Chapter 2 Assessment of knowledge

Dr Steve Capey
University of Swansea

Introduction

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

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

Classification of written questions

Written questions may be categorised on the basis of whether the answer is *constructed* or *selected*. Constructed answer questions call for an answer to

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

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

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

Essay questions

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

Marking is an isolating experience for the assessor; the hours set aside to mark essay questions are a substantial drain on faculty resources. Deriving precise and relevant criteria for the marking schedule is also challenging, and unlike short answer questions or selected answer formats, standardisation of

Box 2.1 Example of an unstructured essay question

Flawed question
'Describe the pharmacology of local anaesthetics.'

Why the question is flawed
The question does not have a precise focus so could be interpreted in a number of ways. Candidates could be drawn to more than one viewpoint, with the trigger question leading to a diverse set of answers on a wide range of topics, example pharmacodynamics, pharmacokinetics, recommendations, indications, contraindications, routes of administration and other nuances related to the use of local anaesthetics in practice. A less discerning student could simply vacillate and write everything they knew about local anaesthetics. Therefore, if the assessor is to make a shrewd judgement, the question should guide the candidates on what is expected within the answer.

To phrase the question another way:
'Describe the pharmacodynamic and pharmacokinetic properties of drugs used for local anaesthesia.'

With this alternative construct, the candidate is steered to cover the essential features of pharmacodynamics and pharmacokinetics at the depth appropriate for their level of academic progression and/or scope of practice.

the assessors is complex, so marking can be subjective. Therefore, it is good practice for the criteria to be agreed in advance, by a group of assessors, who can standardise a comprehensive marking matrix that covers the expected answer content together with guidance on expectations around the award of marks within different grade boundaries (Jolly, 2010; see also Chapter 9 for a discussion on examiner behaviours, and the different assessment practices used to quality assure consistency in marking).

Modified essay questions (MEQ)

The MEQ was developed in the late 1960s to address the issue of 'case specificity' by increasing the knowledge sampling frequency above that of traditional essay questions (Jolly, 2010). An MEQ requires highly structured short paragraphs to answer a series of questions posed by a brief clinical scenario on a precise clinical area. Being shorter than an essay, this test format affords greater sampling of knowledge across the curriculum and, as

model answers are easier to articulate, the commensurate marks awarded can be more clearly defined. Test papers that include a structured marking scheme help candidates to prioritise their time.

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

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

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

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

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

Planning the mark scheme

1. To include the legal and ethical issues related to end of life care and should be constructed in alignment with the teaching on the programme and the current regulatory and legal guidance.
2. To include the different management options for pain relief in Stage IV cancer therapeutics and holistic therapies.

Short answer questions (SAQs)

These are concise questions that require an answer in response to triggers and associated data, such as pathology reports and radiological images. Essentially, they permit the sampling of specific knowledge where the response might range from a single answer construct, a short paragraph of text, data interpretation or an explanatory or annotated diagram. The answer cueing effects that can occur with MCQ-type questions are not generally such a problem because the response is constructed (self-generated). The same diligence as with MEQs is needed to ensure all questions are unequivocal, because a disadvantage of SAQs is that they are open to interpretation by the candidates (and examiners) and, accordingly, post-assessment appeals.

Multiple-choice questions (MCQ)

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

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

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

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

True–false questions

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

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

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

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

Single best answer questions (SBAs)

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

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

Box 2.3 Single best answer

Theme: Dermatology

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

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

Answer options

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

Answer = B

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

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

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

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

Extended matching questions (EMQ)

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

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

EMQ-type assessment papers may be difficult to construct, particularly for some of the more reflective disciplines, and require staff training. A large number of EMQs are required to sample a suitable breadth of knowledge, and the more item statements or scenarios included as part of an individual EMQ serve to increase the difficulty in construction. Nonetheless like SBAs they exhibit reliability when written effectively, and can be marked objectively and efficiently with optical marking equipment. Additional benefits of EMQs are that they:

- Allow the testing of a variety of clinical scenarios on a linked theme and so test knowledge in greater depth than a single MCQ;
- Avoid some of the cueing effects of MCQs because a larger matrix of answers is used;
- The quality of questions can be determined simply by the 'cover up' test, that is candidates should be able to select their response 'without looking' at the answer options – this also applies to SBAs.

A well-constructed EMQ set should include four components:

1. A theme,
2. A list of option/responses in a matrix,
3. A lead-in statement,
4. Item stems or clinical scenarios (see also Box 2.4).

