THE AUSTRALIAN AND NEW ZEALAND COLLEGE OF VETERINARY SCIENTISTS

Measuring what we want to measure.

A guide to examination construction for examiners.

2020

Liz Norman BVSc(Hons) MVM EdD MANZCVS
Massey University, New Zealand

Contents

Blueprinting

Examinations cover only a sample of the curriculum. In other words, they only cover some of what could be covered. It is important that we ensure that what is covered is sufficient and representative of the whole curriculum. Doing so enables us to generalise from performance in this examination and make inferences about the ability of the candidate in the entire subject.

A blueprint is a plan or template that specifies what the examination covers. The blueprint shows how sufficient and representative the examination is of the whole domain of practice. It therefore provides an essential documentation of the validity of generalising performance in the examination to determine whether a candidate can be recognised as a Member or Fellow.

Constructing blueprints

Blueprints are usually constructed as a table or grid that categorises the content of the examination in at least two dimensions. They can be more or less complex and the dimensions shown can vary to suit the purpose and subject matter (Table 1). Constructing more than one blueprint may be easier than trying to put all the dimensions onto one grid.

At their simplest, content is simply indicated by an X in the relevant location of the grid (Table 2). Other blueprints give some detail in each cell that provides information about other dimensions of the examination (Tables 3-5).

Blueprinting begins with consideration of the curriculum. What knowledge, skills, and attitudes make up the domain of practice and are examinable? These should be specified in the learning outcomes for the subject[1]. Blueprinting involves tabulating these as one of the dimensions to blueprint against. Then use other dimensions (Table 1) to characterise the examination. Examples of blueprints are provided in Tables 2-5.

Table 1: Examples of examination dimensions that could be used to construct blueprints

- Learning outcomes
- Diseases or presenting problem
- Body system
- Species
- Knowledge domain: anatomy, pathophysiology, diagnosis, management, prognosis, epidemiology, investigations, ethics and law
- Skills: interpretation, reasoning, decision making, evaluation, problem solving
- Cognitive level: factual recall, analysis and interpretation of data, problem solving
- Degree of abstraction: principles (theoretical), applied (concrete)
- Location of question in the examination components

Examples of blueprints

Table 2: Example of a blueprint using dimensions of learning outcomes and examination component. Note this is a cut-down version for illustration. More rows and columns might be needed.

Learning outcome	Written paper 1	Written paper 2	Oral examination
LO1.1 aetiology, pathogenesis, pathophysiology	X		X
LO1.2 diagnosis, treatment, management			X
LO1.3 diagnostic tests & procedures	X	X	X
LO1.4 preventive medicine		X	

Table 3: Example of a blueprint using dimensions of body system and knowledge domain, that also shows which examination component the question appears in. Note this is a cut-down version for illustration. More rows and columns might be needed.

	Pathophysiology	Investigation and diagnosis	Treatment and management
Gastrointestinal	P1Q1	P1Q1, P2Q4	OQ3
Cardiovascular	P1Q4	P2Q2, OQ1	P2Q2
Nervous		P1Q3, P2Q1	
Endocrine	P1Q3	OQ2	P2Q3
Respiratory			P2Q5

Table 4: Example of a blueprint using 3 dimensions of body system, knowledge domain, and cognitive level that also shows which examination component the question appears in. Note this is a cut-down version for illustration. More rows and columns might be needed.

	Pathophysiology			Investigation and diagnosis			Treatment and management		
	recall	**analyse, interpret**	**problem solve**	**recall**	**analyse, interpret**	**problem solve**	**recall**	**analyse, interpret**	**problem solve**
Gastrointestinal	P1Q1				P1Q1 P2Q4				OQ3
Cardiovascular		P1Q4			P2Q2	OQ1		P2Q2	
Nervous				P1Q2 P2Q1					
Endocrine	P1Q3				OQ2			P2Q3	
Respiratory							P2Q5		

Table 5: Example of a blueprint for a single examination component using dimensions of body system and knowledge domain that also briefly characterises the subject of the question. Note this is a cut-down version for illustration. More rows and columns might be needed.

Written paper 2	Pathophysiology	Investigation and diagnosis	Treatment and management
Gastrointestinal		Investigation of exocrine pancreatic insufficiency	
Cardiovascular		Investigation of atrial fibrillation	Management of cardiomyopathy
Nervous		Diagnosis of L2-3 spinal neoplasia	
Endocrine			Management of diabetic ketoacidosis
Respiratory		Diagnosis of feline asthma	Management of feline asthma

Weighting questions

Some parts of the curriculum are more important than others and need to be given more emphasis in the examination. One way to decide an appropriate relative weighting of parts of the curriculum is to assign scores based on the frequency of occurrence and importance. A scoring system in use for the blueprinting of medical curricula at the University of Calgary (McLaughlin, Lemaire, & Coderre, 2005) is shown in Table 6.

Table 6: Example of scores assigned to curriculum areas in order to determine their relative weighting. Scores for impact and frequency are multiplied to give a total weighting for each curriculum area. From McLaughlin et al. (2005).

Impact	Weight	Frequency	Weight
Less important	1	Rarely seen	1
Essential	2	Relatively common	2
High impact	3	Very common	3

Evaluating the blueprint

Blueprints can be used to plan the content of questions, before questions are drafted. They can also be used to check the content of an already-designed examination to highlight areas of overlap and gaps in the content. They help plan the examination component (Paper 1, Paper 2, Oral examination, Practical examination) within which a question is best placed. The subject guidelines specify the type

of content that will appear in each examination and should guide the blueprint. Remember that it is the examination as a whole that needs to be representative, rather than each component.

Ensuring that the examination content is representative of the subject does not mean that every cell of the blueprint's matrix needs to be filled or that coverage is even across the curriculum. There are two reasons for this. Firstly, it is not always possible to cover all combinations of dimensions, as there is a limited number of questions that can be asked in the available examination time. Blueprints can be useful for tracking the content assessed over years, to ensure that various parts of the curriculum are not consistently missed, therefore making them effectively not part of the curriculum. Secondly, as discussed above, some parts of the curriculum are more important than others and need to be given more emphasis in the examination.

Drafting questions

Examination questions are actually tasks. The candidate performs the task and we evaluate their performance which may be, for example, a written answer, or an oral response. We want the task to be a relevant as possible to the domain of practice we are assessing. Ideally, the task would mirror practice, but in reality, we are always simulating some aspect or another of it. The closer the skills utilised to do the exam task are to the skills needed to practice in the domain, the more valid it is for us to extrapolate from performance on the examination tasks to performance in the domain of practice for that Membership or Fellowship subject. Although we cannot assess all the necessary skills,[2] a case for extrapolation can be made if the skills assessed are relevant and important ones in practice. Tasks that are relevant and important for the domain we are assessing are referred to as *construct relevant*. Tasks that are irrelevant or unimportant for the domain we are assessing are called *construct irrelevant*.

Construct irrelevance is a major threat to the validity of our pass-fail decisions. If the examination score is dependent on skills that are not relevant for practice in that Membership or Fellowship subject, then our decision to award Membership or Fellowship is flawed and questionable.

A major source of construct irrelevance arises when the task required is not clear. This will be discussed in a following section on clarity. Other sources relate to the task complexity and cognitive skills required to answer questions, and to whether there is enough time for candidates to complete the task. These sources will be discussed here.

Creating appropriate tasks

Appropriate tasks require candidates to use skills that are relevant for the practice of the subject. This means that the emphasis shouldn't just be on recall of information, but on assessing the candidate's understanding; skills in analysing and interpreting information; skills in problem solving; and their judgement. Candidates utilise base knowledge in order to perform higher-order skills such as analysing complex problems and making sound judgements. Therefore assessing higher-order skills also assesses base knowledge in a way that requires it to be understood, not just reproduced.

How do you tell what sort of skills a task requires? Various taxonomies may be of use. A classic one you may have heard of is Bloom's taxonomy (Table 7) and four others that I find helpful are presented in Tables 8-10.

