Writing and marking examinations for paediatrics

Graham Clayden

Examiner's responsibility
There are some general principles in assessment that should always be considered irrespective of the assessment tool and the subject matter. The elements being assessed should obviously be those that should have been learned. This presupposes that the candidate has been guided by a list of clear learning objectives, has had teaching and clinical experience based on these objectives, and the knowledge that the examination is based on these. This sounds easy but is very difficult to achieve. Aims and objectives of courses are very straightforward to write even though an onerous and tedious task at times. However, the depth of acquisition of knowledge or the level of skill required at a particular stage is much more difficult to define or about which to achieve a consensus. This is the challenge for the question designer. The healthy recognition of the need to provide a statement of core knowledge and skill at the different stages of undergraduate medical education and later postgraduate training is helping to focus many minds on this problem. The bare minimum in paediatrics for undergraduates is that we should insist that every student should have reached a level where they are able to recognise a potentially ill child, know the most appropriate immediate action, be aware of the changing susceptibility and vulnerability to disease, trauma and treatment with age, respect the child's relationship with the family and with the society within which the child lives. This acquisition must then be so indelible as to last the professional life of the doctor assuming he or she will be one of those who might have no further training in paediatrics or child health. We need to worry less about those who eventually pursue a career involving further training in our subject, as they will be taught and tested again. However, undergraduate course design should have the recruitment of future children's doctors in mind. Blocking the further progress of an undergraduate because of failing to reach the minimal level in paediatrics is a serious step. We must be confident for the sake of the candidate that we are testing vital areas. On this reasoning it should be expected that anyone completing their undergraduate course in paediatrics, and who is taught effectively and who works reasonably diligently, should pass the core examination easily. If they fail it should be because they have become mentally ill and/or have had insufficient motivation to attend the course provided they were appropriately selected for medical school initially. It must not be because they have been subjected to tests that miss the points for which they have prepared. So questions must be prepared with care and be rigorously reviewed or discarded according to changes in knowledge and importance.

The demands for examining postgraduate trainees are no less challenging. Careers may be blighted by an unfair failure at a critical stage in a candidate's life. Equally an unjustified pass allowing a candidate to be promoted into a position of responsibility for which he or she is ill equipped may have even longer devastating effects. The risks to the child patients are obvious but the lasting damage to the trainee's confidence, as well as the potential legal consequences make this a very real factor for examiners to balance against their natural sympathy for the anxious candidate with whom every examiner can identify as a result of their own experience.

Choosing the type of paediatric examination questions
It is important to decide on the area of knowledge/ability and the depth to which it must be tested. Table 1 covers the range of competencies that most doctors need to use in practice. Each area is also subdivided into increasing depth. For every stage of medical education and paediatric training the appropriate level for these should be agreed.

Ideally, the examination should be constructed of a mixture of methods and questions. These should test as wide a range of these competencies within the practical constraints of the examination while recognising the limitations of the mental and physical stamina of the candidate and examiner. The balance between the subjects is usually a compromise between the availability of relevant clinical material and the time available to examine a particular number of candidates.

A key task for the examiner is to choose the assessment tool to match the content. Table 2 shows how the commonly used assessment tools fit in with the competencies and their levels in table 1 above.

Written papers
The problem of the traditional essay question is the subjective nature of the marking. This may be improved by the provision of detailed
marking schemes especially if essential key points are listed. Failure to mention an agreed number of these points in an answer would lead to a low mark. Double marking by examiners should also help especially if blind to each other's marks.

This is a cumbersome method of testing factual recall but does provide a medium to test written communication skills. The ability to explore complicated subjects and to use clear arguments can be tested by essay questions such as writing a letter to a parent, teacher, or general practitioner about an ethical problem or explaining complex physiology in simple language. Marking essays is very time consuming for the examiners. Short essay questions are reasonably reliable, and structured short answer questions are also preferable to the long essay. Modified essay questions (MEQ) allow the examiner to present sections of information such as parts of a case history and to ask questions requiring short statement answers at these stages. It is easier to achieve a consensus by examiners for the answers to the section questions than to find an agreed model answer for essay questions. All forms of modified essay questions require a great deal of time in their preparation and their value is still under debate.

