Designing and Managing Multiple Choice Questions (2nd Ed)
Carneson, John; Delpierre, Georges; Masters, Ken
2016

Contents

1. Introduction

1.1 A Note on the 2nd edition

The first edition of this document was originally published as a hard copy handbook and then a set of web pages on the University of Cape Town's (UCT) Web site, and was designed to familiarise UCT staff with the benefits and limitations of using Multiple Choice Questions (MCQs).

Unfortunately, these pages removed from the site, and the document is no longer available in its original form. As a result, this edition has been created with a DoI so that it can easily accessed. This document does not differ substantially from the original; in most cases, the edits are minor corrections and updates and link verification. In other cases, some references to activities and programmes (e.g. The Technology-Assisted Testing (TAT) Programme that existed at UCT) have been removed. This document may be freely downloaded and distributed, and should be cited as:

Carneson J, Delpierre G, Masters K. 2016. *Designing and Managing Multiple Choice Questions, (2nd ed.)*. [The DoI is also recommended.]

This edition is dedicated to the memory of Georges Delpierre.

1.2 Original Front Page

This handbook [was] aimed chiefly at the staff of the University of Cape Town, South Africa. As a result, some of the information is relevant only to lecturers at the University.

A large amount of the information, however, will be useful for anybody who wishes to implement MCQs, or simply to find out a little more about them.

The hard copy version of **Designing and Managing Multiple Choice Questions** was written by:
- John Carneson
- Georges Delpierre
- Ken Masters

Date of original Production: 1996

1.3 Acknowledgements

The following have made valuable contributions to this document, either in the form of comments, or by supplying examples of questions:
- N. Bak
- J. Head
- A. Leiman
- D. Lincoln
- D. Steyn
- N. Yeld

Thanks also to Allan Kent for help with Tables in HTML.

2. What are Multiple Choice Questions?

Contents of This Section

2.1 Definition

A "multiple choice question" (MCQ) is a question in which students are asked to select one alternative from a given list of alternatives in response to a "question stem."

For example,

Question stem:
"Which one of the following is the currency unit used in Albania?"

Alternatives:
1. Dinar
2. Rouble
3. Kwacha
4. Lek
5. Rupee

The student answers the question by selecting a number 1-5, and entering this either on a form provided for the purpose ("off-line," in computer jargon), or at a computer keyboard ("on-line"). In this example, the correct alternative (the "answer") is 4. The other incorrect alternatives are sometimes called "distractors".

One feature of this type of question is that there is a finite possibility of selecting the correct alternative by purely guessing. In fact, in this example you have 1 chance in 5 of answering the question correctly if you know absolutely nothing about the currencies in use in various countries. This problem is not insurmountable, and will be discussed further in Section 4.

The so-called "TRUE/FALSE" question, in which the student chooses between two alternatives "True" or "False", is a special case of an MCQ. "True/False" questions are used extensively in the UCT Medical School.

Certain MCQs have more than one correct response in the set of alternatives. Such questions should not be used, not only because they tend to confuse students, but also because not all computer programmes available at UCT cater for such a question type. For example, the programme "MCQ", which marks the responses of students who have written the test "off-line", does not admit this type of question. On the other hand, the online test creating packages which are designed principally for "on-line" testing, do allow a much broader range of question types. These questions, however, are not, strictly speaking, MCQs. As a result, they fall outside the scope of this document.

Return to Contents of this Section

2.2 Advantages of Multiple Choice Questions

- MCQs can be easily marked/scored, and this scoring can be both accurate and objective. In particular, MCQs can be *machine-scored* (This means that a computer will do all the marking for you.)

- MCQs can be set at different cognitive levels. For example, a question may simply challenge a student's ability to recall facts, while another may test a student's ability to apply factual knowledge to given situations; or, at a higher level, a question may test a student's ability to evaluate given information. For a discussion of this, please turn to Appendix C.

- MCQs can be designed with a diagnostic end in mind, in order to find out whether specific areas of a given subject are adequately known, or in order to detect misconceptions. This can provide feedback on the effectiveness of instruction.

- Students with poor reading skills and second-language learners need not be disadvantaged, provided the tests are designed appropriately Misreading of a question stem or an alternative may result in the loss of a few marks only; conversely, the misreading of an essay-type question, followed by the poor writing skills normally associated with a second language, can cause the student to lose an inordinate number of marks.

- Statistical information on performance can be readily obtained. One can not only find out how the class performed on a particular question, but determine whether the question was suitable in the context in which it was presented. The statistics that are gathered may also be used to rank questions with respect to their difficulty and their ability to discriminate between students of different competences.

- Tests made up of MCQs can be administered frequently, and thus provide regular information on student performance, not only to the instructor but also to the students.

- Tests made up of MCQs can provide a better coverage of content than essay-type questions, thus broadening the scope of the test. Such tests can be structured to include questions of defined grades of difficulty and discriminating power (See Section 4).

- In MCQ testing, the instructor "sets the agenda", and there are no opportunities for the candidate to avoid complexities and concentrate on the superficial aspects of the topic, as is often encountered in essay-type questions.

2.3 Problems associated with Multiple Choice Questions

- The construction of good test items (questions) requires special care and is therefore time-consuming.

- Instructors tend to favour "recall" type questions, as they are the easiest to design. This manual will provide you with examples which should convince you that it is possible to go far beyond this type of question.

- Experience has shown that in some environments, it may be necessary to win the acceptance of students to this type of testing. Proper communication with the students is therefore essential in order to ensure that they feel comfortable with this type of testing. Feedback from students can be very important and can only lead to an overall improvement of the existing question, for possible inclusion into a question bank.

- Creativity cannot easily be tested (the *Synthesis* level of Bloom's Taxonomy is very difficult to test by using MCQs). Creativity is best tested by discursive questions such as the "Essay-type" question. Bear in mind however that essay questions tend to have a poor sampling of content, have a poor scoring reliability (particularly if several persons are marking large numbers of scripts), and are notoriously time consuming to mark. (Bloom's Taxonomy is discussed further in Appendix C).

- Candidates cannot justify choices that have been deemed "incorrect" by the examiner. Two ways (apart from the necessary clarity of language in the stem and alternatives) of removing this problem are:

  [a] building the argument into the alternative. Examples of this are given in later sections of this document.

  [b] allowing students to discuss the questions with the lecturers and tutors after the test. In addition to allowing issues to be clarified, the questions might possibly need to be ignored in the calculation of test scores, and corrected for or eliminated from future tests.

