



Australian Government
Department of Health and Ageing

National Learning Objectives and Assessment Procedures for the Pharmacological Management of Opioid Dependence



*National
Drug Strategy*

*National
Drug Strategy*

National Learning Objectives and Assessment Procedures for the Pharmacological Management of Opioid Dependence

Project Team

Steve Allsop, Angela Corry and Liz Ernst

Expert Group

Robert Ali, Steve Allsop, James Bell, Roger Brough, Liz Farmer, Tony Gill, Sue Henry-Edwards, Nick Lintzeris, Moira Sim and Michael Tedeschi

© Commonwealth of Australia 2004

ISBN 0 642 82192 5

This work is copyright. Apart from any use as permitted under the Copyright Act 1968, no part may be reproduced by any process without prior written permission from the Commonwealth available from the Department of Communications, Information Technology and the Arts. Requests and inquiries concerning reproduction and rights should be addressed to the Commonwealth Copyright Administration, Intellectual Property Branch, Department of Communications, Information Technology and the Arts, GPO Box 2154, Canberra ACT 2601 or posted at <http://www.dcita.gov.au/cca> .

Publication approval number: 3189 (JN 7348)

Publications Production Unit
Australian Government Department of Health and Ageing



Contents

Section 1.	Introduction	1
	Background	1
	Overview	2
Section 2.	Assessment of clinical competence: Literature review	5
2.1	Overview	5
2.2	Introduction	6
2.3	Evolution of clinical competence assessment	8
2.4	An introduction to measurement properties	9
2.5	Blueprinting	13
2.6	Review of assessment methods	13
2.6.1	Written Tests	13
	Multiple Choice Questions (MCQs)	13
	Extended Matching Questions (EMQs)	14
	Essays	17
	Modified Essay Questions (MEQ)	18
2.6.2	Tests of Clinical Decision-Making	19
	Patient Management Problems (PMPs)	19
	Key Feature Problems	19
	Objective Structured Clinical Examinations (OSCE)	21
	Long Case Exams (e.g. viva)	23
2.6.3	Performance in Practice	24
	Audit	24
	Peer Review Ratings	25
	Observation of Practice (e.g. clinical placement)	25
	Videotaped Consultations	26
	Standardised Patients in Practice Settings	26
	Work Logs	27
2.6.4	Patient-Based Assessment	28
2.7	Conclusion	29
2.8	References	30
Section 3.	Learning objectives	33
Section 4.	Blueprint and decision matrix	41
Appendix.	Writing key feature problems	51





1 Introduction

Background

The National Expert Advisory Committee on Illicit Drugs is concerned to develop and maintain safe and effective treatment for opioid dependence. This committee commissioned the formation of an expert group to identify learning objectives and assessment procedures for medical staff who become involved in pharmacotherapy for opioid dependence. The final document and procedures have relevance for the Chapter of Addiction Medicine within the Adult Medicine Division of the Royal Australasian College of Physicians. Thus, the learning objectives and assessment procedures will be lodged with the Chapter, to maintain relevance and currency.

The project was funded by the Australian Government Department of Health and Ageing and managed by the Drug and Alcohol Services Council of South Australia. The Next Step Specialist Drug and Alcohol Services (WA) were commissioned to facilitate the project.

The project team consisted of individuals with medical expertise, especially in responding to drug related harm, and in clinical education.

A/Professor Steve Allsop	Next Step Specialist Drug and Alcohol Services (WA) and (Chair) Centre for International Health, Curtin University of Technology
A/Professor Robert Ali	Drug and Alcohol Services Council (SA)
Dr James Bell	The Langton Centre (NSW)
Dr Rodger Brough	Alcohol and Drug Physician (VIC)
Ms Angela Corry	Next Step Specialist Alcohol and Drug Services (WA) (Senior Project Officer)
Ms Liz Ernst	Next Step Specialist Drug and Alcohol Services (WA) (Project Officer)
A/Professor Liz Farmer	Department of General Practice, Flinders University (SA)
Dr Tony Gill	Drug Programs Bureau, NSW Health Department
Dr Nick Lintzeris	Turning Point Alcohol and Drug Centre Inc. (VIC)
Dr Moira Sim	Next Step Specialist Drug and Alcohol Services (WA)
Dr Michael Tedeschi	The Canberra Hospital (ACT)

Permission has been obtained from relevant authorities, in all Australian jurisdictions, to include assessment protocols developed as part of individual jurisdictional processes.



Overview

Over the last 20 years, the use of pharmacotherapies in treating opioid dependence has greatly expanded. Until recently, methadone maintenance was the major treatment selected. Recently buprenorphine has been added as an important treatment choice and the viability of LAAM is currently being explored. Naltrexone is another treatment option for a small number of opioid dependent people.

In 1997, the Australian Government published Learning Objectives for Methadone Prescribers (Allsop et al., 1997) to guide training and assessment procedures. The expansion of research evidence and treatment options requires that the Learning Objectives should be updated, and that assessment procedures should be developed to facilitate learning and certification of capability and authorisation to practice, where this is required. The aim of this document is to respond to this requirement, by ensuring that medical practitioners both commencing and continuing to manage patients with pharmacotherapies do so safely and effectively. It should be considered in the context of the relevant national and jurisdictional policies and clinical guidelines (e.g. Lintzeris et al., 2001; National Drug Strategy, 1997).

The Learning Objectives and Assessment Procedures are intended to facilitate the implementation of safe and effective treatment, using pharmacotherapies, for opioid dependence. They can be used to assist in the design of curricula and course content, in summative assessment (e.g. as part of a process of authorisation to use methadone and buprenorphine treatment for opioid dependence) or in formative assessment (e.g. providing feedback to build knowledge and skills). Summative assessment is directly associated with authorisation processes for engaging in treatment using methadone and buprenorphine. It is not a requirement for treatment using naltrexone.

The Learning Objectives and Assessment Procedures were developed through consultation with a group of clinicians and clinical educators with expertise in this domain. Members of the expert group, key professional groups and key clinical stakeholders from around Australia have reviewed the final document.

The expert group identified the Learning Objectives, and related Tasks, associated with safe and effective management of opioid dependence using pharmacotherapies. They subsequently identified and reviewed existing assessment procedures and developed new protocols to cover the range of Learning Objectives and Tasks and related knowledge, skills and attitudes that contribute to capable and effective clinical practice.

The document consists of several sections:

1. Introduction
2. A literature review of clinical assessment
3. The Learning Objectives and Tasks involved in safe and effective management of opioid dependence using pharmacotherapies
4. A Blueprint and Decision Matrix to help guide selection of assessment procedures for knowledge, skills and attitudes related to the Learning Objectives

As already indicated, the intention was to develop assessment processes that can be used for valid, reliable and feasible formative assessment (e.g. enabling the medical practitioner to learn from tests and receive feedback on which to build knowledge and skills) and summative assessment (e.g. for certification/authorisation to prescribe buprenorphine and methadone for the treatment of opioid dependence). Assessment should be able to provide an indication that the medical practitioner has the capability (knowledge, skills and attitudes) to undertake certain tasks (RACGP, 1998), in this case to safely and effectively treat opioid dependent patients using a range of pharmacotherapies. Ideally, assessment would be based on objective criteria, requiring measurement. However, sometimes measurement is not possible and judgments, involving more subjectivity, may need to be made (RACGP, 1998). This means that the different areas of capability and effective practice will need a variety of different assessment methods. A multiple-choice examination, for example, could be a more valid test of knowledge than of communication skills, which might be best assessed with an interactive test. Due to the complexity of clinical competence, many different testing formats should be used (Wass et al., 2001). Also, as new knowledge and new treatment procedures emerge, these assessment procedures will need to be updated. Section 2 of this document describes the history and purpose of assessment, followed by a review of the strengths and limitations of the various assessment procedures (see 2.6).

Section 3 of this document identifies the Learning Objectives that have been identified as being central to safe and effective management of opioid dependence using pharmacotherapies. The Learning Objectives consist of six competencies, each with their own tasks. Effective practice is inferred through performance criteria for each task, relating to knowledge, skills and attitudes. They are important in terms of identifying the factors that contribute to capability in this domain and inform the content of learning programs and assessment procedures. The six competencies are:

- **Attitudes and Professionalism**
- **Assessment**
- **Developing a Treatment Plan**
- **Management of Co-existing Conditions**
- **Patient Management**
- **Quality Assurance**

Conceptual frameworks against which to plan assessments are important. As such, assessment programs need to match the competencies being learnt and the teaching formats being used. Tests need to be planned against agreed learning objectives of the competencies essential to safe and effective practice. This task has been referred to as 'Blueprinting'. A Blueprint is a document providing specifications for selecting a sample of clinical problems for inclusion on an examination from the domain of all possible problems (Wass et al., 2001). A Blueprint can also help identify the most appropriate method(s) of assessment for a particular domain. A Blueprint for Assessment is provided in Section 4 of this document. In addition, a Decision Matrix is provided to indicate the various situations in which different assessment procedures may be used.

The project team also developed a range of different assessment procedures, and it is intended that they will be lodged with the Chapter of Addiction Medicine within the RACP. The bank of assessment procedures will allow a variety of methods to be used, matched to the variety of Learning Objectives and Tasks. The current bank is by no means comprehensive. Clearly, the history in this area means that some assessment procedures are well advanced (e.g. Multiple Choice Questions) while others are less well developed. The methods that are covered range from the more formal methods of cognitive testing that are well supported by a science of psychometrics, to the less scientific methods of assessing physician performance, such as clinical placement and peer review. It is anticipated that those involved in assessment in this field will contribute to continual development of this initial bank of procedures, by submitting new protocols to the Chapter of Addiction Medicine within the RACP for review and inclusion. It is expected that jurisdictions may need to adapt some of the assessment procedures to meet jurisdictional requirements.

References

1. Allsop S., Bell, J., Brough, R., Dwyer, P., Edmonds, C., Lintzeris, N. and Mohindra, V. (1997) Learning Objectives for Methadone Prescribers. AGPS, Canberra.
2. Lintzeris, N., Clark, N., Muhleisen, P., Ritter et al (2001) National Clinical Guidelines and Procedures for the use of Buprenorphine in the Treatment of Heroin Dependence. National Drug Strategy, Commonwealth Department of Health and Aged Care, Canberra.
3. National Drug Strategy (1997) National Policy on Methadone Treatment, Commonwealth Department of Health and Family Services, Canberra.
4. The Royal Australian College of General Practitioners (1998) The College Examination - A Handbook for Candidates and Examiners, The Royal Australian College of General Practitioners, Melbourne.
5. Wass, V., van der Vleuten, C., Shatzer, J. & Jones, R. (2001). Assessment of clinical competence. *Lancet*, 357, 945-49.

2 Assessment of the clinical competence: Literature review

Produced by Angela Corry¹, Liz Ernst¹, Liz Farmer² & Steve Allsop¹

Next Step Specialist Drug and Alcohol Services, Western Australia¹
Department of General Practice, Flinders University, South Australia²

2.1 Overview

Clinical competence is clearly a multi-dimensional construct, and as such, emphasis on exact interpretations varies. As a consequence, measures developed to assess competence in the medical field have historically focused on a variety of conceptualisations. There is some consensus around the apparent differences between the concepts of **competence** – “what a physician is capable of doing” and **performance** – “what a physician actually does in his/her day-to-day practice”. A level of agreement exists that the relationship between the assessment of competence and performance is important from a quality assurance perspective. As such, there is increasing consensus that the most important consideration in selecting methods of assessment is that they measure performance and not the ability to perform *per se*.

Traditional measures were based on an implicit conception of competence. However more recent measures have attempted to assess the more complex cognitive tasks associated with competent clinical performance. Assessment of actual competency is a challenge for all involved in clinical competence testing.

There are several key measurement issues that need to be considered when designing assessments of clinical competencies. The first issue requires a clear view of the purpose of the test. That is, does it need to have a summative or a formative function? Within the context of assessing competence related to the pharmacological treatment for drug and alcohol dependence, summative assessment usually refers to the accreditation process. Formative assessment within this context refers to assessment activities related to continuing professional development. With an increasing focus on the performance of doctors and on public demand for assurance that doctors are competent, assessment needs to have a summative function. Yet if assessment focuses only on certification and exclusion, the all-important influence on the learning process will be lost. Unfortunately, tests that have both formative and summative function are hard to design (Wass, van der Vleuten, Shatzer & Jones, 2001).

Secondly, conceptual frameworks against which to plan assessments are essential. As such, assessment programs must match the competencies being learnt and the teaching formats being used. Therefore, tests need to be planned against agreed learning objectives of the competencies essential to the specialty, in this instance, addiction medicine. This task is referred to as 'Blueprinting'. A Blueprint is a document providing specifications for selecting a sample of clinical problems for inclusion on an examination from the domain of all possible problems (Wass et al., 2001).

Issues related to test reliability and validity require due consideration in the development and selection of appropriate instruments. Validity issues need to be addressed and this requires selecting appropriate test formats for the competencies to be tested. This issue is of greater importance in formative testing, as the test needs to be attractive and meaningful to doctors. Reliability issues also need to be considered. The most important issue to consider here is the requirement of adequate sampling and test length. As clinical competencies are inconsistent across different tasks, consideration of these aspects is crucial if high stake decisions are required.