**Essentials of the multiple choice question**

The multiple choice question (MCQ) is often the brunt of criticism by examiners and nearly overt hatred by candidates. The examiners may feel that they are constrained by the narrow structure and scope of the MCQ, although few doubt the need to assess the basic knowledge that should be resident in the memory. No one can read even this page without the knowledge of words being used. The candidates' antipathy would evaporate if they were to compare the objectivity of the MCQ with any other knowledge test. A great deal of effort and analysis has shown that MCQs have high reliability if constructed properly. There are essentials for constructing MCQs such as:

- Using a succinct stem where each item creates a complete sentence
- Avoiding cueing the answer with items or other questions
- Avoiding mutually exclusive answers
- Avoiding testing more than one fact per item
- Avoiding vague words like usually, maybe, sometimes
- Avoiding unwitling conventions like ‘characteristic’ always being used for ‘true’ answers
- Avoiding ‘always’ or ‘never’ which must be false in medicine
- Avoiding repetitive words in items where it is possible to refine the stem.

For example that list could have been written as follows. MCQs can be improved by avoiding:

- Grammatical mismatch of the stem with each item which should form a complete sentence
- Cueing answer with items or other questions
- Mutually exclusive answers
- Testing more than one fact per item
- Vague words like usually, maybe, sometimes
- Unwitling conventions like characteristic always being used for ‘true’ answers
- ‘Always’ or ‘never’ which must be false in medicine
- Repetitive words in items where it is possible to refine the stem.

Probably the commonest error in setting MCQs is the overestimation of the abilities of the candidate; there is seldom too easy a question. If examiners tried to write questions that they would be shocked if a student at that stage were to get wrong if asked orally on an informal ward round, the level of difficulty would be about right. For example an undergraduate attending a ward round might be expected to know the following about a child who has recently been diagnosed as having diabetes mellitus:

- Polyuria presenting as secondary enuresis in this child is not unusual. Infection may have been a precipitator of the onset of diabetic ketoacidosis at this stage. Diabetes was diagnosed by the general practitioner discovering glycosuria and ketonuria and the hospital confirmed the hyperglycaemia and the dehydration. The child was tachypnoeic as a consequence of the acidosis. Careful rehydration with respect for the potassium concentration was carried out intravenously in view of the child's vomiting. Insulin is given parenterally and care is taken to avoid hypoglycaemia. Abdominal pain is a frequent finding in

### Table 1 Paediatric competency and levels of depth

<table>
<thead>
<tr>
<th>Competencies</th>
<th>Level</th>
<th>Increasing depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge:</td>
<td>Recall Interpretation</td>
<td>Judgment</td>
</tr>
<tr>
<td>Tests</td>
<td>MCQ</td>
<td>Case commentary</td>
</tr>
<tr>
<td>Short note questions</td>
<td>Oral examinations</td>
<td></td>
</tr>
<tr>
<td>Recognition:</td>
<td>Naming Linking Weighing evidence</td>
<td>Best of five choice</td>
</tr>
<tr>
<td>Tests</td>
<td>Photo/video</td>
<td>Case commentary</td>
</tr>
<tr>
<td>Investigation:</td>
<td>Using Interpreting</td>
<td></td>
</tr>
<tr>
<td>Tests</td>
<td>MCQ</td>
<td>Data interpretation questions, essays and MEQ</td>
</tr>
<tr>
<td>Questioning:</td>
<td>Recall Listening</td>
<td>Ensuring feedback</td>
</tr>
<tr>
<td>Examination:</td>
<td>Technique Completeness</td>
<td>Relevance Style Interpretation</td>
</tr>
<tr>
<td>Planning:</td>
<td>Recall Synthesis</td>
<td>Audit Adaptation</td>
</tr>
<tr>
<td>Explanation:</td>
<td>Recall Clarity Accuracy Sensitivity</td>
<td></td>
</tr>
<tr>
<td>Tests</td>
<td>Long clinical case</td>
<td>Surrogate OSCE station</td>
</tr>
<tr>
<td>Empathising:</td>
<td>Respect Motivation</td>
<td>Sympathy</td>
</tr>
<tr>
<td>Tests</td>
<td>Observed history and actual practice</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Paediatric competency, depth, and assessment method

