an audit project was felt to be outwith trainers' experience and confidence. The final assessment instrument[29] consisting of five criteria with suggestions for rather than evaluation of change is shown in Box 10.1.

Box 10.1: Summative assessment audit – five-criteria marking schedule

Please tick the box provided if the criterion for answering each question is/are present.

Question	*Criterion*	*Criterion present*
Why was the audit done?	*Reason for choice* Should be clearly defined and reflected in the title. Should include potential for change.	□
How was the audit done?	*Criteria chosen* Should be relevant to the subject of the audit. Should be justified, e.g. literature.	□
	Preparation and planning Should show appropriate teamwork and methodology in carrying out the audit. If standards are set they should be appropriate and justified.	□
What was found?	*Interpretation of data* Should use relevant data to allow appropriate conclusions to be drawn.	□

What next?	*Detailed proposals for change* Should show explicit details of proposed changes.	□

A satisfactory trainee audit report should include all five criteria to pass.

Please enter your opinion in the box provided.

Pass □
Refer □

If 'refer', please comment on your reasons.

Using three assessors was found to give the optimum balance of sensitivity and specificity for an assessment system with referral for further assessment if one or more of the three felt that a project should be referred.

A factor analysis carried out on 333 5-criteria audit projects showed that 69% of the total sample variance was explained by two factors, namely how the audit was done, i.e. the methodology and why the subject for the audit was chosen, i.e. the reason for the audit. The consequential validity or educational impact confirmed the positive experience for trainees in submitting an audit project given that for most this was their first ever experience of audit.[30]

Moving to the completed audit cycle – evaluating rather than proposing change

Between 1996 and 1997 an increasing number of registrars was evaluating the change they had proposed in their audit project for summative assessment. By completing an audit cycle they were going beyond what was expected from the five criteria against which their project was being assessed. Thus within four years of its implementation in the west of Scotland the confidence which previous registrars had expressed from submitting an audit project was being translated – at least by a sizeable minority – into a peer motivated rising of standards in the completion of their project. As a result of legislation implemented in 1998 a total of seven competences were required to be achieved through summative assessment.[31] The submission of a criterion audit project now required that a registrar had to demonstrate that he/she had acquired the ability to review and critically analyse the practitioner's working practice and manage any necessary changes appropriately.

The five-criteria marking schedule in use satisfied the first part of the competency definition but fell short of managing the change process beyond suggesting proposals. The implication in the definition was that a registrar should be able to complete an audit cycle and, as a proportion of registrars in the west of Scotland was attempting to demonstrate, the assessment process would need to be modified to assess more closely the competence defined in law.

When the original 14 elements were again discussed with the trainers in the west of Scotland it was found that two collections of data were now felt to be achievable, as a result of increased confidence by the trainers. Adequate sensitivity and specificity of the assessment system was achieved using two assessors rather than three and validity and reliability checks on both the instrument and the assessment system were reassuring. The resulting eight-criteria marking schedule is shown in Box 10.2.[32]

Box 10.2: Summative assessment audit – eight-criteria marking schedule

Please tick the box provided if the criterion for answering each question is/are present.

Criterion		*Criterion present*
Reason for choice of audit	Potential for change Relevant to practice	□
Criterion/criteria chosen	Relevant to audit subject and justifiable (e.g. current literature)	□
Standards set	Targets towards a standard with a suitable timescale	□
Preparation and planning	Evidence of teamwork and adequate discussion where appropriate	□
Data collection (1)	Results compared against standard	□
Change(s) to be evaluated	Example supplied	□
Data collection (2)	Comparison with data collection (1) and standard	□
Conclusion	Summary of main issues (e.g. bullet points)	□

A satisfactory registrar audit project report should include all eight criteria to pass.

Pass □
Refer □

If 'refer', please comment on your reasons overleaf.

Pragmatism in the light of trainers' varying ability to teach audit therefore allowed for a move to incremental change from a five-criteria definition of audit with suggestions but no actual change to an eight-criteria definition where change had to be effected and evaluated. Registrars who were more innovative and confident led this change.

National implications

Between September 1 1996 and March 31 2005, 15,442 registrars in general practice in the UK had submitted an audit project for summative assessment. 1,716 registrars (11%) required two attempts to demonstrate their competence in this process. 98 of these failed after re-submission and had to undergo extra training the length of which was at the discretion of the director of a deanery. A project which failed had had six (eight criteria) or seven (five criteria) independent assessments within the deanery and a further two outwith the deanery before further training was recommended. The implications of this suggested that experience in the west of Scotland was not an isolated phenomenon and that the teaching of audit methods based on actual experience could not be assumed.

The research underpinning the development of assessing an audit project was recognised by Hutchinson *et al.*[33] Of the 55 papers identified from 1985 to 2000 in a systematic review only two had tested consequential validity, one of which formed part of the work described earlier. A study by Bowie *et al.* considered the predictive validity using GP non-principals in the west of Scotland.[34] He showed that there was a significant difference in knowledge of audit method and skills for those who entered general practice before and after the introduction of summative assessment.

JCPTGP and criteria for audit for training practice

In recognition of the problems addressed between 1990 and 2000 the JCPTGP decided that two new criteria for audit should be implemented from September 2000. More practical audit was required to be in place in training practices if the GMC competence covering clinical audit was to be addressed as it would be required for revalidation and clinical governance. The criteria were as follows.

- Training practices must demonstrate that the audit process is being taught.
- Training practices must have in place an active programme of audit which demonstrates the full audit cycle and the application of both standards and criteria.

Reflections for the future

The one concern which has not been addressed since the introduction of the audit project as part of summative assessment in 1996 is the fact that a stubborn 12% of submissions required further work. A recent attempt to stop a re-submission and ensure that adequate teaching was being carried out in the training practices concerned was not able to be carried through.

It is likely therefore that some trainers still lack the confidence in basic criterion audit method to ensure that a registrar has a fair chance of passing this component of summative assessment with minimum effort. At the time of writing it is uncertain whether criterion audit will continue as part of a wider workplace based assessment forming part of the new MRCGP. It is certainly the opinion of these authors that without the introduction of an audit submission for summative assessment, criterion audit method would not have been subjected to such scrutiny and, given its importance in appraisal and revalidation, prospective general practitioners would not have had the opportunity to ensure they understood the basic criterion audit method. Given the increasing move to computerisation with the quality and outcomes framework it is uncertain whether this degree of scrutiny will continue.

References

1. Irvine DH, Gray DJD, Bogle IG. Vocational training for general practice: the meaning of 'satisfactory completion' (letter). *BJGP.* 1990; **40**: 434.
2. Campbell LM. The Development of a Summative Assessment System for Vocational Trainees in General Practice. MD Thesis. Glasgow: University of Glasgow; 1997.
3. Calman KC (Chairman). *Maintaining Medical Evidence: review of guidance on doctors' performance. Final report.* London: DoH; 1995.
4. Campbell LM, Howie JGR, Murray TS. Summative assessment: a pilot project in the west of Scotland. *BJGP.* 1993; **43**: 430–4.
5. Campbell LM, Murray TS. Assessment of competence. *BJGP.* 1996; **46**: 619–22.
6. Livinstone SA, Zieky MJ. *Passing Scores*. Princeton, NJ: Education Testing Service; 1982.
7. De Gruijter DNM. Compromise models for establishing examination standards. *J Educ Meas.* 1985; **22**: 263–6.
8. Rethans JJ, Sturmans F, Drop R, van der Vleuten C. Assessment of the performance of general practitioners by the use of standardized (simulated) patients. *BJGP.* 1991; **41**: 97–9.
9. Rethans JJ, Drop R, Strumans F, van der Vleuten C. A method of introductory standardized (simulated) patients into general practice consultations. *BJGP.* 1991; **41**: 94–6.
10. Hays RB. Assessment of general practitioner consultations: content validity of rating scale. *Med Educ.* 1990; **24**: 110–16.
11. Cox J, Mulholland H. An instrument for assessment of videotapes of general practitioners' performance. *BMJ.* 1993; **306**: 1043–6.
12. Fraser RC, McKinley RK, Mulholland H. Consultation competence in general practice: establishing the face validity of prioritised criteria in the Leicester assessment package. *BJGP.* 1994; **44**: 109–13.
13. Ferrell BG. Clinical performance assessment using standardized patients. *Fam Med.* 1995; **27**: 14–19.
14. Williams RG, Barrows HS, Vu NV *et al.* Direct, standardized assessment of clinical competence. *Med Educ.* 1987; **21**: 482–9.
15. Campbell LM, Sullivan F, Murray TS. Videotaping of general practice consultations: effect on patient satisfaction. *BMJ.* 1995; **311**: 236.
16. Campbell LM, Murray TS. Summative assessment of vocational trainees: results of a three-year study. *BJGP.* **46**: 411–14.
17. Alton J, Evans A, Foulkes J, French A. Simulated surgery in the summative assessment of general practice training. Results of a trial in the Trent and Yorkshire regions. *BJGP.* 1998; **48**: 1219–23.
18. Johnson N, Hasler J, Toby J, Grant J. Consensus minimum standards for use in a trainer's report for summative assessment in general practice. *BJGP.* 1996; **46**: 140– 4.
19. Johnson N, Wasler J, Toby J, Grant J. Contents of a trainer's report for summative assessment in general practice: views of trainers. *BJGP.* 1996; **46**(404): 135–9.
20. Johnson N, Hasler J. Content validity of a trainer's report: summative assessment in general practice. *Med Educ.* 1977; **31**(4): 287–92.
21. Kelly MH, Campbell LM, Murray TS. Clinical skills assessment. *BJGP.* 1999; **44**: 947–50.
22. Joint Committee on Postgraduate Training for General Practice. Assessment Working Party. *Interim Report*. London: Joint Committee on Postgraduate Training for General Practice; 1992.
23. Joint Committee on Postgraduate Training for General Practice. *Report of Summative Assessment Working Party.* London: Joint Committee on Postgraduate Training for General Practice; 1992.

24. Coles C. How students learn: the process of learning: In: Jolly B, Rees L (eds). *Medical Education in the Millennium*. Oxford: Oxford University Press, 1988. p. 63–82.
25. Baker R. Problem solving with audit in general practice. *BMJ.* 1990; **300**: 378–80.
26. General Medical Council. *Good Medical Practice.* London: GMC; 1996. para 6.
27. Irvine DH, Irvine S. *Making Sense of Audit*. Oxford: Radcliffe Publishing; 1991.
28. Crombie IK, Davies HTO. Towards good audit. *Brit J Hosp Med.* 1992; **48**:182–5.
29. Lough JRM, McKay J, Murray TS. Audit and summative assessment: a criterion-referenced marking schedule. *BJGP.* 1995; **45**: 607–9.
30. Lough JRM, McKay J, Murray TS. Audit and summative assessment: two years' pilot experience. *Med Educ.* 1995; **29**: 101–3.
31. National Health Service (vocational training for general medical practice). *Regulation 9(2). Statutory Instrument 3150*. London: HMSO; 1997.
32. Lough JRM, Murray TS. Audit and summative assessment – a completed audit cycle. *Med Educ.* 2001; **35**: 357–63.
33. Hutchinson L, Aitken P, Hayes T. Are medical postgraduate certification processes valid? A systematic review of the published evidence. *Med Educ.* 2002; **36**: 73–91.
34. Bowie P, Garvie A, Oliver J, Lough M. Impact on non-principals in general practice of the summative assessment audit project. *Educ Prim Care*. 2002; **13**: 336–9.

Chapter 11

Roles of self-assessment tools and work-based learning in personal development

Jonathan Burton and Neil Jackson

Introduction

In this chapter we will start by discussing self-assessment in general: we will then move on to a discussion of self-assessment in medical practice. In this section we will cover the question of motivation, the standards against which self-assessment takes place, contrast the daily practice of self-assessment with the occasional application of third party assessments, and discuss the relationship between self-assessment and reflective practice.

We will then go on to discuss the shortcomings of self-assessment, its openness to self-deception and other operational errors and what is understood about these. Having discussed the problems of validity, we will move on to suggesting how students can be trained in such a way that they learn the technique of making self-assessments more objective.