Reliability and validity criteria as well as the feasibility and acceptability of assessment methods have been the focus of much research and subsequent conjecture around the concept and assessment of clinical competence. Development of reliable measures of performance with predictive validity of subsequent clinical competency and a simultaneous educational role is a gold standard yet to be achieved (Wass et al., 2001).

In this paper a range of assessment methods are reviewed, including discussion of strengths and weaknesses and the implications for assessment selection. All of the approaches to evaluating professional competence have some strengths and weaknesses. For performance testing, evaluation and generalisation are the weak links, for objective testing, extrapolation is the weak link and for simulations, any of the links may be strong or weak depending on the simulations. It may be concluded that arguments for and against a particular testing method should be weighed against each other, and trade-offs made. Norman, van der Vleuten and de Graff (1991) suggest these trade-offs be made as a result of careful consideration of issues resulting from the purpose of the testing situation –practicality, educational impact and acceptability (p26).

2.2 Introduction

Over the course of time, many authors in the area of medical research and medical education have struggled to define clinical competence. The relevant literature includes a large vocabulary incorporating terms such as ability, behaviours, performance, clinical judgement, and clinical reasoning, habitual performance, problem-solving, clinical competence and the combination of knowledge, skills and attitudes. Fundamentally, this lack of clarity means that different authors mean different things when using the word competence. Inherently related to the use of the concept of competence is the use of measurement instruments. Consequently, measures developed to assess competence have historically focused on one or another interpretation of exactly what competence is and inevitably this has led to disparate views about the efficacy of the different methods (van der Vleuten, 1996).

Clearly clinical competence is a multidimensional set of attributes. There does appear to be some agreement around the apparent difference between the concepts of competence and performance. There is consensus that competence generally means 'what a (physician) is capable of doing' and performance 'what a (physician) actually does in his/her day-to-day practice'. In spite of this, there are

divergent views about the most effective methods of assessing one concept while predicting the other (Rethans, van Leeuwen, Drop, van der Vleuten & Strumans, 1990). There is a plethora of literature outlining studies attempting to determine whether measures of competence are good predictors of performance, much of which fails to establish this relationship. Nevertheless, the relationship between the assessment of competence and performance is important from the perspective of quality assurance.

A fundamental belief within education is that the skills of the learner assessed in one domain or discipline can be transferred or generalised to others. For example, the individual who demonstrates excellence in multiplication will also do well in division. Within medical education the concept of a generalisable skill has been challenged and ample evidence suggests that performance of an examinee on one case or clinical situation is not a good predictor of performance on another case or situation (Turnbull, Danoff & Norman, 1996). This phenomenon is referred to as 'case specificity' or 'content specificity', and has important implications for testing as it impacts on the reliability score of a test. A number of authors have concluded that this is inherently a psychometric problem, in that the instruments used were just not good enough. Conversely, there are those that argue there is little correlation between implicit definitions of what is to be measured and the subsequent selection of measurement instruments. The problem is generally not due to poor quality of a test method but rather the fact that performance seems to be determined more by an examinee's specific knowledge and experience in relation to each case or situation than by his/her general problem-solving skills (Turnbull et al., 1996). As such, the tradition of assessing skills from one or a small number of cases or clinical encounters is of limited value in assessing clinical competence. Newble and Swanson (1988) suggest this dictates sampling from a broad range of clinical situations and increased testing time in order for adequate reliability to be achieved. Figure 1 (below) provides a diagrammatic representation of the relationship of content specificity to sampling and reliability when 3 hours of testing time is available (L. Farmer, personal communication, July 2001).

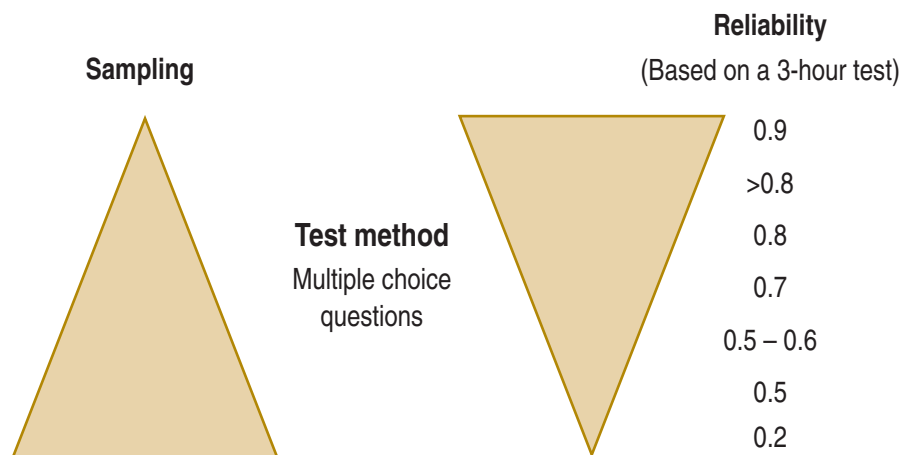


Figure 1. Relationship of content specificity to sampling and reliability (Farmer, July 2001, national workshop presentation).

2.3 Evolution of clinical competence assessment

First assessments undoubtedly involved oral examinations (Case, 1997). Concerns about the inherent subjectivity of oral tests led to the development of written tests. The first objectively scorable standardised written test, Army Alpha, was used during WWI to classify masses of military personnel (Frederiksen, 1984). Essay-type examinations were replaced by multiple choice questions (MCQs) in 1937 after the WHO had demonstrated their application in testing school children (Frederiksen, 1984). As the feasibility of MCQs over written tests became more obvious, their use became widespread. So too did their popularity as they were able to cope with the increased logistical demands (Frederiksen, 1984).

Traditionally, implicit conceptions of the nature of professional competence were used. Competence was seen as an aggregate of different components, which were considered relatively distinct from each other and were thought to be relatively stable across situations and time. The development of competence was considered equal to the development in each component with incremental progress resulting from learning experiences.

In the sixties, attempts were made to measure clinical reasoning ability or problem solving, leading to the development of written simulations where examinees were presented with a patient problem and then asked for management decisions. The decisions and answers to questions were taken as an index of an examinee's problem-solving ability. As computer technology evolved, pencil and paper formats were often replaced by computer simulations. However, three significant consistent empirical outcomes cast doubt on the existing conceptual framework of problem solving. One of these outcomes was 'case specificity'. Secondly, level of experience was not predictive of performance such that those with greater levels of experience did not necessarily perform better, and thirdly, that reliable scores on problem-solving tests correlated highly with other measures including MCQs, which were thought to measure only knowledge. The practical implications of these findings meant that tests needed to be lengthy to achieve minimum reliability and as a logical consequence most institutes abandoned these resource-intensive and therefore expensive simulations.



Educational reform in the seventies and eighties prompted new approaches to teaching and learning. With this came an increasing expectation for a more active role in knowledge acquisition and subsequent concern that conventional assessment methods tended to reinforce old and unwanted learning behaviour (van der Vleuten, 1996). For example, MCQs with their inherent inability to assess complex cognitive tasks were felt to reinforce passive consumption of learning material. The belief that assessment drives learning led to the development of assessment methods that measured the process of learning more directly. While these were not widely introduced, they explicitly highlighted the educational value of assessment. To promote learning, it was considered necessary for assessment to be educational and formative, enabling students to learn from tests and receive feedback on which to build their knowledge and skills. This had particular relevance for continuing medical education where methods such as MCQ, which encourage rote learning, are less meaningful (van der Vleuten, 1996).

At the end of the seventies and eighties, previous simulation-based assessments were advanced to include the assessment of actual performance using standardised live simulations of clinical situations. Pioneered by Harden in 1975, this multistation test of clinical skills using direct observation to assess performance is commonly referred to as an Objective Structured Clinical Examination (OSCE). Multiple station assessments can use simulated patients (lay people trained to simulate clinical scenarios), real patients and/or dummies for assessing procedural skills (van der Vleuten, 1996).

OSCEs provide a flexible approach to testing in which a variety of methods can be embedded to assess clinical skills (van der Vleuten & Swanson, 1990). Individual OSCE stations may use any type of assessment method, and can assess a wide variety of clinical skills demonstrated on a patient (real, simulated or a dummy). This multiple station examination has become an increasingly popular approach to evaluating clinical skills, even being hailed as the “new gold standard in assessing clinical competence” (Sloan, Donnelly, Schwartz & Strodel, 1995, pg. 737).

Most medical schools all around the world use OSCEs in their assessment programs. However, problems with case specificity still apply, therefore sampling must be wide (e.g. approximately 20 stations, or three hours testing time) for reliable high stakes testing (van der Vleuten, 1996). In instances where testing is time-limited, and therefore the number of testing stations possible is relatively small (such as assessing for authorisation to prescribe pharmacotherapies for opioid dependence), test results are typically unreliable.

Key Feature Problems (KFP) were developed in Canada in the 1990's to overcome problems of case specificity. This format resulted from a six-year project commissioned by the Medical Council of Canada in 1996 (Page & Bordage, 1995; Page, Bordage & Allen, 1995; Bordage, Brailovsky, Carretier & Page, 1995) The KFP format is considered the cornerstone of new written assessment formats of clinical decision-making skills (Page, Bordage & Allan, 1995) and is currently used in the RACGP Fellowship Examination (Spike, 1997).

2.4 An introduction to measurement properties

There are certain test standards or criteria that need to be met before an assessment instrument can be used. Such criteria include conventional standards of reliability and validity and, perhaps more importantly, feasibility or acceptability. Further, educational considerations should also be taken into account. Table 1, adapted from Neufeld (1985), provides a checklist of such criteria. As the aim of any clinical competence assessment is to obtain a true measure of an attribute, the checklist is useful in that it can help identify sources of variation that may contribute to a score.

Table 1. Measurement Properties of Tests of Clinical Competence

Properties	Synonym(s)
<p>1. Reliability A quantitative expression of the reproducibility with which the instrument measures the same event on different occasions, with different observers, etc.</p>	Objectivity Repeatability Consistency Reproducibility
<p>2. Validity Non-numerical or quantitative expressions to indicate the degree to which an instrument “truly” measures what is intended.</p> <p><i>Non-numerical measures:</i></p> <p><u>Face validity</u> A non-numerical judgement of the degree to which a test appears to measure the attribute of interest.</p> <p><u>Content validity</u> The extent to which an instrument “covers” or samples all aspects of the attribute of interest (e.g. an educational objective).</p> <p><i>Quantitative measures:</i></p> <p><u>Concurrent Validity</u> Statistical association with the “best” external measure available.</p> <p><u>Predictive Validity</u> Association with some relevant outcome measure obtained some time in the future.</p> <p><u>Construct Validity</u> The demonstration of expected (hypothetical, theoretical) differences, using the test in question.</p>	Common sense Credibility Comprehensiveness
<p>3. Feasibility This includes consideration of costs, scheduling, and logistics.</p>	Criterion-related validity Concurrent criterion validity Accuracy Predictive criterion validity Prognostic accuracy Substitution (direct measurement of attribute not available; a hypothetical approximation is used instead)
<p>4. Educational Considerations / Appropriateness Does the test selected match with its intended educational purpose? Such as: Student learning Certification Decision making Program modifications</p> <p>Use Are the users of the information provided by the test aware of its strengths and limitations?</p> <p>Side Effects Are there ways in which the test does more harm than good?</p>	Applicability Acceptability Test mismatch Test anxiety

Adapted from Neufeld, 1985.

A common way to test reliability is 'internal consistency' (a method of determining if the test is 'balanced') using Cronbach's alpha. On a scale of 0 - 1.0, 0.8 is seen as the minimum requirement for reliable measurement (Wass et al., 2001, pg. 947). Rater variability (which either compares how the same examiner rates the same performance over time, known as "intrarater reliability" or how the same test performance is rated by different examiners, known as "interrater reliability"), contributes a very small amount of unreliability. Case specificity, or "intercase reliability" actually drives the majority of test bias (L. Farmer, personal communication, July 2001; Wass et al., 2001). The term 'objectivity', implying freedom from bias, is commonly used in referring to the reliability of a test. The more objective a test the better it is considered to be; the more subjective the 'softer' it is (Wilson, 1975). Because human judgement is involved in selection of the test-setter, selection of the content areas, selection and wording of the items and designation of the most correct response, bias is therefore inherent in assessment. The challenge is to recognise the bias and minimise it (Case, 1997).

Validity is the other common expression in measurement properties although possibly the most confusing. By definition, validity is the degree to which a test "truly" measures what it is intended to measure. Face validity and content validity are terms that tend to be used interchangeably with credibility and comprehensiveness respectively. Face validity or credibility simply requires the application of reflective commonsense; "Does this instrument appear to measure the attribute of interest?". Wrongly, many licensing boards rely on performance on an MCQ examination as a general reflection of competence when in fact there is an enormous difference between performance on this type of test and the competent care of patients (Neufeld, 1985).