Return to Contents of this Section

2.4 The multiple uses of Multiple Choice Question tests

The most obvious use of MCQs is in formal class tests and examinations, and this use is covered elsewhere in this handbook. In this section we want to draw your attention to a number of other uses, such as:

- Designing a tutorial around a few MCQs. They can serve to guide the tutor; stimulate discussion (perhaps in groups) and test understanding. Students who are normally reticent to speak may do so in defence of their choice of answer. The tutor prepare explanations prepared for both the right answer and distractors. Again, the reaction of students should be used to improve the question data bank where necessary.

- MCQs can be used in a variety of ways in lectures. For example, MCQs can be put on overhead (OHP) transparencies and students individually or in pairs can come up with answers. This can help to facilitate interactive lecturing and usefully 'break' up lectures. If the distractors are carefully selected to 'catch' misunderstandings, then the response of the class can be used as a diagnostic indicator. MCQs could also be used to test prior understanding and to break the ice when starting a new topic.

- Short tests - e.g., of 5 MCQs - can be administered and marked by computer to provide a quick diagnostic feedback on the progress of a class, or as a revision exercise to prepare for exams, or as a bridge into "essay" tasks. They need not carry marks or have the students' names on them: the lecturer could follow up with a general explanation of why one answer was correct and the others were not.

The above suggestions represent just a few of the possible uses of MCQs. Their use in wider contexts than just formal testing is important because it serves to familiarise the students with a range of answering techniques and the MCQ process in general. The methods needed to answer different kinds of MCQs should be made explicit to students and discussed with them. For a variety of reasons there may be considerable resistance from students to MCQs and it is essential that the benefits of MCQs are clearly demonstrated to them.

2.5 Managing the use of MCQs at departmental and course team level

Limiting the use of MCQs to one or two individuals in a department is problematical for three basic reasons:

- the use of MCQs should be taken into account when designing courses, which implies the course team should be involved;

- it is unlikely that an individual lecturer can effectively develop a bank of properly evaluated questions which are used in conjunction with a range of assessment methods;

- winning acceptance of MCQs among the students is crucial and this entails action at the level of the department. Each department will have its own problems and possibilities with regard to the setting up of MCQ banks and integrating MCQs into the design of courses.

3. Selected examples of Multiple Choice Questions.

Contents of This Section

This manual is aimed at giving you an overview of the general principles used in the construction of MCQs. The "nuts and bolts" are to be found in Appendix A. In this Section, we aim to give you an insight into the process whereby lecturers have attempted to apply these general principles in the context of three different disciplines.

3.1 Philosophy of Education (First year B.Ed. course)

3.1.1 Example A. Context: First year course test

```
If we understand learning a school subject on the model of learning a
language we can explain
a)  why teachers do not articulate clearly.
b)  why it is not possible to explain the point of the subject to someone
    who does not understand the subject.
c)  why truth is important.
d)  why subjective experiences are important.
e)  why school subjects are so important.
f)  why school subjects are taught verbally.
```

Lecturer's approach: Bak gives this example in her paper "How to test insight and understanding of philosophical issues by means of multiple-choice questions." (Bak 1991) She explains that in order to answer this kind of question, the students must first understand not only the concepts but "their relation to one another as well." In the course, students were introduced to the "model of learning a language" (see question stem) in order to illustrate the educational implications of the philosophical stance that "we come to understanding only through a process which is not given, but through which we must work." The lecturer's intention was that in arriving at the correct answer (b), the students must have compared each of the statements with their essential understanding of the part of the course dealing with a model of "learning a language"; and they need to be able to understand and relate:

- the concept of a "model of learning a language";
- the general concept of understanding being a process, which in the course was linked to a particular philosophical stance, and
- the argument presented in option (b) "why it is not possible to explain the point of the subject to someone who does not understand the subject."

Comments: The lecturer who set this question knew what concepts they gave prominence in her course and what relationships between the concepts she emphasised in lectures, tasks, etc. Although the lecturer's strategy appears sound, there are technical problems with the question which might hinder some students from dealing with the question in the manner which was intended. Firstly, the correct option (b) is much longer than the distractors. For any student with experience in answering MCQs, this is a powerful clue. Secondly, the use of the double negative ("not...not") in option (b) is confusing, even to first language speakers. Thirdly, one should not repeat the word "why" in front of each response, but include it as part of the stem.

Such problems should be picked up when the questions are reviewed by the lecturing team and the question edited as required, or when the student answers are analysed. (See Section 4)

Return to Contents of this Section

3.2 Sociology (First year course)

3.2.1 Example B. Context: This is in a first year block dealing with organisations, bureaucracy and democracy, using trade unions as an example. This question was answered through group discussion in a supplementary tutorial designed for examination revision.

```
The following statements are all answers to the question
```

Why have workers formed trade unions in South Africa?

Which is the best answer?

a) `They needed to resist the powerful organisations the employers could call on.`
b) `They needed organisations which could advance their struggle for political rights.`
c) `They needed to be come together to discuss what todo about their common problems.`
d) `They needed to be organised in order to improve their working and living conditions.`
e) `They needed to be organised because it was the only way they could be united.`

Lecturer's approach: The lecturer who designed the question and oversees the supplementary tutorial programme worked closely with the lecturer who taught the block. In the course, students were shown how and why workers formed the first trade unions, with biographies of trade unionists used to illustrate a historical process. All the statements used as responses reflect some stages or aspects of this process, but response (d) is the "best" answer because it

- is the most fundamental (i.e. it has a higher order of generality so that it includes all the others);
- logically preceeds the others in causal terms;
- follows the structure of the argument and emphases that the lecturer presented in his block.

The intention of the person who set the question was that the students

- evaluate each statement;
- order them according to one or more or the above criteria.

The question was designed to promote maximum discussion around the responses in order to facilitate the revision process. Nine out of the twelve groups of (mainly second language)

students considered to be "at risk" arrived at the correct answer within 5-8 minutes of group discussions.