Then we will discuss the tools that can be used in self-assessment and comment on their strengths and weaknesses. We will then deal with the question of the public's input into GP assessment.

Finally, we will discuss workplace-based assessment, dealing with new approaches to the workplace-based assessment of individual and team performance.

Self-assessment

There are so many influences on our lives. So many things make us change and develop. Amongst these factors are our personality, our drive, our curiosity, our capacity to recognise problems and solve them and, most of all, how well we are able to build on those capabilities that we gained from our initial training. This latter quality is what Claxton has called 'learnacy'.[1]

In this age of accountability, there is a strong temptation to under estimate these self-driven characteristics. In part, this is because each of us understands them but vaguely. Such self-driven characteristics are rarely expressed when people describe their learning and their practice. Rather people talk in terms of how the outside world (lectures, courses and so on) can help them make up for their deficiencies. Do not think that we are belittling the role of such arranged education; this chapter begins by asking what it is about self-assessment that does make a contribution to the development of our lives.

Self-assessment is not simply an introspective process. It is a process where introspection, that inner running commentary on what we are doing, is guided by our contact with the outside world.

Writing in his autobiography *Childhood, Boyhood and Youth*, Tolstoy described his boyhood discussions with Katya, a poorer relative of his, and of whom he was very fond.[2] He remembered as he wrote down an account of events many years later how she had made him understand the concept of poverty, by contrasting the circumstances of her family with those of her own.

> Has it ever happened to you, dear reader, at any point in your life to become aware all at once that your outlook on things has completely changed, as though all the objects that hitherto been before your eyes had suddenly presented to you another, unfamiliar side? Such a volte-face occurred for me for the first time during that journey of ours, from which I date the beginning of my boyhood.
>
> For the first time I envisaged the idea that we – that is, our family – were not the only people in the world, that not every conceivable interest was centred in ourselves but that there existed another life – that of people who had nothing in common with us, cared nothing for us, had no idea of our existence even. I must have known all this before, but I had not known it as I did now – I had not realised it; I had not felt it.

Tolstoy says that he would not have undergone this major transformation in his thinking and feeling without his conversation with Katya to guide him. We can all think of similar major transformations in our lives. Important as such major transformations are, daily working life is full not of major events like this, but of myriad, minor ones and the self-awareness which is the foundation of self-assessment is key to these.

Self-assessment in medical practice

Motivation and other personal factors

Motivation and other personal factors are to be recognised as the great drivers to self-improvement in professional practice. GPs start their working lives having had one of the longest periods of training of anyone in society. But, thereafter, there is not much formal training and much of the change and development that occurs in our professional lives after qualification as GPs is down to a number of personal factors. For most doctors self-assessment is a continuous and daily process, a self-critical reflection on what has been done in the course of work. Manning and DeBakey interviewed outstanding doctors in a variety of settings.[3] They showed that a driving motivation was, 'a desire never to be (or be seen as) professionally inadequate.' In the physician change study Fox *et al.* showed that the desire to be ever more competent was the prime motivator.[4]

Self-assessment against standards

It is widely recognised that self-appraisal has to be undertaken against standards and benchmarks of practice. Whilst self-motivation is the driver for the continuous improvement in a doctor's competence and performance, common sense dictates that self-assessment must be balanced by other forms of assessment, such as peer review, and also must be informed by a knowledge of and respect for the standards of the day.[5] As Boud says, assessment can be conducted only against benchmarks and criteria.[6] Brown *et al.* highlight the distinction between the self-appraisal which is undertaken to fulfil the expectations of a formal appraisal and that which is undertaken solely in the interest of self-development through reflection.[7] In other words they distinguish between the self-appraisal which submits itself to public accountability and the self-appraisal which is, largely, a private activity for the individual doctor.

Self-assessment is continuous; external assessment is brief

Whilst there are many hours of practice each year, each hour offering opportunities for learning, peer appraisal or third party appraisal only takes place for a short period – in the case of the GP annual appraisal for one or two hours per year. For this reason, self-assessment is a vital tool to ensure that daily practice is put under the spotlight of critical self-reflection.[8] Evans *et al.* say, 'a successful CPD programme demands awareness of remediable weaknesses through continual self-appraisal.'

Self-assessment and reflective practice

Schon, writing about mature professionals, analysed aspects of reflective practice.[9] He suggested that there is an aspect of practice which is not based on book knowledge and which can be observed in the way that professionals handle actual cases. Schon cites the case (page 64) of the ophthalmologist whose patient had uveitis, due to a parasitic condition, and glaucoma, the treatment of each being different. In fact the treatment of the uveitis would make the glaucoma worse and *vice versa*. How to solve such a conundrum was not described in the textbooks and the ophthalmologist had to discover a way of helping this patient, which was:

- to remove all treatment
- at which point he discovered that the glaucoma went into remission – and had in fact been caused by the treatment for the uveitis
- titrate upwards the treatment for the uveitis so that it was satisfactorily dealt with without causing the glaucoma to re-start.

Schon described this aspect of professional practice, this use of judgement and trial and error, as a form of artistry. He compared this form of artistry with that of the tight-rope walker: a unique skill which is performed using a whole series of minor adjustments in order to attain the end in view. To Schon, this vision of professional practice depended on the capacity of the professional to think on his or her feet – what Schon called 'reflection-in-action'.

Schon also defined another form of reflection, something he called 'reflection-on-action'. This was a form of reflection that occurred after an episode of professional practice. Perhaps the professional met with colleagues and discussed his or her cases and was able to reflect on how he or she had handled them, this reflection being helped by the input of peers. Or, the same post-event reflection might occur in the writing up of a professional portfolio.

Self-appraisal, then, can be seen as an innate and necessary part of daily professional practice. It should be promoted widely as a vital part of professional development. It has, however, to stand in relation to benchmarks and external criteria.

Why can't we leave doctors to their own devices, or can we?

Self-assessment is part and parcel of professional life. But, clearly, self-assessment has its limitations, if the purpose of assessment is to show that the doctor is adequate in his or her practice.

Errors in self-assessment

In a study into self-assessment, undertaken by Caputo and Dunning, but not in a medical community, participants were asked to define a number of solutions to different problems, such as word games.[10] Caputo and Dunning showed that participants could not judge their own ability accurately. Participants were able to find some solutions to the problems, but they had little insight into their errors of omission. An error of omission was defined as a solution they could have generated to problems but missed. The authors then showed the participants all the solutions to the problems (the ones found and the ones missed). The participants agreed that the solutions they had missed were just as important as the solutions they had found. The authors proposed that accurate self-evaluation did depend in part on such third-party feedback.

It is the weaker candidates in medical school who tend to overrate themselves.[11] Others have shown that brighter medical students produced more conservative self-evaluations.[12,13] These findings are open to a number of interpretations. Evans *et al.* suggest that it may be that high achievers hold themselves to more stringent standards.[14] They list a number of other reasons for inaccuracy in self-assessment:

- misapprehension – not understanding what is expected of one
- self-deception
- scoring of potential or ideal performance rather than actual performance
- scoring of effort rather than achievement
- compensation for poor performance as a defence mechanism.

Evans *et al.* discussed how students might be helped to develop skills in self-assessment, in a way which brought into play some awareness of these problems and discrepancies.[8] They suggested that medical students should be trained to calibrate their own self-assessments against objective assessments and then think about the discrepancies. They also suggest that the students should be trained to become aware of their emotional reactions to self-assessments – in a way which would make them more aware of when they were self-deceiving or self-compensating. Evans *et al.* summarised their ideas about how students should be trained in self-assessment under four headings (*see* Box 11.1).[8]

Box 11.1: Students should be trained in self-assessment to:

- promote reflection on personal performance
- identify reactions to self-assessment
- evaluate the reliability of the self-assessment
- identify the reasons for the discrepancies between scores of outside assessors and those of self assessors.

Purported errors

Some of the so-called errors of self-assessment are caused by problems with the methodology of the assessment. All forms of assessment are fraught with problems of validity (does it test what it claims to test?) and other methodological constraints. A paper was published which cast self-assessment in a bad light.[15] The participating GPs were firstly asked to make an assessment of their own competence in a number of clinical areas of practice, including the management of thyroid disease. Subsequently they undertook a test of factual knowledge about thyroid disease, a test which had been devised by a panel of GPs and endocrinologists. The authors showed that there was poor correlation between the subjects' self-assessment of their knowledge and their later performance in the knowledge test. The authors of the paper felt that they were able to make the following claims.

- Doctors' perception of knowledge in areas of common practice is no indication of actual knowledge.
- Continuing medical education and other professional development activities that rely on the doctors' self-perception to assess their needs are likely to be seriously flawed.
- To make professional development activities more efficient and effective a more objective assessment of needs is necessary.

There was an interesting discussion of this paper in the *BMJ*. A salient question in the discussion was whether the knowledge test measured what was important for daily practice. It could be said, and this is what the critics of the paper argued, that the knowledge tests, of which an example was given in a paper, were inappropriate for workday GP practice. What the critics of this paper were questioning was the validity of the assessment method – did it test what a GP usually needs to know to do his or her job capably? This is an argument about the validity of the assessment instrument. An assessment instrument is only valid if it tests what it sets out to test – in the case of this paper the critics argued that the knowledge test did not test the sort of knowledge that GPs need to do their job capably.

In summary, self-appraisal is constrained by the difficulties individuals have in making objective assessments of their own performance. A way round this problem is to ensure that doctors develop the sophistication of their self-assessments and, as Evans *et al.* argue, doctors in training should be taught the tools of accurate self-assessment.[8] Self-assessment is also constrained by problems of methodology.

The use of tools in self-assessment

Whilst self-assessment is vital to professional development, its faults and weaknesses, as discussed above, are likely to be fairly obvious to those who have an interest in making sure that GPs are safe and capable. The objectivisation of the assessment process is key to this debate.

Self-assessment can be made more objective. There is already a broad range of methods of assessment.[16] These are, 'used extensively and systematically to determine clinical competence and performance in healthcare professionals.'

We listed a variety of assessment methods that can be applied in the work context to enhance clinical competence and performance (*see* Box 11.2).

Box 11.2: Assessment methods that can be applied in the work context to enhance clinical competence and performance[16]

- Confidence rating scales
- Attitudinal questionnaires
- Sitting in
- Review of consultation record
- Video analysis of consultations
- Feedback from patients, colleagues and staff
- Random case analysis
- Problem case analysis
- Preparation of teaching sessions
- Various forms of knowledge testing, increasingly done on line
- Project and audit work
- Practice exchange visits

Strengths and limitations

By and large, these tools are useful in helping GPs to make their self-assessments more objective. Their strengths lie in their convenience, and in the amount of change that is reported by those that use them.[17] But many of these approaches are difficult to organise and some practitioners will be nervous about using the tools in which their practice is exposed, in ways to which they are unaccustomed. Even in the more widely undertaken approaches, there are some limitations. For example, there is no evidence that the widespread use of on line knowledge testing is being translated into changes in practice.[18] Some of the assessment methods measure knowledge, but not performance (what actually is done at work). Some of the assessment methods included in this list are based on self-assessment only, with all of its strengths and weaknesses already discussed in this chapter. Some of the methods involve input from professional colleagues, but only one involves feedback from the users of healthcare.

Public input into the assessment of GPs

The general public as consumers of medical care would like appraisal to have input from lay experts. A recent Mori poll (www.mori.com/polls/2005/doh.shtml) in the UK shows that the general public firmly believes that the regular appraisal of GPs should be undertaken by a mixture of lay experts and other doctors. Indeed, many of those interviewed in the course of the poll could choose aspects of practice in respect of which *they* would be able to give feedback to their GPs (*see* Table 11.1).

In a patient-led NHS, the views of patients and the willingness of patients to comment on their GPs in the ways set out in Table 11.1 are going to be taken seriously by government.

Table 11.1: Aspects of GP practice about which the lay public would like to comment

Aspect of practice the general public would like to give feedback on:	*% of respondents who would comment on this aspect*
communication skills	53%
has kept up with new developments	36%
involving patients in treatment decisions	36%
according dignity and respect to patients	35%
knowledge and/or technical ability	33%

Workplace based assessment or the assessment of performance at work?

