

ABC of learning and teaching in medicine

Written assessment

Lambert W T Schuwirth, Cees P M van der Vleuten

Some misconceptions about written assessment may still exist, despite being disproved repeatedly by many scientific studies. Probably the most important misconception is the belief that the format of the question determines what the question actually tests. Multiple choice questions, for example, are often believed to be unsuitable for testing the ability to solve medical problems. The reasoning behind this assumption is that all a student has to do in a multiple choice question is recognise the correct answer, whereas in an open ended question he or she has to generate the answer spontaneously. Research has repeatedly shown, however, that the question's format is of limited importance and that it is the content of the question that determines almost totally what the question tests.

This does not imply that question formats are always interchangeable—some knowledge cannot be tested with multiple choice questions, and some knowledge is best not tested with open ended questions.

Five criteria can be used to evaluate the advantages and disadvantages of question types: reliability, validity, educational impact, cost effectiveness, and acceptability. Reliability pertains to the accuracy with which a score on a test is determined. Validity refers to whether the question actually tests what it is purported to test.

Educational impact is important because students tend to focus strongly on what they believe will be in the examinations. Therefore they will prepare strategically depending on the question types used. Whether different preparation leads to different types of knowledge is not fully clear, however. When teachers are forced to use a particular question type, they will tend to ask about the themes that can be easily assessed with that question type, and they will neglect the topics for which the question type is less well suited. Therefore, it is wise to vary the question types in different examinations.

Cost effectiveness and acceptability are important as the costs of different examinations have to be taken into account, and even the best designed examination will not survive if it is not accepted by teachers and students.

“True or false” questions

The main advantage of “true or false” questions is their conciseness. A question can be answered quickly by the student, so the test can cover a broad domain. Such questions, however, have two major disadvantages. Firstly, they are quite difficult to construct flawlessly—the statements have to be defensibly true or absolutely false. Teachers must be taught thoroughly how to construct these question types. Secondly, when a student answers a “false” question correctly, we can conclude only that the student knew the statement was false, not that he or she knew the correct fact.

“Single, best option” multiple choice questions

Multiple choice questions are well known, and there is extensive experience worldwide in constructing them. Their main advantage is the high reliability per hour of testing—mainly

Choosing the most appropriate type of written examination for a certain purpose is often difficult. This article discusses some general issues of written assessment then gives an overview of the most commonly used types, together with their major advantages and disadvantages

Reliability

- A score that a student obtains on a test should indicate the score that this student would obtain in any other given (equally difficult) test in the same field (“parallel test”)
- A test represents at best a sample—selected from a range of possible questions. So if a student passes a particular test one has to be sure that he or she would not have failed a parallel test, and vice versa
- Two factors influence reliability negatively:

Sample error—The number of items may be too small to provide a reproducible result

Sample too narrow—If the questions focus only on a certain element, the scores cannot generalise to the whole discipline

Validity

- The validity of a test is the extent to which it measures what it purports to measure
- Most competencies cannot be observed directly (body length, for example, can be observed directly; intelligence has to be derived from observations). Therefore, in examinations it is important to collect evidence to ensure validity:
 - One simple piece* of evidence could be, for example, that experts score higher than students on the test
 - Alternative approaches* include (a) an analysis of the distribution of course topics within test elements (a so called blueprint) and (b) an assessment of the soundness of individual test items.
- Good validation of tests should use several different pieces of evidence

True or false questions are most suitable when the purpose of the question is to test whether students are able to evaluate the correctness of an assumption; in other cases they are best avoided

Multiple choice questions can be used in any form of testing, except when spontaneous generation of the answer is essential, such as in creativity, hypothesising, and writing skills

because they are quick to answer—so a broad domain can be covered. They are often easier to construct than true or false questions and are more versatile. If constructed well, multiple choice questions can test more than simple facts. Unfortunately though, they are often used to test only facts, as teachers often think this is all they are fit for.

Multiple true or false questions

These questions enable the teacher to ask a question to which there is more than one correct answer. Although they take somewhat longer to answer than the previous two types, their reliability per hour of testing time is not much lower.

Construction, however, is not easy. It is important to have sufficient distracters (incorrect options) and to find a good balance between the number of correct options and distracters. In addition, it is essential to construct the question so that correct options are defensibly correct and distracters are defensibly incorrect. A further disadvantage is the rather complicated scoring procedure for these questions.

“Short answer” open ended questions

Open ended questions are more flexible—in that they can test issues that require, for example, creativity, spontaneity—but they have lower reliability. Because answering open ended questions is much more time consuming than answering multiple choice questions, they are less suitable for broad sampling. They are also expensive to produce and to score. When writing open ended questions it is important to describe clearly how detailed the answer should be—without giving away the answer. A good open ended question should include a detailed answer key for the person marking the paper. Short answer, open ended questions are not suitable for assessing factual knowledge; use multiple choice questions instead.

Short answer, open ended questions should be aimed at the aspects of competence that cannot be tested in any other way.

Essays

Essays are ideal for assessing how well students can summarise, hypothesise, find relations, and apply known procedures to new situations. They can also provide an insight into different aspects of writing ability and the ability to process information. Unfortunately, answering them is time consuming, so their reliability is limited.

When constructing essay questions, it is essential to define the criteria on which the answers will be judged. A common pitfall is to “over-structure” these criteria in the pursuit of objectivity, and this often leads to trivialising the questions. Some structure and criteria are necessary, but too detailed a structure provides little gain in reliability and a considerable loss of validity. Essays involve high costs, so they should be used sparsely and only in cases where short answer, open ended questions or multiple choice questions are not appropriate.