Comments: This question must be seen in the context of the course, and must be considered to be on the difficult side, requiring the highest cognitive skills (see Appendix C). As such, it is ideal to use in a tutorial situation as it will elicit much discussion. The repetition of the phrase "They needed..." in front of each alternative is normally undesirable, but can be defended here on the grounds that one is asked to select the BEST answer, from a list of plausible complete statements

3.2.2 Example C. Context: The question could be used in a first-year sociology tutorial or
       test.

```
Read the following statement and select the answer which you think best
sums up the argument being made.
```

"Traditionally, we divide ourselves into races by the twin criteria of geographic location and visible physical characteristics. But we could make an equally reasonable and arbitrary division by the presence or absence of a gene, such as the sickle-cell gene, that confers resistance to malaria. By this reckoning, we'd place Yemenites, Greeks, New Guineans, Thai and Dinkas in one "race", Norwegians and several black African peoples in another."

```
a.  Norwegians and some black African peoples belong to the same race.
b.  Race is determined by the presence or absence of the sickle-cell gene.
    Thus there are two races in the world.  Those with the sickle-cell gene
    and those without it.
c.  We all belong to a race.  Which race we belong to depends on where we
    come from and our skin colour.
d.  "Race" as we commonly understand it is an arbitrary means of classifying
    human beings.  We could choose many other random criteria and get all
    sorts of different "races".
```

Lecturer's approach: The lecturer has a double agenda. Firstly, she wants to test and develop the students' ability to extract the central argument from a text. Secondly, she wants to test their grasp of a specific concept: that there is no credible biological basis on which to divide humans into "races". (The correct answer is "d").

Comments: The distractors need to be made more equal in terms of length. This can usually be done fairly easily by "padding" the shorter distractors. There are only four distractors - a fifth could make the question more challenging. Students for whom English is a second language could find the phrasing in the text difficult to penetrate. Students who have some knowledge about genes and sickle-cell anaemia might be at an advantage.

Return to Contents of this Section

3.3 Economics (First year course)

3.3.1 Example D. Context: The question was used in an Economics first year quarterly test.

```
If planned aggregate expenditure is greater than National Income and the
short run aggregate supply curve is horizontal, the economy:

a. cannot be on the aggregate demand curve.
b. cannot be on the short run aggregate supply curve.
c. cannot be on the long run aggregate supply curve.
d. will experience rises in both GDP and prices.
e. both (b) and (d) are correct.
```

Lecturer's approach: The correct answer is (a). Although the answer will not be apparent to non-economists, the question does show how MCQs can be used to test students' ability to apply technical terms and concepts. The test consisted of about one third recall questions, one third of questions aimed at the "average" student and one third questions - like the one above - designed to discriminate between the more advanced students and the rest. To answer the question the students would have to:

1. understand a number of assumptions (eg., that prices are fixed);
2. understand a number of definitions (eg., of a horizontal supply curve);
3. analyse a particular case and apply their understanding to it.

Students were told that they could arrive at right answers by eliminating the wrong choices - for example, if prices are fixed then option (d) must be wrong - but they could not do this reliably without some understanding of basic terms and assumptions. They were also told that in cases where a problem had to be visualised then they should use scrap paper to draw the diagram.

Comments: This is a demanding question, not only on account of the subject knowledge required, but because of the complex structure of the stem and alternatives. The conditional ("if...") combined with the negative in "cannot" (alternatives a, b, and c) contribute in making the syntax challenging, especially to students for whom English is not the mother-tongue. Students are expected to go through a number of fairly involved logical reasoning steps and they should be warned of this in advance of the test. Bearing the above points in mind, sufficient time should be allocated to allow a reasonable range of students to tackle such questions. Alternative "e" should be revised, as (i) it wrongly suggests that more than one alternative could be correct (ii) can mislead students into thinking that "b" or "d" may be correct and (iii) adds to the overall complexity of the question.

4. Scoring and statistics

Contents of This Section

In this Section we will consider the issues of scoring of individual MCQs and whole tests, as well as the statistics which pertain to students and questions.

4.1 Negative marking. To use or not to use?

The subject of "negative marking" is unnecessarily emotive, and the explanation that follows is aimed at enabling one to discuss the subject dispassionately, if at all possible. By negative marking, one means a scoring scheme whereby marks are subtracted every time an incorrect alternative has been selected. This "penalty" is frequently viewed as unfair by students, and is such a contentious issue in some quarters that its use is rare.

What can be said to defend its use?

We must approach this by remembering that it is an inevitable consequence of the structure of MCQ testing that random answering of the alternatives to a large number of questions will result in some being answered correctly by chance alone. For example, if a spin of a coin were to dictate the answers to 1000 true/false questions, one can expect that close to one half of the questions would be answered correctly, and one half incorrectly. If one mark were to be given for a correct response, and zero for an incorrect one, the final score would be 500/1000, or 50%, which is a pass mark! Clearly, this is undesirable.

Return to Contents of this Section

4.2 Applying negative marking to each question

We can solve this problem by awarding 1 mark for a correct response, and deducting 1 mark for an incorrect response. In this way, a totally prepared candidate will achieve 100%, a totally unprepared candidate 0%, and the other candidates will achieve a percentage score that should be a reflection of their degree of preparation - as far as any test can. This, after all is surely the prime goal of any test.

The above example deals with a question with two alternatives. Let us look at a question with three alternatives.

If the question has three alternatives, it would be unfair to the candidates to award 1 mark for the correct alternative, and deduct 1 mark for each incorrect alternative. Rather, the total marks of the incorrect alternatives should off-set the marks awarded for a correct response. Therefore, if the correct answer is awarded 1 mark, then 0.5 marks should be deducted for each incorrect alternative. (Instead of working with half marks, it is probably easier to work

with whole marks, where the answer is awarded 2 marks, and 1 mark is deducted from each incorrect alternative).

We apply the same logic to a question with four alternatives - one correct and three incorrect. In this case, the value of the correct alternative should be 3 marks, with 1 mark deducted for each incorrect alternative. Here, the value of using whole marks becomes apparent. This is the system we will apply from here.

Instead of using words, we can apply a few symbols to make the explanation easier to follow:

The total number of alternatives offered in each question, will be denoted by the letter "n". Because the number of correct alternatives in each question is always 1, the total number of **incorrect** alternatives will be denoted by "n-1".

The number of marks awarded for a correct response will be denoted by the letter "C".

The number of marks deducted for each incorrect alternative will be denoted by the letter "I".

As we increase our alternatives, a pattern develops, and can be laid out as in this table:

1	2	3	4	5	6
Total number of alternatives (n)	Number of correct alternatives (C)	Mark for correct alternative	Number of incorrect alternatives (n-1)	Deduction for each incorrect alternative	Tot. deductions for incorrect alternatives
2	1	1	1	1	1
3	1	2	2	1	2
4	1	3	3	1	3
5	1	4	4	1	4

By looking at Column 3 and Column 6, we can see that in all cases, the mark given to the correct alternative will be off-set by the total marks given to the incorrect alternatives; this is achieved by ensuring that the mark awarded to each incorrect alternative is adjusted as the number of incorrect alternatives is increased.