Norcini has said that the shift of assessment away from education and towards work is a reflection of the drive towards public accountability and quality improvement.[19] Workplace based assessment as a universal approach is in its early years. The assessment of performance at work does not have to take place at work, although it may do. Just as work based learning can take place away from work (for example, most doctors who keep written portfolios write these up at home), so the assessment of performance at work does not only rest on the inspection of evidence in the workplace.

Indeed it is only recently that *all* GP practices have been subjected to practice inspection visits in the UK. Previously, practice inspection visits which measured work based performance would have been undertaken for those who volunteered to do something extra – to become trainers of young GPs or to be tested for excellence by a professional body (for example, in the Fellowship by Assessment of the Royal College of GPs).

Now, however, all individual GPs and all primary care teams are subject to routine assessments or appraisals. At the time of writing (autumn 2005) these approaches to assessment or appraisal are still in their infancy and may well be subject to considerable change in the next few years. Table 11.2 shows how these assessments or appraisals relate to self-assessment.

Table 11.2: The role of self-assessment in workplace-based assessment in UK general practice

Type of assessment	*Role of self-assessment*	*Is there external assessment against practice records?*
GP appraisal	The preparation is by self-assessment, the discussion is by peer assessment	No
Practice assessment under new General Medical Services contract	The process is entirely by third party assessment	Yes

Workplace based assessment assesses or appraises what is done at work, or learned for and from work. In respect of the new GMS contract the assessment or appraisal will occur at work. The checking of written or computerised evidence is, in the main, the principle of the assessment under the new General Medical Services. Here the practice team puts forward a report as to how it has performed in a number of aspects of patient care. The assessors have various ways of checking the evidence put forward by each practice team. They can examine practice logs and records, both written and computerised. They can select a number of patient records to decide whether claims made with regards to aspects of preventative practice are borne out by the evidence held by the practice records system. They can interview practice staff. This is a third party assessment of the practice team's performance and self-assessment has little part to play in it.

But, in respect of GP appraisal, the annual appraisal of the individual GP, the aim has, up to now, been that the process should be about facilitating the professional development of that GP and this is a peer discussion of a prepared self-assessment. What is discussed is based on activities that may have occurred at work, or away from work but related to work. For example, GP appraisees may present a reflective diary, which they have prepared at home during many evenings, but which is based on their thoughts or reports on aspects of their daily work. At present, appraisers undertaking the appraisal of individual GPs do not usually have the option of looking at the individual GP's records to confirm evidence of his or her statements as to the development of practice. The preparation of the evidence for the appraisal is still truly a self-appraisal process, whilst the appraisal itself is a peer-appraisal process.

Conclusion

The features of self-assessment in human life in general and in medical practice in particular have been discussed. We have emphasised how personal qualities such as drive and pride in what is done are essential motivators in medical practice. We have emphasised too that self-assessment is nothing if it is not measured against standards and benchmarks. Each year a doctor will work for thousands of hours, and in each of those hours will have opportunities for self-assessment – a self-motivated questioning of what he or she has done at work. But external assessments, however powerful they are in the rigor of their enquiry, will only take up a few hours. To reject self-assessment because of its problems as to accountability would be to reject an essential and valuable part of human development, 'an awareness of remediable weaknesses through continual self-appraisal'.[8]

We have also discussed the limitations inherent in self-assessment and discussed how these may be addressed. We have described some of the tools which may be used to make self-assessment more objective, and commented on their strengths and weaknesses. We have discussed how assessment in medicine may be affected by patients' views. And finally, we have discussed the concept of workplace based assessment and the role of self-assessment within it.

References

1. Claxton G. *Not Learning More, but Learning Better: tracking the development of learnacy.* Presentation at the Beyond the Exam conference Bristol: Watershed Media Centre; 2003. www.nestafuturelab.org/events/past/be_pres/gc01.htm
2. Tolstoy LN. *Childhood, Boyhood and Youth.* Harmondsworth: Penguin; 1974. p. 118.
3. Manning PR, DeBakey L. Life-long learning tailored to individual practice. *JAAMA* 1992; **268**: 1135–6.
4. Fox RD , Mazmanian PE, Putnam RW. *Changing and Learning in the Lives of Physicians*. New York: Praeger; 1989.
5. Holm HA. Quality issues in continuing medical education. *BMJ.* 1998; **316**: 621–4.
6. Boud D. *Enhancing Learning Through Self Assessment.* London: Kogan Page; 1995.
7. Brown G, Bull J, Pendlebury M. *Asssessing Student Learning in Higher Education.* London: Routledge, 1997.
8. Evans AW, McKenna C, Oliver M. Self assessment in medical practice. *JRSM.* 2002; **95**: 511–13.
9. Schon DA. *The Reflective Practitioner.* Aldershot: Ashgate Publishing Limited; 1996.
10. Caputo D and Dunning D. What you don't know: the role played by errors of omission in imperfect self-assessments. *J Exp Soc Psychol.* 2005; **41**: 488–505.
11. Antonelli MA. Accuracy of second year medical students' self-assessment of clinical skills. *Acad Med.* 1997; **72**: 563–5.
12. Arnold L, Willoughby TL, Calkins EV. Self-evaluation in undergraduate medical education: the longitudinal perspective. *Med Educ.* 1985: **60**: 21–8.
13. Wooliscroft JO, TenHaken J, Smith J, Calhoun JG. Medical students' clinical self-assessments: comparisons with external measurements of performance and students' self-assessments of overall performance and effort. *Acad Med.* 1993; **68**: 285–94.
14. Evans AW, McKenna C, Oliver M. Self assessment in medical practice. *JRSM* 2002; **95**: 511–13.
15. Tracey J, Arroll B, Barham P, Richmond D. The validity of general practitioners' self assessment of knowledge: cross sectional study. *BMJ.* 1997; **315**:1426–8.
16. Burton J, Jackson N. *Work Based Learning in Primary Care.* Oxford: Radcliffe Publishing; 2003. p. 157.
17. Morris P, Burton K, Reiss M, Burton J. The difficult consultation. An action learning project about mental health issues in the consultation. *Educ Prim Care.* 2001; **12**: 19–26.
18. Lacey Bryant S, Ringrose T. Evaluating the doctors.net.uk model of electronic continuing medical education. *Work Based Learn Prim Care.* 2005; **2**: 129–42.
19. Norcini J. Current perspectives in assessment: the assessment of performance at work. *Med Educ.* 2005; **39**: 880–9.

Chapter 12

Practice based assessment

Penny Trafford and Sanjiv Ahluwalia

Introduction

This chapter sets out to define practice accreditation, its uses, and to explore aspects of organisational assessment in general practice relevant to training practice accreditation. The chapter will review existing UK general practice accreditation schemes and how they attempt to improve quality of care.

The history of training practice accreditation is briefly described along with current changes. Issues in assuring the quality of accreditation processes are also discussed.

What is practice accreditation?

Buetow and Wellingham describe accreditation as a voluntary but formal process of self-assessment and/or external and independent peer review.[1] Accreditation reviews assess 'measurable' performance, or capacity to perform, against pre-determined and explicit standards that GPs and other stakeholders have produced.[1] Results may include recommendations for continuous improvement of safety and quality in the practice. Certification is typically the end point of an accreditation process. Accreditation gives official approval or endorsement. In general practice, it typically applies to GPs' work settings in recognition of delivery of general practice services, and to accreditation agencies in recognition of competency to accredit general practices.[1]

What is accreditation meant to achieve?

Practice accreditation can have at least five purposes:

- quality control
- regulation
- quality improvement
- information giving
- marketing.

Quality control

This purpose protects public safety and meets demands for increased openness and accountability to the public (patients and tax payers), government and other stakeholder groups.[2] Quality control seeks to, 'assure or even better improve trust of external parties such as patients, financiers and government'. It highlights the importance of medicine as a profession and service. It also reflects the need to eliminate unnecessary and inappropriate interventions, increase equity of access,[3] monitor health outcomes, and demonstrate that practices function efficiently, offering value for money.

Regulation

Regulation assesses practices' adherence to government contractual and other legal requirements. This regulatory process offers rewards for practices that demonstrate such adherence. For example, there are significant financial rewards for general practices that achieve administrative and organisational standards defined in the new General Medical Services (nGMS) contract,[4] negotiated between the British Medical Association (BMA) and the UK government.

Quality improvement

Accreditation of practices enables their entry into or development of elements of a framework of continuous quality improvement. According to this framework, the whole practice team can improve over time with the quality of the organisation and delivery of its services. Scope to improve practice quality and safety can be identified by comparing individual practices against accreditation standards, measures of their own past performance, and/or rates or norms based on accreditation results from other practices (benchmarking).[1]

Information giving

Stakeholders in general practice care can use information from accreditation processes to support comparisons between practices, show levels of adherence to standards, highlight opportunities for improvement, inform and guide decision making, and enhance confidence.[1]

Marketing

In a competitive healthcare environment accreditation can have a marketing benefit for accredited practices until they account for a high proportion of all practices. A marketing benefit may put competitive pressure on practices to gain accreditation and develop programmes for quality improvement.[1]

Organisational assessment in general practice

Organisational assessment is an integral part of quality assurance and quality improvement activity in general practice. Externally led quality assurance and internally led quality improvement are not distinct activities and can be viewed as two end points along a spectrum.[5]

External and internal assessment

Organisational assessment for the purpose of quality assurance lies at one end of the spectrum. It is reliant on external assessment, based on evidence and primary stakeholders are typically governments and health insurance companies.

The middle ground of organisational assessment is occupied by professionally led assessment mechanisms. In the UK, accreditation mechanisms are used both to recognise past achievements and to catalyse future quality improvement.

At the other end of the spectrum, organisation assessment is conducted for the purpose of practice-driven quality improvement. The emphasis is on continual development, self-assessment, local identification of problems and their likely solutions. The team usually instigates organisational assessment. The team then matches the skills and resources of team members with local initiatives or opportunities. The purpose is to foster collaboration and to motivate team members to try new ways of doing things. The need for a structured approach to making changes is still important, but both planning and structure for achieving improvement are driven and owned by the practice.

The problem with external assessment is that it may stifle the potential for internally led quality improvement.[1] On the other hand an over-reliance on internally led quality improvement does not enable practices to compare with and learn from each other, nor does it reassure external stakeholders.[6] The solution appears to be keeping quality assurance and quality improvement as separate activities within a coordinated systems based framework of assessment.

Methods of assessment

The assessment of organisational aspects of general practice is high on policy agendas, both as a means of stimulating quality improvement and achieving accreditation.[7] In most contexts the systems of assessment are summative in that judgements are made against preset standards for deciding levels of achievement. [1] Assessment methods are conceptualised as accreditation type processes in that they are based on inventories of indicators or items. The standards applied typically cover a wide range of organisational issues from premises to equipment to delegation, communication and leadership.

Problems with organisational assessment

Organisational measurement processes seem to be conceptually grounded on a regulatory concept rather than on a formative aim of providing information to motivate developmental change. They seldom involve people from different roles in organisations in the process of assessment. It is known that assessments that respect historical restraints and incentives, are sensitive to different starting points, engage teams, identify developmental needs, and help to set priorities for future change are much more in tune with the internal workings and motivation of those who work in most organisations.[8]

Overt summative approaches risk the loss of a formative development feedback approach that could inform quality improvement strategies. These systems also have other disadvantages for practices. Systems that judge against minimal standards can often fail to inspire movement towards improvement. Likewise, systems that judge against gold standards (based on leading edge practice) can sometimes discourage practices with substantial development needs embarking on quality improvement activities.[8]

General practice accreditation schemes in the UK

The Royal College of General Practitioners has been at the forefront of developing practice accreditation systems in the UK. Currently available to individual practices are two schemes: Quality Team Development (QTD) and Quality Practice Accreditation (QPA).

