[✓ YES] NO COND
MH

TH 7/14

AAA

168363162

CUY Downloaded on: 2016-7-13
Needs Before: 20160802

STATUS: ~~COND PEND~~
LENDERS: *CUY CUI TMB VTM VMC

BORROWER: AAA
ILL #: 168363162
SHIP TO ADDRESS: ILL Borrowing 231 Mell St. US Auburn 36849-5606 US-AL Auburn University Libraries

SHIP VIA: Best Method
EMAIL:
MAXCOST: IFM 65.00
VERIFIED: <TN:659266><ODYSSEY:131.204.73.182/ILL> OCLC

LENDER: CUY
RETURN ADDRESS: Interlibrary Services
University of California, Berkeley
133 Doe Library
Berkeley, CA 94720
EMAIL (LOANS): lending@library.berkeley.edu
EMAIL (SCANS): illpmail@library.berkeley.edu

RESTRICTIONS: Do not affix tape, labels, or adhesives to our material. ****Prefer returns sent boxed and traceable****
LENDING NOTES: Two chapters requested--we need a request for each. Do you want to proceed with this request for Chapter 1?

OF ITEM(S): ____*** MUST RETURN TOGETHER *** DAMAGED PRIOR TO LOAN? YES NO
ITEM DETAILS:

TO BE INVOICED: YES [✓ NO]
SHIP DATE:
DUE DATE:

ILL #: 168363162
OCLC #: 74525534
ISSN:
ISBN: 9781846190469

PHOTOCOPY

AUTHOR:
TITLE: Assessment in medical education and training : a practical guide /
IMPRINT: Oxford ; New York : Radcliffe, ©2007.
EDITION:

ARTICLE AUTHOR:
ARTICLE TITLE: Val Wass, et al (Ch 1) and Kamila Hawthorne (Ch 2): The principles of assessment design (Ch. 1) and Assessment in the undergraduate curriculum (Ch 2)
VOLUME:
NUMBER:
DATE: 2007
PAGE(S): 11-~~40~~ 26 Chapter 1 only on this request.
Must submit a separate request for chapter 2.
BORROWING NOTES: We DO NOT charge loan or copy fees to other academic or public libraries. Please ARIEL if no add'l fee. (maxCost: 65.00)
BILLING NOTES: Same; UMI D980003; BRI Customer Code 51-8658 CISTI DD718556

RECORD 3 of 3
Library: Optometry/Health Sciences, CallNo: R837.E9 A75 2007

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

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.

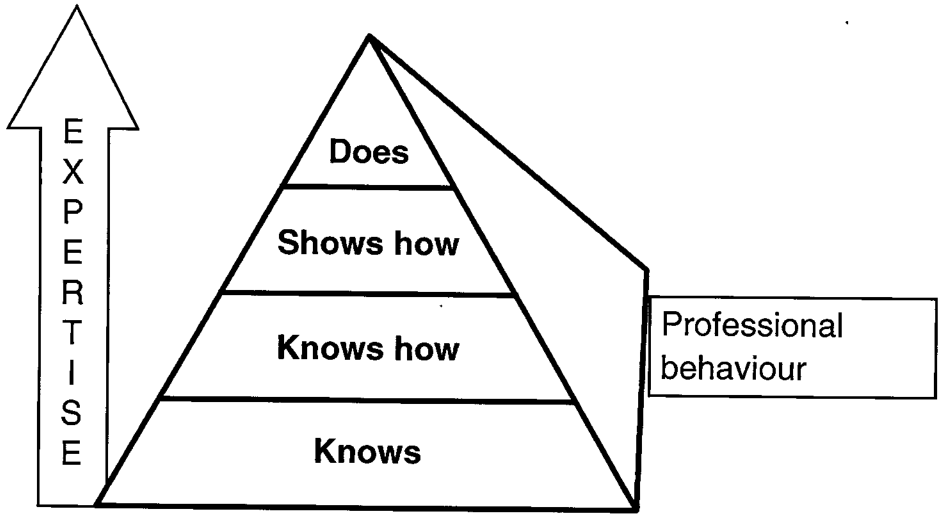

Figure 1.1: Miller's pyramid of clinical competence.[9]

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.

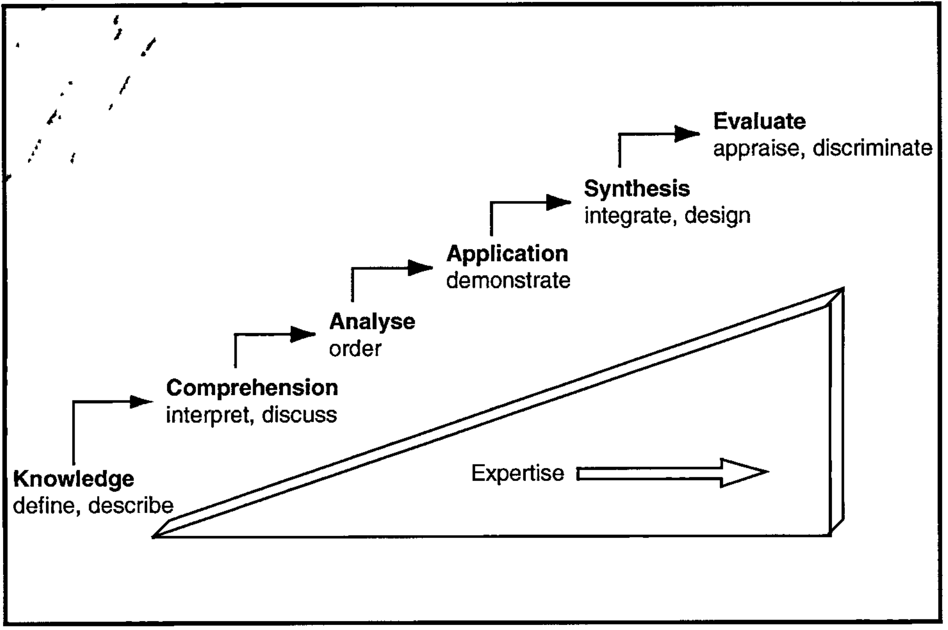

Figure 1.2: Hierarchy of knowledge: Bloom's taxonomy.[13]

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.

utility = reliability × validity × feasibility × acceptability × 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/