Finally, we calculate the value of I (Column 5) in the following way:

$$I = \frac{\text{Mark awarded for correct answer (Column 3)}}{\text{Number of incorrect alternatives (Column 4)}}$$

Using the notation above, this translates into

$$I = \frac{C}{n - 1}$$

written more easily as $I = C/(n-1)$

4.3 Applying negative marking to the entire test

We may wish to allow each question to count four marks, and have no negative marking on each question, but apply the correction to the final result of the test. Even if the method outlined in 4.2 is adopted, it is useful to read through the rest of 4.3, 4.4 and 4.5. While the amount adjusted is exactly the same as it would have been for individual questions, it will have the added advantage of easily illustrating that when the candidate has taken an educated guess at a question's answer, by eliminating one or more possible alternatives, a higher score results than if he/she had skipped that question.

We have seen that the totally unprepared candidate has a chance of scoring a number of marks unfairly - that is, due to simply guessing. This mark could just as easily have been obtained by a monkey by randomly guessing, and is often referred to as the "monkey score" (Koeslag et al. 1983) (and why MCQ tests are often disparagingly called "monkey puzzles"), or the "undeserved score." As we have seen, in a test consisting of questions with only two alternatives, this monkey score is as high as 50%. In a test consisting of more alternatives, this monkey score is reduced. In fact, if we look at the table and formula above, we will realise that the total monkey score is simply the number of questions multiplied by I.

Therefore, we see that if negative marking is NOT applied, the final test score of a candidate consists of

Score = Deserved score + undeserved score

In order to correct this, we introduce a correction in the same way as we introduced a correction I above, so that the score now is

Score = Deserved score + undeserved score - I

where I is again calculated as $I = C/(n-1)$

Let us look at how this is applied to a final test mark.

More typically, students will answer some questions correctly because they actually know or can work out the correct answer, and some questions correctly on account of some "partial knowledge", which can enable them to guess more judiciously that would be the case of pure random guesswork. They will score higher than the "monkey score" on a number of questions, and this represents a partial score awarded for partial knowledge.

The following example will make this clear

A student takes a test where 100 MCQs with 5 responses are presented. A correct answer attracts 4 marks, and an incorrect answer -1 mark. The student answers 50 questions correctly by deliberately selecting the correct answer as a result of applying knowledge (the "deserved score"), 25 questions by random guesswork (leading to an "undeserved score" which will be corrected by the application of negative marking) and 25 questions by first eliminating 2 out of 5 responses and then guessing on the other three (leading to a "partial score"). Of these, assume that 8 are correctly and 17 incorrectly answered. The student's final score will be:

Deserved	j	= 200 marks
Undeserved	5 x 4 - 20 x 1	= 0 marks
Partial Score	8 x 4 - 17 x 1	= 15 marks
Total		= 215 marks

At this point, however, we should note that it is essential that the test questions should have the identical number of alternatives.

The 15 marks can be regarded as a bonus for partial knowledge. Note that if the student had abstained from answering all the questions where he or she harboured any doubts as to the correct answer, this bonus would have been thrown away.

If negative marking had not been in effect, (by awarding 0 for an incorrect answer), the student would have obtained 200 + 20 + 32 = 252 marks.

Two points must be emphasized:
- The correction should **NEVER BE GREATER THAN C/(n-1);**
- It is (contrary to popular belief) in the student's advantage to answer all questions (Harden *et al*, 1976), since students are not totally ignorant of the subject matter, and they will tend thus to score higher than the monkey score on those questions where they make "informed" guesses.

The fact that the student receives a statistical credit for partial knowledge by being able to eliminate what to him or her seems to be obviously incorrect answers is perfectly fair. This is the equivalent of "educated waffling" in an essay, which scores some marks, but not the full number of possible marks.

Return to Contents of this Section

4.4 The percentage score

Final scores are commonly expressed as a percentage. When negative marking is in effect, there are three ways of expressing the percentage (M is the "monkey score"):

- based on the range of available marks: in the above example, the student obtained a score of 215 in a scoring range stretching from -100 to +400.

```
                    M                     215
      |_____|_____|_____|_+____|_____|
    -100       0       100      200     300     400
```
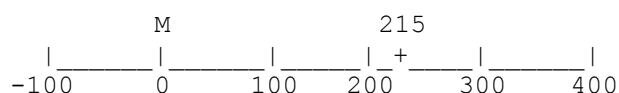
This is equivalent to a score of 315 in a range 0 to +500 marks:

```
                    M                     315
      |_____|_____|_____|_+____|_____|
      0       100      200      300     400     500
```
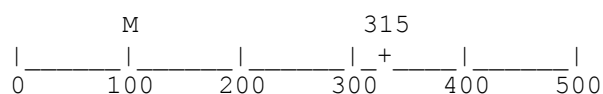
i.e., the student obtained 63.0% of the available marks

```
                   M                    63
     |_____|_____|_____|_+____|_____|
     0%       20%      40%      60%      80%     100%
```

- based on the maximum score only: here our student would have obtained 215 out of 400, or 53.8%. The advantage of this is that the "monkey score" of 0 marks is in fact 0%, which is not unreasonable. The disadvantage is that students who fare less well than our putative monkey, will get a negative percentage score, and this can play havoc with certain individuals' self-esteem.

```
                   M                   215
     |_____|_____|_____|_+____|_____|
     -100      0       100      200     300      400
     -25%      0%      25%      50%      75%     100%
```
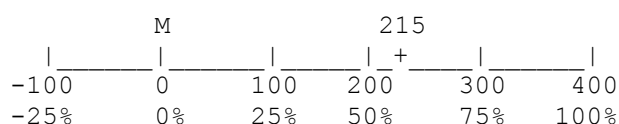
- based on the maximum score only, but assigning all scores less or equal to 0 a mark of 0. Here our monkey will get 0, and our student with 215 marks still gets 53.8%.

```
                   M                   215
     |_____|_____|_____|_+____|_____|
     0         0       100      200     300      400
     0%        0%      25%      50%      75%     100%
```
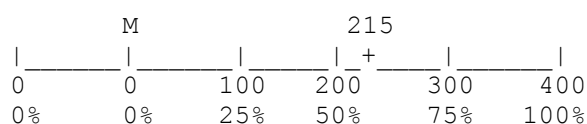
Which method will actually be used should be carefully thought out before administering the test, and students should be informed as to the method of scoring and its implications. They should also be cautioned against skipping questions.