Common flaws in question items using multiple-choice formats

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

Box 2.4 Extended matching questions

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

Theme: Mechanisms of anti-microbial drugs.

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

List of options: Each option may be used once, more than once, or not at all.

A. Amphotericin B
B. Clarithromycin
C. Clofazimine
D. Doxycycline
E. Isoniazid
F. Mefoloquine
G. Penicillin V
H. Selegiline
I. Sulfadiazine
J. Trovafloxacin

Items or stems

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

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

- The order in which the possible answers are written could inadvertently give away the answer; this is easily avoided by listing the distracters for

Box 2.5 Five steps to writing multiple-choice questions Adapted from Case and Swanson (2002)

Step 1: Identify the topic for the MCQ: The topic could be one of a number of areas within the curriculum, but it is good practice to group questions together that are around a similar theme or specialty.

- Patient treatment; for example drug treatments for hypertension;
- A feature of clinical treatment; for example management, diagnosis, investigations;
- Non-clinical studies, for example ethical issues.

Step 2: Write the clinical vignette:

The scenario or vignette should be concise, providing only essential information needed to answer the question, with loquacious and irrelevant information avoided.

***Step 3:* Prepare the list of answers:**

List all possible responses as:

- Either a few words or short sentences;
- In alphabetical order.

Step 4: Review the question and list of answers:

Ensure that the following are reviewed as quality assurance regarding:

- The relevance of the featured question in the whole test item;
- There is only one 'most appropriate' answer for the question. All other answers should be 'possible' responses and relevant to the vignette, otherwise any overtly inappropriate responses would simply reduce the number of potential distracters and confer an answer-cueing effect. For example, if a candidate is asked to select a drug therapy, then all possible responses should be drugs.

***Step 5: Peer review*:**

The final part of assessment preparation is the review of assessment items:

- By an experienced colleague with a credible knowledge base to critique the questions for content accuracy, technical construction quality and to check for any ambiguities;
- Any essential information required to answer the question is provided.

Additional considerations for writing EMQs

Lead in:

An essential component when writing the lead-in question is that it needs to:

- Indicate the relationship between the scenarios/vignettes and the options in the response matrix;
- Clarify the question posed for candidates.

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

Box 2.6 Example of a bad MCQ

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

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

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

each question in alphabetical order, which provides uniformity of presentation that does not point to the answer;

- Avoiding the use of negative questioning;
- Grammatical errors are common, arising when the test writer has to provide 5 possible responses, with less attention given to the distracters than the correct response (see Box 2.6). A clear pointer might be using any one of the following:
  - Singular or plural terms;
  - A word or phrase that is given in both the stem and the correct response;
  - The correct response has more details than the other distracters;
  - When asked to identify the correct response, ALL are correct (see Box 2.7 for examples).

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

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

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

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

Confidence assessment of multiple-response items

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

knowledge but severely punishes poor students who are unaware of their ignorance, thus eliminating students likely to be dangerous in practice.

Specific tests of clinical reasoning

Key feature problems (KFP)

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

Box 2.8 Example of key feature problems

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

Metformin, 500 bd
Bendrofluaziade 2.5 mg one a day
Aspirin 75 mg
Simvisatatin 40 mg

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

Na 118 mmol/L	(Normal range 136–146 mmol/L)
K 3.2 mmol/L	(Normal range 3.4–4.4 mmol/L)
Urea 10 mmol/L	(Normal range 7.9–16.4 mmol/L)
Cr 265 µmol/L	(Normal range 60–110 µmol/L)
eGFR 45 ml/min/1.73 m^2	(Normal range 100–130 ml/min/1.73 m^2)

Questions:

1. From the information you have what is the likely cause of David's confused state? *List 2 only* (2 marks)
2. What further tests would assist you in your diagnosis? *List 2 only* (2 marks)
3. What are the major contra-indications for bendrofluazide? *List 3 only* (3 marks)
4. What are the main beneficial actions of metformin? *List 2 only* (2 marks)
5. What three (3) immediate interventions would you take to improve David's confused state? *List 3 only* (3 marks)

Model answer and scoring plan:

Qu1. Dehydration (1 mark)
Hyponatremia (1 mark)
Hypoglycaemia (1 mark)
Hypokalemia (0.5 mark)

Qu.2. Serum osmolality (1 mark)
Urine osmolality (0.5 marks)
Osmolality (0.5 marks)
Urine dipstick, BN stick (0.5 mark)
Blood glucose (1 mark)
Urine sodium (0.5 mark)
Urine FENA (0.5 mark) forced excretion of sodium
Full Blood Count (0 marks)