The overall principles of these two schemes are the same. Both require specific primary care quality criteria to be met. The clinical criteria include health promotion, women and children health, mental health and chronic disease management. Organisational criteria include team working and communication. These schemes also support evidence based practice by encouraging implementation of clinical guidelines.[9]

Continuous quality improvement is promoted by encouraging teams to look objectively at what they do, for example through audit and significant event analysis, to identify areas where care is not as good as would be expected. Teams are then expected to use such knowledge to improve working practice and therefore patient care.[9]

Quality criteria contained within the RCGP schemes have been designed to reflect core primary care services. For a team to meet the criteria, different disciplines must work together. Team working is particularly important for pre-assessment, when staff are preparing documentation to demonstrate that the criteria have been met. They are designed to be achievable, but some work will be required to meet them, although preparation will vary. Other criteria require development of new protocols or completion of clinical audits and will be more challenging. Although participation is greatest pre-assessment, members of staff report this phase as benefiting them most, usually through personal learning and team building.

The QPA is an example of an assessment system that aims to reward excellence and/or minimum standards of care.[10] Such schemes are attractive to practices that seek accreditation of minimal standards or to those that are able to achieve high standards with a manageable degree of work.

How do accreditation systems improve quality of care?

Much previous research concentrated on individual components of quality improvement, such as significant event auditing,[11] conventional auditing,[12] and patient feedback.[13] However, multi-level strategies for change that combine education, audit, research and clinical effectiveness in unified multi-professional educational strategies lead to the changes in behaviour that enhance quality improvement.[14]

Primary care organisations (PCOs) use myriad approaches to improve quality. These include audit, significant event analyses, team based education and training events, sharing of comparative data, personal and practice learning plans, the setting and monitoring of standards, and the use of quality indicators.[15] PCOs are also advocating collaborative and corporate learning (all practices learning together) and team-based learning (all staff within practices learning together). Such strategies, highlighting the concept of learning organisations, are appropriate, as quality improvement requires fundamental changes in organisational and behavioural cultures, which are far from straightforward and take time to achieve.[16]

Challenges to practice accreditation

It has been suggested that the organisation-wide benefits of quality improvement as a method of improving outcomes and lowering costs have not been consistently demonstrated in healthcare organisations. [17] A further problem is that accreditation does not necessarily provide assurance of quality. This is because in general practice much illness is undifferentiated and the relationships between organisational structures, processes and outcomes are poorly understood (*see* Box 12.1). [18]

Box 12.1: Challenges to practice accreditation

- Developing an evidence base to support practice accreditation as a means for improving outcomes
- Developing an evidence base to support practice accreditation as a valid means for assuring quality
- Demonstrating the cost-effectiveness of practice accreditation
- Tension between the philosophies of quality assurance and improvement

As previously discussed, there is a tension between the philosophies of assurance of quality and improvement. Furthermore, accreditation schemes are costly in terms of monetary, time, effort and other staff costs.[1]

History of training practice accreditation

The postgraduate education committee was formed by the Royal College of General Practitioners (RCGP) in 1952. Its express function was, 'to encourage postgraduate instruction as a prerequisite for practice and training a qualified doctor for a career in general practice'.[19] The Joint Committee for Postgraduate Training in General Practice (JCPTGP) was formed in 1976 to monitor the quality of general practice training. It was given responsibility for assessing the training and experience of doctors applying to work as GPs.[20] In 1997, the legal powers of the JCPTGP were extended to include approval of all training posts in general practice.[20]

In April 2003, legislation was passed to create the Postgraduate Medical Education Training Board (PMETB). There were many reasons for this. The government and medical colleges affirmed the principle of professional self-regulation and independence from government, protecting patients through more robust and transparent improvements in medical education, providing consistency and integration to a diverse range of historical training arrangements, encouraging greater multi-professional education, engaging patients in medical education, and raising the profile of medical education (the JCPTGP and the Specialist Training Authority (STA)) were thought not to be powerful enough to make change happen when required.[21]

In September 2005, the PMETB assumed statutory powers taking over responsibilities from the JCPTGP and STA. The PMETB's responsibilities include establishing and assuring standards of medical education as well as promoting medical education across the UK.

In line with other medical colleges in the UK, the RCGP has taken on a greater responsibility in setting standards for assessment of GP registrars. It is intended that from August 2007, the new membership examination of the RCGP shall replace summative assessment as the exit standard for independent practice as a general practitioner.

The deaneries act on behalf of the PMETB in accrediting trainers and training practices. The PMETB in turn ensures standards of assessment are applied uniformly by a system of regional visits and accreditation of deaneries, training schemes and training practices. The PMETB publishes guidelines to ensure UK-wide comparability in the training and education leading to vocational training certificates.[20]

The accreditation of practices is based upon a collation of information related to the trainer, the practice and any previous or ongoing educational activity in the practice. The collated information is assessed against a list of criteria set out by the JCPTGP and adopted by PMETB.[21]

The practice assessment takes place by means of a visit. The JCPTGP specifically stated the purpose of the visit as being, 'to assess the suitability of the practice for training and to assist the applicant in identifying areas for change' as an educationalist.[20] Assessment in a practice visit is a multi-faceted intervention, in which one or more observers come to a practice to assess and discuss the quality of care or services against guidelines and criteria. Since the 1950s practice visits have been acknowledged as a powerful means of achieving change and a great possible asset in quality improvement.

Ensuring the quality of accreditation of training practices

A set of performance indicators form part of the accreditation process. These can be related to the outputs from educational activities, educational processes, learning resources and access to services for trainees. In developing a set of indicators the aim is to find a balance between measurability (reliability) and relevance (validity).[22] When making assessments regarding training practices several important aspects need to be considered.

- Outcomes to be assessed should be clearly defined and available to the assessor and the assessed.
- Assessment technique should be an appropriate means of making the assessment.
- The wider the range of assessment methods used and number of people on the team visiting, the greater the accuracy of the conclusions.
- Assessment process should always be used constructively to encourage development and change.
- Assessment techniques should be valid and reliable in what they set out to achieve.

Validity

Content validity is defined as the extent to which an assessment measures the intended content area, i.e. does the assessment cover the area appropriately and the necessary content of what is to be assessed?[23] The empiric literature has focused on the competencies required of a good GP trainer.[24–30] However, studies reviewing the characteristics of the ideal learning climate and the training practice are much more limited.[31,32]

Smith provided evidence that there are significant differences in the way learners perceive their learning environment and current accreditation processes failed to take these perceptions into account.[31] Recommendations were made for changes to the practice accreditation visit to enhance the validity of the process.

There is a need for greater research to determine the characteristics of training practices that promote quality education for general practice registrars and inform the accreditation process so that training practice accreditation can be seen to have high validity.

Reliability

In relation to practice accreditation, there are several potential threats to reliability.[32] These include inter-observer variation (the tendency for one observer to consistently mark higher than another), intra-observer variation (the variation in an observer's performance for no obvious reason), and case specificity (the variation in the candidate's performance from one challenge to another, even when he/she seems to test the same attribute).

Concern has been raised about reliability in training practice accreditation visits (personal correspondence). These include variable interpretation of the criteria by external visitors and different regions within a deanery implementing the criteria differently.

The accreditation visit has multiple visitors that should enhance reliability. However, it has also been suggested (personal correspondence) that this can be undermined where a visiting member predominates the decision making process. Further work is needed to determine the extent and role of inter-rater variation in training practice accreditations, as well as strategies to reduce this.

Capacity

Changes to medical education in the UK have placed significant pressure on the primary care medical education workforce to expand its capacity. The shift of undergraduate medical education from ward based secondary care to general practice,[33] the need for all foundation doctors to have experience of general practice after 2008,[34] and government policy to increase the number of general practitioners in the UK[4,35] have contributed to this pressure for increased capacity.

The assessment of training practices, or those wishing to become a training organisation, requires substantial resources in terms of time and manpower. In our own deanery, London, each training practice assessment requires four visitors; an associate director, local course organiser, established trainer and practice manager or nurse. The practice requires that the trainer, trainee and other members of the primary healthcare team take time out of clinical activity for four hours.[19]

To alleviate this pressure, deaneries have considered the use of the new General Medical Services (nGMS) contract[4] within the overall accreditation process. The nGMS quality and outcomes framework (QOF) is designed to raise and enhance the clinical and organisational standards in primary care. Activity is rewarded subject to the level of points achieved. The assessment of achievement with QOF is externally led by a panel of visitors from the local primary care trust (PCT) through a panel consisting of a lay member, clinician (usually a GP) and a PCT manager. The practice aspirational targets and actual achievement are assessed against robustly developed criteria which are subject to substantial scrutiny and review.[33] Whilst QOF achievement does not formally accredit practices as learning organisations, there is substantial overlap between the criteria assessed by QOF, the JCPTGP criteria (*see* Box 12.2) and other quality schemes such as the RCGP Quality Practice Award (QPA).[34]

Box 12.2: Training practice criteria not covered by QOF

Physical resources:

- Sufficient consulting rooms or space for GPR
- Access to secondary care services
- Access to computer and IT resources for GPR
- Library
- Access to teaching aids, e.g. video

Education and training:

- Involvement of the practice team in teaching
- Involvement of the GPR in practice meetings
- Appropriate GPR timetable and supervision
- Arrangements for out of hours (OOH) training

It has been suggested, both by trainers in general practice and departments of postgraduate general practice education, that QOF could be used as an integral part of the training practice selection process. This, it is argued, would avoid the duplication of monitoring visits, reduce paperwork and enable the trainer re-approval process to focus on educational rather than clinical or organisational aspects of practice activity. Some deaneries, e.g. Manchester and Leicestershire, Northamptonshire and Rutland, have already proposed certain score profiles as appropriate for training practice approval. Research is currently under way to look at its usefulness in practice visit accreditations.

References

1. Buetow SA, Wellingham J. Accreditation of general practices: challenges and lessons. *Qual Saf Health Care*. 2003; **12**(2):129–35.
2. Nichols A, Schilit R. Accreditation of human service agencies: costs, benefits, and issues. *Admin Social Work*. 1992; **16**: 11–23.
3. Wilkinson JR, Murray SA. Assessment in primary care: practical issues and possible approaches. *BMJ*. 1998; **316**(7143): 1524–8.
4. British Medical Association. *New GMS Contract: Investing in General Practice*. London: BMA; 2003. www.bma.org.uk/ap.nsf/Content/NewGMSContract
5. Rhydderch M, Edwards A, Elwyn G, Marshall M *et al.* Organizational assessment in general practice: a systematic review and implications for quality improvement. *J Eval Clin Pract*. 2005; **11**(4): 366–78.
6. Crabtree BF, Miller WL, Stange KC. Understanding practice from the ground up. *J Fam Pract*. 2001; **50**(10): 881–7.
7. Huntington J, Gillam S, Rosen R. Clinical governance in primary care: organisational development for clinical governance. *BMJ*. 2000; **321**(7262): 679–82.