Table 7: Revised Bloom's Taxonomy (Krathwohl, 2002)

Remember	Retrieving relevant knowledge from long-term memory.	Recognizing, Recalling
Understand	Determining the meaning of instructional messages, including oral, written, and graphic communication.	Interpreting, Exemplifying, Classifying, Summarizing, Inferring, Comparing, Explaining
Apply	Carrying out or using a procedure in a given situation.	Executing, Implementing
Analyze	Breaking material into its constituent parts and detecting how the parts relate to one another and to an overall structure or purpose.	Differentiating, Organizing, Attributing
Evaluate	Making judgments based on criteria and standards.	Checking, Critiquing
Create	Putting elements together to form a novel, coherent whole or make an original product.	Generating, Planning, Producing

Table 8: Recall-higher order task classification (Peitzman, Nieman, & Gracely, 1990)

Fact recall	Questions capable of being answered by reference to one paragraph in a text or notes (or several paragraphs for questions requiring recall of several facts)
Applied (higher order)	Questions that require the use of facts or concepts, the solution of a diagnostic or physiologic problem, the perception of a relationship, or other process beyond recalling discrete fact

Table 9: Knowledge – understanding, adapted from Wiggins and McTighe (2005).

Knowledge	knowing about; a body of coherent facts; can be thought of as right or wrong.
Understanding	knowing how and why; the meaning of facts; the theory that links facts and provides meaning; how sense is made of facts to enable them to be applied to analysis, synthesis, evaluation; to be able to explain why particular facts or skills are applicable to a particular situation; to know which fact to apply when; to be able to create new knowledge or modify or adapt an idea to a new situation.

Table 10: Structure of the observed learning outcome (SOLO) taxonomy, adapted from Biggs and Tang (2011).

Knowledge structure	Response characteristics	Instructional verbs	Signs to look for in answers (not all signs need to be present)
Prestructural	Lack of knowledge, or common knowledge only		• Not answered • Incorrect • Off topic or irrelevant • Just rephrases the question
Unistructural	Answer addresses a single aspect	Paraphrase, define, identify, count, name, recite, follow simple instructions, calculate, reproduce, arrange, recognise, find, note, seek, sketch, pick	• Addresses one aspect or concept or process or carries out one procedure • Relevant and correct.
Multistructural	Answer addresses multiple aspects. Connections simple or lacking	Combine, classify, structure, describe, enumerate, list, do algorithm, apply method, account for execute, formulate, solve, conduct, prove, complete, illustrate, express, characterise	Includes multiple pieces of correct/relevant material but also shows some/all of: • No particular order to aspects presented • Inclusion of irrelevant material • Inconsistencies or incorrect aspects • Superficial or oversimplified • Replication of material from sources – rote learned or reproduced without significant transformation • Includes correct material related to the topic but does not answer the question
Relational	Answer addresses multiple aspects. Makes connections between aspects	Analyse, compare, contrast, integrate, relate, explain causes, apply theory (to its own domain), argue, implement, plan, summarize, construct, design, interpret (some senses), structure, conclude, substantiate, exemplify, derive, adapt	Includes multiple pieces of correct/relevant material but also shows some/all of: • Aspects explained relative to one another • Logically organised answer • Analysis and or synthesis • Compares similarities and differences • Integrates multiple levels (eg: molecular, biochemical, systemic) • Evaluates inconsistencies • Expresses reasons • Explains implications or reaches a conclusion • Expresses relative importance, value, significance of aspects. • Selective answer that addresses the point of the question and may be shorter than a multistructural answer • Uses the language of the discipline - terminology and phrasing • Relates answer to examples or experience • Relates answer to organising principles of the discipline
Extended abstract	Goes beyond a relational answer but has more originality, creativity, metaconnections, and utilisation of overarching abstract principles	Theorise, generalise, hypothesise, predict, judge, transfer theory (to new domain), assess, evaluate, interpret (some senses), critically reflect, predict, criticise, reason	• Hypothesises • Reaches original conclusions • Deduces principles • Incorporates knowledge from other domains • Creates new knowledge • Appreciates different possibilities and the role of context

Table 11: Webb's depth of knowledge (Webb, 2007)

recall	Recall information or perform a simple step, identify, measure, describe, explain simple ideas
skill/concept	Requires more than one step, comparing, interpreting, estimating, making observations, explaining, organising and displaying data
strategic thinking	Requires planning and using evidence, explaining reasoning, make conjectures, draw conclusions, solving problems
extended thinking	Complex reasoning, planning developing and thinking over an extended period of time, making multiple connections, synthesis of ideas into new concepts

Specifying the task and its scope

Your exam question needs to specify the task you want performed and indicate to the candidate the type of response required and the scope of the task.

The task to be performed and the type of response are often specified using one instructional verb (Table 12). Examples include name, explain, discuss. Be sure to include all aspects of what is required. For example, you may wish candidates to explain how and also justify why.

Sometimes the task instructions are separated from the instructions about how to respond. For example, you may require candidates to evaluate some data and then discuss options based on their evaluation. In this kind of situation, think carefully about whether you really also want them to explain or justify their evaluation before they discuss options, and if so, say so.

Try to be clear about the scope. Do you want a general answer about this condition that applies to all possibilities or one that specifically applies to this case? Do you want an answer to include all possible options, or those that are currently available in this region of the world, or those that are available under cost constraints?

Writing a marking scheme while you are drafting a question helps you see what type of response you are looking for, so that you can adjust your instructions to candidates accordingly. It is very important not to leave candidates guessing. If they are not sure, candidates may try to guess what you want by the marks available, and may be wrong. Another strategy used by candidates if they are not sure what you want is to write a "just in case" answer, where they include aspects in their response just in case you want them. This type of answer can seem off-topic and poorly focussed and may cause candidates to lose marks. In addition, they will waste time on this instead of concentrating on other questions in the paper. Therefore, it is important that you specify the exact nature of the task and scope as clearly as possible.

Table 12: Examples of instructional verbs

Compare: to find similarities between things, or to look for characteristics and features that resemble each other.
Contrast: to find differences or to distinguish between things.
Discuss: to present a detailed argument or account of the subject matter, including all the main points, essential details, and pros and cons of the problem, to show your complete understanding of the subject.
Define: to provide a concise explanation of the meaning of a word or phrase; or to describe the essential qualities of something.
Describe: to portray, in words, the characteristics, significant qualities, features, or details of something.
Explain: to clarify, interpret, give reasons for differences of opinions or results, or analyse causes.
Illustrate: to use a picture, diagram or example to clarify a point.

Controlling task difficulty and demand

The tasks you set need to be the appropriate level of demand but without being unnecessarily difficult. You do not need to purposely make questions tricky in order for them to be difficult for candidates. Remember that what seems easy to you reflects years of experience. Candidates for examination are at a much earlier stage of their expertise than you are. Your task as an examiner is really only to determine if they meet the passing standard (or not). You do not need to determine how good they are above that.

Many factors affect how demanding and difficult an examination task is (Table 13). It is important to tailor the demand to the requirements of the level (Membership/Fellowship) and the subject domain. You should also look out for factors that will increase the difficulty of questions in ways that are irrelevant to what you are assessing. An example would be using a case presentation that is of Fellowship complexity in a Membership examination. Another example is requiring a life-like drawing.

You should also consider things that will affect the ease of a question that are not relevant to the subject. An important example of this is advance knowledge of the question. If candidates have advance knowledge of questions then the demands of the tasks become those of recall, even if the original task required higher order skills. Because of this it is best to design novel tasks for each examination.

Table 13: Some aspects of exam questions that affect their demand and difficulty. Derived from Hughes, Pollitt, and Ahmed (1998) and Crisp and Novakovic (2009).

