Multiple Choice Questions Not Considered Harmful

Karyn Woodford, Peter Bancroft
School of Software Engineering and Data Communications
Queensland University of Technology
George Street, Brisbane QLD 4000
k.woodford@qut.edu.au, p.bancroft@qut.edu.au

Abstract

Increasingly, academics are confronted with issues associated with assessment in large classes, arising from a combination of factors including higher student enrolments and the introduction of a trimester of study in many universities. The resulting increased time pressures on marking are causing many academics to search for alternative forms of assessment. University teachers are making more frequent use of multiple choice questions as a matter of expediency and in some cases, the quality of the assessment is being neglected. This describes the current situation in Information Technology. The aim of this paper is to provide practical guidelines in the form of a checklist for lecturers who wish to write tests containing multiple choice questions. Some of the points raised may be considered common knowledge for those teachers with a background in Education, however not all Information Technology lecturers would fall into this category. While the intended users of the checklist are Information Technology lecturers who, in general, are unlikely to be familiar with many of the matters discussed, teachers in other disciplines may find it a useful reference. In addition to the checklist, this paper also discusses the major criticism of multiple choice questions (that they do not test anything more than just straight recall of facts) and examines ways of overcoming this misconception.

Keywords: multiple choice questions, assessment, Bloom, large class assessment.

1 Introduction

In Information Technology faculties, even with recent downturns in enrolments, teachers of many first year units are faced with classes in excess of 250 students. Whilst marking each assessment item for this number of students takes a considerable amount of time, there is usually significantly more pressure during end of semester examination marking. This is due to a number of factors including the required turnaround time being shorter for examination marking, and the need for consistency in marking (usually resulting in one marker having to complete a particular question for every student).

In other disciplines, alternative testing techniques such as multiple choice questions have long been used to help alleviate these problems and there is an awareness of the extensive body of research in the area. As lecturers in Information Technology are now making widespread use of multiple choice questions (Lister 2000, Lister 2001, Carter, Ala-Mutka, Fuller, Dick, English, Fone and Sheard 2003), this paper will utilise this extensive body of research to provide some practical guidelines to assist in the construction of well written questions in an IT education environment.

In Section 2.1 we review the terminology used to describe multiple choice questions and then in Section 2.2 we suggest methods for measuring the effectiveness of this type of question. In Section 3 we discuss a range of issues that should be considered when composing questions, including the grammar and wording, the optimal number of options, ordering, and commonly used options such as "none of the above" and "all of the above". These issues will be discussed in some detail, and will be summarised in the form of a handy checklist in Appendix A.

Multiple choice questions face criticism due to the belief that they do not test anything deeper than a superficial memorising of facts. However we contend that it is possible to construct multiple choice questions that are able to test higher levels of cognition. The following diagram represents the levels within the cognitive domain as identified by Bloom (1956). The simple recall of facts is at the lowest level, increasing to the evaluation skills at the top.

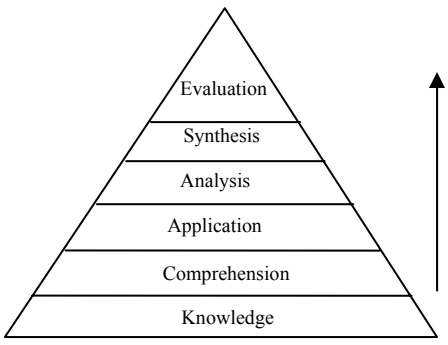


Figure 1: Bloom's levels of cognition

In section 4 of the paper we address the problem of how multiple choice questions can test more than just straight recall of facts. Specifically we discuss the comprehension, application and analysis levels of cognition, and give examples of multiple choice questions to test students at these levels.

In conclusion, it is our aim to provide a practical checklist to assist IT teachers who want to set multiple choice questions while maintaining the integrity of their assessment. Teachers in other disciplines who have experience with this kind of testing may also find the checklist to be a useful reference.

2 Writing effective multiple choice questions

2.1 The parts of a multiple choice question

Common terminology (Isaacs 1994) for describing the separate parts of a multiple choice question is illustrated in the following example:

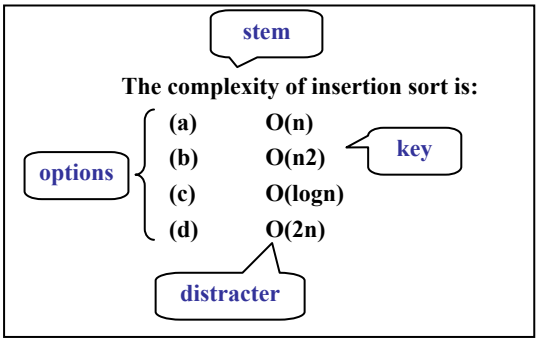


Figure 2: The parts of a multiple choice question

A single multiple choice question, such as the one above, is known as an *item*. The *stem* is the text that states the question, in this case "The complexity of insertion sort is". The possible answers (correct answer plus incorrect answers) are called *options*. The correct answer (in this case b) is called the *key*, whilst the incorrect answers (a, c and d) are called *distracters*.