Return to Contents of this Section

4.5 Adjustment of final percentages

If negative marking is not used, the random guessing factor will result in inadequate students scoring relatively more than their knowledge should allow. Hence it is common practice to make some adjustment to the final percentage scores.

If F% is the final corrected percentage score, and R% the uncorrected percentage score, then
$$F\% = (nR\% - 100)/(n-1)$$
where n is the number of responses in a test consisting of MCQs with a uniform number of responses and scoring.

From the above, we see that a raw score of 75% in a True/False test, where 1 mark is awarded for a correct answer and 0 for an incorrect answer is equivalent to a corrected score of 50%.

This impacts on the "pass mark" which must be achieved in MCQ tests. If the final pass mark is held to be 50%, then the uncorrected scores which are equivalent to a 50% pass mark will be:

- 75% in a test consisting of true/false questions;
- 67% in a test consisting of MCQs with 3 responses;
- 63% in a test consisting of MCQs with 4 responses;
- 60% in a test consisting of MCQs with 5 responses.

Note that other criteria, such as the minimum score for "fist class", "upper second" and so on should also be adjusted.

Return to Contents of this Section

4.6 Question attributes

The quality of a testing instrument, be it an essay or collection of MCQs, should be a matter of great concern to all involved. In the first place, students should feel that the test has been "fair", both in terms of the type of questions that they were asked to answer, and that the marking of their scripts should be as objective and consistent as possible. Secondly, the lecturers should monitor feedback from the test in order to ensure that the test was of a high quality and that the results reflect the real ability of the students.
Statistical treatment of the results obtained by a class will provide a window on the quality of a test consisting of MCQs.

4.6.1 How difficult/easy is a question?

If everyone in a class were to get a particular question right, it is clear that that question might have been too easy. Conversely, if everyone in the class had answered it incorrectly, the question might have been too difficult. The facility question is an attribute of that question which can be measured in terms of the performance of the class on that question. It is simply calculated by

  facility = number of correct answers/total answers

For example, if 100 students attempted the question, and 60 students entered a correct answer for it, the facility will be $60/100 = 0.6$. (Sometimes, the difficulty of a particular question is used. It is simply 1 - facility, or 0.4 in this case). The facility is sometimes expressed as a percentage, i.e. the percentage of those students answering the question who gave the correct answer, in this case 60%.

Questions which are too easy or too difficult should not be included in a test unless there are very good reasons to do so. Exceptions would be in order to identify students who lack certain "entry-level" knowledge to a course, or for some other diagnostic purposes. In general, one should aim for questions with a facility in the range 0.3 to 0.7. Some of the key reasons for this are discussed below.

4.6.2 The discrimination of a question

One fatal inadequacy of questions with excessively low or high facilities is that they do not differentiate between students of widely differing abilities. The manner in which a question is answered by students of different abilities is called the discrimination or discriminating power of the question. It may be calculated in different ways, one of them based on the

performance of two student groups of equal size, one group being of "low performance" and the other group being of "high performance":

$$\text{discrimination} = (Ch - Cl)/Nh$$

where Ch is the number of correct answers returned by the high performing group, and Cl the number of correct answers returned by the low performing group, and Nh is the number of students in the high (or low) performing group.

In common practice, the groups are defined as the **upper and lower quartiles** of the class. These quartiles are obtained by ordering the test scores from lowest to highest, and separating these scores into four parts. The part scoring the least marks is the **lower quartile** and the part scoring the highest marks is the **upper quartile**.

(Never mind if the class is not divisible by 4, the computer will do the necessary calculations and adjustments! Statisticians will be quick to recognize that this method is only approximate, and that there are more reliable methods. (Ward, p. 215-218). It is given here for the purpose of information only.

Suppose we have a class of 158 students, divided into quartiles with 40 students each in the lower and upper quartiles. If 35 students in the upper quartile and 17 students in the lower quartile answered the question correctly, then the discrimination will be $(35 - 17)/40 = 0.45$.

This question has a satisfactory discrimination.

Suppose that the situation is reversed. 17 students in the upper quartile and 35 students in the lower quartile answered correctly. The discrimination is now negative, indicating that students who did poorly o**n the test as a whole** fared better on the question than those who did well in the test! This question should be eliminated from the test and looked at carefully. Perhaps it was poorly formulated, or the correct response flagged as incorrect.

The minimum acceptable discrimination level depends *inter alia*, on the number of students taking the test, but can be taken as 0.3 for practical purposes.

The above discussion is far from being exhaustive. If you would like to know more, consult the references which are listed in Appendix A of this document.

Return to Contents of this Section

5. Appendix A: Where to get Help

Contents of This Section

5.1 A1. Books and articles

- Airasian, P. W. (1994) *Classroom assessment*, 2nd Ed McGraw- Hill
- Bak, N, (1990) "How to test insight and understanding of philosophical issues by means of multiple-choice questions." *S.A. J.Ed.*, 10, pp. 103-108.
- Bloom B S. (Ed.) (1989) *Taxonomy of educational objectives. Handbook I: cognitive domain* (New York, McKay)
- Delpierre G R (1991) "The degradation of higher levels of the cognitive domain and its implication for the design of computer-based question episodes" *Studies in Higher Education*, 16, pp. 63-71.
- Gronlund, N. E. (1982) *Constructing achievement tests*, 3rd Ed. Prentice- Hall.
- Harden R McG, Brown R A, Biran L A, Dallas Ross, W P and Wakeford R E (1976), "Multiple choice questions: to guess or not to guess." *Med. Educ.*, 10, 27-32.
- Koeslag, J., C. W. Melzer, and S. R. Schach, "Penalties in Multiple-Choice and True-False Questions," *S. Afr. Med. J.* 63 (1983), pp. 20-22.
- Mehrens W A and Lehman I J (1980) S*tandardized tests in education*, Holt Rhinehart and Winston
- Mehrens W A (1991) *Measurement and evaluation in education and psychology*, Harcourt Brace Jovanovitch
- Rust, W. B. (1973) *Objective testing in education and training*, Pitman Educational Library.
- Ward. C. (1981) *Preparing and using objective questions*, Stanley Thornes.