<table>
<thead>
<tr>
<th>Competencies</th>
<th>Level</th>
<th>Increasing depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge:</td>
<td>Recall Interpretation</td>
<td>Judgment</td>
</tr>
<tr>
<td>Tests</td>
<td>MCQ</td>
<td>Case commentary</td>
</tr>
<tr>
<td>Recognition:</td>
<td>Naming Linking Weighing evidence</td>
<td>Best of five choice</td>
</tr>
<tr>
<td>Investigation:</td>
<td>Using Interpreting</td>
<td></td>
</tr>
<tr>
<td>Tests</td>
<td>MCQ</td>
<td>Data interpretation questions, essays and MEQ</td>
</tr>
<tr>
<td>Questioning:</td>
<td>Recall Listening</td>
<td>Ensuring feedback</td>
</tr>
<tr>
<td>Examination:</td>
<td>Technique Completeness</td>
<td>Relevance Style Interpretation</td>
</tr>
<tr>
<td>Planning:</td>
<td>Recall Synthesis</td>
<td>Audit Adaptation</td>
</tr>
<tr>
<td>Explanation:</td>
<td>Recall Clarity Accuracy Sensitivity</td>
<td></td>
</tr>
<tr>
<td>Tests</td>
<td>Long clinical case</td>
<td>Surrogate OSCE station</td>
</tr>
<tr>
<td>Empathising:</td>
<td>Respect Motivation</td>
<td>Sympathy</td>
</tr>
<tr>
<td>Tests</td>
<td>Observed history and actual practice</td>
<td></td>
</tr>
</tbody>
</table>

MEQ=modified essay questions
Writing and marking examinations for paediatrics

children with diabetic ketoacidosis but the differential diagnosis must always include appendicitis and urinary tract infection. Children with diabetes are prone to urinary tract infections. The prognosis is good for survival but the diabetes will be permanent. Control during childhood is likely to be relatively straightforward but problems of compliance may arise in adolescence. The long term complications of diabetes are unlikely to occur in childhood but good control may delay their onset.

If this set of information bites is converted into MCQs a series of questions of appropriate difficulty can be constructed.

MCQ 1 – Recognised presenting features of diabetes mellitus in childhood include:
A. Secondary enuresis
B. Excessive weight gain
C. Abdominal pain
D. Reluctance to drink
E. Tachypnoea

Key: ACE are true or often expressed as TFTFT.

MCQ 2 – Diabetes mellitus in childhood:
A. Is often temporary
B. Is effectively treated with insulin given orally
C. Leads to renal failure in adolescents in the majority of cases
D. Predisposes to urinary tract infections
E. Becomes easier to treat during adolescence

Key: D is true or often expressed as FFFTF.

MCQ 3 – In a 6 year old boy who presents with increasingly severe abdominal pain the following would suggest a diagnosis of diabetes mellitus rather than acute appendicitis:
A. A three week history of nocturnal enuresis
B. Increasing frequency of vomiting over two days
C. A trace of ketones on urine testing
D. Polyuria
E. Periumbilical abdominal pain

Key: AD are true or often expressed as TFTTF.

MCQ 3 is testing at a deeper level where judgment as well as factual recall are necessary. Knowledge of diabetes and appendicitis are needed but each item tests a single element in this problem.

The examples given here are the multiple true/false type of MCQ, which is the type used mainly in the UK. They have the advantage that each MCQ tests five elements of knowledge whereas the 'best of five’ MCQ tests one. However, these may be very useful in testing decision making and the interpretation of information. For example, keeping to our 6 year old boy presenting with vomiting:

Question: A 6 year old boy who presents with increasingly severe abdominal pain over the preceding two to three days is noted to have a trace of ketones and +4 glycosuria on urine testing. He has been vomiting large volumes of bile stained liquid for the last 24 hours.

He is now drowsy and both pulse rate and respiratory rate have steadily risen over the last two hours. Initial results: plasma glucose concentration 14 mmol/l, urea 17 mmol/l, sodium 129 mmol/l, and potassium 4.9 mmol/l.

In addition to intravenous hydration, choose from this list the single most valuable next action:
A. Given intravenous insulin
B. Perform an erect and supine abdominal x ray
C. Give broad spectrum antibiotics intravenously
D. Perform a lumbar puncture
E. Proceed to laparotomy

Answer key A0, B+4, C+2, D0, E+1.