8. Elwyn G, Rhydderch M, Edwards A, Hutchings H *et al.* Assessing organisational development in primary medical care using a group based assessment: the Maturity Matrix. *Qual Saf Health Care.* 2004; **13**(4): 287–94.
9. Ring N. The RCGP Quality Practice Award for primary care teams. *Br J Community Nurs.* 2003; **8**(3): 112–5.
10. Royal College of General Practitioners. *Quality Practice Award.* [online cited 16 October 2005]. www.rcgp.org.uk/faculties/scotcoun/qpa.asp
11. Westcott R, Sweeney G, Stead J. Significant event audit in practice: a preliminary study. *Fam Pract.* 2000; **17**(2): 173–9.
12. Hopeyiow K, Morley S. Putting confidence into audit: using confidence intervals to set objective standards in primary care audits. *J Clin Gov.* 2001; **9**: 67–70.
13. Greco M, Brownlea A, McGovern J, Cavanagh M. Consumers as educators: implementation of patient feedback in general practice training. *Hlth Commun.* 2000; **12**(2): 173–93.
14. Cantillon P, Jones R. Does continuing medical education in general practice make a difference? *BMJ.* 1999; **318**(7193): 1276–9.
15. Campbell SM, Sweeney GM. The role of clinical governance as a strategy for quality improvement in primary care. *BJGP.* 2002; **52**(Suppl): S12–17.
16. Davies HT, Nutley SM, Mannion R. Organisational culture and quality of health care. *Qual Health Care.* 2000; **9**(2): 111–19.
17. Grol R. Between evidence-based practice and total quality management: the implementation of cost-effective care. *Int J Qual Hlth Care.* 2000; **12**(4): 297–304.
18. Klein R. Can policy drive quality? *Qual Hlth Care.* 1998; **7**(Suppl): S51–3.
19. Royal College of General Practitioners. *Full Chronology of The RCGP.* www.rcgp.org.uk/history_and_heritage/history_heritage__archives/history__chronology/chronology/detailed_chronology.aspx
20. Joint Committee for Postgraduate Training in General Practice. 2001. *Recommendations on the Selection of General Practice Trainers.* [online cited 10 August 2005]. www.jcptgp.org.uk/policy/selection.pdf
21. Postgraduate Medical Education and Training Board. 2004. *PMETB – The first three years.* [online cited 4 December 2005]. www.pmetb.org.uk/media/pdf/n/g/First_three_years_1.pdf
22. McMimm J. 2002. *Evaluating teaching and learning.* [online cited 6 December 2005]. www.clinicalteaching.nhs.uk/site/Docs/UTL-4%20QTY%20GLOSS.pdf
23. Rhodes M. 2002. *Assessment.* [online cited 6 December 2005]. www.clinicalteaching.nhs.uk/site/ShowModule.asp?l=next&SlideID=3006
24. McKinstry B, Todd M, Blaney D. What teaching skills do trainers think they need to improve? The results of a self assessment questionnaire in South East Scotland. *Educ Prim Care.* 2001; **12**(4): 412–20.
25. Peile EB, Easton GP, Johnson N. The year in a training practice: what has lasting value? Grounded theoretical categories and dimensions from a pilot study. *Med Teach.* 2001; **23**(2): 205–11.
26. Boendermaker PM, Schuling J, Jong BM-d, Zwierstra RP *et al.* What are the characteristics of the competent general practitioner trainer? *Fam. Pract.* 2001; **7**(6): 547–53.
27. Munro N, Hornung R, McAleer S. What are the key attributes of a good general practice trainer: a delphi study. *Educ Gen Pract.* 1998; **9**: 263–70.
28. Bligh J, Slade P. A questionnaire examining learning in general practice. *Med Educ.* 1996; **30**(1): 65–70.
29. Freeman J, Roberts J, Metcalfe D, Hillier V. The influence of trainers on trainees in general practice. *J R Coll Gen Pract Occas Pap.* 1982; **21**: 1–17.
30. Donner-Banzhoff N, Merle H, Baum E, Basler HD. Feedback for general practice trainers: developing and testing a standardised instrument using the importance-quality-score method. *Med Educ.* 2003; **37**(9): 772–7.

31. Smith V. A learner-centered model of training practice inspection: in-depth interview study of GP registrars' perceptions of the learning climate of their training year. *Educ Prim Care*. 2004; **15**: 361–9.
32. Pringle M. Minimum standards for training practices. *Br Med J Ed.* 1984; **88**(6427): 1353–4.
33. General Medical Council. 1999. *Implementing Tomorrow's Doctors*. [online cited 26 October 2006]. www.gmc-uk.org/education/undergraduate/tomorrows_doctors_implementation.asp#appendix%20a
34. Department of Health. 2002. *Unfinished Business: proposals for reform of the Senior House Officer grade*. [online cited 10 August 2005]. www.mmc.nhs.uk/download_files/Unfinished-Business.pdf
35. Department of Health. *The NHS Plan: a plan for investment, a plan for reform*. London: HMSO; 2000.

Chapter 13

Assessment when performance gives rise to concern

Debbie Cohen and Melody Rhydderch

Introduction

Understanding underperformance in doctors is complex. Once recognised it must be assessed and managed. The assessment of underperformance in doctors has been well researched and models of assessment are well developed. Models for the management of underperformance, however, are not so well researched or documented.

Performance assessments for doctors can be said to exist on three levels.[1] The first involves screening a population of doctors, the second the selective assessment of those doctors thought to be at risk and the third the targeted assessment of underperforming doctors. In the UK, the mechanisms for assessing at these three levels include appraisals, local trust-led performance procedures and referral to the National Clinical Assessment Service (NCAS) or General Medical Council (GMC) respectively. Data from the first 50 cases seen by the NCAS have confirmed that causes of underperformance are multifactorial.[2] Factors associated with clinical care, behaviour and attitude, health and wellbeing as well as organisational issues were all cited as reasons for referral.

The interaction between these factors is important. There is a clear link between the health and wellbeing of both a doctor and the organisation they work within.[3] Whilst only 1% of doctors referred to the GMC health committee had a physical health problem, mental health disorders predominated. Stress in health professionals is high, with 28% showing above threshold symptoms compared to 18% of workers as a whole in the UK. If the individual is David, then the organisation is Goliath. Factors such as high workload, shift systems, work patterns, poor leadership, team working, all have the potential to impact negatively on an individual's wellbeing and to distort patterns of behaviour and ability to perform.

Remediation if it is to be successful and sustainable must be able to respond to the assessment process and offer effective interventions. Remediation must be sensitive to the problems identified and provide a flexible response. It should offer support and direction to both the organisation and the individual. It is important to recognise that the problem may lie in the organisational structure and culture rather than solely with the individual referred.[4] Health and wellbeing must be considered alongside personality, motivation to change, organisational and social issues.[5]

The Individual Support Programme (ISP) sits within the Communication Skills Unit at the Department of General Practice at Cardiff University. The ISP provides assessment and remediation for doctors who are struggling with their performance. Since its inception in 2001, we have received over 100 referrals ranging from undergraduates through to consultants. Case studies from the first two years of the service were published in 2005.[6] In this chapter we describe the service provided by the ISP, discussing the advantages and disadvantages before drawing conclusions about implications for policy, practice and research.

Individual Support Programme (ISP)

Developing the right environment for change

The ISP is led by an occupational health physician and provides a 'needs' assessment and structured remediation for clients. As the service has grown, additional staff have been recruited to provide the necessary breadth of skills and services required to manage the range of complex problems presented. The team now includes a general practitioner with occupational health experience, two occupational psychologists and a language specialist.

The ISP has been constructed using motivational interviewing (MI) methods to provide an environment conducive to change.[7] The theories underpinning motivational interviewing suggest that ambivalence to change is normal and that confrontational interviewing increases resistance. Therefore, providing an individual with 'space' within the discussion and honouring their autonomy around how they might change engenders the right environment for constructive engagement. Providing a menu of opportunities around possible strategies and interventions gives ownership to the individual and enhances motivation. The independent nature of the service engenders confidence and is important for the process of engagement. Providing a service that is contained within a transparent, explicit, repeatable process that is capable of being independently evaluated enhances credibility for referring organisations.

The overall process of referral, assessment and remediation is illustrated in Figure 13.1. Referrals are made by the undergraduate and postgraduate deaneries or trusts and follow an agreed format. Confidentiality and the independent nature of the service are highlighted at the first interview. Initial assessments are used to build an individually tailored programme addressing clients' needs. In some cases onward referral to other specialist services for assessment may be suggested. Integrated case notes for each referral are maintained. Regular case conferences are held to ensure continuity of care and structured support for the clients. Contact with the referring body is also maintained to allow feedback and mediation between both parties.

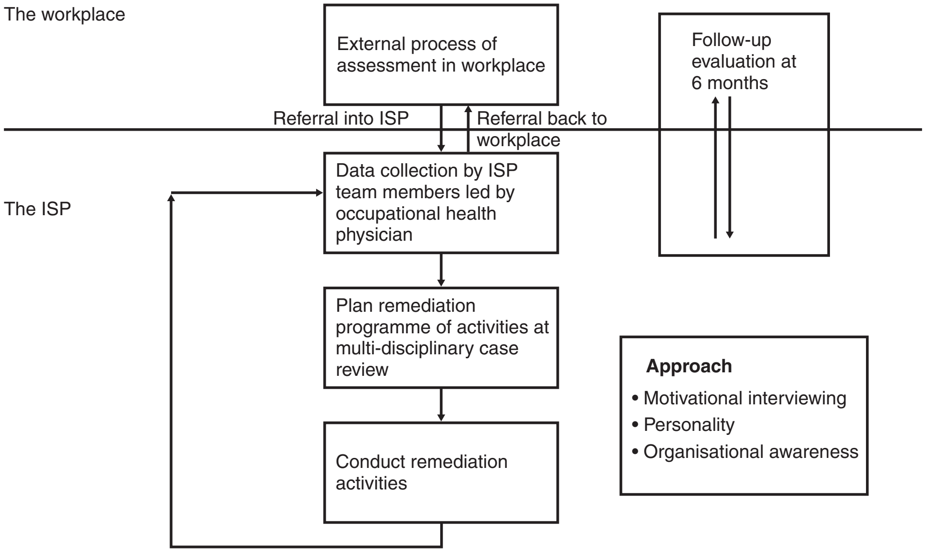

The workplace
External process of assessment in workplace
Follow-up evaluation at 6 months
Referral into ISP
Referral back to workplace
The ISP
Data collection by ISP team members led by occupational health physician
Plan remediation programme of activities at multi-disciplinary case review
Approach
• Motivational interviewing
• Personality
• Organisational awareness
Conduct remediation activities

Figure 13.1: The individual support programme (ISP) referral, assessment and remediation process.

Figure 13.2 describes the activities undertaken to support assessment and remediation. The majority of the assessments take place within the unit although in some cases workplace assessments are also made. Referral data varies, but may include critical incident reports, 360 appraisals and copies of Records of In-Training Assessments (RITAs). Data collected during the ISP assessment uses a standardised format.

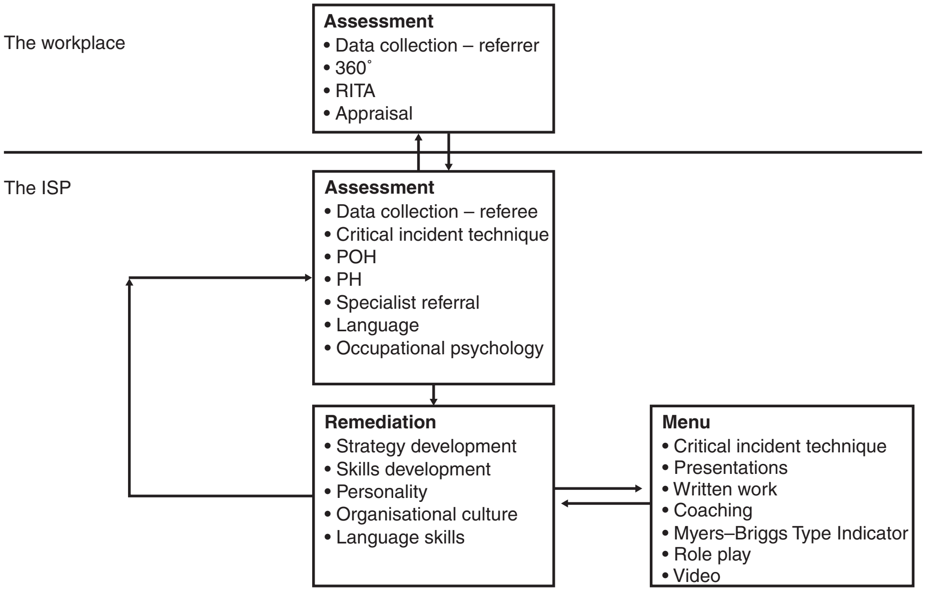


Figure 13.2: The individual support programme (ISP) referral, assessment and remediation activities.

Initially, information is gathered by asking the individual to recall work experiences using the critical incident technique.[8] A detailed past occupational history and past medical history are included and an occupational psychology assessment is usually undertaken. For individuals where language may be an issue, a language assessment is conducted. The remediation activities are constructed in collaboration with the individual and have a clear framework, with timelines. Simulated patient consultations, video feedback and critical incident analysis are woven into a programme that fits the individual's needs and preferred learning styles.

Who is referred to the service?