Return to Contents of this Appendix

5.2 A2. Internet

- APA's "Code of Fair Testing Practices in Education."
- Buros Center for Testing
- ERIC Clearinghouse on Assessment and Evaluation
- Computer Aided Assessment (Hertfordshire University, Learning and Information Services)
- National Board of Medical Examiners (U.S.A.) has a useful pdf on Constructing Written Test Questions for the Basic and Clinical Sciences

Return to Contents of this Appendix

6. Appendix B: Designing MCQs - Dos and don'ts

Contents of This Section

6.1 B1. General considerations

The task of designing good MCQs may be simplified by referring to past student work such as tutorials, class tests and examination answers, as these are rich sources of the type of errors and misconceptions which students frequently expose. MCQs designed with this material at hand will be relevant to the subject area and will be taken seriously by the students.

- **Test for significant learning outcomes.**
  The questions should be designed to test the learning objectives of the course, and not trivia associated with the subject matter. Questions should be recognized as being relevant to the goals of the course.

- **Test for the intended intellectual skills.**
  A question may unintentionally be a "test within a test." Look at the following stem:
  `"How many permutations are possible in a bridge hand?"`
  A student who knows how to calculate permutations and combinations (which is assumed to be the skill which is tested for in the context of an elementary Statistics course) will not be able to answer this question if he/she has never played cards, let alone bridge.

- **Pay special attention to the language used.**
  The level of the language should be within reach of the students, bearing in mind that there may be many in the class whose home language is not English. Use correct grammar throughout and avoid the use of jargon, unless you are specifically testing terminology. Remember that "second-language" students will take longer to read and understand a question. This, of course, is true for other types of examinations, in particular, those based on essay-type questions, where mis-reading the question by the student may lead to the writing of a good essay totally off the point envisaged by the examiner.

- **Be sensitive to cultural and gender issues.**
  Avoid turns of phrase and figures of speech that could reasonably be construed as racist or sexist, or which may have a cultural bias.

- **Avoid giving clues.**
  The correct answer may sometimes be deduced by "test-wise" students from unintentional clues linking the stem to the answer:

```
A group of islands is called an

   a. peninsula
   b. archipelago
   c. moraine
   d. polder
```

Here, there is a grammatical clue linking the stem to the correct answer b) in that the "an" at the end of the stem is the only article acceptable for the word "archipelago". Rather end the stem with "a/an".

```
Which one of the following publications is considered to be a prime
source for research articles on achievement testing?

    a. Journal of Educational Psychology
    b. Journal of Educational Measurement
    c. Journal of Clinical Psychology
    d. Review of Educational Research
```

The correct answer d) is the only alternative to include the word "research", which is also found in the stem. In addition, it differs from the distractors in that it is not a "Journal," but a "Review."

- **Avoid using interrelated items.**
  Questions should not overlap.

6.2 B2. The question stem

- State the problem clearly, simply and as concretely as possible. Avoid vague generalizations and do not include irrelevant information. It is essential that the students should know exactly what is expected of them.

- The stem should contain most of the wording in order to reduce the reading load.

- Avoid using negative stems. If this cannot be done, the negative words should always be highlighted by underlining or capitalisation:

  ```
  "Which one of the following is NOT an inert gas?"
  ```

6.3 B3. The alternatives

- They should be equally plausible and they should be internally consistent.

  ```
  The internal friction due to molecular cohesion in fluids is called

  a. surface tension
  b. viscosity
  c. fluidity
  d. ductility
  e. absorbance
  ```

Alternative "d" has nothing to do with fluids and so can be eliminated by a student who is otherwise uninformed.

- Only one of them should be unambiguously correct. Critical examination by a colleague is essential.
- Try and ensure that they are approximately the same length.

- Avoid "All of the above" (Analysis shows that these questions tend to be the easiest. This response shows that the question designer has run out of ideas).

- Use "None of the above" with care (Analysis shows that these questions are more difficult. On the plus side, they reduce the effect of guessing, but give no diagnostics on misconceptions).

- The position of the answer should vary.

Return to Contents of this Appendix

Appendix C: Bloom's Taxonomy

Contents of This Section

---

7. C1. Bloom's Taxonomy

Following the 1948 Convention of the American Psychological Association, B S Bloom took a lead in formulating a classification of "the goals of the educational process." Three "domains" of educational activities were identified. The first of these, named the Cognitive Domain, involves knowledge and the development of intellectual attitudes and skills. (The other domains are the Affective Domain and the Psychomotor Domain, and need not concern us here).

Eventually, Bloom and his co-workers established a hierarchy of educational objectives, which is generally referred to as **Bloom's Taxonomy**, and which attempts to divide cognitive objectives into subdivisions ranging from the simplest behaviour to the most complex. It is important to realise that the divisions outlined above are not absolutes and that other systems or hierarchies have been devised. However, Bloom's taxonomy is easily understood and widely applied.

7.1 C1.1. Knowledge

Knowledge is defined as the remembering of previously learned material. This may involve the recall of a wide range of material, from specific facts to complete theories, but all that is required is the bringing to mind of the appropriate information. Knowledge represents the lowest level of learning outcomes in the cognitive domain.

Examples of learning objectives at this level are: know common terms, know specific facts, know methods and procedures, know basic concepts, know principles.

7.2 C1.2. Comprehension

Comprehension is defined as the ability to grasp the meaning of material. This may be shown by translating material from one form to another (words to numbers), by interpreting material (explaining or summarizing), and by estimating future trends (predicting consequences or effects). These learning outcomes go one step beyond the simple remembering of material, and represent the lowest level of understanding.

Examples of learning objectives at this level are: understand facts and principles, interpret verbal material, interpret charts and graphs, translate verbal material to mathematical formulae, estimate the future consequences implied in data, justify methods and procedures.

7.3 C1.3. Application

Application refers to the ability to use learned material in new and concrete situations. This may include the application of such things as rules, methods, concepts, principles, laws, and theories. Learning outcomes in this area require a higher level of understanding than those under comprehension.

Examples of learning objectives at this level are: apply concepts and principles to new situations, apply laws and theories to practical situations, solve mathematical problems, construct graphs and charts, and demonstrate the correct usage of a method or procedure.

7.4 C1.4. Analysis

Analysis refers to the ability to break down material into its component parts so that its organizational structure may be understood. This may include the identification of parts, analysis of the relationship between parts, and recognition of the organizational principles involved. Learning outcomes here represent a higher intellectual level than comprehension and application because they require an understanding of both the content and the structural form of the material.

Examples of learning objectives at this level are: recognize unstated assumptions, recognises logical fallacies in reasoning, distinguish between facts and inferences, evaluate the relevancy of data, analyse the organizational structure of a work (art, music and writing).