Type of operation: simple steps are less demanding than synthesis/evaluations/interpretation
Degree of novelty: familiar concepts are less difficult than novel or uncommon ones
Number of components or ideas involved: the greater the number of ideas or relationships that have to be considered, the more difficult. It is even more difficult if the ideas are ones that are not usually considered together.
Whether resources are provided or need to be generated by the candidate: tasks in which the candidate must generate their own resources are more difficult than those where the resource is provided. An example of a resource in this context might be a radiographic interpretation.
The question wording and any images/diagrams/tables provided: the way a question is worded and information in it is provided and the amount of information provided can make parts of the question more or less obvious or distract candidates, changing the difficulty of the task. Images are particularly distracting and should be included only if necessary for the task.
Degree of abstraction: In general, abstract concepts (principles, generalisations, and things that can't be seen) are more difficult than concrete ones. Concepts are more abstract if not within the candidate's experience. Generalising to a principle when given a specific example is much more demanding than giving an example when presented with a principle. Abstract concepts for which there is no shared definition are particularly difficult (for example concepts such as "professionalism").
Response strategy – simple, stepwise, integrated: The number of steps involved in responding affects the difficulty, as well as whether these can be performed stepwise or must be integrated. For example, candidates may need to interpret information before formulating a response and then writing that down. If they need to respond while they are formulating a response (say in an oral exam where they feel under pressure to begin speaking) the same task is more difficult.
Guidance provided: Questions can be worded to provide more or less guidance to the candidate about how to approach the task or how to formulate a response. For example, questions may divide sections into parts (i), (ii), (iii) which provide a response structure that simplifies the task.

Pacing examinations

It is important to allow enough thinking time in examinations. Since we want to assess higher order learning outcomes such as the candidate's ability to make judgements and solve problems we need to give them time to do this. In addition, time stress increases the difficulty of task in irrelevant ways. It decreases working memory processing capacity, increases the use of schemas, and decreases the ability to focus on relevant information and suppress irrelevant information. These aspects are discussed further in the section on clarity.

It is preferable for the examination to be worth total marks that are in multiples of 60, so that the time that should be spent on each question is easily calculated. This helps examiners during examination construction and also helps candidates, when answering questions.

Reading time

Reading time needs also to be accounted for and should be timed at 40 words per minute. This allows time for candidates to read carefully for understanding of complex material (Klatt & Klatt, 2011).

Remember also to allow time for interpreting clinical information such as images, laboratory results etc, if this is included in your questions.

Writing time

Candidates are capable of writing much less than we might expect during examinations and therefore care needs to be taken to pace the exam such that the answer can be given in the available time. Researchers have found writing speeds in examinations to be surprisingly low (Summers & Catarro, 2003), and these figures have been confirmed in an analysis of candidate answers conducted by the Board of Examiners.

For a rough guideline use 20 words per minute. Table 14 gives a guide for different length questions.

Table 14: Number of words that candidates can write in examination questions of different durations

Duration of question	Number of words in answer	No of pages in answer
5 minute	100	0.3
10 minute	200	0.6
20 minute	400	1.3
30 minute	600	2
1 hour	1200	4

Grading criteria

In constructed-answer examinations (those in which the candidate must construct an answer rather than just selecting one from a list), a lack of agreement between different examiners about the marks to award for an answer affects reliability. Poor reliability leads to uncertainty about the accuracy of the total mark, giving wide confidence intervals and uncertainty about the correct classification (pass or fail) of a borderline candidate. It thus affects the validity of our decision to pass or fail a candidate. Two ways that inter-examiner reliability and agreement can be increased, are examiner training, and the use of marking schemes. The College also monitors inter-examiner agreement (which is why it is important you record your independent decision before conferring), and re-evaluates cases where there is significant disagreement.

Marking schemes identify the criteria that will be assessed and how marks will be awarded for different standards of achievement against each of those criteria. They should be written during question construction, because they help you to refine the question. They force you to think through what you are expecting candidates to do. They provide a focus for you and other examiners to discuss your expectations of candidates while there is still time to reconstruct the examination, and for you to come to a mutual agreement about how candidate responses will be marked. If examiners cannot come to agreement about the marking scheme, then the question is unsuitable for use in an examination and should be discarded. It is much better to know this ahead of time.

During question construction, marking schemes help to check the question content to ensure that it is something important in the curriculum and meets the specifications of the blueprint. They help to check that the response required is at the right level of knowledge, skills, and judgement for Membership or Fellowship. They help to check question wording to ensure it clearly specifies the task that you actually want and that the question is answerable. They also help you to check that there is enough time for candidates to respond in a way that will attract full marks. Finally, marking schemes allow a moderator to check the same things.

During question marking, marking schemes help you to decide on the marks to award and to justify your marks in a defensible manner. They also reduce bias, by prespecifying criteria that all candidates must meet. They will help you provide feedback for failing candidates about areas they need to work on.

Types of marking scheme

Marking schemes identify the important aspects of a candidate's response and specify how marks will be awarded for different types of response. Types of marking scheme include:

- Model (ideal) answers
- Points-based schemes
- Criteria- and level-based schemes
- Schemes with incorporated principle(s) for discriminating levels

Model answers

Model answers give an example of an ideal answer. They have a number of uses but need to be accompanied by other information to form a marking scheme. By themselves, they give no indication of how marks will be awarded for alternative correct and partially correct answers.

Model answers are, however, worthwhile writing to accompany a marking scheme. They give you a chance to have a go at the question yourself and think about what it asks of candidates. They also allow you to check the length and structure of the type of answer you think is ideal, so you can make sure your instructions about the response format are clear. Lastly, they allow you to count the words in your ideal response and check that the candidate would have time to write that answer in the time they have available. See the section on pacing examinations for more details on how to do this.

Examiners sometimes use model answers as a summary to update their own and other examiners knowledge on the area. Such summaries often discuss all that is known about an area and therefore may be more than could be expected from a candidate under examination conditions. So, while producing a summary for yourself and other examiners is a good idea, it would be best for this to be an accompanying document rather than being included in a marking scheme.

Points-based marking schemes

Points-based marking schemes are those that assign a mark (or a few marks) for identifiable and relevant points made in the answer. These seem to be extremely common and are easy to use, however they have a number of problems which make them unsuitable for use in College examinations.

The problems centre around the fact that points-based marking schemes are based on an assumption that saying more (correct) things equates to higher level performance, irrespective of what else is said or how the response is formed. They are focused on correctness rather than quality. When thought about in terms of any of the learning taxonomies shown in Tables 8-10, they reward performance at lower levels and do not reward understanding, connected knowledge, or the candidate's ability to apply knowledge. Such marking schemes can reward unfocussed answers in which facts are presented in no specific order or relative importance, and in which there is irrelevant, inconsistent, contradictory, or incorrect material as well as correct material. They may also penalise expertise, because experts often give more limited but pertinent responses which focus on the most important aspects. Points-based marking schemes encourage candidates to say all they know about a topic, relevant or not, in case they score a point.

Where possible you should use a marking scheme that considers the quality of the answer and the evidence of higher order thinking. The only situation in which points-based marking schemes are acceptable is in situations where the question does not allow a higher-order response. For example, if the candidate is asked to provide an unordered list or a one-word answer.

Criteria- and standard-based schemes

Here criteria refers to different dimensions of performance and standard refers to different levels of quality of the performance on a given criterion. Descriptors help to define the quality standards. These might be generic (in that they can apply to many different types of task) or specific for the question. Criteria- and standards-based schemes can be analytic or holistic.

An analytical marking scheme divides the available marks between different criteria, with weightings according to their relative importance. The sum of the marks awarded for each criteria gives the total mark. A holistic scheme does not divide the marks between the criteria. Instead, how the criteria contribute to the quality of the whole is considered. There are various advantages and disadvantages to both systems and both are equally valid.

Analytical marking systems can make marking easier when the work being assessed is long and involved and or there are many criteria that need to be considered. It can be hard to sustain the level of concentration required to assess such work holistically and dividing the task into smaller components simplifies the task. However, ideally, application of such a marking scheme requires us to examine the work separately against each criterion which may mean re-reading the work several times. We often do not do this, as it is time-consuming. Instead, we assess against several criteria at once, and therefore, in actual practice, we may lose the benefit of an analytic system.