2.2 What is an effective question?

A simple measure of the effectiveness of a question is provided by the distribution of student responses amongst the options. If too many students select the correct answer, then perhaps the distracters are not convincing. If very few students answer correctly, then the question may not be clear or a deliberately misleading distracter may have been used. The proportion of students answering a question correctly is called its *facility*. Whilst there are no hard and fast rules about an item's facility, it may be appropriate to have a range somewhere between 0.4 and 0.6 when the goal of the test is to rank students in order. However for examinations aiming to test whether a student can answer a question, a facility of 0.8 or higher may be appropriate (Isaacs 1994).

Item discrimination provides another way of assessing a question's effectiveness. Rather than simply looking at the proportion of students who answer the question correctly, discrimination is concerned with whether a student with a high total on the exam is more likely to get the question correct than a student who does poorly on the exam. As a rough guide, the higher the discrimination index, the better. A discrimination index of 0.40 or over indicates that an item is very well written, while a score of below 0.20 indicates that the item is probably quite poor. Questions in this range are either so difficult that almost no students answer correctly, or so easy that almost all students will be correct – thus the item fails to discriminate students who score well from those who do not (Isaacs 1994).

Another measure of a question's effectiveness is whether the question tests the desired level of cognition (as described in Figure 1 above).

2.3 Limitations of multiple choice questions

The traditional style of multiple choice questions - a simple stem question with a key and distracters - has its limitations. A student may select the correct answer by knowing that answer is correct or by eliminating all of the other options. While this may initially seem desirable, it does not necessarily test the students' full knowledge of the subject - knowing one option is correct doesn't guarantee they know that the others are incorrect. Similarly, working out the correct answer by a process of elimination doesn't demonstrate that the student necessarily knows the solution - if faced with that single answer in a true / false environment, they may not have known that it was correct. This limitation may be overcome by using a different style of multiple choice question (see section 3.4).

3 Factors affecting the validity of multiple choice questions

When writing good multiple choice questions there are several factors to consider - some relate to the actual question whilst some relate to the options (key and distracters).

3.1 Correct grammar and wording

The use of incorrect grammar in the stem of a question can often allow students to exclude an option immediately. Consider the following question.

> When outputting binary data to a file, the primitive type that uses 8 bytes of storage is a
> (a) char.
> (b) int.
> (c) double.

A test-wise student may identify option (b) as being incorrect as it starts with a vowel and the stem ends with "a" and not "an". (Similarly, if b was the correct answer, the incorrect grammar may mislead students). To avoid this, the options should include the article:

> When outputting binary data to a file, the primitive type that uses 8 bytes of storage is
> (a) a char.
> (b) an int.
> (c) a double.

There are several other grammatical considerations (Wilson and Coyle 1991):

- ensuring the stem and options are worded in the same tense;
- avoiding additional qualifying words or phrases to the key (a test-wise student will often identify

a longer, more precise answer as the correct option); and
- using similar wording in all options, particularly making sure that the key doesn't sound like it is directly from a text book.

3.2 Number of options

The number of options is one of the most fiercely debated issues amongst supporters of the multiple choice question. Strong arguments have been made for 3, 4 and 5 options. Those who argue for 5-option tests believe that 3- or even 4-option tests increase the probability of a student guessing the correct answer to an unacceptably high level. Those who argue for 3-option tests claim that their tests can be as effective as a 4- or 5-option test, as the additional distracters are likely to be less believable. The arguments for 3-option and 4- or 5-option tests are considered below, along with a brief discussion on removing non-functioning options. Once the number of desired options is decided, it is advisable to use this number of options for every item in the examination to reduce the possibility of careless mistakes.

3.2.1 Three Options

A well written multiple choice question with three options (one key and two distracters) can be at least as effective as a question with four options. According to Haladyna and Downing (1993) roughly two thirds of all multiple choice questions have just one or two effectively performing distracters. In their study they found that the percentage of questions with three effectively performing distracters ranged from 1.1% to 8.4%, and that in a 200 item test, where the questions had 5 options, there was not one question with four effectively performing distracters.