Between 2001 and February 2006, 76 doctors had completed an assessment and remediation programme with the ISP.[9] The majority of our referrals have come from the Welsh deaneries with a small proportion of these being self-referrals. However, as the service has grown and received wider attention referrals have also been received directly from trusts both in Wales and England. Our sample of 76 therefore reflects a younger population than those who might be referred directly to the NCAS. In Wales GPs with performance difficulties are usually referred to an Advanced Trainers Network, a service provided by the Post Graduate Department for General Practice Education in Wales and therefore our sample contains only a small percentage of GPs. It is interesting to note that although the age range seen by the ISP differs from the NCAS the spread of gender and specialties referred to the ISP (*see* Figures 13.3 and 13.4) show strong similarities to the first 50 cases referred to the NCAS.

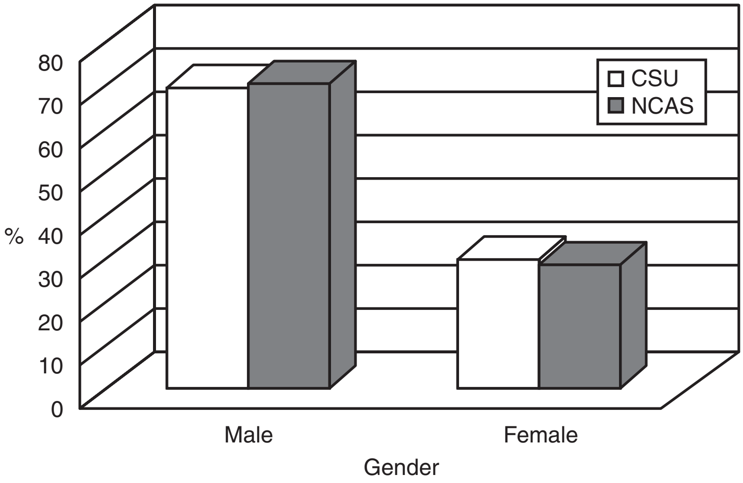


Figure 13.3: Gender of doctors referred to the CSU and NCAS.

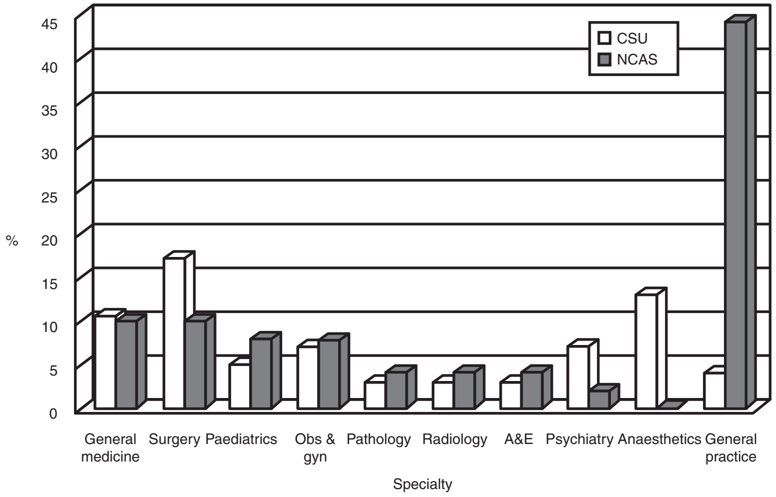


Figure 13.4: Specialties of doctors referred to the CSU and NCAS.

Why are referrals made?

The majority of those referred to the ISP presented with at least two or more areas of concern.[9] Examples of the types of referral include the following doctors that:

- have failed their RITA due to poor team working and clinical prioritisation
- have language difficulties and poor team management skills
- are inaccessible and failing to achieve.

This illustrates the need for multiple remediation activities.

What is known about outcomes of the service?

Success is currently measured by a multidisciplinary review of the individual's progress at the end of the remediation process. In addition, the lead physician remains in contact with the referring organisation and proactively invites feedback regarding progress. For some cases, quantitative data in the form of RITA assessments and 360 appraisals are available as evidence of improvement. Based on this approach to evaluation, there is evidence to suggest the approach adopted by the ISP is effective in creating sustained improvements in individual performance. 65.8% of referred doctors achieved an improvement as noted by the programme director and/or the deanery at the end of the programme. Examples of improvements include receiving an improved RITA grade, passing exams, decrease in patient complaints, positive feedback from colleagues and improved language skills.

Advantages and disadvantages

We have described a programme that provides tailored remediation for doctors struggling with their performance. Our principal finding is to reinforce the view that poor communication skills may be symptomatic of the presence of a more complex picture of performance deficit. Standing back and reflecting on the pros and cons of the service, the main advantages of the current design of the ISP centre around its ability to deliver independent personalised remediation that engages individuals and takes account of organisational issues.

Personalised remediation

The purpose of personalised remediation is to match provision closely to individual need and to do so through one-to-one working over a period of time between the doctor and the ISP team members. It is typical for referred doctors to work with more than one team member. Team working allows a more holistic approach and matches improvements over more than one area. In addition to working on a specific learning need, the benefit of personalised remediation is that it can be tailored to different learning styles. We use the Myers–Briggs Type Indicator to help doctors become more aware of their behavioural habits and how these impact upon their learning styles and general performance. Some doctors have a preference for learning through practical activities such as role plays, whilst others prefer more reflective approaches to skill acquisition. We recognise the need to follow a learning cycle where different learning/teaching styles are used, but are clear about the best starting point into the process for each individual concerned. This has reaffirmed our use of motivational interviewing as a technique for giving people autonomy over their unique process of change and development. One final advantage of personalised remediation is that it is based on a broad multidisciplinary approach. This format provides ability for the assessor to identify issues that may not have been raised in the referral and address these early on.

Engagement and motivation

Experience within the Canadian health system suggests that a doctor's motivation may be low at the beginning of the process, but that the involvement of a licensing authority had a positive effect on the doctor's motivation and co-operation.[10] However, in our experience, whilst the high stakes involved may motivate a person to attend the unit, it is not always enough to stimulate genuine engagement with the process as opposed to compliance. The assessment process is not only important to identify needs, but also to identify an individual's readiness to engage in the process.

Remediation does not begin until an individual is ready to engage. In our experience, the trigger for engagement is different for each person depending on the nature of the presenting problem. For some it occurs when they have an opportunity to reflect on the reasons for their career choice (medicine and specialty), for others it takes place as they become more aware of how their personality interacts with the workplace. For some it might be discovering that language difficulties can be easily overcome. This leads to the second fundamental, the need to take account of wider organisational issues.

Organisational issues

Historically, remediation has been viewed as being provided by or overseen by an educational supervisor. Whilst this is appropriate up to a point, there comes a stage where the causes of underperformance have to be well understood before education or even remediation can be effective. For example, we have seen referrals for poor performance which have related clearly to rapid organisational change resulting in a loss of support and supervision for a doctor. This typically presents as 'poor communication skills' due to inappropriate displays of anger or poor team leadership on the part of the individual. Remediation consisted of skills improvement as well as coaching to identify support networks in the new organisational structure. A wider organisational concern for us as a unit is the need to consider what happens to the individual once they are no longer able to access the support and resources of the ISP. Will the organisational environment erode the individual's capacity to sustain improvements that have been made?

Independent nature of the service

Performance assessment impacts upon three distinct groups: patients, doctors and employers.[11] While these groups may have conflicting beliefs and expectations of assessment, the process must be acceptable for all. This is also true for remediation. Independence is a mechanism for promoting acceptability and this is a key feature of the ISP. Independence means that all parties can assume that regardless of the original reason for referral, the ISP undertakes an independent assessment that takes account of the wider organisational context. There may be times when either party may disagree with our findings. Independence is considered essential to the fair and legitimate delivery of the service.

What is needed to improve the ISP?

Whilst progress with specific remediation models such as the ISP has been made, what seems to be lacking is a clear consensus about which remediation methods are appropriate in different circumstances. This has also been recognised by Leape in the USA.[12] The recognition that the causes are complex and multiple suggests that for remediation to be successful it must address all factors. This is not a simple task and needs more than a prescription to attend anger or time management courses. This approach may be the sticking plaster to make both referrer and referee feel better but it is unlikely to last.

It is worthwhile standing back and taking a wider view of the problem of underperformance. Doctors are no different to the wider population where performance relates to wellbeing as well as to skills and knowledge. Taking a biomedical model to remediation perhaps is unwise. If we are to accept that which is inferred by methods of assessment, then the biopsychosocial model[13] is more appropriate and is now widely accepted as being fundamental to rehabilitation.[14] An example is the successful 'pathways to work pilots' for individuals on long-term incapacity benefits.[15] Here health, social and psychological factors are addressed and a case-by-case approach to provide effective support is developed. Assessors are trained in the holistic management of cases and address motivation to change at the outset. Remediation of doctors should follow the same model. Underperformance requires a holistic approach that also addresses motivation to change at the outset. Understanding an individual's motivation to change and engaging them in the process of change is well established in behaviour change methods[16] and needs careful consideration. Establishing a connection between assessment and provision of remediation is necessary. Health, wellbeing and personality are inextricably linked. Remediation must echo this and to achieve this end demands a closer working connection between occupational health physicians and occupational psychologists.

Conclusions

At present remediation services for underperforming doctors are fragmented. The lessons learnt from the ISP over the last five years suggest that effective remediation is possible but should be personalised, independent, engage the individual and be delivered by a multi-disciplinary team. However, we have also learned that individual performance takes place within a wider organisational context that cannot be ignored. The aim overall must be to develop an evidence based approach to remediation. If remediation is to evolve we need to be able to learn from our collective experiences. The fragmentation of existing provision is not at present conducive to this. In our own data set we are aware that there are many biopsychosocial markers. Some of these might prove in the future to be valuable indicators for underperformance and remediability but a common larger data set is badly needed.

We have to conclude that remediation should not be restricted to an educational model of skills acquisition. At the individual level, it should take account of motivation, personality and organisational awareness. At a system level, there is a certain irony; perhaps we in occupational health, human resources and education also have to be mindful of the need for better communication. At the heart of the process lies the doctor and the stress and distress that they might encounter must not be underestimated.

References

1. Finucane PM, Bourgeois-Law GA, Ineson SL *et al.* A comparison of performance assessment programs for medical practitioners in Canada, Australia, New Zealand, and the United Kingdom. *Acad Med.* 2003; **78**(8): 837–43.
2. Berrow D, Faw L, Jobanputra R. *NCAS, Evaluation, Research and Development*. London: National Clinical Assessment Authority, National Patient Safety Agency; 2005.
3. West M, Spendlove M. The impact of culture and climate in healthcare organisations. In: Cox J *et al.* (eds) *Understanding Doctors' Performance*. Oxford: Radcliffe Publishing; 2006. p.91–103.
4. Harrison J. Illness in doctors and dentists and their fitness for work – are the cobblers' children getting their shoes at last? *Occup Med.* 2006; **56**: 75–6.
5. Barrick MR, Mount MK. The big five personality dimensions and job performance: a meta-analysis. *Person Psychol.* 1991; **44**(1): 1–27.
6. Cohen D, Rollnick S, Smail S *et al.* Communication, stress and distress: evolution of an individual support programme for medical students. *Med Educ.* 2005. **39**(5): 476–81.
7. Rollnick S, Mason P, Mason BC. *Health Behavior Change: a guide for practitioners*. Edinburgh: Churchill Livingstone; 1999.
8. Flanagan FC. The critical incident technique. *Psychol Bull.* 1954; **51**(4): 327.
9. Cohen D, Rhydderch M, Kinnersley P *et al.* Stress and distress: factors associated with doctors identified as struggling with their performance. Submitted for publication. 2006.
10. Goulet F, Jacques A, Gagnon R. An innovative approach to remedial continuing medical education 1992–2002. *Academic Med.* 2003; **80**: 533–40.
11. Finucane PM, Barron SR, Davies HA *et al.* Towards an acceptance of performance appraisal. *Med Educ.* 2002; **36**: 959–64.
12. Leape L, Fromson JA. Problem doctors: is there a system level solution? *Ann Intern Med.* 2006; **144**: 107–15.