No valid assessment methods that measure all facets of clinical competence have been designed. Miller's pyramid of competence (taken from Wass et al., 2001) which is presented in Figure 2 (below) outlines the issues involved in analysing validity. The base represents the knowledge components of competence: 'knows' (basic facts) followed by 'knows how' (applied knowledge). These can be easily assessed with basic written tests such as MCQs. The 'shows how' is a more important facet of competency required by a qualifying doctor, requiring hands on, not in the head, demonstration. However, the ultimate goal for a valid assessment of clinical competence is to test what the doctor actually 'does' in the workplace. Over the past 40 years, research in this area has focused on developing valid ways of assessing the summit of the pyramid, i.e. a doctor's actual performance.

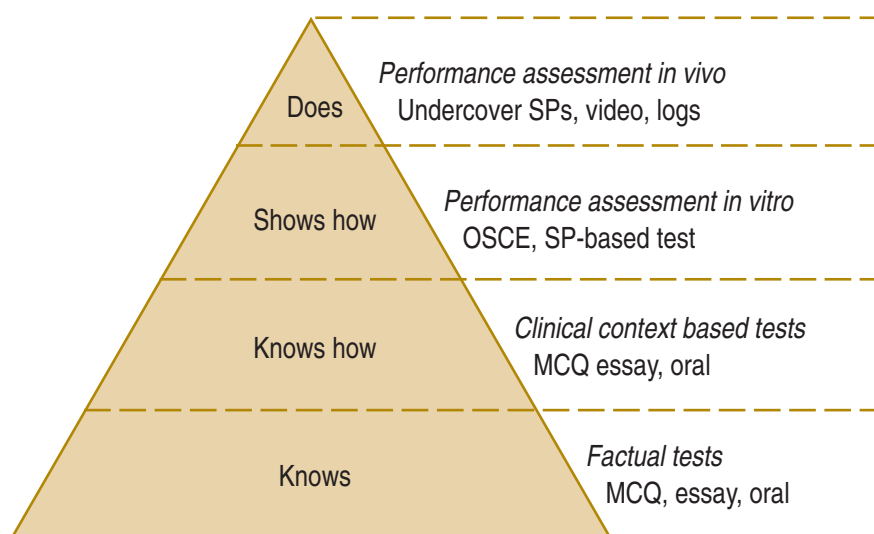


Figure 2. Miller's pyramid of competence (taken from Wass et al., 2001)

Content validity or comprehensiveness describes the extent to which a test covers or samples the area of competence under consideration. The remaining expressions of validity are usually quantitative measures and may involve comparison with some “gold standard” or criterion for measuring an attribute, or where no gold standard exists, a hypothetical construct is substituted. Concurrent validity and predictive validity are forms involving comparison with a criterion measure, but differ in terms of when the measure is obtained. Concurrent validity compares performance on the measure of interest with performance on the best existing external measure available at the time, while predictive validity applies the criterion measure at some point in the future.

Construct validity is used when there is no gold standard with which a new test can be compared. Instead, attempts are made to demonstrate expected differences in performance between individuals or groups differing on some other dimension that is hypothesised to be related to the competence being assessed, such as increased training or experience. For example, it is expected that clinicians with more experience would perform better on a written test of clinical problem-solving ability than less experienced clinicians, with the test being validated if a difference is found. However, if no difference is found the test may either be inadequate or the hypothetical construct may be incorrect. Thus, this approach is relatively weak as concomitant variables which also influence performance are often ignored (Neufeld, 1985).

More recently, authors have emphasised the impact that assessment programs have on the learner. When assessment objectives are poorly matched to educational objectives, there is strong evidence that the objectives of the assessment will prevail. van der Vleuten (1996) suggests assessment may drive learning in a number of ways. Assessment has been shown to influence learning through its content. For example, if a test aims to assess recall of isolated facts then invariably examinees tend to rote learn these facts, thus limiting the validity of such an instrument in assessing performance. Others have argued that assessment should be viewed as a learning exercise in itself and not simply as a decision tool. Commonsense suggests that tests should reflect what you want students to learn. Case (1997) advises of the importance of “developing a clear map of where you want the curriculum to go; drive student learning in the same direction; and finally, develop assessments in which teaching to the test is a valid use of instructional time” (p4). Testing specifically for summative purposes where the focus is on exclusion of sub-standard performers overlooks the powerful influence that formative assessment methods can have on learning (i.e. self-assessment to identify areas for increased learning).

Feasibility is a non-numerical measure of whether a given assessment method can be used in the real world. Also known as applicability or acceptability, this measure needs to consider whether the test is affordable and whether it can be administered in a feasible way in terms of scheduling and logistics. Arrangements for updating the test and maintaining its quality also need to be relatively simple. For formative testing purposes, an assessment method needs to be particularly attractive to the student.

2.5 Blueprinting

Conceptual frameworks against which to plan assessments are essential. An important step to guide the selection of problems for an examination is the development of an examination blueprint. A blueprint is a document providing specifications for selecting a sample of clinical problems for inclusion on an examination from the domain of all possible problems. Assessment programs must match competencies being learnt and the teaching formats being used. All tests should be checked to ensure that they are appropriate for the objective being tested. Many medical curricula define objectives in terms of knowledge, skills and attitudes. These cannot be properly assessed by a single test format. A multiple choice examination, for example, could be a more valid test of knowledge than of communication skills, which might be best assessed with an interactive test such as a clinical viva. Due to the complexity of clinical competence, many different testing formats should be used (Wass et al., 2001).

2.6 Review of assessment methods

A variety of assessment methods will now be discussed in terms of their psychometric properties, feasibility and educational concerns. The methods that are covered in this section range from the more historic formal methods of cognitive testing that are well supported by a science of psychometrics, to the less scientific methods of assessing physician performance, such as peer review. The strengths and weaknesses of each method are highlighted.

2.6.1 Written Tests

Multiple Choice Questions (MCQs)

A MCQ is a written question followed by several (usually 4 or 5) alternative response options. Examinees are required to select the correct response from the list of alternative options, usually by circling their response. In its simplest format, only one response option is correct. However modifications can require examinees to select a number of correct options from the response list.

It is well known that examinations made up predominantly of MCQs have dominated the assessment process in medical education for over 40 years, particularly for licensing and certification. This reliance on MCQs is justified by psychometric studies showing excellent reliability and their ability to handle large numbers of examinees. However, the role of MCQs in the assessment of clinical competence has changed as the constructs of clinical competence have been defined.

Example of a multiple choice question (MCQ)

A suitable starting dose of buprenorphine for a heroin dependent person reporting 2 heroin injections a day, who reports last use 12 hours ago, and who has mild signs of withdrawal, is:

1. 2mg
2. 6mg
3. 12mg
4. 20mg
5. 30mg

Answer: 2

(taken from NSW Buprenorphine Examination, 2001)

Strengths

- Sampling reliability is assured in a well-constructed test, as it is possible to use large numbers of completely independent items.
- A large number of items can be easily marked, with minimal rater error thought to exist (Newble & Swanson, 1988).

Weaknesses

- Familiar criticisms of MCQs include their low face validity – that is - MCQs do not really measure performance. Some imply they are not an accurate measure of performance because clinicians do not routinely answer MCQs as part of their daily practice, whilst others suggest they don't accurately measure performance due to inappropriate selection of content.
- Newble, Baxter and Elmslie (1979) suggest that MCQs measure a combination of what the student knows, partially knows, can guess, or is cunning enough to surmise from cues in the questions. This cueing effect is one of the major criticisms of MCQs and has often led to them being called "multiple guess".
- The constraint of one correct response to every question, is a serious limitation in the domain of medicine where consensus has not been achieved. There is no direct assessment of how the examinee reflected upon or deliberated about the alternatives, which is precisely the point in controversial cases where no single correct answer exists.
- With MCQs designed with questions that involve either establishing a diagnosis or applying principles of management, it is difficult to determine if the answers are generated from rote memory or by a more abstract, purposeful relating of principles to the case at hand, i.e. judgement or problem-solving (Elstein, 1993).
- MCQs are widely disliked at a post-graduate level (L. Farmer, personal communication, September 2001)

Implications

- MCQs can and should be designed to assess 'understanding' systematically and not just to test factual recall ability. They should assess the ability to transfer between basic science concepts and principles and the 'practical wisdom' employed in clinical diagnostics and therapeutics (Elstein, 1993).
- The main challenge to MCQs' validity is in establishing the link between scores on the test and performance in practice.
- It is common practice to supplement MCQs with other techniques in order to obtain valid and reliable data on all essential aspects of competence (Levine, McGuire & Nattress, 1970).
- Their place in testing 'knowledge' still remains alongside other written tests. However attempts to measure more complex cognitive constructs of clinical competence require this method to be supplemented with other techniques.

Extended Matching Questions (EMQs)

The EMQ format refers to a question-matching format with more than five options. As with traditional matching items (i.e. MCQs), extended matching items are grouped into sets, with a single option list used for all items in a set. Well constructed EMQs include four elements: a *theme*, or general topic that

is addressed by the item set; a *lead-in statement*, which provides the directions for the set; an *option list*, which provides the response choices that apply to the items in the set; and two or more *item stems*, which are the questions to be answered.

Work to improve MCQs resulted in the development of EMQs. It is argued that the EMQ format provides a good compromise between free response questions (i.e. essay and short answer) and traditional MCQs, retaining most of the advantages of each and avoiding many of the disadvantages (Fenderson, Damjanov, Robeson, Veloski & Rubin, 1997).

Example of an extended matching question (EMQ)

Theme: Fatigue

- Options:**
- a) Acute Leukemia
 - b) Anemia of chronic disease
 - c) Congestive heart failure
 - d) Depression
 - e) Epstein-Barr virus infection
 - f) Folate deficiency
 - g) Glucose 6-phosphate dehydrogenase deficiency
 - h) Hereditary spherocytosis
 - i) Hypothyroidism
 - j) Iron deficiency
 - k) Lyme disease
 - l) Microangiopathic hemolytic anemia
 - m) Miliary tuberculosis
 - n) Vitamin B₁₂ (cyanocobalamin) deficiency

Lead-in: For each patient with fatigue, select the most likely diagnosis. Each option can be used once, more than once, or not at all.

- Stems:**
1. A 19-year old woman has had fatigue, fever, and sore throat for the past week. She has a temperature of 38.3C (101 F), cervical lymphadenopathy, and splenomegaly. Initial laboratory studies show a leukocyte count of 5000/mm³ (80% lymphocytes, with many lymphocytes exhibiting atypical features). Serum aspartate aminotransferase (AST, GOT) activity is 200 U/L. Serum bilirubin concentration and serum alkaline phosphatase activity are within normal limits.
 2. A 15-year-old girl has a two-week history of fatigue and back pain. She has widespread bruising, pallor, and tenderness over the vertebrae and both femurs. Complete blood count shows hemoglobin concentration of 0.7g/dL, leukocyte count of 2000/mm³, and platelet count of 15,000/mm³.

(taken from Case & Swanson, 1993)

Strengths

- Fenderson et al., (1997) noted that extended matching and un-cued tests had considerable advantages over multiple choice and true/false examinations. They were found to be more reliable and valid, better able to discriminate the well prepared from the marginal student and well suited for testing core knowledge.
- Relative to other multiple-choice formats, there is less cueing and less chance of the examinee guessing the correct response, both because there are more options and because the options list includes all relevant responses (Case & Swanson, 1993).
- The systematic approach to item writing – inherent in the use of extended-matching format – can greatly reduce the technical flaws in item phrasing that are commonly found in one-best-answer questions.
- The long option list allows inclusion of all relevant options, rather than requiring item authors to guess the three or four distracters that they think will be most appealing to examinees (Case & Swanson, 1993).

Weaknesses

- Many people argue that, like MCQs, only trivial knowledge can be tested in this way and the active generation of knowledge is avoided (Wass et al., 2001).

Implications

- Case and Swanson (1993) suggest the format of EMQs actually aids in specifying and organising examination content. For example the utility of extended matching sets allows for the development of several content parallel test formats. A large pool of items can be written for each set and then randomly divided into test forms for use in different rotations, or as pretests, posttests and make-up exams.
- Clearly, there is no capacity to test all components of clinical competence with this method and EMQs are best supplemented with other forms of assessment.

Essays

Essay-type examinations are still popular in the UK and other European countries, despite being excluded for more than 20 years from assessments in North America on grounds of unreliability (Wass et al., 2001).

Strengths

- Essay questions are the ideal mode of evaluating the intellectual skills of reflection and deliberation, revealing better than any other mode of testing how the examinee frames problems, appraises and relies to alternative views, evaluates evidence, and defends conclusions (Elstein, 1993).
- Essay questions are perceived as more valid than MCQs as they require the examinee to generate their own answers rather than select them from a short list of options (Case & Swanson, 1993).
- A variety of components of competence can be explored using essays, including ethical situations and attitudinal issues.