A problem with analytical systems is that examiners may find that the marks awarded do not feel right and are not consistent with an overall judgement of quality. This comes about because the sum of the whole can be more or less than its parts. It also comes about because not all criteria can be accurately prespecified and there may be unanticipated, but valid, aspects to the performance which positively or negatively influence your judgement. It is argued by some that it is impossible to define all aspects of quality in advance – only a sample of possible qualities can ever be specified. Furthermore aspects of quality often overlap and cannot be neatly divided into separate parts for marking. An appropriate weighting of different criteria is often contextual and hard to prespecify. For example, a 30% weighting of marks for the criterion of analysis, may be completely inappropriate if the analysis is of the wrong thing.

Holistic marking systems avoid these issues but are seen by some to be less objective, since the process of awarding marks against criteria is internalised and not so explicit. There may also be less agreement between examiners. However holistic marking schemes are time-honoured and used in many areas where judgement is required and are considered equally valid. In the end, all marking requires a judgement.

Incorporating principles or rules into marking schemes

Some marking schemes include a principle or rule which examiners can apply to discriminate levels of performance. The use of the principle can obviate the need for detailed descriptors of different levels of performance. Principles are also useful when candidates may use different content or examples to address a question. For example, the principle might state that the candidate provides an analysis of their chosen example that captures both (x) and (y) aspects of (z).

Writing marking schemes

The important steps in writing marking schemes are to identify the *key qualities* that you will be looking for in the work. You should think about this during the design of the task as it will influence the instructions you give to candidates in the question wording. The qualities you are looking for should be aligned with the learning outcomes for the subject in both scope and standard.

You then to articulate these qualities as well as you can for you or another examiner to follow later when marking. Table 16 provides generic criteria that apply to many examination questions and can be adapted for use in specific questions. Although Table 16 provides a large number of criteria, you should limit yourself to describing 3-4 important areas that the question addresses and enables candidates to demonstrate. As discussed in the next section and shown in Table 15, not all criteria can be assessed by all types of question.

Choose criteria based on their ability to differentiate passing and failing levels: your criteria do not have to be a comprehensive list of all the knowledge and skills required for the task; only those that are both important in the discipline and will help you tell if the candidate's answer is of a passing or failing level.

You may choose to provide descriptions of various levels of the quality, or just what constitutes a high level, but try to include guidance about how passing answers and failing ones should be distinguished. Consider what parts of an answer are essential and how many errors are tolerated. Remember that an expert answer may be less complete, but more relevant and focussed on the issues of importance, rather than every detail. This is the nature of expertise.

For an analytical marking scheme you then need to assign the value of marks possible for each criteria and each standard if you have specified a series of standards. I recommend that the highest weightings be given to areas other than factual knowledge, because adequate factual knowledge contributes to the quality of other criteria and therefore does not need to be separately weighted.

Marking schemes do not need to be very detailed; some summary points are all that is required. They can also be quite generic, with the same marking scheme suitable for using for many questions of the same type and structure.

Criteria and levels for marking schemes

There are four main types of criteria to consider when writing marking schemes, shown in Table 15 and Table 16.

Factual knowledge refers to "knowledge of" and "knowledge that". It is knowledge that is typically recalled and requires remembering. All types of examination question assess factual knowledge to some extent. Some questions (such as multiple choice and short answer questions) can assess nothing more than factual knowledge unless they are structured in a way that requires the candidate to perform a task before answering (for example to interpret clinical material). Answers to factual knowledge questions or parts of questions are usually it is thought of as being correct or incorrect and can be simply marked as either fully or partly correct or incorrect.

Understanding refers to knowledge of how and why and the meaning of things. Assessing understanding requires us to hear or read a candidate's explanations, rationale, discussions and justifications. Most long answer or oral questions are aimed at assessing understanding (Table 15). Marking requires us to consider how the question asks the candidate to respond and to judge the quality of their response. There are several aspects that could be considered, depending on the task required by the question (Table 16).

Planning or approach refers to the skill exhibited by a candidate in performing a task during an examination. In College examinations, the types of skills that are assessed frequently revolve around interpretation, analysis, problem solving, and planning management of problems. The candidate may be required to explain or articulate their approach, or it may be visible through their workings or by observing them performing the task. There are several aspects that could be considered, depending on the task required by the question (Table 16). Long answer and oral questions can enable us to evaluate a candidate's planning or approach but not all of them do so (Table 15).

Product or output refers to the candidate's response. For some types of questions, the quality of the response is not relevant (for example multiple choice questions where the candidate merely indicates the correct response, or very short answer questions, where only one or a few words are written, Table 15). In those types of question, it is the correctness of the answer rather than the quality of the answer that is assessed. For longer answer questions (whether written or oral), the quality of the response of the candidate can indicate relevant skills, including communicative and cognitive skills and is therefore informs the criteria used to mark the response (Table 15). There are several aspects that could be considered, depending on the task required by the question (Table 16).

Table 15: Types of criteria addressed by different types of examination questions.
Key: x – usually addressed; (x) - sometimes addressed depending on how the question is structured.

Type of criterion	Multiple choice	Short answer	Long answer	Oral
Factual knowledge	x	x	x	x
Understanding	(x)	(x)	x	x
Process or approach			(x)	x
Product or output			x	x

Table 16: Generic criteria and descriptors for examinations, grouped into four types of criteria: factual knowledge, understanding, process or approach, and product or output. Depending on the type of task and response required, examination questions may demand any combination of the criteria.

Factual knowledge

Criterion	High quality descriptor	Low quality descriptor
factual knowledge	Correct and sufficient facts or content (focus may mean it is not comprehensive but includes all that is important).	Errors of fact or content, missing important information.

Understanding

Criterion	High quality descriptor	Low quality descriptor
sense (as in making sense) and extent of explanation	Thorough, coherent, complete, systematic, deep and broad and goes beyond the information provided.	Incomplete or superficial, does not extend beyond what is given.

Criterion	High quality descriptor	Low quality descriptor
connection and association, grasp of subtleties and nuances	Explains or illustrates subtle connections, distinctions and associations, and considers the whole picture and mitigating circumstances. Draws inferences, conceptualises implications and explains assumptions and exceptions. Explanation relates to the wider theories and principles of the discipline or wider disciplines.	Overly generalised, black and white account, little discussion of associations, implications, assumptions, exceptions. Parts treated additively and independently. Explanation may lack connection to the wider theories and principles of the discipline.
degree of support – rationale and justification, understanding of why	Fully supported, justified, verified, by evidence or argument, including explaining counter-arguments and counter-opinions. Answer is qualified and takes account of context.	Knowledge telling with limited support, argument, justification or verification.
understanding of meaning	Thorough and insightful interpretation or analysis of the significance, importance, meaning.	Decodes with sparse or simplistic and reductionist interpretation, restates what is given, does not explain significance, importance or meaning.
contextualisation	Contextualises appropriately, applies meaning to specifics of the situation. Answer is selective - addresses what is important.	Theoretical and general response, not specific to situation. Non selective answer - incorporates irrelevant material.
insight	Illuminates tacit or overlooked assumptions, implications, conclusions.	Overlooks or glosses over tacit assumptions, implications, conclusions.