13. Creed F. Are the patient centred and biopsychosocial approaches compatible? In: White P (ed.) *Biopsychosocial Medicine: an integrated approach to understanding illness*. Oxford: Oxford University Press; 2005. p.187–99.
14. Waddell G, Burton AK. *Concepts of Rehabilitation for the Management of Common Health Complaints*. London: TSO; 2004.
15. Department for Work and Pensions. *Pathways to Work: helping people into employment*. London: TSO; 2002.
16. Miller WR, Rollnick S. *Motivational Interviewing: preparing people for change*. London: Guilford Press; 2002.

Chapter 14

Legal perspectives of assessment

Anthea Lints

Introduction

Examination boards are influenced and guided by the Code of Practice for the Assurance of Academic Quality and Standards in Higher Education, 2004[1] which evolved from the Dearing and Garrick Reports published in 1997.

The Dearing report recommended that, 'arrangements for handling complaints from examinees should reflect the principles of natural justice; be transparent and timely; include procedures for reconciliation and arbitration; include an independent, external element; and be managed by a senior member of staff.'

The general principles, which underpin this code, are as follows.[1]

1 There should be an effective procedure for resolving academic appeals. Those sitting examinations should have opportunity to raise, individually or collectively, matters of proper concern to them without fear of disadvantage and in the knowledge that privacy and confidentiality will be respected.
2 The appropriate governing authority must ratify the process of appeal and the outcome.
3 The appeal procedure must be fair and decisions must be reasonable.
4 Appeals must be dealt with promptly using simple and transparent procedures.
5 Information outlining the appeals process must be easily accessible and understandable.
6 There must be a source of impartial help.
7 There must be a source of authoritative guidance.
8 Investigation and judgement of the complaint must be impartial.
9 The process must allow the complainant to be accompanied by a friend or mentor throughout.
10 There must be the possibility to have the judgement considered at appeal.
11 The outcome must be supported by appropriate remedial action.
12 Reasonable expenses of the complainant should be met.
13 The process must be subject to regular review.
14 There must be arrangements to monitor the volume, nature and outcome of complaints.

An example of the Appeals procedure for the MRCGP examination can be found at www.rcgp.org.uk in the section on Quality Control and an example of the appeals process relevant to Summative Assessment for General Practice in the London Deanery can be found on the London Deanery website www.londondeanery.ac.uk.

Review process

Normally an unsuccessful candidate can ask the examination board to review the conduct or the result of their examination. They may not request a review of matters which relate solely to the examiners' judgements or that challenges the academic content or methodology of the examination.

The sorts of situations, which this would include, fall into four broad categories:

- organisational
- content
- conduct
- determination of result.

The following are examples of the situations described above.

- Organisational failing: missing or wrong documentation or even poor environmental conditions in which the examination was sat.
- Complaint about content: a question which was of no relevance to the subject being examined.
- Complaint about conduct: questions relating to the candidate's age or gender, cultural background or beliefs.
- Complaint about determination of result: rejection by the examiner of a correct answer.

The candidate will be advised of the outcome of the review.

Some examination boards will also consider exceptional personal circumstances which had an adverse effect on performance but, normally, it would be expected that such circumstances would have been made known to the appropriate authority prior to the examination taking place.

Appeal process

If the candidate remains dissatisfied with the outcome, they may make a formal appeal. The examination board is entitled to request a fee, which may be refunded if the appeal is successful. An independent appeals panel will be invited to consider the appeal. The members of the panel may include examiners who have not been previously involved with that particular candidate nor in the previous review. The chairman would normally be non-medical but someone who has experience of examining in postgraduate education. Normally the candidate will be invited to appear to present his or her case.

Outcomes of the appeal process

Outcomes of the appeal process fall into three categories.

- The appeal may be dismissed. No further appeal will be considered.
- The appeal is upheld but the matter does not justify a different conclusion.
- The appeal is upheld and either the candidate is awarded appropriate credit which could lead to a pass mark or the examination is considered void and the candidate is allowed to re-sit the examination with no further financial cost.

The reasons for the final decision will be shared with the candidate.

It is unlikely that the English or Scottish Courts will become involved unless the examination board fails to follow its own published procedures or these procedures are seen to be capricious or arbitrary. The need for a transparency of process with clear standards of conduct, investigation of misconduct and hearing procedures cannot be over-emphasised.

Appeals should be dealt with swiftly and in a manner in which an outside observer would find fair and reasonable. In addition appeals must be lodged within a stated period of time, in writing, clearly stating the grounds for appeal.

Selection and appointment of examiners

To ensure that examinations are fair, most examination boards have a rigorous selection process for examiners which includes essential examiner specifications, training and calibration and, if suitable, formal appointment to a panel of examiners. Some colleges insist that examiners successfully re-sit the examination for which they will ultimately be examining.

Examiners are regularly reviewed to ensure that their performance is reliable and consistent. Normally examiners are selected for a specified period of time, which can be extended or terminated.

Preparation of an appeal

It is the responsibility of the appellant to prepare the grounds and evidence upon which the appeal is based and he or she may need to seek legal advice.

Access to information

The Freedom of Information Act was introduced in 2001.[2] Under section 1.1 of the act, anyone who makes a request for information from a public authority is entitled to be told whether or not the authority has that information and, if it does, to receive the information requested.

The means of making a request is defined in section 8. The request must be legible, describe the information sought and include the name of the applicant and a correspondence address. Requests can be made by e-mail.

The authority from which information is sought may not ask why this is being requested nor can they refuse to comply because of what you might wish to do with the information. The Act does not define precisely what sort of information may be sought, however, there are some exemptions. Some are absolute whilst others must fulfil a public interest test, which must be applied to individual requests and judged on the merit of the situation.

One absolute exemption (Part 2 section 21) is information which is published elsewhere and is accessible to the applicant by another means. Another exception (Part 2 section 36) deals with, 'Prejudice to effective conduct of public affairs.' Where an examination board has a limited pool of questions, for example a written or MCQ examination, this exemption could be relevant. To request access would be unreasonable because to release these questions would degrade the quality of the assessment in the future.

Requests must be responded to within 20 days unless further information is sought by the authority in order to identify the precise information you seek. If your request is refused then a reason must be given. There is a process of seeking review and appeal through the Information Commissioner whose decision is almost always binding.

The responsibilities of the candidate

Examination and assessment boards rightly expect candidates to behave honestly and professionally. Academic dishonesty, cheating and plagiarism will not be tolerated and may result, where this relates to medical examinations, to referral to the GMC. 'There is no correlation between success and cheating; cheaters do not perform better on exams.'[3]

Whilst cheating has been likened to stealing, plagiarism has been likened to forgery. The seriousness of academic dishonesty, a set of deliberate, unacceptable behaviours, cannot be underestimated. 'This is superior work. It was excellent when Saint Thomas Aquinas wrote it, just as it is today. Saint Thomas gets an A. You get an F'.[4]

There are many examples of academic dishonesty.

1 Copying another student's answers.
2 Using notes or study guides brought into an examination when this was explicitly not permissible.
3 Passing off the work of someone else as one's own.
4 A candidate getting a surrogate to sit the examination in his place.

There are also more subtle examples such as communicating with someone else by mobile phone during an examination, using a commercial organisation to prepare answers and downloading entire examination papers or research projects from the Internet.

There have been some highly publicised cases of falsification of results, which can have serious and significant consequences for public safety.

Normally examination boards do publish the consequences of academic dishonesty which would at the very least lead to a lower grade being awarded but may debar the student, unless adequate explanation is offered and accepted, from further attempts or even referral to their particular professional body. Often dishonesty is difficult to prove and therefore punishment can be variable. This can best be dealt with by clear definition of what precisely constitutes cheating or plagiarism, preferably with concise (and including cross-cultural) examples. Policies of how cheating or plagiarism will be dealt with need to be clearly communicated and adhered to.

References

1. Quality Assurance Agency for Higher Education (QAA). 2004: Career Education, Information and Guidance: www.qaa.ac.uk/public/COP/codesof practice.htm
2. Freedom of Information Act 2000. www.dca.gov.uk/foi
3. Dowd SB. *Academic Integrity: A review and case study.* (ERIC Document Reproduction Service No ED 349-060). 1992.
4. Alschuler AS, Blimling GS. Curbing academic cheating through systemic change. *Coll Teaching.* 1995; **43**(4): 123–6.

Chapter 15

Post-modernising medical careers: assessment in an age of uncertainty

John Launer

> The language is no sooner minted than it fractures into different perspectives and simultaneously we sense, somewhere in our bones, that it is certainty itself that has ended.
>
> *Paul Hodgkin*[1]

> And immediately
> Rather than words comes the thought of high windows:
> The sun-comprehending glass,
> And beyond it, the deep blue air, that shows
> Nothing, and is nowhere, and is endless.
>
> *Philip Larkin*[2]

Introduction

In this chapter I want to reflect on the position of medicine in contemporary culture and society, and what this means for the future of assessment in the medical profession. To place these reflections in context, I shall begin by describing briefly my own work and professional background, and my impressions of medical practice in its current state.

For the last three years I have worked at the London GP Deanery as clinical supervision lead. My remit has been to promote a culture of clinical supervision among general practitioners (GPs) in London, bringing them closer to the ways of working that are commoner among some of the mental health professions including clinical psychologists and psychotherapists. I have brought to this work my experience as a GP and GP trainer, but also experience as a family therapist, supervisor, and member of a child and family mental health team. In addition, I have a background in the humanities, having been an English graduate and a teacher before becoming a doctor.

Working as I do in many different contexts, with GP patients and with family therapy ones, with trainees and experienced practitioners in many different disciplines, I have been struck by the following impressions.

1 Most medical practitioners appear to function with a high level of commitment and competence. However, as doctors become older and more experienced, factual knowledge may decline while wisdom and intuition may increase, so that a good 30-year-old doctor is an entirely different professional from a good 60-year-old one. Some practitioners remain all-rounders to the end of their careers while others become more focused on one area of clinical or managerial expertise. Doctors – and particularly GPs – tend to work within

local or organisational micro-cultures, so that it is possible for two practices to operate in the same street, for example, with widely different standards and in complete ignorance of each others' performance. Equally, how one practitioner might define competent practice may vary greatly in its scope and level from how another might do so. Some nowadays subscribe to a very technocratic vision of medicine, while others hold on to a whole person or humanistic approach.

2 Accounts of poor or problematical clinical practice appear to be relatively common but such practice can be hard to pin down. Often this is because the problem really belongs to the network, team or organisation as much as the individual. Alternatively, the team may cover up for a doctor with problem, or avoid confrontation. Sometimes there are also aspects of social discrimination or professional deprivation, and this may be enduring and institutionalised. The preponderance of referrals to the General Medical Council (GMC) of south Asian doctors nearing the end of their careers is probably an example of this. The historical dependence of the National Health Service (NHS) in some places on doctors who are known to be ill-trained and ill-equipped for the job is another. Performance problems are thus inseparable from wider social, economic and cultural issues.

3 The mechanisms for trying to address problematical practice are cumbersome and in some ways arbitrary. Flagrant malpractice may evade notice or punishment because of cleverness, conflict avoidance, power imbalance, legal niceties, a compassionate wish to preserve colleagues' livelihoods and self-respect, or for other reasons. At the same time, minor or equivocal faults may become the focus for raucous publicity, execration and professional ruin. The assessment methods currently available do appear to capture something that has meaning and consistency, but other important aspects of performance – both positive and negative – may be too complex or indefinable to reduce to what is measurable. Every doctor works constantly with the risk that any moment of inattention, or any single fragment of ignorance, can potentially lead to physical catastrophe for a patient. Every single doctor is liable to make significant mistakes from time to time. We all live with the fear of being found out. It is no easy task to construct and to sustain a system that can identify consistently bad practitioners without frightening and demotivating practically everyone else. Assessment and regulation are not neutral activities: they reflect social and political imperatives, and they exert their own influences on the ethos of medicine.