This test an awareness of intestinal obstruction (volvulus), that a trace of ketonuria and a plasma glucose of 14 mmol/l may occur in any ill child, and the risk of performing lumbar puncture on a drowsy child. However the single choice gives no indication on the knowledge of these other areas. If a wide range of knowledge is to be tested then the traditional multiple true/false MCQs are preferred. The best of five can be greatly improved by extending them to 10 or 15 choices -- the extended choice questions. These can be used in combination with more detailed case histories or with illustrative material such as the actual x ray in the case above showing fluid levels and a collection of gas in the splenic area only. These can be part of a written paper, although printing clinical photographs and radiographic images is expensive, or, alternatively, as part of an objectively structured clinical examination (OSCE).

There has been a lively debate whether negative marking should be used. Traditionally the multiple true/false MCQ has been marked awarding +1 for a correct statement of true or false and −1 for an incorrect response with zero for a blank or don’t know response. Those who support negative marking claim that doctors are penalised if they guess rather than admit their ignorance and seek more expert help in clinical practice. The opposition claim that negative marking tests both knowledge and the strategy used for uncertainty in candidates thus damaging the effectiveness of the test. Bright but anxious candidates may do less well than more ignorant but confident (or cavalier) students. Students can be helped with the certainty variable by being advised to put themselves in the position of the examiner who must construct the items. True items are relatively easy to set but false ones must be found by using opposites (for example MCQs 1B and 1D, 2A and 2E), false modifications of a true statement (for example MCQ 2B and 2C), or false dimensions of time or severity for example. It is much easier to find false options when comparing two diagnoses (for example MCQ 3). The candidate can then support their conclusion about the truth of the item beyond reasonable doubt from the basis of their factual knowledge that should offset their anxiety to a degree. However, this is a personality trait that may put some candidates at a permanent disadvantage. The removal of the penalty mark may help them and encourage them to attempt...
more questions where their first instinctive response is likely to be correct.

All examinations are subject to sampling error when trying to estimate a candidate's overall knowledge or skill. The MCQ does allow a very wide range of knowledge testing. Some examinations are composed of 60 MCQs each with five parts but candidates appear to cope with 100 MCQs each with five items without too much fatigue.

The effort for the examiner is in the design and syntax of the question. Marking can be performed by optically scanning candidates' answer sheets directly into the computer database. This allows detailed statistical analysis of each question as well as providing the results for the candidates. Most computer programmes will provide the mean scores for each MCQ as well as reporting the percentage who responded correctly, incorrectly, and who did not attempt each item. The discriminating power of the question is usually assessed by calculating the correlation coefficient for the MCQ by comparing the candidates' rank order for that question with their rank order for the whole MCQ test. The discriminating power of each item can also be checked by comparing the overall MCQ exam performance of those correctly answering the item with the overall performance of those who answered the item incorrectly. The mean score for each MCQ and the percentage of candidates who answered each item correct will indicate how difficult the question was for the candidates. The correlation coefficients indicate how discriminating the test has been. This also facilitates measures of internal reliability for the MCQ test. These data help reviewing committees of examiners to modify the questions, thus building up banks of used but finely honed questions. By using a number of identical questions in subsequent examinations, comparisons can be made of a particular candidate cohort with previous cohorts. This can be one way of using a norm referenced examination fairly provided the results are modified if significant differences are found between cohorts that cannot be expected by growth in knowledge or familiarity over a period of time (for example knowledge of retroviruses over the 1980 decade).

The debate over norm referencing (competitive examination with a predetermined percentage passing) or criterion referencing (non-competitive with predetermined pass mark) will probably be resolved when the core curriculum for each stage in training is established. Criterion referencing will then allow examinations based on the essential facts with some leniency for the odd gap in knowledge or technical mistake in making a response. Currently most MCQ examinations are norm referenced where there is a real risk of unfairness if a candidate is part of a very bright cohort or conversely a risk of passing the unprepared. To avoid this examinations should be taken by reasonable numbers of candidates, should include marker questions from previous examinations, and examiners should be prepared to modify their pass rate in the light of previous results. Another defence against the vagaries of the norm referenced exam is to make this just one part of a number of tests. Multipart or serial examination will allow any cohort effect to be reduced, although it does raise the question of how much summation of results should occur. Is it logical for one test to compensate for another? Before considering this it is important to review the methods available to test clinical skills as it is results in these parts of the examination that are often added to knowledge tests to come up with an overall pass mark. Most examinations in paediatrics put a greater weight on the clinical examination than on the knowledge test and many have veto marks available when examiners fear that clinical competence falls below levels that are safe for future patients.