7.5 C1.5. Synthesis

Synthesis refers to the ability to put parts together to form a new whole. This may involve the production of a unique communication (theme or speech), a plan of operations (research proposal), or a set of abstract relations (scheme for classifying information). Learning outcomes in this area stress creative behaviours, with major emphasis on the formulation of new patterns or structure.

Examples of learning objectives at this level are: write a well-organized theme, gives a well-organized speech writes a creative short story (or poem or music), propose a plan for an experiment, integrate learning from different areas into a plan for solving a problem, formulates a new scheme for classifying objects (or events, or ideas).

7.6 C1.6. Evaluation

Evaluation is concerned with the ability to judge the value of material (statement, novel, poem, research report) for a given purpose. The judgments are to be based on definite criteria. These may be internal criteria (organization) or external criteria (relevance to the purpose) and the student may determine the criteria or be given them. Learning outcomes in this area are highest in the cognitive hierarchy because they contain elements of all the other categories, plus conscious value judgments based on clearly defined criteria.

Examples of learning objectives at this level are: judge the logical consistency of written material, judge the adequacy with which conclusions are supported by data, judge the value of a work (art, music, writing) by the use of internal criteria, judge the value of a work (art, music, writing) by use of external standards of excellence.

There is a great deal of information on Bloom's taxonomy on the Web.   A starting point is:
- https://www.unlv.edu/sites/default/files/page_files/3/BloomsTaxonomy.pdf

Return to Contents of this Appendix

---

8.   C2 Application of Bloom's Taxonomy to the design of MCQs

8.1 C2.1. Level 1: Knowledge

At this level, one simply requires the recall of acquired knowledge. *Warning*: A test at this level can easily become an examination of trivial information.

Example C2.1.1

```
Which one of the following persons is the author of "Das Kapital"?

 a. Mannheim
 b. Marx
 c. Weber
 d. Engels
 e. Michels
```

Note that the responses are internally consistent - they are all the names of Germans whose written work have been major contributions on social issues.

Example C2.1.2

```
In the area of physical science, which one of the following definitions
describes the term "polarization"?

 a. The separation of electric charges by friction.
 b. The ionization of atoms by high temperatures.
 c. The interference of sound waves in a closed chamber.
 d. The excitation of electrons by high frequency light.
 e. The vibration of transverse waves in a single plane.
```

Simple recall of the correct definition of polarization ("e") is required. Internal consistency and plausibility are maintained in that all responses are actual physical phenomena.

Example C2.1.3

```
According to the microgenesis of perception concept, the threshold of
awareness consists of a hierarchy of thresholds.  Which one of the
sequences shown below is correct?

 a. Recognition thresholds > physiological thresholds > detection
    thresholds.
 b. Physiological thresholds > detection thresholds > recognition
    thresholds.
 c. Physiological thresholds > recognition thresholds > detection
    thresholds.
```

```
d. Recognition thresholds > detection thresholds > physiological
   thresholds.
```

In this example, nothing more is required than the recall of the order of certain pieces of related information. The correct answer is "b."

8.2 C2.2. Level 2: Comprehension

At this level, knowledge of facts, theories, procedures etc. is assumed, and one tests for understanding of this knowledge.

Example C2.2.1

```
Which one of the following describes what takes place in the so-called
PREPARATION stage of the creative process, as applied to the solution of a
particular problem?

a. The problem is identified and defined.
b. All available information about the problem is collected.
c. An attempt is made to see if the proposed solution to the problem is
   acceptable.
d. The person goes through some experience leading to a general idea of
   how the problem can be solved.
e. The person sets the problem aside, and gets involved with some other
   unrelated activity.
```

In this question, the knowledge of the five stages of the creative process must be recalled (KNOWLEDGE), and one is tested for an understanding (COMPREHENSION) of the meaning of each term, in this case, "preparation."

Note that this question violates the guideline that the answer and distractors should all be of about the same length. It is difficult to get around this one here, so the text is edited so that each line is about the same length.

8.3 C2.3. Level 3: Application

In order to classify a question into this group, ask yourself if prior knowledge of the background to the question is assumed to be both known and understood, and whether one is merely expected to apply this knowledge and understanding. Calculations based on known formulae are good examples of this, as shown in the example below:

Example C2.3.1

```
Which one of the following values approximates best to the volume of a
sphere with radius 5m?

a. 2000m³
b. 1000m³
c. 500m³
d. 250m³
e. 125m³
```

In order to answer this question, the formula $4\pi r^3 /3$ must be known (recall of knowledge) and the meaning of the various symbols in the formula understood (comprehension) in order to answer this question. The correct answer is "c."

Example C2.3.2

```
Which one of the following memory systems does a piano-tuner mainly use in
his occupation?

a. Echoic memory.
b. Short-term memory.
c. Long-term memory.
d. Mono-auditory memory.
e. None of the above.
```

This is clearly a case of testing for the application of previously acquired knowledge (the various memory systems), which is also understood, as the meaning of each term must be clear before the student can decide whether it is applicable to the given situation. The correct answer is "a." Note that students may not necessarily know what a piano-tuner is or does. Watch out for cultural bias!

The next example is more difficult to classify:

Example C2.3.3

```
You are the sole owner and manager of a small enterprise employing 15
workers.  One of these, Alfred, (who has been working for you for the past
year and has somewhat of a history of absenteeism), arrives late for work
one Wednesday morning, noticeably intoxicated.  Which one of the following
actions is the most appropriate in the circumstances?

a. You terminate Alfred's employment on the spot, paying him the wages
   still due to him.
b. You parade Alfred in front of the other workers, to teach them all a
   lesson.
c. You give Alfred three weeks' wages in lieu of notice, and sack him.
d. You wait until Alfred is sober, discuss his problem, and give him a
   final written warning, should it be required.
e. You call Alfred's wife to take him home and warn her that this must not
   happen again.
```

Note that this this question is classified as APPLICATION as in order to answer it, the relevant labour legislation should be known and understood. One could made a case for it to have a higher classification such as EVALUATION, on the grounds that one is asked to evaluate which one of the proposed actions is the best in the circumstances, or ANALYSIS, on the grounds that in order to select the most appropriate answer, one should analyse the possible outcomes of each decision. For both these levels, one would expect a greater amount of information as to Alfred's situation, the relationship between Alfred and his co-workers, union involvement in the enterprise etc., and have a more sophisticated set of distractors. Here, option "d" is clearly the best both on legal and human terms. Note that the figure of speech "on the spot" may not be understood by second- language students. Use suitable language.