Weaknesses

- Studies by Norcini and Swanson (1989) were not promising. While they found essays could be constructed with public criteria that would probably lead to defensible decisions, they concluded the essay approach to be impractical, as so much time would be required to establish adequate reliability.
- Case and Swanson (1993) argue this type of testing format requires examinees to “guess” what question the author intended and what the graders will reward, suggesting there is indeterminacy regarding the focus and purpose of the question, the detail desired in the response and the level of specificity required. This ambiguity can therefore reduce the reliability and validity of scores.
- Hand scoring is logistically cumbersome, time-consuming and resource intensive and by nature, typically rely on the subjective opinion of the scorer.
- A major weakness of essay questions relates to case specificity. They tend to include fewer items than other pencil and paper tests requiring similar time to complete, resulting in problems with content coverage and therefore score reliability.

Implications

- Essay questions must be efficiently administered, logistically practical, perceived as fair and impartial and cost effective to be warranted.

Modified Essay Questions (MEQ)

The MEQ was developed by Hodgkin and Knox in 1975, initially for the examination of the Royal College of General Practitioners, but have since been used in undergraduate education as a test that is particularly suitable for the assessment of problem-solving skills. MEQs are generally constructed in booklet format where each page carries a separate question, preceded by an item of information. The information is part of a patient problem, and the question asks the examinee to use the data in order to make a decision. For example, examinees may be asked to generate hypotheses or to list what further information they wish to obtain through interview, physical examination or investigation in order to refine their hypothesis. Questions may also ask the examinee to write a referral or the development of a management plan. In all instances, examinees are expected to use their knowledge and understanding in providing a concise answer.

Example of a modified essay question (MEQ)

During a ward round, a week after Helen Lane has been discharged from hospital, you were discussing her case, and the neurosurgeon recounted the story of another girl, of similar age with a head injury and lacerations with blood loss following a road accident. She began passing copious amounts of urine half an hour after admission to casualty. She continued to pass enormous quantities of urine over the ensuing hours and the medical team nearly lost her because of profound shock. The neurosurgeon said that a first year medical student made the diagnosis on physiological grounds and that the diagnosis was confirmed on a skull X-ray. The student also correctly suggested the two crucial in-parallel lines of treatment that proved life saving.

1. Hazard a guess at the diagnosis.
2. What was the mechanism underlying the clinical problem?
3. What was the solution suggested by the student (the two lines of in-parallel treatment)?
4. What was found on the skull X-ray?

(10 minutes)

(taken from Feletti & Engel, 1980)

Strengths

- This format provides an opportunity for examinees to demonstrate competence in applying their knowledge in clinical problem solving when the use of real or simulated patients would be inappropriate.
- The format of MEQs does not contain explicit choices, but rather expects the examinee to apply what they have learnt, therefore any cueing effects are minimised.

Weaknesses

- Despite questions, model answers and examinee answers are generally brief, the design of the paper, the development of model answers and marking of MEQs can be resource intensive and time-consuming.

Implications

- While it is readily acknowledged that the construction and marking of MEQs present considerable intellectual challenges and can be time consuming, Feletti and Engel (1980) reckon that both assessors and students should find this test to be interesting and a valid educational experience.

2.6.2 Tests of Clinical Decision-Making

Patient Management Problems (PMPs)

PMPs, were designed to be an accurate measure of clinical decision-making. Developed by McGuire in the 60's, using pencil and paper, they present a clinical case scenario followed by several sections of items that elicit examinees' responses to steps in the work-up and management of the case. Historically, these methods have tended to be unreliable due to problems with content specificity. However, they achieved widespread usage in high stake examinations in 1970-1980, but are now considered to be outdated due to the following weaknesses.

Weaknesses

- A number of studies investigated performance on PMPs with performance on other "criterion" measures purporting to assess the same skills and behaviours. They found, on average, physicians performed better on the PMPs than in actual practice but those who did best on the PMPs did not consistently do best in actual practice, therefore limiting inferences about test results (Goran, Williamson & Gonnella, 1973).
- The use of PMPs has been found to lead students to pursue significantly more options than on a criterion measure, suggesting the listing of options in PMPs may have a significant cueing effect (Norman & Feightner, 1981). Support has also been demonstrated for the argument that PMPs measure ability to perform and not performance per se.
- Comparison studies of PMPs and performance in the practice setting have found PMPs to be good predictors of what subjects did not do in practice and poor predictors of what they did do.
- Most importantly, concerns were expressed in the early 1980s about the reliability of PMPs, with several investigators reporting that performance on PMPs was highly problem-specific. Correlation of scores across problems typically averaged 0.1, resulting in low test score reliability in the 0.3 to 0.6 range for a three hour examination (Page & Bordage, 1995). Despite the problems encountered with PMPs, interest in testing clinical decision-making increased.

Key Feature Problems

In an attempt to overcome case specificity problems associated with decision-making, a "key features" project was conducted in Canada, based on the rationale that successful resolution of a problem is contingent on effective manipulation of a few key elements of a problem (its "key features"). A new written examination was developed to replace the PMP examination, consisting of many brief problems of mixed question format that required examinees to supply their responses or select their responses from lists, and focused only on the problem's key features. A guide to developing key feature problems is provided in Appendix 1.

Example of a key feature problem (KFP)

For an adult patient complaining of a painful, swollen leg, the physician should:

1. Include deep venous thrombosis in the differential diagnosis;
2. Elicit risk factors for deep venous thrombosis through the patient's history; and
3. Order a venogram as a definitive test for deep venous thrombosis.

Sample key features examination problem:

Paul, a 56-year-old man, consults you in the outpatient clinic because of pain in his left leg which began two days ago and has been getting progressively worse. He states his leg is tender below the knee and swollen around the ankle. He has never had similar problems. His other leg is fine.

Question 1 What diagnosis would you consider at this time? List up to three.

- 1.
- 2.
- 3.

Question 2 With respect to your diagnoses, what elements of his history would you particularly want to elicit? Select up to seven.

- | | |
|----------------------------------|--------------------------------------|
| 1. Activity at onset of symptoms | 16. Palpitations |
| 2. Alcohol intake | 17. Paresthesia |
| 3. Allergies | 18. Paroxysmal nocturnal dyspnea |
| 4. Angina pectoris | 19. Polydipsia |
| 5. Anti-inflammatory therapy | 20. Previous knee problems |
| 6. Cigarette smoking | 21. Previous back problems |
| 7. Colour of stools | 22. Previous neoplasia |
| 8. Cough | 23. Previous urinary tract infection |
| 9. Headache | 24. Recent dental procedure |
| 10. Hematemesis | 25. Recent immobilisation |
| 11. Hormone therapy | 26. Recent sore throat |
| 12. Impotence | 27. Recent surgery |
| 13. Intermittent claudication | 28. Recent work environment |
| 14. Low back pain | 29. Wounds on foot |
| 15. Nocturia | 30. Wounds on hand |

(taken from Page, Bordage & Allen, 1995)

Strengths

- By focusing only on the problem's key features, many brief problems can be included (e.g. 30 to 40), thereby broadening the sampling of the domain and thus increasing the reliability of the examination scores (Page & Bordage, 1995).
- Constructing the examination from a carefully developed blueprint resulted in a representative and adequate sample of problems.
- The conceptual basis for the problems with the issue of critical steps, or key features, is embedded in the current views of medical expertise and its assessment (Page, Bordage & Allen, 1995).
- The problems permit a flexible approach to question format, the number of options, and instructions regarding the number of responses (Page, Bordage & Allen, 1995).
- Sound reliability and validity measures were obtained from this test (Page, Bordage & Allen, 1995). A reliability score of 0.72 was obtained from a 25 item, 3 hr RACGP post-graduate test.

Weaknesses

- Key Feature Problems can be labour-intensive to develop. Special attention needs to be paid to the levels of clinical knowledge and reasoning skills of the examinees and the clarity of the KFPs.
- To date, limited psychometric studies have been conducted on KFPs.

Implications

- Feedback raised by physicians external to the development committee highlighted that KFPs need to be not too vague or too restrictive (Bordage et al., 1995).
- KFPs are now used extensively in both undergraduate and postgraduate medical assessment. They are currently employed for summative assessment purposes by the RACGP.

Objective Structured Clinical Examinations (OSCE)

This method of assessment uses real or simulated patients or specialised technical devices at 'stations' requiring examinees to perform a particular skill or manage a patient. Typically, performance of examinees is scored on pre-coded checklists and/or rating scales by staff examiners or trained patients. Despite a number of psychometric shortcomings and high costs, SP-based tests are receiving increased use, largely due to their hypothesised educational impact. They have been shown to have a dramatic impact on study habits of students, redirecting learning activities to be more relevant to clinical practice (Newble, 1988). OSCEs and all their components have received much interest from researchers and the abundance of literature relating to this method of assessment reflects this.

Strengths

- OSCEs allow observation of performance of complex problems in a realistic setting and as such provide a high degree of fidelity.
- Wass et al., (2001) suggested OSCEs are a potential solution to the difficulties of adequate sampling and standardisation of cases as wide sampling and structured assessment improve reliability.

Weaknesses

- As with all assessment methods, reliability studies have consistently found variation in performance from station to station on an OSCE to be the major source of error, in that examinee performance on one case has been found to be a poor predictor of performance on other cases (case specificity).
- Rater variability and differences between SPs playing the same role also affect reliability although, providing examinees are randomly assigned to raters and SPs, these appear to have little effect on the precision of scores.
- Rater training is resource intensive and may be impractical in settings other than centralised assessment contexts.
- Checklists and/or rating scales can be another source of rater error. A review of a number of studies indicates that interrater agreement is generally better for checklists. However there is an increasing tendency to have an expert observer to view the entire performance of the candidate as overall judgement is enhanced (van der Vleuten & Swanson, 1990).
- This examination format is labour-intensive and expensive.

Implications

- Large numbers of testing stations must be included in OSCEs to obtain stable and reproducible results, with 3 to 4 hours of testing time necessary to obtain even minimally reproducible scores (van der Vleuten, 1996).
- The use of written questions linked to SP presentations has been shown to broaden the meaning of total scores, but reduces generalisability and increases test length requirements. Studies indicate that it is more efficient to focus SP-based tests on assessment of hands-on clinical skills with patients and separately administer written tests for other components of competence.
- The literature implies that practical and educational considerations should be the most important factors in rater selection. If faculty physicians are available to participate, it may be worthwhile incorporating them as raters as they could provide valuable feedback. On the other hand, nonphysicians including SPs as raters may provide a logical, cost-effective and satisfactory alternative.
- A number of studies have investigated the impact of training different types of raters. van der Vleuten et al. (1989) did this using trained and untrained groups of nonphysicians, medical students and physician faculty to rate the videotaped performance of examinees with two standardised patients. They reported that the need for and effectiveness of training varied across groups and that training of raters almost eliminated any differences in accuracy.
- Swanson and Norcini (1989) showed the use of multiple raters per station to have only marginal effects on reproducibility of scores. Therefore, as long as the sample is large enough, one rater per station is adequate. They suggested that if larger numbers of raters are available it is more effective to increase the number of stations and assign one rater to each.
- Sufficient agreement can be obtained from checklists or rating scales if tests involve large enough numbers of stations and are of adequate duration. If both options are viable, the decision to use checklists or rating scales should depend on a number of considerations. For example, checklists are difficult to develop in areas of competence that are less well defined (such as attitudes or aspects of communication skills) without trivialising the aspect of examinee performance.

- Norcini et al. (1993) suggested that the selection of scoring systems and the decision about whether to set standards should not be based on the psychometric characteristics of the test, but should result directly from the purpose of the test and the need for meaningful scores.
- Methods of score interpretation have also been shown to influence the reproducibility of scores. Norm-referenced or domain-referenced frameworks are usually adopted to interpret scores, with the latter method naturally more desirable as there is no reference to the performance of other examinees and scores are interpreted in absolute terms. However, studies on the two methods have found reproducibility of scores to be lower for domain-referenced scores than norm-referenced, which is attributed to differences in station difficulty affecting the former but not the latter. Testing time needs to be increased in domain-referenced frameworks to compensate for the drop in reproducibility (van der Vleuten, van Luyk, Ballegooijenm & Swanson, 1989).

Long Case Exams (e.g. viva)

These type of examinations in medicine have been promoted as well suited for observing a candidate's ability to assemble clinically pertinent information systematically, for testing problem-solving skills, and for observing communication skills and clinical judgement. The traditional format for these oral examinations is to have one or two examiners test the skills of a trainee using a limited number of patients (usually one), in an atmosphere that thrives on spontaneity and tests the examinee to "think on his/her feet" at the bedside (Boudreau, Tamblin & Dufresne, 1994).

Strengths

- Oral exams/long cases provide the ability to observe a wide range of clinical skills and assess them and as such prove reasonably valid.
- Ethical situations and attitudinal issues can be explored using orals.