Tests of clinical competence

The traditional long and short case examinations are increasingly being replaced by the OSCE. These examinations have been rigorously studied since their first description. They have been shown to be both reliable and valid10 provided a reasonable number of 'stations' are used. They compare favourably with other tests of clinical competence.12 13

OSCEs are usually cycles of 15–25 stations of tasks lasting between four and 10 minutes. They may be patient and examiner based stations or non-patient stations. The candidate is asked to examine one bodily system and the examiner uses an agreed checklist to reward each section of that clinical examination that is successfully carried out as well has having the opportunity of rewarding marks for attitude to the patient, recognition of abnormal signs, making diagnoses, and mentioning a plan for further management when time allows. Non-patient based stations may test communication or history taking skills, the ability to understand the use of equipment, observing photographs, radiographs or videos, and answering either extended choice questions or with free text single word or brief statements. The educational advantages of using OSCEs include:

- A potential for a much wider number of different types of clinical skills to be tested14
- The same selection of clinical skills being tested for every candidate
- A series of tests that are objectively marked using standardised checklists thus reducing observer bias
- A method for testing a large number of candidates in a relatively short period.

Some of the problems of OSCEs include:

- Strict time constraint on candidates and examiner
- Potential boredom from repetitive tasks for examiners
- High demand for large number of examinees, although for shorter number of days.

The traditional long clinical case is where a candidate is asked to take a history from a parent and child, perform a detailed examination, and prepare to present the findings to the
Writing and examinations

Some ways of improving traditional long case clinical exams include:

- Observing the candidate taking history
- Structured checklist for examiners
- Providing clear instructions for patients and parents about how much of the diagnosis to tell the candidates
- Double marking by two examiners blind to each others marks with no negotiation.

The traditional short clinical case is where the candidate is taken to a series of three to five children with a variety of physical signs. The candidate is usually tested in three to four body systems and reports the findings as they are discovered or invited to make a diagnosis and support this with evidence from the examination.

Some ways of improving traditional short cases:

- Select cases on educational grounds, ie, testing specific skills rather than conveniently available cases
- Group a series of short cases for examiners to use in a planned way
- Introduce the case to the candidate with a brief but realistic history
- Use standard checklists for clinical skills
- Provide positive feedback to candidates to reduce anxiety.

The oral examination usually lasts about 20 minutes with the candidate being asked questions on a wide range of topics by two examiners. There has been a great deal of debate on its value as a measure of professional competence. Many candidates feel that their anxiety inhibits their performance in this confrontation although anxiety does not seem to affect results.

Some ways of improving traditional oral examination (viva):

- Use a structured series of questions
- Focus on communication and explanation skills
- Use a wider range of subjects such as ethics, management and others which are difficult to test in MCQ format.

Marking schemes

The marking of these rather confrontational examinations is subjective but the use of marking schemes which use words rather than scores may help. For example grading: Outstanding ... Excellent ... Very good ... Good ... Satisfactory ... Barely acceptable ... Poor ... Very poor ... Abysmal rather than scores of 10 to 1 may help examiners to decide on whether compensation between one part of an examination should be permitted to allow the candidate to reach a pass mark. Many examinations merely add the scores for each section of the examination and allow compensation, others have veto marks or allow the candidate to fall below ‘satisfactory’ in only one part of the examination. Compensation does reduce the risk of one very strict or generous scorer leading to a pass or fail by his/her decision alone.

Conclusion

Examinations are an essential element of medical education, which generates vehement debate but unfortunately a relative lack of rigorous critical analysis. There appears to be a background anxiety that research findings that might suggest an examination has been less than material is then contrasted to arguments with candidates who have failed that examination. It is a major responsibility of all those involved in examining to seek evidence of the fairness, reliability, and validity of the methods and the organisation of the tests. Computers have made analysis of results much easier. Access to shared banks of all types of questions and answer sheets should allow examiners to select the subject first and the assessment tool second but from a range of tested and continually modified questions which allow comparison of candidates’ performance both in time and between institutions. The creation of examination content that can be considered as valuable an activity as research in academic life. There is little point in child health research if the advances in knowledge and skills that this generates cannot be shown to have been acquired eventually by present and future paediatricians.