Qu3. Refractory hypokalaemia (1 mark)
Hyponatraemia (1 mark)
Hypercalcaemia (1 mark)
Hyperuricaemia (1 mark)
Addison's disease (1 mark)

Qu4. Inhibition of hepatic glucogeonesis (1 mark)
Increase of peripheral glucose utilisation (1 mark)
Inhibits intestinal glucose absorption (1 mark)
Improves insulin production (0.5 mark)

Qu5. Stop bendroflurazide (1 mark)
Stop metformin (1 mark)
Rehdration therapy, Fluid resuscitation, intravenous infusion, IVI (1 mark)
Dextrose (1 mark)

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

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

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

Script concordance items (SCI)

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

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

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

Box 2.9 Eight steps to preparing key feature problems (adapted from Page, Bordage and Allen, 1995)

***Step 1:* Select a clinical problem.** It is good practice to define the following:

- Age and gender of the patient;
- The setting of the clinical presentation;
- The appropriate clinical data, for example pathology reports, radiological images.

***Step 2:* 'What are the critical steps in the resolution of this problem?'**

- Identify the key steps that would be involved in the clinical decision making to manage or diagnose this clinical presentation.

***Step 3:* Think of different ways that patients present with this problem.**

- The presenting complaint and/or reason for the clinical encounter, for example signs and symptoms and their duration.

***Step 4:* List the essential key features involved in the care of this group of patients.**

- List as many of the essential features that are required for the resolution of the specific clinical presentation.

***Step 5*: Select a typical case presentation and write the clinical vignette for this particular presentation.**

- The vignette should be as detailed as possible without containing unnecessary information that is irrelevant to solving the problem.

***Step 6*: Write the questions and construct scoring keys which test only the key features of the case presentation.**

- The question format – short answers or a choice from a prepared list of responses;
- With short answers, the question format should be direct, for example 'What is your provisional diagnosis?'
- Additional instruction should include the number of allowed responses,
  - list up to five . . . ;
  - select up to three . . . ;
  - when asked to select one response, it suggests a definitive answer is required.

***Step 7*: Scoring criteria.**

- This is essential information for precise assessment which needs to be unambiguous and provides for any marginal difference in formulating the answers;

- Appropriate marks awarded for each element or answer to the questions;
- The relative weighting of responses should be an indication of the significance and potential consequences of the answer given in patient management and decision making;

***Step 8*: The final part of assessment preparation is the review of assessment items.**

- By an experienced colleague with a credible knowledge base to critique the questions for content accuracy, technical construction quality and to check for any ambiguities;
- Check that any essential information required to answer the question is provided.

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

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

You were thinking of: Angina pectoris.

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

What effect would this finding have on your diagnosis?

+2 Almost ruled out
+1 Less probable
0 This finding has no effect on the diagnosis
–1 More probable
–2 Almost certain

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

Projects and dissertations

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

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

Guidance for successful delivery of projects and dissertations

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

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

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

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

Table 2.1 Examples of word length for projects, proposals and dissertations

Assessment task	Length of the assignment
Written assignment, e.g. essay, report	1500–6000 words depending on the ratio in the weighting of different elements of assessment in a named module; e.g. 40% written assessment and 60% examination (2 hours).
Undergraduate research dissertation (e.g. 15 ECTS)	10000 words
Masters research dissertation (e.g. 30 ECTS)	30000 words
Written research proposal	
Science, Engineering; Medicine; Dentistry	Around 500 words
Humanities and Social Sciences	Between 500 and 1500 words.
Postgraduate thesis	
Doctor of Philosophy (PhD)	100000 words
Master of Philosophy (MPhil)	60000 words
Doctor of Medicine (Research) (MD)	50000 words
Doctor of Surgery (Research) (MS)	50000 words

Assessment of projects and dissertations

The important issues that exercise students and academics are:

- The level of guidance and supervision; this can be significant for undergraduates;
- The diversity of subjects covering narrow areas of content in-depth;
- The identification of assessors with the appropriate knowledge to judge them equitably.

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

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

Portfolios

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

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

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

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

There are expectations for portfolios to include documents on reflective practice that contribute to academic progression, together with evidence of mentorship and an action plan for achieving the learning outomes. Like other log book-type assignments, their immediate usefulness is not always acknowledged, and they can be seen as time consuming for students to prepare and the faculty to assess. Non-compliance, or at least procrastination over portfolio maintenance, is common, with a flurry of activity to amass proof of practice around the submission deadline. One concern that surrounds the portfolio-based assessment is the question of equivalence of marking the content, and to avoid any conflicts of interest the mentor for portfolio maintenance should not be the assessor for summative purposes.