8.4 C2.4. Level 4: Analysis

Example C2.4.1

"The story is told of the famous German Organic Chemist Auguste Kékulé who was struggling with the problem of how the six carbon atoms of benzene were linked together. He was getting nowhere with the problem, and one day fell asleep in front of the fireplace while he was pondering on it. He dreamt of molecules twisting and turning around like snakes. Suddenly, one of the snakes swallowed its own tail and rolled around like a hoop. Kékulé woke up with a start, and realized that his problem could be solved if the six carbon atoms of benzene were attached to each other to form a ring. Further work showed that this was entirely correct."

The above passage illustrates a particular phase of the creative process. Which one is it?

```
a. preparation
b. incubation
c. orientation
d. illumination
e. verification
```

In the above example, the student is expected to know and understand the five stages of the creative process, and to apply this knowledge to an important factual example of creative thinking (the elucidation of the chemical structure of the benzene molecule). The ability to analyse the data (i.e. the given text) in terms of each of the five stages is what is being tested. The correct answer, by the way, is "d."

Example C2.4.2 (Assume the question below is asked in a philosophy test.)

Read carefully through the paragraph below, and decide which of the options a-e is correct.

"The basic premise of pragmatism is that questions posed by speculative metaphysical propositions can often be answered by determining what the practical consequences of the acceptance of a particular metaphysical proposition are in this life. Practical consequences are taken as the criterion for assessing the relevance of all statements or ideas about truth, norm and hope."

```
a. The word "acceptance" should be replaced by "rejection".
b. The word "often" should be replaced by "only".
c. The word "speculative" should be replaced by "hypothetical".
d. The word "criterion" should be replaced by "measure".
```

This question requires prior knowledge of and understanding about the concept of pragmatism. The paragraph, seen in this light, contains one word which vitiates its validity, and the student is tested on his/her ability to **analyze** it to see whether it fits with the accepted definition of pragmatism. With this in mind, b is correct. Option "a" would degrade the paragraph further, while c and d would simply result in changing to acceptable synonyms. Note that this question does not address Level 6 (Evaluation), as one is not asked to pass a value judgement on the text. This must be considered as a very difficult question, and will obviously require a high level of reading skills. Bear in mind that there will be a significant time factor involved.

Example C2.4.3

```
Look at the following table and indicate which countries'statistics are
being reported in rows A, B and C.
```

	GNP per capita 1991 ($ USA)	Growth rate of GNP per capita p.a. 1980-91	Population growth rate 1980-91	Structures of total employment 1980-85 (percentages)		
				Agriculture	Industry	Services
A	500	2,5%	1,5%	51	20	29
B	1570	5,8%	1,6%	74	8	8
S.A.	2560	0,7%	2,5%	17	36	36
C	25110	1,7%	0,3%	6	32	32

```
Choose your answer from the following list of possible
answers:

    1.    A is South Korea; B is Kenya; C is Canada.
    2.    A is Sri Lanka; B is Germany; C is Thailand.
    3.    A is Sri Lanka; B is Thailand; C is Sweden.
    4.    A is Namibia; B is Portugal; C is Botswana.
```

In order to answer this question, students must be able to recall the relative economic rankings of various countries (KNOWLEDGE) and understand the basis for such a ranking (COMPREHENSION). They must be able to apply these concepts when information is supplied to them (APPLICATION), and they must be able to ANALYZE the given information in order to answer the question. Students did not like this question when they were faced with it in a class test, as their immediate reaction was that "it was impossible to remember the statistics for all the countries that were discussed in class and given to them in handouts". They were surprised when told that such detailed knowledge was in fact not expected of them, but that they were to examine the table and perform a ranking on the basis of concepts that they should have mastered. The correct answer is 3.

8.5 C2.5. Level 6: Evaluation

At this level, one is asked to pass judgement on, for example, the logical consistency of written material, the validity of experimental procedures or interpretation of data.

Example C2.5.1

```
A student was asked the following question: "Briefly list and explain the
various stages of the creative process".

As an answer, this student wrote the following:

"The creative process is believed to take place in five stages, in the
following order: ORIENTATION, when the problem must be identified and
defined, PREPARATION, when all the possible information about the problem
is collected, INCUBATION, when there is a period where no solution seems in
sight and the person is often busy with other tasks, ILLUMINATION, when the
person experiences a general idea of how to arrive at a solution to the
problem, and finally VERIFICATION, when the person determines whether the
solution is the right one for the problem."

How would you judge this student's answer?
```

```
a. EXCELLENT (all stages correct in the right order with clear and correct
   explanations).
b. GOOD (all stages correct in the right order, but the explanations are
   not as clear as they should be).
c. MEDIOCRE (one or two stages are missing OR the stages are in the wrong
   order, OR the explanations are not clear OR the explanations are
   irrelevant).
d. UNACCEPTABLE (more than two stages are missing AND the order is
   incorrect AND the explanations are not clear AND/OR they are
   irrelevant)
```

In the above question, one is expected to make value judgment on the content of the given text (KNOWLEDGE of the subject is required), the meaning of the terminology used (COMPREHENSION of the subject matter), and its structure (ANALYSIS of the answer for the right order of events. The correct answer here is "a", but suitable modification of the putative student answer could provide a small bank of questions with other correct answers.

Example C2.5.2

Another example is the "Assertion/Reason" question, in which two statements linked by "BECAUSE" have to be evaluated in the light of certain criteria:

```
Judge the sentence in italics according to the criteria given below:

"The United States took part in the Gulf War against Iraq BECAUSE of the
lack of civil liberties imposed on the Kurds by Saddam Hussein's regime."


a. The assertion and the reason are both correct, and the reason is valid.
b. The assertion and the reason are both correct, but the reason is
   invalid.
c. The assertion is correct but the reason is incorrect.
d. The assertion is incorrect but the reason is correct.
e. Both the assertion and the reason are incorrect.
```

The correct answer is "b", since, while it is true that the United States took part in the Gulf War, and it is also true that the Kurds in Iraq did not enjoy an abundance of civil liberties, but the threat to the US's oil supply as a result of Iraq's invasion of Kuwait was a much more pertinent reason for the United States joining in the fray. A knowledge and understanding of Middle East politics is assumed. What is tested here is the ability to evaluate the between cause and effect in the sentence in terms of predefined criteria.