4 Health professionals in Britain work in an opaque and constantly shifting context of administrative structures and regulatory agencies that often relate to each other in ways that are bureaucratic or dysfunctional. There is a pervasive mismatch between the insistently personal nature of clinical encounters on the ground, and the managerialised, procedure-driven culture of health service administration and government. This mismatch is demonstrated by increasingly anxious attempts from above to monitor, count and govern professional activities that by their very nature are subtle, unique, ambiguous and elusive. The same process is insinuating itself into the medical consultation itself, where there is a growing tension between traditional, individualised and responsive care on the one hand, and the insistent intrusions of proactive, epidemiological interventions on the other. While some of these interventions

make an undeniable and major difference to some peoples' lives, taken together they also determine what can no longer be done: for example, out of hours care, continuity of care and perhaps family and palliative care.

In this chapter I want to argue that these things are not coincidences, nor have they arisen in a vacuum. They are the results of cultural forces that have to a lesser or greater extent affected much of the world in the second half of the 20th century, and continue to do so. For convenience, I want to refer to these forces by the umbrella term of 'post-modernism' – a word that puts many doctors off, but actually has quite a simple meaning. Following a discussion of what post-modernism means to medicine, I want to venture some ideas about its implications for assessing doctors in the future, with particular emphasis on assessing established GPs.

Post-modernism and medicine

In its philosophical form, post-modernism is the abandonment of any search for a certain, final form of truth, whether it be scientific, political, or moral. It is a loss of faith in progress, or a belief in belief. To put it another way, post-modernism is an acceptance that all forms of truth are only ever defined provisionally, and through social or linguistic agreement, rather than because anything holds true for ever. In its cultural form, post-modernism expresses itself in terms of pluralism and diversity. For example, certainties that would have gone undisputed a generation or two ago – about homosexuality, for example, or about the right of European countries to colonise non-European ones – have within a very short space of time become antique. We no longer live within boundaried communities of belief (of any kind) that remain unchallenged.

Medicine as a discipline stands in a peculiar relationship with post-modernism. On the one hand, medicine is identified with a progressive search for irreducible knowledge – the kind of knowledge that will hold good across all times and in all cultures. On the other hand, medicine always has to function within the wider social context. In a post-modern world, this has given rise to a wide range of tensions. Here are just some examples.

- Reproductive technology now allows a wide range of assisted conception to take place and this will almost certainly include human cloning in the near future. At the same time, there is no longer any universal moral framework to define what is and is not acceptable in this field. In place of such a framework we have committees that test the waters at any given time and offer provisional guidance accordingly.
- There is now a wide variety of treatments for disorders such as depression, ADHD, erectile impotence, obesity, reduced female libido and premenstrual syndrome. Yet alongside this, social scientists and political theorists are arguing that all these conditions – and many others – are constructions brought about by oppressive social and gender relations and by the vested interests of the pharmaceutical industry.
- Clinical governance is becoming universal within medicine, at least in the UK. It provides many mechanisms for dictating whether or not treatments are evidence-based and whether or not these should be prescribed. Medical activity itself has become subject to a plethora of guidance concerning what must be

> enquired into, measured and done. Yet all of this regulated practice operates within a wider framework where doctors are also told to offer choice and to be sensitive to the wishes of consumers. Paradoxically, these wishes may and often do extend to all kinds of investigations and treatments that are entirely outside the domain of orthodox scientific medicine and possibly outside the sphere of western rationality itself.

On the ground, in everyday practice, post-modernism is played out in many ways. It is shown, for example, in the ambivalence with which parents may demand an immediate appointment for a child with a minor viral infection in the belief that the doctor must have the knowledge to cure it, while at the same time questioning the doctor's judgement in offering routine childhood immunisations. Another manifestation is the presence in surgeries of para-medical disciplines (first-contact care practitioners, physician assistants, primary care mental health workers and so on) who are charged with applying evidence-based, protocol-driven treatment, while being subject to lower levels of assessment and regulation than doctors are.

One way of looking at all these phenomena is in terms of authority. While medicine is being required – by society, by governments and by the profession itself – to demonstrate its own authority in increasingly rigorous ways, that same authority is also being undermined by questions that challenge the dominance of any single body of knowledge, or the possessors of that knowledge.

Within this slippery world of certainty-within-uncertainty, of concrete castles built on shifting sands, it is perhaps no surprise that practitioners, managers, regulators and governments find themselves nervously renegotiating at every turn what is and is not normative practice.

Performance and supervision: the uses of ambiguity

Having briefly stated the predicament in which I believe doctors currently find themselves, alongside those who train and assess them, I want to focus on two concepts that may offer a way forward. One is performance and the other is supervision.

Performance

The word performance in the English language is helpfully ambiguous. In a medical context it means competence, but in a theatrical context it means the ability to act a part. It is a cliché to say that doctors are required to be actors and that the acquisition of a good manner – sustainable in moments of insincerity as well as sincerity – is a prerequisite for credibility and self-preservation. Yet there is a serious and necessary side to this kind of performance. In some profound sense, the doctor who cannot *act* like a competent doctor is indeed not a competent doctor.

To put this in even cruder terms, if a doctor attended the London Deanery as the result of a GMC determination and put the question to us, 'Are you just telling me I have to put on a better performance in front of patients?' it would be legitimate to say, 'Yes, of course.' That answer defines the precise nature of competent performance in a post-modern world. It flags up that medicine in all its aspects, from prescribing the right drugs to behaving correctly towards the opposite sex, is the product of social, cultural and political agreement. Transgressing the norms of that agreement is, quite simply, putting on a poor performance in both senses of the word. Failure to perform at a medical level may also represent a more radical failure to understand how human beings have to 'perform' generally in the social world. This may be why performance problems are often inextricably bound up with conduct problems as well.

Supervision

There is a similar ambiguity in the word supervision and it is equally helpful. Supervision is often understood to mean looking over someone's shoulder. It also means looking after them. While some people have tried to promote the view that supervision is solely a developmental task rather than one that involves monitoring and standard setting, I have come to believe that it inescapably has an element of both. For example, it is impossible to supervise even the most competent doctor without holding in one's mind the idea of acceptable practice. Conversely, supervising a challenging and worrying junior will still involve an attempt to bring forth competence – otherwise the conversation would not be happening in the first place.

What this means for the actual conduct of supervision, whether formal or informal, is that one has to move continually between a position of facilitating new understanding on the one hand and introducing information about good and better performance on the other – through advice or questions or both. In supervising a young and recently trained practitioner, this may mean inviting a deeper understanding of contexts, processes, and feelings. In working with the older practitioner, who may be more experienced but rusty in terms of knowledge, it may be more important to establish the contemporary parameters of good practice. The whole enterprise of supervision can be conceptualised as a collective activity by which professions establish, promulgate, refine and develop their standards of practice, both technical and affective. In other words, supervision is a regulatory activity, but it is also one in which the regulations themselves are in a continual state of conversational evolution.

If we are prepared to live with the ambiguities inherent in words like performance and supervision, then we have the potential to navigate a world that is both insistently certain in some of its aspects and yet uncertain in its foundations. We can learn, as it were, to take positions that will protect both ourselves and our patients, while recognising that those positions will nearly always be provisional. We will also be agile enough intellectually to move to different positions as conditions change, or as contexts are reframed. Without becoming cynics or ironists, we will be able to hold on to the importance of maintaining an appearance of professional authority, while having no illusions about how insubstantial that authority really is.

What will this mean for assessment?

Having set out what I consider to be the cultural context for practising medicine for the foreseeable future, I now want to offer some predictions about the likely direction of assessment, specifically in relation to established GPs, since that is my own main interest currently. It should go without saying that in a time of such uncertainty I would place no money on any specific predictions; these are meant more as signposts than as clairvoyance.

First, I see no reason to doubt that the tension or dialectic between scientific medicine and the surrounding cultural context will continue to intensify. Medicine will continue to demand of itself that it can justify itself in statistical and apparently incontestable terms. Governments and managers will seize on this form of discourse more and more because it gives them an administrative and (to some extent) an economic handle on a profession whose autonomy is seen as threatening in many ways. Alongside this rationalistic, or seemingly rationalistic enterprise, there will be a simultaneous breakdown of consensus about where medicine stands as a significant determinant of cultural reality. Globalisation and privatisation are likely to contribute to this process by upping the stakes on both sides. On the one hand they will offer cheaper and competitive forms of conventional medicine such as intercontinental packages for elective surgery. On the other hand they will contest the underlying belief system of medicine itself by promoting alternative remedies and services in the market place, either in pseudo-medical form or through explicitly anti-medical rhetoric.

One option for survival within this fragmented and contradictory reality will no doubt be for doctors to identify themselves entirely with whatever limited and provisional aims are defined at any moment by the state and its corporate allies in the form of targets, desirable actions and outcomes and so forth. There will almost certainly be a convergence between much that is currently considered to be organisational audit (e.g. successful operations performed, patients started on statins) with activities currently considered to be part of professional regulation, including appraisal and revalidation. Given this scenario, doctors who can content themselves with satisfying certain arithmetical measures and who can equate these with good medical practice – either unreflectively or with enthusiasm – will probably pass muster. Alongside this, it seems equally certain that all doctors sooner or later will face regular tests of their knowledge and possibly their technical skills. Certainly, the absence of such tests is becoming increasingly indefensible within a discourse of public accountability.

Thinking beyond this context, however, there will be wider challenge for doctors. In order to reduce the probability of complaints from a public that may become more assertive and possibly litigious – and will certainly have access to more information than is currently imaginable – doctors will need to see themselves more and more as cultural mediators. In other words, they will need to understand their role not as telling people what to do, nor just as good and sympathetic listeners, but as interpreters whose duty is to make sense of one culture (medicine) to people with a wide range of beliefs and expectations drawn from other cultures and micro-cultures. They will need to become far more tolerant of other peoples' constructed realities including those that appear fundamentally incompatible with the medical world view. They will need the skills to create a space in the consultation for an intelligent dialogue between the language of medicine and the many other languages that are brought into the surgery – including languages that are literally as well as metaphorically foreign. Doctors who can do so will be less likely to come under scrutiny by a state that is immensely concerned to represent itself as protective of citizens' rights to self-determination.

In this cultural and political context, doctors will increasingly require the skill of *reflexivity*, namely the skill of observing themselves in interaction with others, together with the ability to note and correct sources of misunderstanding at the level of basic assumptions. A pre-requisite for any kind of reflexivity among doctors will be a humility in relation to the truth and universal applicability of medicine itself and an ability to take a more neutral position in relation to what is right and desirable for patients. Another aspect of this skill will be the capacity to note the power relations inherent in any medical encounter – whether this power belongs to the doctor, the consumer, the team, the government, or some other agency – and to remain both respectful of these power relations, but also willing to bring them into question.

No doubt attempts will be made – and perhaps they should – to test this kind of communicative capacity. This may involve such methods as 360-degree feedback, video review, simulated surgeries, patient satisfaction surveys and so forth. To judge by current trends, these assessment technologies are likely to grow into something of an industry. There is always the risk that such approaches to assessment will become dumbed down to a reductionist set of boxes for ticking, in which the complexity, authenticity and subversiveness of self-awareness and situational awareness are lost. The challenge of professionalism, as always, will be to transcend this risk. We will need to collaborate in forms of testing that define and assess doctors' abilities to respond to patients' needs and wishes, and to function within complex professional and interprofessional networks. We will also need to remain radically sceptical about such testing, recognising that it inevitably has its own effect in recolonising and institutionalising ourselves. The most important assessment we can be involved in as medical practitioners may be a continual self-assessment of our place within the intersecting systems in which we find ourselves.

Although I have painted a generally astringent picture of medical culture and medical assessment in the future, I believe that challenges should also excite doctors. Ultimately, I do not believe that individual doctors as moral agents will lose a vision of medicine that rises above the politically fashionable or expedient, and above a form of practice that is driven by the state, the organisation, consumerism or the market. While respecting all these forces and taking them seriously, a vision of doctors as cultural mediators restores to us a kind of medicine that is fundamentally interactional, interpretative or hermeneutic, one where the 'stuff' we know as doctors has no existence in its own right, but attains meaning only when it enters into dialogue with the other: the patient.

References

1. Hodgkin P. Medicine, postmodernism and the end of certainty. *BMJ.* 1996; **313**: 1568.
2. Larkin P. *High Windows.* London: Faber; 1974.

Index