Weaknesses

- Subjectivity of examiners is the major criticism. Attempts have been made to improve examiner variability through the use of structured cases or protocols, standardised grading criteria and administration procedures, and examiner training.
- Studies of the inconsistency among examiners support the premise that examiners have unique individual perspectives that influence candidate performance. Even after training and evaluation experience, these unique qualities are not removed.
- In common with many British medical schools, the final-year clinical examinations in medicine at the University of Western Australia for many years consisted of a combination of a 'long' case and several 'short' cases. However, the long case examination was abandoned on a number of grounds of: (1) it involved a large organisational burden; (2) it was considered to provide little information in addition to that given by the short case examination; and (3) the large variability in both patients and examiners led to considerable loss of objectivity (Nowotny & Grove, 1982).

Implications

- This format of assessment has limited usefulness due to inherent psychometric problems.
- Reliability problems of content specificity mean testing time must be long.
- In a study by Turnbull et al. (1996) inter-rater reliability was acceptable, though generalisability was

very low, presumably reflecting content specificity. They concluded that by having repeated, shorter clinical cases with a single observer, increases in sampling content could be obtained with the same expenditure of examiner resources.

- Methods for understanding and then controlling for the multiple facets of an oral or other examiner-mediated examination are needed for reliable objective measurement.
- In 1995 modifications were made to the long/short case component of the College of Physicians' clinical examination to consist of two patients in a long case format extending over one hour and thirty-five minutes each, and four short case patients for fifteen minutes each. Further, the long-cases were triple-weighted and four different pairs of judges were employed to assess each candidate across the six cases. Paget (1997) reported that these modifications resulted in significant improvements in the reliability and validity of the test.

2.6.3 Performance in Practice

The measurement of a physician's performance in practice has been simulated by the need to ensure quality of care delivered to patients, and to examine the effectiveness of educational programs. Performance measurement appears to be most useful when it is used as a formative tool as part of a more complex set of quality-improvement activities (Weiss & Wagner, 2000).

Audit

For decades, health-care systems have used clinical audits as a tool for quality assessment. Audits of this type usually seek to characterise cases through the systematic review of a series of patient experiences (Weiss & Wagner, 2000). Most often, the information is obtained by a retrospective audit of the charts or medical records for documentation of specific clinical practices/procedures. This is often conducted using peer review methods with implicit criteria determined by the reviewers or review by nonmedical personnel using explicit, condition-specific criteria (Norman, Neufeld, Walsh, Woodward & McConvey, 1980). In Australia, the audit is used by the Royal College of General Practitioners as a compulsory Quality Assurance and Continuing Education (QA&CE) requirement. It would appear that this method would also prove significant in the ongoing accreditation to prescribe pharmacotherapies for opioid dependence.

Strengths

- A number of studies have been able to demonstrate substantial improvements in performance of both residents and faculty physicians following feedback from clinical audits (Weiss & Wagner, 2000).

Weaknesses

- The reliability of using medical records as a form of assessment has been shown to be a major problem, as information present in the encounter is often not recorded on the chart. From a variety of methods, including the use of videotaped consultations, a number of investigators have found that approximately only half of all information is recorded in the medical record, with as great as 90% omission rates in areas such as "patient education" (Norman et al., 1980).
- Some areas of clinical competence such as "interpersonal skills" are unlikely to be assessable by chart audit (Norman et al., 1980).

- There is conflicting evidence regarding whether or not audits are effective in changing provider behaviour, with several investigators demonstrating little change in targeted prescribing patterns for various clinical conditions as a result of audit and feedback (Weiss & Wagner, 2000).

Implications

- While clinical audit with feedback is an attractive approach to changing physician behaviour, its efficacy is unclear (Weiss & Wagner, 2000).

Peer Review Ratings

Despite the apparent face validity of peer ratings, this evaluation strategy has received little systematic attention due to the perception that these ratings are unreliable and represent little more than personal recommendations from friends. However, confidence in using peer ratings to assess performance has been supported by a number of studies.

Strengths

- Ramsey et al. (1993) investigated clinical performance of physicians using written questionnaires mailed to randomly selected professional associates (physicians and nurses). They concluded this method to be feasible to obtain the number of ratings needed for use in identifying outlying physicians in areas such as clinical skills, humanistic qualities and communication skills.
- Carline, Wenrich and Ramsey (1989) reported that ratings by physician peers and nurses with whom the physician worked could differentiate certified and noncertified internists and even in small towns a sufficient number of ratings could be obtained.
- The Department of General Practice at the University of Nijmegen in The Netherlands carried out a project to train GPs for peer review (Grol, Mookink & Schellevis, 1988). Definite standards and criteria for GP care were formulated before the start of the project. Although labour-intensive, this training was found to have a positive impact on the clinical performance of GPs with reports that they were working more in agreement with the developed criteria, with the greatest effect noticed in those GPs who had conformed least with the criteria beforehand. Following the training, participants reported they had less fear of their practice being criticised and judged.
- Pooled results are often available for individual GPs to compare their practice pattern with other GPs performance.

Weaknesses

- The absence of criterion standards makes it difficult to study the validity of this assessment method. Standardised criteria to measure performance in areas accessible by peer review methods need to be developed (Ramsey et al., 1993).

Observation of Practice (e.g. clinical placement)

Assessing performance via clinical placement is commonly employed. This involves placement of a practitioner at a suitable site that will enable him/her to assume treatment and prescribing responsibilities under supervision of an accredited assessor. Appraisals are usually performed using structured assessment forms. This method is well suited to assess a broad range of skills, as well as aspects of practitioner style (such as attitude) that are often difficult to assess with other methods. Problems with interrater variability exist, affecting the reliability of this method. Hence it is important not to rely solely on this method for assessing practitioner competence.

Videotaped Consultations

The importance of psychosocial issues and building a therapeutic alliance is well recognised by the medical profession. Inadequate interviewing performance has been shown to be associated with patient dissatisfaction, noncompliance with prescribed treatments, doctor shopping, and even malpractice litigation (Mumford, Schlesinger, Cuedon & Scully, 1987). In Australia, the Royal College of General Practitioners utilise videotaped consultation as a method of both formative and summative assessment.

Strengths

- Videotaped recording of consultations in general practice is considered by many to be the ideal method to analyse real life interview skills. The trainee and trainer can watch the consultation together and stop the videotape for discussion as they please.
- As a direct method of assessment, videotapes have high face validity.
- This format may have advantages for both formative and summative assessment where other methods are not practical (e.g. practitioners in rural or remote locations).
- Ram et al. (1999) found video assessment of GPs in daily practice to be a valid and reliable method, and one that is useful for education and quality improvement. They reported there to be a trade-off between feasibility on the one hand, and validity, reliability and credibility on the other hand. They also suggested the cost was acceptable compared to investments in observation methods in standardised settings.
- Ram et al. (1999) found assessment for quality improvement of family physicians' practices by video observation in daily practice to be superior to video assessment in a simulated setting using standardised patients.

Weaknesses

- A number of studies have highlighted that acceptability to both physicians and patients and ethical considerations due to informed consent are areas of concern.
- Given the requirement for assessor training, this format may prove labour-intensive.

Standardised Patients in Practice Settings

An alternative approach to assessing physician performance and quality of care is the use of standardised patients (simulated or actual patients) who are trained to present a clinical problem repeatedly and consistently. Standardised patients have been used in practice undetected and either videotaped or trained to accurately recall the encounter and report or judge the behaviour of the physician based on fixed criteria.

Strengths

- A number of authors have demonstrated the feasibility of introducing standardised patients into physicians' offices, with relatively low rates of detection (Norman et al., 1980; Beullens, Rethans, Goedhuys & Buntinx, 1997).
- The bias of the testing situation is removed, as standardised patients are generally believable. Usually less than 15% are detected, indicating this method is valid (Rethans et al., 1990).

- The patients can assess comprehensively the full range of physician competence and can be trained to recall the encounter accurately so that errors of omission are not an issue (Norman et al., 1980; Tamblyn et al., 1992).
- Cohorts of physicians can be assessed using the same problem (Norman et al., 1980).
- Intra-standardised patient as well as inter-standardised patient reliability have been shown to be 0.85 or more (Beullens et al., 1997, Rethans et al., 1990).

Weaknesses

- A limited range of symptoms and syndromes can be simulated on physical examination. Usually the SP technique is used for the first contact with the patient only, therefore this method is less appropriate (compared to written simulations for example) to assess the medical decision-making process (Beullens et al., 1997).
- Tamblyn et al. (1992) also reported 'first-visit bias' finding physician performance to be underestimated when only the first visit was considered.
- Problems remain with scoring the data from patient encounters and developing performance criteria (Norman et al., 1980).
- This method of assessment is also susceptible to many of the problems common to other assessment methods such as low generalisability across problems (Norman et al., 1980).
- Technically and logistically very difficult to do. This method is time and work demanding, requiring thorough selection and training of SPs. This limits the number of physicians that can be measured and measurement is usually limited to one consultation (Beullens et al., 1997).
- Patients can often be reluctant to make comments about the service that they receive for fear of appearing ungrateful.

Implications

- In Australia, physicians must be informed that this method of assessment is being used. However in the USA this is not the case. This would have implications for the validity of the measure.
- To obtain sufficient reliability and validity, a thorough selection and training of SPs is required, as is careful organisation with an eye for detail (Beullens et al., 1997).
- The SP technique should be supplemented with other methods of assessment. Single-visit SP assessment should be limited to cases that require definitive clinical action during the first visit (Beullens et al., 1997).

Work Logs

This form of practice assessment requires physicians to record all interactions with their patients on a standardised form for a specified period of time. A number of ABMS boards have employed this method as part of their recertification procedures. The American Board of Obstetrics and Gynecology requires applicants to submit a list of all cases treated and procedures performed over the last 6 months, along with information on complications and outcome. Directors of the board review this information and additional measures are taken if practice appears substandard (Langsley, 1991).

2.6.3 Patient-Based Assessment

The adoption of patient-based assessment (e.g. patient feedback) is useful in that it can provide a real measure of the effect of teaching (i.e. the effect on real patients) and provide physicians with useful feedback about their interpersonal skills (Greco, Francis, Buckley, Brownlea & McGovern, 1998). Increasingly, physicians are seeing the value of asking for the patients' perspective on the quality of clinical care that they receive. The literature shows this method of assessment to be sound in terms of reliability and validity provided large enough samples are included.

Strengths

- High levels of acceptability have been reported by a number of authors. Jenkins and Thomas (1996) found the use of a patient satisfaction questionnaire that had been developed by the GPs themselves based on a published prioritised list of what patients wanted from their doctor to be highly acceptable to both patients and GPs.
- Borgiel et al. (1985) included patient questionnaires amongst other methods in a feasibility study to assess the quality of care in family physicians' practices. They found patients could offer a valuable perspective on the quality of care their physician provides and physicians reported that this method was acceptable.
- Kinnersley, Stott, Peters, Harvey and Hackett (1996) were able to demonstrate support for patient satisfaction questionnaires for use in primary care in terms of acceptability, reliability and validity.
- Patient-based assessments are assumed to provide a direct measure of quality of care. Patients experience first-hand the complex processes and systems that work in health care organisations and are likely to know where these work smoothly and where they are repetitive, inefficient and wasteful.
- Patients can feel more of a sense of belonging to or supporting a local health care organisation when they understand the staff's intent to improve care.
- They provide a valid measure of humanistic qualities and interpersonal skills from a consumer perspective.
- Patient information and insight can play an important role in the audit process, where the purpose is to monitor to what degree predetermined standards for any healthcare activity are met, to identify reasons why they are not met and to identify and implement changes to practice to meet those standards.
- Feedback from patients can confirm much of what physicians and managers suspect, and can provide a catalyst for action.

Implications

- Kinnersley et al. (1996) found that levels of satisfaction were lower if patients completed the questionnaires at home rather than in the general practitioners' surgeries, suggesting timing of feedback can impact on the results.
- Patient questionnaires need to be designed to provide valid information. Poorly designed questionnaires may not allow patients to raise important issues or suggest ways to act on problems. Questionnaires should include an open-ended question that allows patients to raise issues that may have been missed by the survey.

2.7 Conclusion

As no single method of assessment is considered optimal to assess all tasks, a comprehensive assessment procedure should combine a number of methods. For nearly all the methods of assessment that have been described, investigations into the psychometric properties of the test suggest problems of case specificity. Irrespective of the skill that is being assessed (e.g. clinical decision-making, problem-solving ability) an examinee's performance on this skill tends to be specific to the case or problem encountered. This problem can be overcome if sampling is broad enough, however this creates logistical problems. As testing time will be limited for practitioners wishing to engage in pharmacotherapy treatment for opioid dependence, the feasibility and reliability of the assessment methods will be deduced. Therefore, it is important that methods are selected that have good reliability (e.g. MCQ or KFP) to balance the low reliability of other methods (e.g. chart audit).

All of the approaches to assessing clinical competence have some strengths and weaknesses. The arguments for and against a particular assessment method need to be weighed against each other, and trade-offs made. Norman et al. (1991) suggest these trade-offs be made as a result of careful consideration of issues resulting from the purpose of the testing situation – practicality, educational impact, acceptability – rather than on the dogma that objective methods, like Orwell's four legged animals, are inherently superior. In selecting appropriate methods of assessment for physicians to engage in pharmacotherapy treatment for opioid dependence, a number of issues must be considered. Such issues include the assessment situation (i.e. authorisation of a first time prescriber or when competence has been brought into question), the operating context (for example, a small remote practice or a large city clinic) and other circumstances (such as the availability of assessors at the time of assessment). Thus, it is important for the assessment procedures to be flexible to allow application in a variety of circumstances and that the assessment methods selected are valid for the assessment situation at hand.