Process or approach

Criterion	High quality descriptor	Low quality descriptor
efficacy of process or approach	Effective application even in difficult contexts, performs the task completely and correctly.	Does not perform the task or does so incorrectly, or can only do so in simple contexts, outcome not produced or incomplete.
planning and approach	Methodical, logical, systematic and thorough plan for approaching the problem. Approach fully accommodates all requirements, purpose(s), and contextual matters.	Approach is unplanned or is disorganised, not thorough, illogical or unsystematic. Does not fully accommodate all requirements, purpose(s) or contextual matters.
customisation of process	Performance adapted to suit the situation, performs appropriately to context, constraints, purpose and audience. May be innovative in application.	Scripted, procedural, algorithmic or predetermined performance, follows steps irrespective of context or situation.
critical stance	Adopts a dispassionate, circumspect, and critical stance that examines an issue from different perspectives, considers other points of view, acknowledges other ways to approach a problem.	Singular perspective adopted, dismisses or unaware of alternative points of view or approaches.
awareness of limitations	Acknowledges the boundaries of own and others' understanding.	Unaware of the bounds of own understanding, overconfident.
recognition of personal prejudice and bias	Acknowledges role of own prejudices and bias on understanding.	Does not acknowledge role of projections and prejudice in opinions presented and approaches taken.

Criterion	High quality descriptor	Low quality descriptor
responsive, self-adjusting thinking or approach	Employs approaches that account for or limit effects of own limitations, prejudices and biases. Willing to change thinking or approach when finds inconsistencies.	Employs approaches that do not account for or limit effects of own limitations, prejudices and biases. Holds to thinking or approach even in the face of inconsistencies.

Product or output

Criterion	High quality descriptor	Low quality descriptor
fluency, efficiency, elegance, confidence of product or output	Fluent, efficient, elegant, confident performance. Writing is coherent, organised, engaging and persuasive. Uses appropriate technical language and ways of phrasing.	Mechanical or tentative performance, falters, slow, inefficient, lacks confidence. Writing lacks coherence, may be disordered, incomprehensible, repetitive or contain irrelevancies.

Revising marking schemes during marking

Devising questions and marking schemes is an inexact science. While there will be fewer problems if you follow the guidelines in this booklet, there are still times when marking schemes need to be modified during marking. Your approach should be to mark a number of candidate responses, and then re-evaluate the range of marks the marking scheme is producing against your holistic judgement. This can be done before you have marked all candidate responses to save you remarking everything after any adjustments. If you think a marking scheme might need revising, you should raise the issue with the Head Subject Examiner early, so that a new marking scheme can be agreed, and applied by all examiners before marking is complete.

Marking schemes may need to be adjusted because they seem to be scoring candidates overly leniently or stringently than a holistic judgement suggests is appropriate. Sometimes this involves only small adjustments of the weighting of marks. In other situations, you may find that your predetermined marking scheme does not capture all aspects of performance (good and bad) that you find in the candidate answers. This is relatively common in open-ended and higher order questions because it is impossible to characterise every aspect of quality of responses. When these situations occur, revising the marking scheme to incorporate a principle for examiners to apply may help clarify how to award marks to a range of different approaches made by candidates.

Marking schemes may also need to be adjusted during marking if candidate responses reveal unexpected problems with the question, such as unclear wording. Candidate responses may suggest that they have interpreted the question in a different way to that intended. This happens frequently enough that examiners should always be on the lookout for this issue when marking. On other occasions, problems with question wording do not become apparent until an item analysis is done—statistical analysis of the individual question performance amongst the group of candidates. When you find a candidate answer that does not address the question as you anticipated, your first thought should be that the fault may lie with the question wording rather than the candidate's level of knowledge and ability. Using the principles articulated in this booklet may help you to determine the nature of the problem with the question. Table 17 lists some causes of problems you should look out for.

Table 17: Construct-irrelevant causes of candidates providing a lower-level response than you expected. A candidate may not answer as we expect, not because they cannot, but because of problems we caused with the question. This table shows possible causes of candidates providing a response that is lower on the SOLO taxonomy (see Table 10) than the question was designed to elicit.

Level of response intended	Level of response produced	Possible causes unrelated to candidate ability.
Any level question	Prestructural answer	• The question asked has no answer • The question itself is irrelevant • The task was not made explicit • The question wording is ambiguous or unclear • The question evoked an inappropriate schema or stereotype
Multistructural or higher level question	Unistructural answer	• The question did not ask for more than one aspect
Relational or higher level question	Multistructural answer	• The question did not ask for more than a list of items • The question did not ask for an integrated and coherently structured answer (eg write short notes on) • There was not enough time to answer the question
Extended abstract question	Relational answer	• The question did not involve a complex problem or require a novel solution, alternative thinking or problem solving • There was not enough time to answer the question

Example marking schemes

Discipline principles question

This is a sample marking scheme for a question that requires discussion of principles of the discipline worth 25 marks. It can be adapted to a variety of questions where the main response requires explanation and discussion.

Question

Discuss the pathogenesis of disease xyz, including the current thinking on the role of initiating factors and the evidence supporting this thinking. (25 marks)

Marking Scheme

Category	Criterion	High quality descriptor	Low quality descriptor	Weighting
Factual knowledge	Knowledge of pathogenesis	Correct and comprehensive coverage of content, but focussed on most important aspects.	Errors of fact or content, missing important information.	10% 2.5 marks
Understanding	sense and extent of explanation, connection and association, understanding of why	Explanations are coherent and deep, incorporating principles at multiple levels (molecular, cellular, systemic etc).[1]	Superficial and generic explanations. Lacks connection to the wider theories and principles of the discipline.	30% 7.5 marks

[1] At a Fellowship level, the high level descriptor may include the use of references to key literature

Category	Criterion	High quality descriptor	Low quality descriptor	Weighting
Understanding	understanding of meaning, grasp of subtleties and nuances.	Importance and significance or implications discussed. Exceptions discussed. Subtle distinctions made clear.	Little or no discussion of significance, importance, implications or exceptions.	20% 5 marks
Product or output	coherent organised answer using appropriate technical language	Discussion is coherent, organised and persuasive. Uses appropriate technical language and ways of phrasing.	Lacks coherence, may be disordered, incomprehensible, repetitive or contain irrelevancies.	20% 5 marks

Clinical scenario question

This is a sample marking scheme for a written paper clinical scenario question worth 25 marks in total. It could be adapted for use with a variety of scenario questions.

Question:

[Clinical scenario including information to interpret.]

(a) Interpret the results of xyz tests presented and explain your interpretation. (5 marks)

(b) List 5 differential diagnoses that warrant consideration in this case. For each, explain what aspects of the scenario justify the consideration of it as a differential. (10 marks)

(c) Discuss the options for investigating this case. (10 marks)

Marking scheme

Part (a): Interpret the results of xyz tests presented and explain your interpretation. (5 marks)

Category	Criterion	High quality descriptor	Low quality descriptor	Weighting
Process or approach	Interprets correctly	Interprets all results completely and accurately	Incomplete or inaccurate interpretations	30% 1.5 marks
Understanding	Explains meaning and significance of results	Thorough and insightful explanation of the meaning and significance of all results. Interprets results in the light of each other and the clinical scenario presented. Explains subtle distinctions and exceptions.	Does not explain meaning or significance, or simplistic, reductionist, explanations. Explanations overly generalised and not interpreted in the light of the scenario.	50% 2.5 marks
Product or output	Coherent organised answer using appropriate technical language	Explanation is coherent, organised and persuasive. Uses appropriate technical language and ways of phrasing.	Lacks coherence, may be disordered, incomprehensible, repetitive or contain irrelevancies.	20% 1 mark

Part (b): List 5 differential diagnoses that warrant consideration in this case. For each, explain what aspects of the scenario justify the consideration of it as a differential. (10 marks)

Category	Criterion	High quality descriptor	Low quality descriptor	Weighting
Factual knowledge	Knowledge of differentials	5 appropriate differentials included, all are plausible.	Inappropriate differentials or less than 5 presented	30% 2 marks
Understanding	Justification for inclusion[2]	Choice of differentials fully supported, justified, verified, by evidence or argument, including explaining counter-arguments. Answer is qualified and takes account of context of the scenario. Answer is selective - addresses what is important.	Justification not provided or superficial, generic (not specific to case) or not well explained. Answer not contextualised and specific for the case.	50% 5 marks
Product or output	Coherent organised answer using appropriate technical language	Explanation is coherent, organised and persuasive. Uses appropriate technical language and ways of phrasing.	Lacks coherence, may be disordered, incomprehensible, repetitive or contain irrelevancies.	20% 2 mark