“Key feature” questions

In such a question, a description of a realistic case is followed by a small number of questions that require only essential decisions; these questions may be either multiple choice or open ended, depending on the content of the question. Key feature questions seem to measure problem solving ability

Teachers need to be taught well how to write good multiple choice questions

Which of the following drugs belong to the ACE inhibitor group?

(a) atenolol	(h) metoprolol
(b) pindolol	(i) propranolol
(c) amiloride	(j) triamterene
(d) furosemide (frusemide)	(k) captopril
(e) enalapril	(l) verapamil
(f) clopamide	(m) digoxin
(g) epoprostenol	

Example of a multiple, true or false question

Open ended questions are perhaps the most widely accepted question type. Their format is commonly believed to be intrinsically superior to a multiple choice format. Much evidence shows, however, that this assumed superiority is limited



“Key feature” questions aim to measure problem solving ability validly without losing too much reliability

validly and have good reliability. In addition, most people involved consider them to be a suitable approach, which makes them more acceptable.

However, the key feature approach is rather new and therefore less well known than the other approaches. Also, construction of the questions is time consuming; inexperienced teachers may need up to three hours to produce a single key feature case with questions. Experienced writers, though, may produce up to four an hour. Nevertheless, these questions are expensive to produce, and large numbers of cases are normally needed to prevent students from memorising cases. Key feature questions are best used for testing the application of knowledge and problem solving in "high stakes" examinations.

Extended matching questions

The key elements of extended matching questions are a list of options, a "lead-in" question, and some case descriptions or vignettes. Students should understand that an option may be correct for more than one vignette, and some options may not apply to any of the vignettes. The idea is to minimise the recognition effect that occurs in standard multiple choice questions because of the many possible combinations between vignettes and options. Also, by using cases instead of facts, the items can be used to test application of knowledge or problem solving ability. They are easier to construct than key feature questions, as many cases can be derived from one set of options. Their reliability has been shown to be good. Scoring of the answers is easy and could be done with a computer.

The format of extended matching questions is still relatively unknown, so teachers need training and practice before they can write these questions. There is a risk of an under-representation of certain themes simply because they do not fit the format. Extended matching questions are best used when large numbers of similar sorts of decisions (for example, relating to diagnosis or ordering of laboratory tests) need testing for different situations.

Conclusion

Choosing the best question type for a particular examination is not simple. A careful balancing of costs and benefits is required. A well designed assessment programme will use different types of question appropriate for the content being tested.

Further reading

- Case SM, Swanson DB. Extended-matching items: a practical alternative to free response questions. *Teach Learn Med* 1993;5:107-15.
- Frederiksen N. The real test bias: influences of testing on teaching and learning. *Am Psychol* 1984;39:193-202.
- Bordage G. An alternative approach to PMPs: the "key-features" concept. In: Hart IR, Harden R, eds. *Further developments in assessing clinical competence; proceedings of the second Ottawa conference*. Montreal: Can-Heal Publications, 1987:59-75.
- Swanson DB, Norcini JJ, Grosso LJ. Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education* 1987;12:220-46.
- Ward WC. A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement* 1982;6(1):1-11.
- Schuwirth LWT. *An approach to the assessment of medical problem solving: computerised case-based testing*. Maastricht: Datawise Publications, 1998. (Thesis from Department of Educational Development and Research, Maastricht University.)

BMJ 2003;326:643-5

Example of a key feature question

Case

You are a general practitioner. Yesterday you made a house call on Mr Downing. From your history taking and physical examination you diagnosed nephrolithiasis. You gave an intramuscular injection of 100 mg diclofenac, and you left him some diclofenac suppositories. You advised him to take one when in pain but not more than two a day. Today he rings you at 9 am. He still has pain attacks, which respond well to the diclofenac, but since 5 am he has also had a continuous pain in his right side and a fever (38.9°C).

Which of the following is the best next step?

- Ask him to wait another day to see how the disease progresses
- Prescribe broad spectrum antibiotics
- Refer him to hospital for an intravenous pyelogram
- Refer him urgently to a urologist

Example of an extended matching question

- (a) *Campylobacter jejuni*, (b) *Candida albicans*, (c) *Giardia lamblia*, (d) *Rotavirus*, (e) *Salmonella typhi*, (f) *Yersinia enterocolitica*, (g) *Pseudomonas aeruginosa*, (h) *Escherichia coli*, (i) *Helicobacter pylori*, (j) *Clostridium perfringens*, (k) *Mycobacterium tuberculosis*, (l) *Shigella flexneri*, (m) *Vibrio cholerae*, (n) *Clostridium difficile*, (o) *Proteus mirabilis*, (p) *Tropheryma whippelii*

For each of the following cases, select (from the list above) the micro-organism most likely to be responsible:

- A 48 year old man with a chronic complaint of dyspepsia suddenly develops severe abdominal pain. On physical examination there is general tenderness to palpation with rigidity and rebound tenderness. Abdominal radiography shows free air under the diaphragm
- A 45 year old woman is treated with antibiotics for recurring respiratory tract infections. She develops a severe abdominal pain with haemorrhagic diarrhoea. Endoscopically a pseudomembranous colitis is seen

Using only one type of question throughout the whole curriculum is not a valid approach

Lambert W T Schuwirth is assistant professor and Cees P M van der Vleuten is professor and chair in the department of educational development and research at the University of Maastricht in the Netherlands.

The ABC of learning and teaching in medicine is edited by Peter Cantillon, senior lecturer in medical informatics and medical education, National University of Ireland, Galway, Republic of Ireland; Linda Hutchinson, director of education and workforce development and consultant paediatrician, University Hospital Lewisham; and Diana F Wood, deputy dean for education and consultant endocrinologist, Barts and the London, Queen Mary's School of Medicine and Dentistry, Queen Mary, University of London. The series will be published as a book in late spring.