2.8 References

- Beullens, J., Rethans, J.J., Goedhuys, J. & Buntinx, F. (1997). The use of standardized patients in research in general practice. *Family Practice*, 14; 58-62.
- Bordage, G., Brailovsky, C., Carretier, H. & Page, G. (1995). Content validation of key features on a national examination of clinical decision-making skills. *Academic Medicine*, 70(4), 276-281.
- Borgiel, A.E.M., Williams, J.I., Anderson, G.M., Bass, M.J., Dunn, E.V., Lamont, C.T., Spasoff, R.A. & Rice, D.I. (1985). Assessing the Quality of Care in Family Physicians' Practices. *Canadian Family Physician*, 31; 853-862.
- Boudreau, D., Tamblyn, R. & Dufresne, L. (1994). Evaluation of a Consultative Skills in Respiratory Medicine Using a Structured Medical Consultation. *American Journal of Respiratory and Critical Care Medicine*, 150; 1298-1304.
- Carline, J.D., Wenrich, M.D. & Ramsey, P.G. (1989). Characteristics of ratings of physician competence by professional associates. *Evaluation & the Health Profession*, 12; 409-423.
- Case, S.M. (1997). Assessment Truths that We Hold as Self-Evident and Their Implications. *Advances in Medical Education*; 2-6.
- Elstein, A.S. (1993). Beyond multiple choice questions and essays: The need for a new way to assess clinical competence. *Academic Medicine*, 15; 244-9.
- Feletti, G.I. & Engel, C.E. (1980) The modified essay question for testing problem-solving skills. *Medical Journal of Australia*, 1; 79-80.
- Fenderson, B.A., Damjanov, I., Robeson, M.R., Veloski, J.J. & Rubin, E. (1997). The virtues of extended matching and uncued tests as alternatives to multiple choice questions. *Human Pathology*, 28(5); 526-532.
- Frederiksen, N. (1984). The Real Test Bias: Influences of Testing on Teaching and Learning. *American Psychologist*, 39(3); 193-202.
- Goran, M.J., Williamson, J.W. & Gonnella, J.S. (1973). The Validity of Patient Management Problems. *Journal of Medical Education*, 48; 171-177.
- Greco, M., Francis, W., Buckley, J., Brownlea, A. & McGovern, J. (1998). Real-patient evaluation of communication skills teaching for GP registrars. *Family Practice*, 15(1); 51-57.
- Grol, R., Morkink, H. & Schellevis, F. (1988). The effects of peer review in general practice. *Journal of the Royal College of General Practitioners*, 38; 10-13.
- Jenkins, M. & Thomas, A. (1996). The assessment of general practitioner registrars' consultations by a patient satisfaction questionnaire. *Medical Teacher*, 18(4); 347-350.
- Kane, M.T. (1992). The assessment of professional competence. *Evaluation and the Health Professions*, 15; 163-82.
- Kinnersley, P., Stott, N., Peters, T., Harvey, I. & Hackett, P. (1996). A comparison of methods for measuring patient satisfaction with consultations in primary care. *Family Practice*, 13(1); 41-49.
- Langsley, D.G. (1991). Medical Competence and Performance Assessment. *JAMA*, 266(7); 977-980.
- Levine, H.G., McGuire, C.H. & Nattress, L.W. (1970). The Validity of Multiple Choice Achievement Tests as Measures of Competence in Medicine. *American Educational Research Journal*, 7(1); 69-82.
- McGuire, C.H. & Babnott, D. (1967). Simulation technique in the measurement of problem solving skills. *Journal of Educational Measurement*, 4; 1-10.
- Mumford, E., Schlesinger, H., Cuerdon, T. & Scully, J. (1987). Ratings of Videotaped Simulated Patient Interviews and Four Other Methods of Evaluating a Psychiatry Clerkship. *American Journal of Psychiatry*, 144(3); 316-322.
- Neufeld, V.R. (1985). An Introduction to Measurement Properties. In: Neufeld, V.R editor. *Assessing Clinical Competence*. New York: Springer; 39-50.
- Newble, D.I. (1988). Eight years' experience with a structured clinical examination. *Medical Education*, 22; 200-4.
- Newble, D.I., Baxter, A. and Elmslie, G. (1979). A comparison of multiple choice and free response tests in examinations of clinical competence. *Medical Education*, 13; 263-268.

- Newble, D., Dawson, B., Dauphinee, D., Page, G., Macdonald, M., Swanson, D., Mulholland, H., Thomson, A. & van der Vleuten, C. (1994). Guidelines for Assessing Clinical Competence. *Teaching and Learning in Medicine*, 6(3); 213-220.
- Newble, D.I. & Swanson, D.B. (1988). Psychometric characteristics of the objective structured clinical examination. *Medical Education*, 22; 325-334.
- Norcini, J.J., Stillman, P.L., Sutnick, A.I., Regan, M.B., Haley, H.L., Williams, R.G. & Friedman, M. (1993). Scoring and Standard Setting with Standardized Patients. *Evaluation & The Health Professionals*, 16(3); 322-332.
- Norcini, J.J. & Swanson, D.B. (1989). Factors influencing testing time requirements for measurements using written simulations. *Teaching and Learning in Medicine*, 1; 85-91.
- Norman, G.R. & Feightner, J.W. (1981). A comparison of behaviour on simulated patients and patient management problems. *Medical Education*, 15; 26-32.
- Norman, G.R., van der Vleuten, C.P.M. & de Graff, E. (1991). Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Medical Education*, 25; 119-26.
- Norman, G.R., Neufeld, V.R., Walsh, A., Woodward, C.A. & McConvey, G.A. (1980). Measuring Physician's Performance by Using Simulated Patients. *Journal of Medical Education*, 60; 925-934.
- Nowotny, R.E. & Grove, D.I. (1982). Description of an examination for the objective assessment of history-taking ability. *Medical Education*, 16(5); 259-263.
- Page, G. & Bordage, G. (1995). The Medical Council of Canada's Key Features Project: A More Valid Written Examination of Clinical Decision-making Skills. *Academic Medicine*, 70(2); 104-110.
- Page, G., Bordage, G. & Allen, T. (1995). Developing Key-feature Problems and Examinations to Assess Clinical Decision-making Skills. *Academic Medicine*, 70(3); 194-201.
- Page, G.G. & Fielding, D.W. (1980). Performance on PMPs and performance in practice: are they related? *Journal of Medical Education*, 61; 529-7.
- Ram, P., Grol, R., Rethans, J.J., Schouten, B., van der Vleuten, C. & Kester, A. (1999). Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of validity, reliability and feasibility. *Medical Education*, 33(6); 447-454.
- Ram, P., van der Vleuten, C., Rethans, J.J., Grol, R. & Aretz, K. (1999). Assessment of practicing family physicians: comparison of observation in a multiple-station examination using standardized patients with observation of consultations in daily practice. *Academic Medicine*, 74(1); 62-69.
- Ramsey, P.G., Wenrich, M.D., Carline, J.D., Inui, T.S., Larson, E.B. & LoGerfo, J.P. (1993). Use of Peer Ratings to Evaluate Physician Performance. *JAMA*, 269(13); 1655-1660.
- Rethans, J.-J., van Leeuwen, Y., Drop, R., van der Vleuten, C. & Sturmans, F. (1990). Competence and Performance: Two Different Concepts in the Assessment of Quality of Medical Care. *Family Practice*, 7(3); 168-174.
- Sanazaro, P.J. & Worth, R.M. (1985). Measuring Clinical Performance of Individual Internists in Office and Hospital Practice. *Medical Care*, 23; 1097-1114.
- Sloan, D.A., Donnelly, M.B., Schwartz, R.W. & Strodel, W.E. (1995). The Objective Structured Clinical Examination: The New Gold Standard for Evaluating Postgraduate Clinical Performance. *Annals of Surgery*, 222(6); 735-742.
- Spike, N. (editor). (1997) The Royal Australian College of General Practitioners The College Examination: A handbook for candidates and examiners 1998, Examination and Assessment Department (RACGP), Melbourne.
- Swanson, D.B. & Norcini, J.J. (1989). Factors Influencing Reproducibility of Tests Using Standardized Patients. *Teaching and Learning in Medicine*, 1(3); 158-166.
- Tamblyn, R.M., Abrahamowicz, M., Berkson, L., Dauphinee, W.D., Gayton, D.C., Grad, R.M., Isacc, L.M., Marrache, M., McLeod, P.J. & Snell, L.S. (1992). First-visit bias in the measurement of clinical competence with standardized patients. *Academic Medicine*, 67, S22-S24.
- Turnbull, J. Danoff, D. & Norman, G. (1996). Content specificity and oral certification examinations. *Medical Education*, 30; 56-59.

- van der Vleuten, C.P.M. (1996). The Assessment of Professional Competence: Developments, Research and Practical Implications. *Advances in Health Sciences Education*, 1; 41-67.
- van der Vleuten, C. & Swanson, D. (1990). Assessment of clinical skills with standardized patients: state of the are. *Teaching and Learning in Medicine*, 2; 58-76.
- van der Vleuten, C., van Luyk, S., Ballegooijenm A. & Swanson, D. (1989). Training and experience of medical examiners. *Medical Education*, 23; 290-6.
- Wass, V., van der Vleuten, C., Shatzer, J. & Jones, R. (2001). Assessment of clinical competence. *Lancet*, 357, 945-49.
- Weiss, K.B. & Wagner, R. (2000). Performance measurement through audit, feedback, and profiling as tools for improving clinical care. *Chest*, 118(2 Suppl); 53S-58S.
- Wilson, D.R. (1975). Assessment of clinical skills: from the subjective to the objective. *Annals: The Royal College of Physicians and Surgeons of Canada*, 8; 109-118.



3 Learning objectives

The Learning Objectives consist of six competencies that have been identified as central to safe and effective management of opioid dependence using pharmacotherapies. Each competency is briefly described followed by relevant tasks. Effective practice is inferred through performance criteria for each task, relating to knowledge, skills and attitudes. They identify the factors that contribute to capability in this domain and inform the content of learning programs and the development and implementation of assessment procedures. They relate to, and should be considered in the context of, standards of effective clinical practice identified by relevant professional bodies (e.g. RACGP; RACP).

Competency 1. Attitudes and Professionalism

Description:

Patient Focus – The practitioner conveys an accepting, non-judgmental attitude that allows him/her to develop effective therapeutic relationships. Practitioners accept that working with drug dependent people requires respect for the individual's autonomy, while working to develop and maintain a thoroughly professional therapeutic relationship. This will often require the practitioner to skillfully set limits and negotiate conditions of ongoing management to minimise risks to the drug user and those around him/her. The practitioner is aware of and sensitive to issues of ethnicity, culture, gender, age and sexuality. In addition, the practitioner recognises the importance of the patient's family and significant others in supporting the patient, as well as the potential difficulties the family and significant others may experience in caring for a drug-using person.

Professional Role – The practitioner behaves with courtesy, responsibility and accountability towards patients, their families and significant others, and towards other health professionals. A competent practitioner understands the extent and limitations of their competence. He/she recognises and respects the contributions and roles of other medical practitioners in the process of care and consults and/or refers with due regard to their level of competence and experience.

Interdisciplinary Management – The practitioner recognises the necessity of interdisciplinary team management, and respects the key role of other professionals – including nurses, pharmacists, and counsellors – in the care of opioid-dependent people. He/she understands the specific skills of each professional involved in patient care, and develops a close professional relationship with these workers. The practitioner appreciates the synergistic effect of interdisciplinary management and supports all professionals involved in joint care of patients.

Patient Advocacy – Practitioners involved in the care of opioid-dependent people support their patients in overcoming stigma and in receiving appropriate medical care and psychosocial support. They also recognise the importance of ensuring their patient's independence and dignity.

Task	Performance Criteria
1.1 Engage patients in a respectful and non-judgmental manner	1.1.1 Details sought from patient are directly relevant to assessment and treatment
	1.1.2 Define and work within professional boundaries
	1.1.3 Patient autonomy is fostered
	1.1.4 Listens, understands and communicates understanding
1.2 Identify importance of and limits to patient confidentiality	1.2.1 Generally patient information divulged only with patient's consent
	1.2.2 Patient details and records are only accessible to people who need to know information
	1.2.3 Patient details and records are stored securely
1.3 Communicate with family members and significant others to ensure optimal care	1.3.1 Obtain from, and provide appropriate information to, family and significant others to ensure optimal care
1.4 Communicate with other health professionals to ensure optimal care and support	1.4.1 Communicate treatment plans and procedures to others involved in treatment
	1.4.2 Clarify roles of others involved with patient care
	1.4.3 Maintain contact with other health professionals
1.5 Refer patient appropriately according to the need for comprehensive care	1.5.1 Appropriate referral and or consultation procedures implemented

Competency 2. Assessment

Description:

The cornerstone of all treatment of dependency is assessment. This is the process of clarifying why the patient is motivated to present for treatment, the nature of their problems, their particular supports and difficulties, and includes the development of a treatment plan. It is based on a comprehensive approach to a patient's medical, psychological and social circumstances.