Part (c): Discuss the options for investigating this case. (10 marks)

Category	Criterion	High quality descriptor	Low quality descriptor	Weighting
Factual knowledge	Knowledge of options for investigation	Comprehensive discussion of all plausible options, including all most appropriate ones, and ones that would suit different contexts (owner constraints etc)	Options presented not comprehensive or includes some inappropriate options. Options presented do not suit a range of contexts.	30% 3 marks
Understanding	sense and extent of discussion, degree of justification – understanding of why[2]	Thorough discussion of pros and cons of a variety of approaches to investigation. Recommendations fully supported and justified, including explaining applicable contexts. Answer is qualified and takes account of context.	Incomplete discussion of pros and cons or misconceptions. Limited support or justification for options presented. Differing contexts not considered.	50% 5 marks
Product or output	coherent organised answer using appropriate technical language	Discussion is coherent, organised and persuasive. Uses appropriate technical language and ways of phrasing.	Lacks coherence, may be disordered, incomprehensible, repetitive or contain irrelevancies.	20% 2 marks

[2] At a Fellowship level, the high level descriptor may include the need to include references to key literature or organising principles of the discipline

Clarity

All our efforts in designing a set of appropriately challenging tasks for an examination will be wasted if the candidates were to do different task that we had not intended. We need them to do the task we envisaged. Therefore, it is very important that it is clear to them what they are to do.

Clarity is always important, but it is especially important during examinations because candidates are anxious and time-stressed. Both of these conditions affect thinking. Working memory is reduced, which results in reduced processing capacity. The use of schemas and stereotypes increases. The ability to concentrate on relevant information and suppress irrelevant information decreases. These effects can lead candidates to misunderstand the intent of the question or the task required of them.

When candidates read a question, they must form a mental model of the task required and begin to plan their response. They begin to form a model before they have even finished reading the question. The model they form is influenced by their expectations and their pre-existing mental schemas and sterotypes.

Schemas and expectations

All of us develop schemas or stereotypes which help us categorise and process complex information quickly. Particular features of questions trigger certain schemas and hence expectations. They do so without us being aware of it and before the reading of the text reaches consciousness, they affect our interpretation of the question.

There are five important implications for question writing

1. You may not see the problems with questions yourself because your schemas are different from those of candidates and informed by your expertise in the discipline and your own expectations for examination questions. Someone without your discipline expertise may see problems you don't see.

2. The text that you place at the very beginning of a question will have more influence on the schema elicited and the expectations developed by candidates than the text you place at the end. Images have more influence on the expectations developed than text (Crisp & Sweiry, 2006).

3. Anxiety and time stress (as occur in all candidates during examinations) increase the use of schemas. They can also make us "close" on a certain schemas too quickly and not look for others. Therefore, exams can be measures of a propensity to anxiety or writing speed rather than a measure of what we want to measure.

4. Each candidate will have different schemas based on their experience. This includes their relevant discipline experience, and this influence on performance may be considered relevant to what we are trying to assess. However it will also include general life-experience and experience with examinations that is irrelevant to what we are trying to assess and can lead to bias. If we are

not careful, our examinations can be more a measure of particular cultural backgrounds, or experience with examinations of this type, than the subject matter and abilities we intend to assess.

5. Problems with questions may not affect all candidates equally and therefore may be hard to detect.

Contextualising questions

Contextualising questions brings relevance and authenticity. It allows the assessment of concrete or specific examples rather than abstract concepts or generalisations. It also allows assessment of applied learning—doing not just knowing. However contextualising questions also brings the potential for bias, meaning that the question is easier for some candidates than others because of differences unrelated to the learning outcomes.

Contextualising questions can lead to difficulties in several different ways (Ahmed & Pollitt, 2007):

- By requiring extra words to be used. This increases reading time and load on short term memory which decreases cognitive processing capacity.
- By containing colloquial expressions or culturally-specific terms. These may elicit different schemas in different candidates and cause bias.
- Through the context being familiar to some candidates and not others. This differentially affects the difficulty of questions for individual candidates
- By containing information that is both relevant and irrelevant and requires sifting. This increases working memory load and makes questions more difficult.
- By activating schemas that may interfere with thinking (for example of a case that was similar in some ways but not in another important way).

Focussed contextualisation, however, can improve questions and result in fewer candidates misunderstanding the question (Ahmed & Pollitt, 2007). Contextualised questions are well-focused when the most salient aspects of the context are also the main issues addressed in the question. This makes them more likely to elicit helpful schemas than unhelpful ones in candidates with relevant knowledge. It does not make them easier for candidates that lack relevant knowledge.

Recommendations for maximising clarity

The effects of candidate schema use and expectations on their response to examination questions and the effects on thinking of contextualising questions, have implications for question design in order to maximise clarity.

Use clear language that is as simple as possible while still being precise.

- Ensure that the kind of answer required and the level of detail required is clear
- Use language that is not overly emotive
- Do not use humour
- Remove extraneous details unless you are trying to assess the candidate's ability to sift relevant information from irrelevant

Contextualise purposefully for what it brings to the task.

- Include only relevant and authentic scenarios
- Only include images if they are required for answering the question
- Focus the scenario by including only the salient aspects that the question is addressing.

Avoid contradicting expectations or switching schemas part way through a question. If you must, then ensure very very clear signalling.

- Examples of expectations being contradicted include:
    - asking what something is not (candidates expect to be asked what something is);
    - asking for candidates to address only one (or a few) aspects;
    - overly easy questions (candidates expect questions to be hard); or
    - when the marks available seem to indicate a length of answer that differs from the instructions.
- Examples of switching schemas part way through include:
    - asking about a particular case and then switching to a general situation about any case;
    - asking candidates to role play as if speaking to clients followed by questions on the same scenario that are not as if speaking to a client.

Steps for checking clarity

The steps in this section will take you through the process of checking the question wording once you have drafted a question and marking scheme. It is often helpful to have the wording checked by someone outside your discipline. They will spot problems not obvious to discipline-experts.

1. Ensure the wording gives an instruction (see page 8).
2. Ensure that the question scope is clear (see page 8). Specify it if necessary.
3. Check the question asks for what you are rewarding in your marking scheme—all of it.
4. Check that understanding the task does not rely on information that might not be common knowledge in different parts of the world. For example, do not assume candidates know the climate in a particular area unless this is part of the curriculum.
5. Consider how the phrasing of the question may distract or misdirect candidates and rephrase if necessary.
6. Use the simplest language possible while still being precise. This includes word choice and sentence structure. Remove all unnecessary wording.
7. Check that the discourse of the question mirrors the discourse of the response you require (see page 28). For example, do write as if an owner was speaking unless you want the candidate to write an answer as if speaking to a client (which you almost certainly do not).
8. Check spelling, grammar, and punctuation. Explain all abbreviations unless knowledge of them is considered part of the curriculum.

Oral examinations

Oral examinations are a traditional form of examination used in many disciplines. Although there are reasons to be critical of their use, oral exams can be better than other forms of assessment for determining the depth of a candidate's knowledge and their higher order skills in applying knowledge (Table 18).

Table 18: Strengths of oral assessment.

Oral assessment is a valuable technique for assessing: • depth of knowledge of candidates, • analytical, problem solving, and clinical reasoning abilities, • cognitive processes, • decision-making abilities, • judgement, • oral communication, and • professionalism, values, and ethics.

In the past, these benefits were often not realised. Surveys of oral examination questions in medicine have demonstrated that the majority assessed only recall, something that could be better and more reliably assessed in written examinations. Attention to the design of questions is important in order to maximise the benefits of oral examinations for assessing candidates.

An important disadvantage of oral examinations is their unreliability, which casts doubt on the accuracy of scores and threatens the validity of pass-fail decisions. Unreliability arises in two main areas in oral examinations:

1. case-to-case unreliability;
2. examiner-to-examiner unreliability.