The argument for three options therefore is that the time taken to write a third and possibly a fourth distracter (to make a 4- or 5-option test) is not time well spent when those distracters will most likely be ineffective. In Sidick and Barrett (1994) it is suggested that if it takes 5 minutes to construct each distracter, removing the need for a third and fourth distracter will save ten minutes per question. Over 100 questions, this will save more than 16 hours of work. Supporters of 4- or 5-option tests would argue that any time saved would be negated by a decrease in test reliability and validity. However Bruno and Dirkzwager (1995) find that although reliability and validity are improved by increasing the number of alternatives per item, the improvement is only marginal for more than three options.

3.2.2 Four or five option

The most significant argument against three option multiple choice tests is that the chance of guessing the correct answer is 33%, as compared to 25% for 4-option and 20% for 5-option exams. It is argued that if effective distracters can be written, the overall benefit of the lower chance of guessing outweighs the extra time to construct more options. However a distracter is non-functioning (if less than 5% of students choose it) then that distracter is probably so implausible that it appeals only to those making random guesses (Haladyna and Downing 1993).

3.2.3 Removing non-functioning options

Removing a non-functioning distracter (i.e. an infrequently selected one) can improve the effectiveness of the test. In Cizek and O'Day (1994) a study of 32 multiple choice questions on two different papers was undertaken. One paper had 5-option items, whilst the other paper contained 4-option items, a non-functioning item from the identical 5-option item having been removed. The study concluded that when a non-functioning option was removed, the result was a slight, non-significant increase in item difficulty, and that the test with 4-option items was just as reliable when compared to the 5-option item test.

3.3 "Not" and the use of double negatives

Asking a student to select which option is not consistent with the stem can be an effective test of their understanding of material. However teachers should be very careful when using "not" to ensure that it is very obvious to the student. A student who is reading too quickly may miss the "not" keyword and therefore the entire meaning of the question. It is suggested that when "not" is used, it should be made to stand out, with formatting such as bold, italics or capitals.

Whilst the use of "not" can be very effective, teachers should avoid the use of double negatives in their questions, as it makes the question and options much more difficult to interpret and understand.

3.4 Multiple correct answers

As discussed previously, multiple choice questions have some limitations - specifically that a student may be able to deduce a correct answer, without fully understanding the material. Having multiple correct answers helps eliminate this issue, however it is generally agreed that multiple choice questions with more than one key are not an effective means of assessment (Kolstad and Briggs, 1990), as they often lead to incorrect answers and confusion (Kolstad, Goaz and Kolstad 1982).

A hybrid of the multiple answer and the conservative formats can be achieved (Kolstad and Kolstad, 1994), by listing the "answers" then giving possible combinations of correct answers, as in the following example:

Which of the following statements describe a Java interface?

(i) It defines what, not how
(ii) It must specify the default constructor
(iii) It must declare all methods to be abstract
(iv) It cannot extend another interface
(v) It cannot extend another (non-interface) class
(vi) It requires any non-abstract implementation to implement all methods

(a) i, v, vi
(b) iii, v, vi
(c) i, iii, v
(d) i, ii, iii

In this format the student has to know the correct combination of answers. There is still a possibility that if they know one of the answers is incorrect then this may exclude one (or more) options, however by applying this hybrid format, a more thorough test of their knowledge is achieved.

3.5 Order of questions

At issue here, is whether questions should be in the same order as the material was taught, or scrambled. In Geiger and Simons (1994), the results of studies indicate that the ordering of questions makes no difference to the time taken to complete the examination, or to the results, however it may have an effect on student attitude. The authors suggest that the reason why question ordering doesn't have much impact is that most students seem to employ their own form of scrambling, answering the questions they are confident with, and going back to others later.

3.6 Order of options

It is recommended that options be arranged in some logical pattern – however patterns among the keys within a multiple choice test should be avoided (for example, having a repeating ABCD sequence). To ensure that there is no pattern to the keys, it might be advantageous to apply some kind of constraint on the options (for example, put them in alphabetical order) (Wilson and Coyle 1991).

3.7 Use of "all of the above" and "none of the above"

The option "all of the above" should be used very cautiously, if not completely avoided. Students who are able to identify two alternatives as correct without knowing that other options are correct will be able to deduce that "all of the above" is the answer. In a 3-option test this will not unfairly advantage the student, however in a 4 or 5-option test a student may be able to deduce that the answer is "all of the above" without knowing that one or even two options are correct. Alternatively, students can eliminate "all of the above" by observing that any one alternative is wrong (Hansen and Dexter, 1997). An additional argument against the use of "all of the above" is that for it to be correct, there must be multiple correct answers which we have already argued against.