Initial assessment is often best undertaken over several interviews. Throughout treatment, there will be occasions for reassessment, particularly when problems are appearing or a change of treatment is being contemplated. Assessment is a critical ingredient in establishing a therapeutic alliance. Comprehensive documentation is a critical component of assessment.

Task	Performance Criteria
2.1 Conduct a comprehensive assessment	2.1.1 Take a detailed drug history
	2.1.2 Identify the presenting problem and motivations for treatment
	2.1.3 Take a medical and psychiatric history
	2.1.4 Take a psychosocial history
	2.1.5 Obtain corroborative evidence of dependence and drug use history where required
	2.1.6 Perform a focused mental state examination
	2.1.7 Perform a focused physical examination
	2.1.8 Undertake appropriate investigations
	2.1.9 Provide feedback to patients
	2.1.10 Discuss treatment options with patient
	2.1.11 Document assessment and provide appropriate written reports and letters
2.2 Identify problems and diagnose	2.2.1 Estimate degree of dependence
	2.2.2 Estimate the current level of tolerance and withdrawal
	2.2.3 Identify drug related harm and risk behaviours
	2.2.4 Diagnose medical and psychiatric problems
	2.2.5 Identify significant psychosocial factors likely to influence management

Competency 3. Developing a treatment plan

Description:

Competence in medical management of opioid dependence requires the practitioner to plan, initiate, administer and review appropriate and comprehensive long-term management.

The practitioner demonstrates knowledge of those approaches to treatment that have been demonstrated to improve outcomes.

Task	Performance Criteria
3.1 Establish suitability for the range of treatment options	3.1.1 Identify indications, contraindications and precautions for the range of treatment options
	3.1.2 Establish suitability of the individual patient for the range of pharmacotherapies
	3.1.3 Provide accurate, balanced information (both verbal and written) about treatment options
3.2. Plan ongoing management in conjunction with the patient	3.2.1 Identify and document agreed treatment goals
3.3 Obtain informed consent	3.3.1 Rationale, duration and problems of pharmacological treatment explained
	3.3.2 Pharmacology of drug treatment explained including side effects and drug interactions explained
	3.3.3 Consent documented
3.4 Facilitate safe induction to treatment	3.4.1 Identify risk factors during induction
	3.4.2 Ensure compliance with local jurisdictional requirements
	3.4.3 Apply jurisdictional recommendations for induction

Competency 4. Management of co-existing conditions

Description:

The practitioner regularly monitors the patient's medical, psychiatric and social condition. The practitioner identifies both risks and protective factors to assist the individual to develop resilience in preventing the uptake of hazardous behaviour(s). The practitioner responds appropriately to co-existing medical, psychiatric and social conditions.

Task	Performance Criteria
4.1 Identify and initiate management of medical, psychiatric and social problems	4.1.1 Medical and psychiatric and social conditions managed
	4.1.2 Identify and manage complications of other medical problems including complications of IDU e.g. BBVs
	4.1.3 Identify and manage concomitant pain and opioid dependence
	4.1.4 Diagnose and respond appropriately to intoxication and withdrawal from alcohol, benzodiazepines, opioid and other psychoactive drugs
	4.1.5 Provide accurate information and appropriately manage drug use in pregnancy
	4.1.6 Identify and refer or manage psychiatric factors affecting rehabilitation management including adjustment disorders, depression, anxiety and psychosis
	4.1.7 Identify and refer or manage social problems affecting rehabilitation management including employment issues, living arrangements, family and other relationship issues.
4.2 Integrates drug and alcohol rehabilitation into the wider framework of the patient's medical care	4.2.1 Liaise, including consultation, with other medical practitioners and health services
	4.2.2 Identify and refer or manage circumstances affecting rehabilitation

Competency 5. Patient Management

Description:

The practitioner reviews patients regularly, seeks to maintain an effective therapeutic relationship, and in discussion with patients, updates a treatment plan in light of evidence regarding factors associated with better treatment outcomes. The practitioner manages (or advises on management) problems (including intoxication, dosing errors, non-compliance and disruptive behaviour), so as to minimise the harm to patients and to others.

Task	Performance Criteria
5.1 Communicate the pharmacology of each pharmacotherapy	5.1.1 Rationale of methadone, buprenorphine, naltrexone and LAAM and duration of treatment explained
	5.1.2 Pharmacology of each medication explained
	5.1.3 Side effects and drug interactions of each pharmacotherapy explained
5.2 Implement safe induction to pharmacotherapy treatment (according to jurisdictional guidelines)	5.2.1 Safe initial dose prescribed
	5.2.2 Incremental changes prescribed in accordance with patient needs and safety
	5.2.3 Regular frequent review of patient progress in initial stages
	5.2.4 Adjust maintenance doses to adequate level
	5.2.5 Treatment plan negotiated and documented
	5.2.6 Side effects and other related adverse events managed effectively
5.3 Manage transfer to other pharmacotherapies	5.3.1 Transfer plan implemented safely and effectively
5.4 Manage withdrawal from pharmacotherapy treatment	5.4.1 Establish withdrawal plan
	5.4.2 Rate of reduction specified
	5.4.3 Supportive care provided

(Continued)

Task	Performance Criteria
5.5 Regular review, monitoring and documentation of patient's progress	5.5.1 Regular review of patient's progress documented, including the use of standardised outcome measures
	5.5.2 Revision of treatment plans and goals, as necessary
	5.5.3 Monitoring the adequacy of the patient medication
	5.5.4 Determining and reviewing the appropriate setting for treatment
	5.5.5 Appropriate use of urine drug monitoring
5.6 Deliver appropriately structured treatment	5.6.1 The practitioner manages or advises on management problems including intoxication, dosing errors, non-compliance and disruptive behaviour so as to minimise the harm to patients and to others
	5.6.2 Deliver treatment in ways consistent with research findings
	5.6.3 Act appropriately as required on legal and professional responsibilities including Mental Health Act and child protection concerns
	5.6.4 Identify and detail appropriate review timeframe with other members of the treatment team
	5.6.5 Appropriate use of take-home medication
	5.6.6 Frequency of appointments specified

Competency 6. Quality Assurance

Description:

The principles of quality clinical practice in treating opioid dependence are not unique. Maintaining involvement in professional development has relevance for quality assurance in all aspects of clinical practice. To this end, safe and effective treatment of opioid dependence will be enhanced if the practitioner engages in the quality assurance standards determined by the relevant professional bodies (e.g. RACGP, RACP).

Task	Performance Criteria
6.1 Engage in the quality assurance program of the relevant college	6.1.1 Evidence of engaging in the quality assurance program of the relevant college



4 Blueprint and decision matrix

The development of tests of clinical competence requires careful and thorough planning. As part of the process of defining what is to be tested, a number of steps are suggested. Following the identification of clinical problems and the description of related tasks, it is usually recommended that a Blueprint should be developed to guide the selection of problems to be included in the assessment procedure. A Blueprint provides a valid basis from which to “sample” for assessment of knowledge, skills and attitudes. Based on the Learning Objectives, the following assessment Blueprint has been developed for both formative and summative assessment of medical practitioners who respond to opioid dependence using pharmacotherapies.

In addition, a decision matrix is provided to indicate the various situations in which different assessment procedures may be used. The Blueprint and matrix are intended to be used as a reference to guide and describe the choice of assessment method(s). They are not intended to prescribe or restrict use of particular methods and they are not comprehensive.

To facilitate interpretation of the Blueprint, the reader is referred back to the literature review. In addition, a brief summary of each assessment method is provided as a key.

Key to Assessment Methods

MCQ Multiple choice questions are written questions followed by several (usually 4 or 5) alternative response options from which examinees are required to select the correct response(s).

EMQ Extended matching questions are similar to MCQ but have more than five options for response. Usually extended matching items are grouped into sets, with a single option list used for all items in a set.

KFP Key feature problems are usually brief problems of mixed question format. Examinees are required to supply their responses or select their responses from lists, and focus only on the problem’s key features.

MEQ Modified essay questions are generally constructed in booklet format where each page carries a separate question, preceded by an item of information on a patient problem. The questions may ask the examinee to use the data in order to make a decision, or to list what further information they wish to obtain through investigation, or to write a referral or the development of a management plan.

OSCE Objective structured clinical examinations consist of 'stations' which require examinees to perform a particular skill or manage a patient. Typically, performance is scored on pre-coded checklists and/or rating scales by staff examiners or trained patients.

Observation of practice (e.g. clinical placement/case history) This should involve placement of a practitioner at a site where the practitioner can assume prescribing and treatment responsibilities, under supervision, of patients in various stages of treatment. In remote practice locations this may involve "on-site" supervision. In circumstances where this is not possible, assessment of a case study may be employed.

Work logs provide a brief overview of a practitioner's management of several patients, including reason for presentation, action taken and outcome. A record is made on a standardised form at the end of the consultation process for consecutive patients over a given time period and the log is then sent to the assessor.

Long case (e.g. viva) This usually involves an examiner orally presenting a clinical case to the practitioner. The practitioner is required to make clinical decisions and is assessed on a wide range of clinical skills using standardised criteria. Several shorter cases (short case) may be presented as an alternative.

Audit An audit is an examination of prescriber practice often undertaken by an accredited assessor. Usually a retrospective audit of the medical record is done using a standardised rating form. Audits may also be carried out as a self-assessment tool.

Patient based assessment (e.g. patient feedback) This method employs direct assessment, by the patient, of a practitioner's quality of care, usually via patient questionnaires.



Competency 1. Attitudes and Professionalism

Assessment Methods

Task	Assessment method		
	Knowledge	Skills	Values/Attitudes
Engage patients in a respectful and non-judgmental manner	None specified	Patient Based Assessment Observation of practice (e.g. clinical placement)/case history Role Plays	Observation of practice (e.g. clinical placement/case history)
Identify importance of and limits to patient confidentiality	KFP MEQ	Audit	Observation of practice (e.g. clinical placement/case history)
Communicate with family members and significant others to ensure optimal care	KFP MEQ	Observation of practice (e.g. clinical placement/case history) Audit	Observation of practice (e.g. clinical placement/case history)
Communicate with other health professionals to ensure optimal care and support	KFP MEQ	Audit	Observation of practice (e.g. clinical placement/case history)
Refer patient appropriately according to the need for comprehensive care	KFP MEQ EMQ MCQ	Work Logs Audit	None specified

Competency 2. Assessment

Assessment Methods

Task	Assessment method		
	Knowledge	Skills	Values/Attitudes
Conduct a comprehensive assessment	Long case (e.g. viva) OSCE KFP MEQ	Observation of practice (e.g. clinical placement/case history) Patient Based Assessment Audit OSCE	Observation of practice (e.g. clinical placement/case history)
Identify problems and diagnose	MCQ EMQ KFP MEQ OSCE	Long case (e.g. viva) Observation of practice (e.g. clinical placement/case history) OSCE	Observation of practice (e.g. clinical placement/case history)

Competency 3. Developing a Treatment Plan

Assessment Methods

Task	Assessment method		
	Knowledge	Skills	Values/Attitudes
Establish suitability for the range of treatment options	MCQ EMQ KFP Long case (e.g. viva) OSCE	Observation of practice (e.g. clinical placement)/Case history Role playing OSCE	Observation of practice (e.g. clinical placement)/Case history Audit Patient Based Assessment
Plan ongoing management in conjunction with the patient	KFP Long case (e.g. viva) OSCE	Long case (e.g. viva) Observation of practice (e.g. clinical placement/case history) OSCE	Audit
Obtain informed consent	KFP OSCE	Observation of practice (e.g. clinical placement/case history) Audit OSCE	Observation of practice (e.g. clinical placement/case history)
Facilitate safe induction to treatment	MCQ EMQ KFP OSCE	Observation of practice (e.g. clinical placement/case history) Audit OSCE	Observation of practice (e.g. clinical placement/case history) Patient Based Assessment

Competency 4. Management of Co-existing Conditions

Assessment Methods

Task	Assessment method		
	Knowledge	Skills	Values/Attitudes
Identify and initiate management of medical, psychiatric and social problems	MCQ EMQ KFP Long case (e.g. viva) OSCE	Long case (e.g. viva) Audit OSCE	Observation of practice (e.g. clinical placement/case history) Patient Based Assessment
Integrates drug and alcohol rehabilitation into the wider framework of the patient's medical care	KFP	Audit Patient Based Assessment	None specified