Case-to-case unreliability refers to the variation in candidate ability (and hence scores) over different cases or clinical problems. This means that to some extent, a candidate's score will depend on the sample of cases they happen to be presented. Case-to-case unreliability can be reduced by increasing the number of cases each candidate is assessed on: the more the better. In order to facilitate this, questions should be focussed to get straight to the point, and the candidate should be moved along and kept to time so that they progress through all the cases planned.

Examiner-to-examiner unreliability refers to differences in the scores that different examiners award the same performance. This can be reduced through a variety of methods, including examiner training; advance preparation of questions, prompts, and marking schemes; and allowing only preplanned questions and prompts.

Additional methods that will reduce unreliability and increase reliability include:

- assessing all candidates on the same cases, and using the same prompts for each;
- assessing all candidates on the same number of cases;

- using well-designed questions that are aimed at assessing appropriate skills and are clear and straightforward for candidates to understand regardless of their culture or background;
- using well-structured marking schemes which are easy to apply and well-understood by examiners;
- planning and preparation to ensure that the oral examination conditions and process are as standardised as possible.

Good reliability is achievable with a high degree of structure in the questioning, marking, and conduct of oral examinations, however increased structure also reduces the opportunity to assess the very abilities that oral examinations are well placed to assess (Table 18). In addition, if not carefully designed, very short questions, while improving reliability by increasing the number of cases examined, may not enable assessment of integrated complex skills. Therefore, a balance must be found in order to provide adequate reliability while assessing important skills. The following sections will give some suggestions for how to find a good balance.

Preparing oral examinations

Blueprints

The oral exam content should complement the content of the other exam components used to assess candidates (written exams, practical exams). The emphasis should be on assessing the candidates' abilities the oral examination is most suited to assess (Table 18), rather than content knowledge. This means that the oral exam content will not necessarily be representative of the whole curriculum.

Plan the content to cover various areas of competence, for example, diagnosis, management, preventative care, communication, professional behaviour. It is important that all that is assessed in the oral examination lies within the scope of professional practice within your discipline and are specified in the subject guidelines. Preparing oral examination questions may suggest revisions to subject guidelines for future years, so that they more clearly reflect professional practice in the discipline.

Questions

Topics for questions should be centred on clinical scenarios. Plan to start a new topic about every 5 minutes. Make your questions short, straightforward, and clear. They should be able to be asked in a few sentences. Get straight to the point. You might even announce what the point is, before asking the question. For example, "This question is going to be about (*disease x*) and I want to explore your approach to management." Avoid very detailed scenarios, because you need to cover sufficient different scenarios in the time.

Questions should be open-ended. Open-ended questions require candidates to problem-solve, evaluate, make decisions, or make judgements using the knowledge they recall and the interpretations they make of the clinical information you provide. Open-ended questions do not have one right answer, but allow a range of responses. They are less predictable and do not involve simple recall of learned material. Examples include asking candidates to discuss various interpretations and

the evidence supporting them, or to discuss the strengths and weaknesses of various options, or to suggest a course of action and explain their reasoning.

Do not ask questions that are phrased in a way that asks for personal experience ("how is this usually managed in your practice?") or opinion ("what is your opinion on…"). The answer to these questions cannot be marked because the answers are things you cannot make a judgement about.

It is useful to classify draft questions as one of three types: recall, interpretive, and problem solving (Table 19). You should be aiming for the majority of questions to be interpretive and problem solving.

Table 19: Classification of question types (McGuire, 1966)

Recall: Questions that ask candidates only to recall facts, principles, or other information.

- "List the most common causes of (x)."
- "Define (x)."
- "What is the most common disease associated with (x)?"
- "Explain two mechanisms by which (x) results in (y)."
- "Describe the differences between Type 1 and Type 2 (x)."

Interpretive: Questions that require interpretation of information, and using principles to deal with new situations.

- "Describe the changes on this radiograph."
- "Interpret the results of this biochemistry panel."
- "Describe the abnormalities you see (in this image)."

Problem solving: Questions that require finding a solution.

- "What is the list of differentials for these findings?"
- "Outline your management of this case."
- "What would be the best next step in this case?"
- "Can you account for the clinical findings in this case?"

Follow-up prompts and probes

Follow-up prompts and probes are designed to explore the candidate's abilities and depth of knowledge on the topic presented in the main question. They should be carefully planned in advance based on likely responses from the candidate, but they can never be entirely standardised, because they depend on the answers given by the candidate to the main question. Because you need to move to a new topic within 5 minutes or so, there should be few follow-up prompts.

Examples of the types of prompts and probes that are useful are given in Table 20.

Table 20: Types of prompts and probes

Exploring the depth of knowledge and higher order abilities

- "Can you explain your reasoning?"
- "What are the implications of each of these options?"
- "What is the justification for your decision?"
- "What evidence supports your judgement in this case?"
- "Explain how you arrived at that conclusion"
- "Because?"
- "Why is that?"
- "What would you do if your approach does not work?"

Seeking more detail from the candidate

- "Any other differentials?"
- "Can you tell me which specific drug you would use?"
- "Which particular suture pattern would be most appropriate?"
- "Which particular signs are leading you to conclude that?"
- "Can you elaborate on that?"

Clarifying the question being asked or redirecting the candidate

- "What I am getting at here is…"
- "Can I get you to focus on (x) in this case?"
- "Can you tell me all the options rather than what you would do?"
- "Just coming back to something you mentioned earlier about (x), can you tell me…"

Asking about things that the candidate has missed out

- "Can you tell me any other implications? What about implications for the whole farm?"
- "One thing you didn't mention was (x). Can you tell me about that?"

Discourse

There are different ways of speaking which are used for different purposes or to signal different frames or reference or perspectives. Three different types of discourse are found in use in oral examinations (Roberts, 2000). Personal experience discourse is talk that deals with personal experience and feelings and often involves narratives and stories. Professional discourse is talk of professionals to clients and colleagues to colleagues. Institutional discourse is abstract and analytic and depersonalised.

It is important to frame your question in the same discourse that you want an answer. So if you want an answer to be framed in terms of principles (as you most often will), then don't ask candidates to speak as if to a colleague or a client, and don't ask about their own experiences.

Consider the difference between the answer to "How would you explain it to an owner?" and "Imagine I am an owner. Explain to me (x)." The first expects an answer in an institutional discourse, for example "Well I would make sure I gave all the options, because that ensures owners can make an informed decision, and …". The second is asking for a role play in which the discourse will be

completely different. The second question is very difficult for candidates to answer, because you are clearly not the owner. Most candidates will revert to an institutional or professional discourse when asked a question such as this and most examiners will be satisfied or even expect a response framed in this way. However having to consider how to respond in these situations can cause candidates to seem hesitant and may add to their anxiety. Candidates from ethnic minorities and those trained in different countries may not realise that a question framed in one discourse should actually be answered in another, leading to equity problems in examinations (Roberts, 2000).

Ancillary materials

Ancillary materials such as images, results, radiographs, need to have a clear function. Do not use them as decoration or scene-setters because they are distracting to candidates (see the section on clarity for more detail on the influences on candidate expectations). Images also need to clearly show what they are intended to show. Check that someone who is unfamiliar with the case can also see what you are seeing in the image.

Marking schemes

A structured marking scheme can improve reliability and agreement between examiners. It is also a useful basis for discussion between examiners and agreement ahead of time about the qualities of a good performance in that question. See the section on grading criteria for more detail about marking schemes.

It is best to keep the grading system simple and use only a few criteria. For example, a marking scheme may rate candidates' performance on a few criteria such as factual recall, analysis and interpretation of data, problem solving, and verbal presentation, plus assign an overall holistic grade. Each main question or topic should be separately graded. Using descriptors for different levels of performance is helpful.