It is worth mentioning that support for writing in a reflective style is essential, particularly for students from a predominantly 'hard science' education, to make sure that they have the skills to cope with the attitudinal and professionalism domains assessed via the portfolio. A more accurate longitudinal view of performance is afforded when portfolios are prepared systematically. An electronic format has the added dimension of flexibility of access for students and assessors and so ought to encourage concordance.

References

Case, S.M., and Swanson, D.B. (1993). Extended-matching items: a practical alternative to free-response questions. *Teaching and Learning in Medicine*, 5(2) 107–115.

Case, S.M., and Swanson, D.B. Constructing Written Test Questions for the Basic and Clinical Sciences 3rd Edition (revised) NBME 2002 http://www.nbme.org/pdf/itemwriting_2003/2003iwgwhole.pdf.

Case, S.M., Swanson, D.B., Ripkey, D.R. (1994). Comparison of items in five-options and extended matching format for assessment of diagnostic skills. *Academic Medicine*, 69(10 Suppl):S1–3.

Charlin, B., Brailovsky, C., Leduc, C. and Blouin, D. (1998). The diagnosis script questionnaire: a new tool to assess a specific dimension of clinical competence. *Advances Health Science Education Theory and Practice*, 3: 51–58.

Feletti, G. (1980). Reliability and validity studies on modified essay questions. *Journal of Medical Education*, 55: 933–941.

Fournier, J.P., Demeester, A. and Charlin, B. (2008). Script concordance tests: guidelines for construction. *BMC Medical Informatics and Decision Making*, doi: 10.1186/1472-6947-8-18.

Gagnon, R., Charlin, B., Coletti, M., *et al.* (2005). Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? *Medical Education*, 39: 284–291.

Gardner-Medwin, A.R. (2006) Confidence-based marking: Towards deeper learning and better exams. In *Innovative Assessment in Higher Education*, C. Bryan and K Clegg (eds), pp. 141–149. Routledge-Taylor & Francis.

Haladyna, T.M. and Downing, S.M. (2003). How many options is enough for a multiple-choice test item. *Psychological Measurement*, 53: 999–1009.

Hammond, E.J., McIndoe, A.K., Sonsome, A.J. and Spargfo, P.M. (1998) Multiple-choice examinations: adopting an evidence based approach to examination techniques. *Anaesthesia*, 53:1105–1108.

Jolly, B. (2010). Written examinations. In *Understanding Medical Education*, T. Swanwick, pp. 208–321. Wiley-Blackwell.

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical Teacher*, 26: 709–712.

Muijtjens, A.M.M., van Mameren, H., Hoogenboom, R.J.I. *et al.* (1999). The effect of a 'don't know' option on test scores: number-right and formula scoring compared. *Medical Education*, 33: 267–275.

Page, G. and 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: 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.

Palmer, E.J. and Devitt, P.G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. *BMC Medical Education*, 7: 49.

Palmer, E.J., Duggan, P., Devitt, P. and Russell, R. (2010). The modified essay question: Its exit from the exit examination? *Medical Teacher*, 32: e300–e307.

Van Tartwijk, J. and Driessen, E.W. (2009). Portfolios for assessment and learning: AMEE Guide no. 45. *Medical Teacher*, 31: 790–801.

Further reading

Bland, A.C., Kreiter, C.D. and Gordon, J.A. (2005). The psychometric properties of five scoring methods applied to the script concordance test. *Academic Medicine*, 80: 395–399.

Fischer, M., Kopp, V., Holzer, M., *et al.* (2005). A modified electronic key feature examination for undergraduate medical students: validation threats and opportunities. *Medical Teacher*, 27: 450–455.

Swanson, D.B., Holtzman, K.Z., Allbee, K. and Clauser, B.E. (2006). Psychometric characteristics and response times for content-parallel extended-matching and one-best-answer items in relation to number of options. *Academic Medicine*, 81 (10 Suppl): S52–55.

Tigelaar, D., Dolmans, D., Wolfhagen, I. and van der Vleuten, C. (2004). Using a conceptual framework and the opinions of portfolio experts to develop a teaching portfolio prototype. *Studies in Educational Evaluation*, 30: 305–321.

Wass, V. and Archer, J. (2010). Assessing learners. In *Medical Education: Theory and Practice*, T. Dornan, K. V. Mann, A. J. Scherpbie and J. A. Spencer, pp. 230–255. Churchill Livingstone.