Competency 5. Patient Management

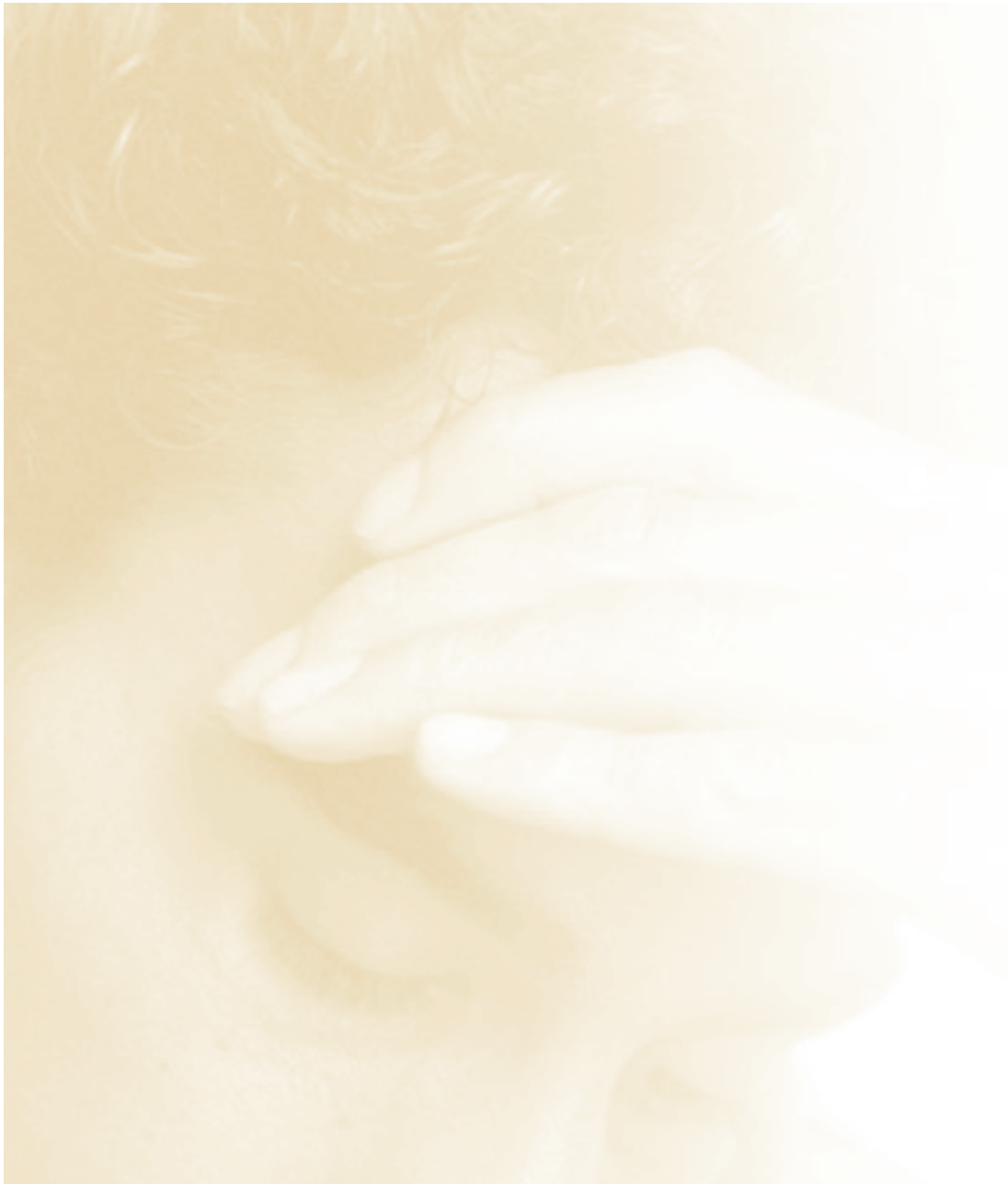
Assessment Methods

Task	Assessment method		
	Knowledge	Skills	Values/Attitudes
Communicate the pharmacology of the range of pharmacotherapies	MCQ EMQ KFP MEQ	Observation of practice (e.g. clinical placement/case history) Role Plays OSCE	Patient Based Assessment Observation of practice (e.g. clinical placement/case history)
Implement safe induction to pharmacotherapy treatment (according to jurisdictional guidelines)	MCQ EMQ KFP MEQ	Audit Observation of practice (e.g. clinical placement/case history) OSCE	Observation of practice (e.g. clinical placement/case history)
Manage transfer to other pharmacotherapies	MCQ EMQ KFP MEQ	Audit OSCE	Patient Based Assessment
Manage withdrawal from pharmacotherapy treatment	MCQ EMQ KFP MEQ	Audit Observation of practice (e.g. clinical placement/case history) OSCE	Observation of practice (e.g. clinical placement/case history) Patient Based Assessment
Regular review, monitoring and documentation of patient's progress	MEQ	Audit	None specified
Develop appropriately structured treatment	MCQ EMQ KFP MEQ	Audit Observation of practice (e.g. clinical placement/case history)	None specified

Competency 6. Quality Assurance

Assessment Methods

Task	Assessment method		
	Knowledge	Skills	Values/Attitudes
Engage in the quality assurance program of the relevant college (e.g. RACGP, RACP)	N/A	N/A	N/A



Decision Matrix for Summative and Formative Assessment Formats in Various Assessment Situations

As already indicated, assessment procedures may be employed in numerous situations, for a variety of purposes. The following format is provided as a descriptive and indicative guide to the use of a range of assessment formats for summative assessment in different circumstances. This suggested matrix is not comprehensive. It does not include all formats of assessment or all potential situations where they may be used.

Situations	Written Test (e.g. MCQ, KFP, MEQ)	Observation of Practice (e.g. clinical placement/ case history)	Work Logs	Audit	Patient Based Assessment	Other
(i) Aim to become authorised prescribers for the first time	●	●				
(ii) Have been involved in prescribing but not had competence assessed	●	●		●	●	
(iii) Have gone through the process of authorisation some time ago and wish to maintain status and therefore require re-assessment	●		●	●	●	
(iv) Have their competence brought into question subsequent to authorisation		●		●	●	Re-examination or as determined by authorising body
(v) Require “fast track” authorisation for “emergency” situations (e.g. when another prescriber suddenly ceases to provide this service)	●	●		●		Prescriber review committee makes recommendation, authorises on restricting/ contexting service
(vi) Wish to change the conditions of practice (e.g. increase number of patients under their management past level determined by jurisdiction/ increase authorisation for other pharmacotherapies)		●	●	●	●	
(vii) Wish to transport accreditation from one jurisdiction to another	●					All could be deemed to have some relevance, but jurisdiction may focus on assessment of knowledge of jurisdictional requirements



Appendix

Writing key feature problems

© Associate Professor Liz Farmer
National Co-ordinator Key Feature Problems
Royal Australian College of General Practitioners

What are Key Feature Problems?

Key Feature Problems test clinical decision making and problem solving skills. They are different from Multiple Choice Questions (MCQ or EMQ) which primarily test knowledge. A Key Feature Problem focuses on the critical steps in handling any clinical scenario.

Critical Steps

Typically, critical steps are the essential steps in decision making and/or resolution of a problem. These steps may occur in any part of a clinical scenario, for example:

- In diagnosis (certain features of the history, physical examination or investigation may be essential).
- In management (critical steps in decision making may include the patient's past history, allergies, current drug therapy and intercurrent illnesses).
- In prevention (critical steps may include risk factors relating to age, gender etc).

Critical steps may also include:

- Steps most likely to lead to errors
- The difficult aspects of identifying and managing problems in general practice.

Critical steps may not occur in every part of a clinical scenario.

Key Feature Problems involve case presentations that consist of a clinical scenario followed by two or more questions that focus **only** on the critical steps. The scenario may vary in length. A diagnostic scenario may be brief, whereas with critical steps focussing on interpreting test results, the scenario would be longer and contain information regarding history and physical examination.

Writing critical steps can be divided into two easy parts.

Part 1:

First choose a *clinical area* that a general practitioner (or subject at the appropriate level being tested) may encounter and be expected to handle competently. Some examples may include:

- diagnosing non-insulin dependent diabetes mellitus
- management of high cholesterol
- immunisation protocols and side-effects
- interpretation of test results
- managing acute trauma
- travel medicine advice
- approaching an undifferentiated problem e.g. lethargy
- assessing the sick child
- diagnosing mental illness
- emergency procedures

The next step is to write a *clinical scenario* that illustrates the problem as you might expect to encounter it in general practice (or other appropriate setting). Use the grid (see page 150) to assist with this stage.

Remember that the scenario should provide a challenge that requires **decision making** not simply knowledge. It should be as realistic a presentation as possible.

The clinical scenario does not have to begin with questions to be obtained in the history. It may start for example at the point where a patient returns with an abnormal test result, with a complication of a chronic condition or a well patient seeking preventative advice.

The common elements of any case scenario include:

- patient's name age and gender
- the setting of the consultation
- reason for attendance
- whatever clinical details are required prior to the first question or set of questions.

An example:

Clinical Area: Back pain

Clinical scenario:

Mrs K, a 74 year old widow, presents to your consulting rooms complaining of sudden onset of severe mid-thoracic back pain which began after a fall in the garden. Her general health has been good apart from a left mastectomy for carcinoma 5 years previously. She has had intermittent backache for several years but has not required any regular medication for this.

Examination reveals her to be in severe pain and you notice a thoracic kyphosis. Her back is tender to palpation in the mid-thoracic region with associated muscle spasm.

Part 2:

Now it is necessary to identify the *critical steps* in handling this scenario, wherever they occur. There may be any number of critical steps in each problem. They should all be recorded in a logical order. To continue the above example:

Critical steps: The GP should

1. Consider 3 diagnostic possibilities – osteoporotic crush fracture, malignant infiltration and facet joint pain related to the fall.
2. Admit the patient to hospital for pain relief and further investigation
3. Order a plain X-ray of the spine and bone scan.

Validate and Reference

The next step is to *validate and reference* the question. This may be done by discussion with colleagues and using the literature. The case is now ready to be formatted into a Key Feature Problem.

Formatting Questions

Once the critical steps have been decided (as above), the cases can be formatted into test questions. More than one set of questions can often be written using the one clinical scenario. The steps are as follows:

- Create the questions
- Select the format for the answer
- Create the scoring key.

Key Feature Problems are flexible, but there are two main answer formats used:

The *Completion* item. This is typically a list of possible responses, usually **at least double, preferably treble** the number of available options to correct options. That is, if five responses are to be selected, the list should contain at least 10-15 options in total. A large number of options reduces cueing.

When writing a completion item the writer must select the number of options required i.e.

- select only one
- select up to x number
- select x number
- select as many as are appropriate

Completion items can be directly data entered or scanned and are therefore logistically easy to mark.

The *Write in* (WI) response. This requires the candidate to write in the answer(s). This format requires hand marking using the scoring key and is therefore more time consuming. However, it reduces cueing to nil and has been shown to be highly discriminating in identifying the range of candidates' abilities.

Thus each question begins with a clinical scenario based on the 'critical steps', followed by one or more questions based on the critical steps, with or without additional clinical information.

In pencil and paper tests, any subsequent questions must not cue the candidate to review the answer(s) to the original question. In practice, this can be difficult to achieve and must be carefully monitored.

Examples of some typical questions are:

Completion

1. With respect to your possible diagnoses, what elements of the history would you particularly want to elicit? Select up to *five* (5) elements from the list.
2. Which first line investigations should be included? Select as many from the list as are appropriate.
3. What are the most likely interpretations of this test? Select the three (3) most likely interpretations.

Write- in responses are best used for diagnoses and/or treatment questions, as these are most prone to cueing. They are also easier to hand mark, as there are fewer ways of writing the same response.

Write in format:

1. What is your probability diagnosis? Write, in note form only, your single (1) diagnosis
2. What diagnoses are most likely? List up to three (3) most likely diagnoses.
3. What critical steps would you include in your immediate assessment? List, in note form only, as many steps as are required.



Scoring Keys

The final step is to develop the individual *Scoring Key*.

Each Key Feature Problem can have any total value. All problems are converted to percentages so that all questions contribute equally to the total mark.

Examples of each scoring key are in the practice papers.

Responses may attract unequal marks i.e. 3 marks may be given for a more important response, and 1 mark for a less important response. This rewards candidates who are able to focus on the critical steps.

Penalty marking and negative marking are possible when using this test format i.e.

- If a candidate selects more options than requested, i.e. 7 options when 5 were requested, then no marks may be awarded or marks can be deducted.
- Essential items can be marked as 'must be present' otherwise score 0 for the question.
- No marks may be awarded for the entire problem when a dangerous or life-threatening selection or answer has been given despite the value of other answers.

GRID FOR WRITING CRITICAL STEPS

<p><i>Clinical Area:</i></p> <p>Age group:</p> <p>Patient's name:</p> <p>Patient's age:</p> <p>Patient's gender:</p> <p>Setting for the consultation:</p> <p>Reason for attendance / presenting complaint:</p> <p>Clinical details provided prior to first question / set of questions:</p> <p>Critical steps (as many as required):</p> <ol style="list-style-type: none">1.2.3.4.5. <p>© Associate Professor Liz Farmer, National Co-ordinator Key Features Problems RACGP</p>
--