Conducting oral exams

Creating the right climate

Getting the right climate or atmosphere in the examination helps reduce candidate anxiety and allows them to perform at their best. Begin with introductions and non-threatening conversation, such as about the weather.

Throughout the exam convey a sense of being genuinely interested in the candidate. Be attentive to how your behaviour is influencing the behaviour of the candidate. For example, some candidates feel confronted by you maintaining eye contact, but others need eye contact as a mark of your interest in what they are saying. Be aware that feedback such as nodding, or saying "yeah", "mhm", or "uhuh", can be interpreted by some candidates as agreement, by others as encouragement to go on, and by others as a signal that they should stop talking so you can ask the next question. There is therefore no one right way to encourage candidates. Be attentive and adjust your body language or behaviour if necessary.

It is best to avoid humour in the examination. Humour can be easily misunderstood. A candidate may think they are being laughed at. Other candidates may think that the use of humour means they have passed.

Finding the middle ground in structure: not too much and not too little

Oral examinations should be well planned, including prompts and probes and how these should be worded. However, the examination should not be run as if reading from a script. Candidates may, for example, give an answer to the main question that encompasses all planned probes so that these are not needed. Or they may answer in such a way that it is better to change the planned order of probes. Clarifications or redirections are only used if needed. Remember the important principle is that all candidates should be asked for the same knowledge, interpretations, solutions, or judgements and be provided with the same clinical information from which to base their answers. While it is important to maintain the planned discourse (or way of speaking), the exact words used can vary, especially the words used for metacommunication (communicating about the examination structure and processes). Examples of metacommunications include “we are going to move on to the next question now” or “now the next question is…”.

Dealing with problems

When candidates can’t answer or say too little

Firstly ensure that you have given the candidate enough time to think and respond. Do not rush in to clarify when there is a pause in the conversation. There are cultural and regional differences in how long people wait before responding, even when they know what they are going to say.

Do not give hints as they require the candidate to guess what you are thinking, and they mean that some candidates get more information than others with which to formulate an answer. Ask the candidate if they would like the question clarified and if they do, be careful not to provide additional information. Also be careful not to change the discourse or requirements of the answer. For example if the original question asked about a general situation, do not change to ask for an example from their experience.

If in doubt it is better to simply move on to the next question, or to directly ask for omitted information as illustrated in Table 20.

When candidates go off track or take too long

Use clarifying or redirecting questions (Table 20) to keep candidates on track. Remember you need to get through all the questions in the time.

When candidates are exceedingly nervous or anxious

All candidates are anxious to some extent, but when candidates are exceedingly nervous, trying to help them with redirecting or clarifying questions can make things worse. Although anxiety may be the cause of poor performance, the opposite can also be true, where poor performance is the cause of the anxiety. Therefore, sometimes it is best to simply move on to another question.

Scoring candidates

Record your grade for each main question/topic as you go. It is best to assess each question separately and if you wait until the end to score the questions, your opinion will be influenced by the candidate's performance in the later questions. Take brief notes that help justify your grading decision.

Examiners are often reluctant to award the very highest grades to a candidate. Considering what candidates would need to have done to get a higher grade is a useful step and helps you use a wider range of grades than you might otherwise. This improves reliability. For borderline performances, it is often helpful to ask yourself if you would trust this candidate to manage a case such as has been presented. This can help you decide on a clear pass or fail score for the candidate.

At the end of the oral examination, grades for individual questions may be summed to obtain an overall mark. You should also look over your grades and comments about the candidate's performance and make a holistic judgement about their level of performance, and whether you consider them to be a solid pass, borderline pass, borderline fail, or solid fail. This holistic judgement should also be recorded along with comments justifying your decision. It is possible that your holistic judgement is different than the sum of marks suggests. You should record it in any case.

It is tempting to make comments to the other examiner(s) about the candidate's performance before you have completed these steps, but you should not. Once you have completed your evaluation you should confer with your fellow examiner(s) about the grade and discuss any discrepancies in judgement, in order to come to a consensus decision.

Notes

[1] In working with the learning outcomes, Examiners may find that there are ways they could be improved to suit their use for characterising the domain of practice being assessed. Feedback from Examiners to the Chapter is needed to revise learning outcomes and keep them up-to-date and useful.

[2] Note that important areas of the curriculum are assessed through the training documentation provided by Fellowship candidates, to support inferences in relation to practical skills not easily assessable in examinations.

Further reading and references

Ahmed, A., & Pollitt, A. (2007). Improving the quality of contextualized questions: An experimental investigation of focus. *Assessment in Education: Principles, Policy and Practice, 14*(2), 201-232. doi:10.1080/09695940701478909

Biggs, J. B., & Tang, C. S.-K. (2011). *Teaching for quality learning at university* (4th ed.). Maidenhead UK: McGraw-Hill.

Crisp, V., & Novakovic, N. (2009). Are all assessments equal? The comparability of demands of college-based assessments in a vocationally related qualification. *Research in Post-Compulsory Education, 14*(1), 18p. doi:10.1080/13596740902717366

Crisp, V., & Sweiry, E. (2006). Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research, 48*(2), 139-154. doi:10.1080/00131880600732249

Crisp, V., Sweiry, E., Ahmed, A., & Pollitt, A. (2008). Tales of the expected: The influence of students' expectations on question validity and implications for writing exam questions. *Educational Research, 50*(1), 95-115. doi:10.1080/00131880801920445

Hamdy, H. (2006). Blueprinting for the assessment of health care professionals. *The Clinical Teacher, 3*(3), 175-179. doi:10.1111/j.1743-498X.2006.00101.x

Hughes, S., Pollitt, A., & Ahmed, A. (1998). The development of a tool for gauging the demands of gcse and a level exam questions. *BERA, Queen's University Belfast*.

Joint Committee on Intercollegiate Examinations (JCIE). (2016). Marking descriptors. *Policy No. G15 v1.0*. Edinburgh, Scotland UK.: Joint Committee on Intercollegiate Examinations, The Royal College of Surgeons of Edinburgh. Retrieved from https://www.jcie.org.uk/content/content.aspx?ID=8

Klatt, E. C., & Klatt, C. A. (2011). How much is too much reading for medical students? Assigned reading and reading rates at one medical school. *Academic Medicine, 86*(9), 1079-1083. doi:10.1097/acm.0b013e31822579fc

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory Into Practice, 41*(4), 212-218. doi:10.1207/s15430421tip4104_2

McGuire, C. H. (1966). The oral examination as a measure of professional competence. *Journal of Medical Education, 41*(3), 267-274. doi:10.1097/00001888-196603000-00011

McLaughlin, K., Lemaire, J., & Coderre, S. (2005). Creating a reliable and valid blueprint for the internal medicine clerkship evaluation. *Med Teach, 27*(6), 544-547. doi:10.1080/01421590500136113

Peitzman, S. J., Nieman, L. Z., & Gracely, E. J. (1990). Comparison of "fact-recall" with "higher-order" questions in multiple-choice examinations as predictors of clinical performance of medical students. *Academic Medicine, 65*(9), S59-60. doi:10.1097/00001888-199009000-00044

Roberts, C. (2000). Oral examinations---equal opportunities, ethnicity, and fairness in the MRCGP commentary: Oral exams---get them right or don't bother. *BMJ, 320*(7231), 370-375. doi:10.1136/bmj.320.7231.370

Scholten, I., Keeves, J. P., & Lawson, M. J. (2002). Validation of a free response test of deep learning about the normal swallowing process. *Higher Education, 44*(2), 233-255. doi:10.2307/3447458

Summers, J., & Catarro, F. (2003). Assessment of handwriting speed and factors influencing written output of university students in examinations. *Australian Occupational Therapy Journal, 50*(3), 148-157. doi:10.1046/j.1440-1630.2003.00310.x

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*(1), 7-25. doi:10.1080/08957340709336728

Wiggins, G., & McTighe, J. (2005). *Understanding by design* (2nd ed.). Alexandria, VA, USA: Association for Supervision and Curriculum Development.