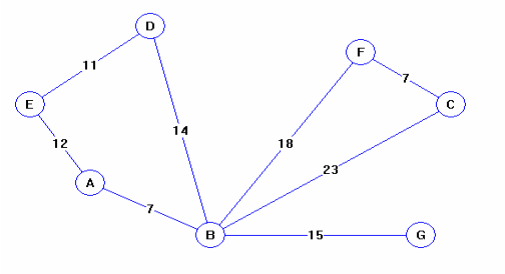
The use of "none of the above" however is more widely accepted as an effective option. It can make the question more difficult and less discriminating, and unlike "all of the above", there is no way for a student to indirectly deduce the answer. For example, in a 4-option test, knowing that two answers are incorrect will not highlight "none of the above" as the answer, as the student must be able to eliminate all answers to select "none of the above" as the correct option.

In Knowles and Welch (1992) a study found that using "none of the above" as an option does not result in items of lesser quality than those items that refrain from using it as an option.

3.8 Writing plausible distracters

An important consideration in writing multiple choice questions is that the distracters are plausible. Poorly written distracters could easily cue a student to the correct answer. For example, if a question asked:

> Given this undirected graph, what would be the result of a depth-first iterative traversal starting at node E?



(a) EABCFDG
(b) EDBFCG
(c) EDBGFCA
(d) EADBCFG
(e) EGDCFBA

certain distracters would be ineffective - a distracter that didn't include every node would be clearly wrong (option b). Most students would also realise that the second node in a traversal would usually be one close to the starting node, so writing an option that jumps suddenly to the other "end" of the graph may also be easily discarded (option e).

When writing distracters for this question, a teacher should consider the types of mistakes associated with a poor understanding of the algorithm and attempt to offer distracters that include these errors. Additionally, an option containing the answer to a similar type of question could be a good distracter - for example, in this traversal question a distracter could contain the correct result for a depth-first recursive traversal (option a) or a breadth-first traversal (option d). Only a student who knows the correct algorithm and is able to apply it to the graph will be able to determine which of the plausible options (a, c and d) is the actual key.

4 Testing More Than Just Recall

The main advantage of multiple choice tests is obvious - they result in a significant reduction of marking for teachers. However one of the greatest criticisms of using this type of questioning is that it only tests facts that students can learn by rote. An extension of this argument is the contention that whilst multiple choice questions may be useful for formative assessment and perhaps even mid-semester examinations, they have no place in examinations where the student should be tested on more then just their ability to recall facts. We believe however that well written multiple choice questions can test up to the sub-synthesis levels of cognition that is, knowledge, comprehension, application and analysis. It should be noted that whilst we are arguing in favour of using multiple choice questions to test more than just recall,

there is always a place for testing knowledge, including fundamental facts that every student of a subject should know.

4.1 Testing Comprehension

To test students at the comprehension level, we should present questions that require them to understand information, translate knowledge into a new context, interpret facts and predict consequences. In IT, we could ask students to predict the result of a particular change to the way a data structure is implemented or interpret code for example, as in the following question:

Consider the following fragment of C code:

```
int x = 2;
float y = 2.0;
float z = 3.0;
z = 1/x*y;
```

What value does z take?

(a) 0.0
(b) 0.25
(c) some other int value

4.2 Testing Application

The application level requires solving problems by applying acquired knowledge, facts, techniques and rules. To test a student's application of knowledge in a subject, they could be asked, for example, to apply a known algorithm to some data.

In Computer Science subjects, there are many opportunities to test at the application level, for example asking the student to apply:

- searching and sorting algorithms,
- ADT-specific algorithms (eg AVL-Tree rotations, Hash Table insertions)
- other algorithms (eg Dijkstra)

The question below tests application of knowledge by asking the student to apply a known algorithm.

The following closed hash table uses the hash function, H(key)= key % 11. The keys are of type *int*.

0	*1*	2	3	4	5	6	7	8	9	10
mt	*del*	310	419	mt	357	258	mt	481	383	del

mt = empty bucket;
del = deleted value;

What is the average number of probes to find a value that is in the table?

(a) 1.0
(b) 1.5
(c) 6.0

Whilst there are many opportunities to write questions testing a student's ability to apply knowledge such as algorithms, it is important to ensure that the distracters are plausible (this topic has been discussed above).

4.3 Testing Analysis

Analysis requires the examination of information, breaking it into parts by identifying motives or causes; identifying patterns; making inferences; finding the underlying structure and identifying relationships.

Asking a student to analyse the effect of some code on a given data structure, or identify patterns in the way an ADT processes information are a good way to test their ability to analyse. However asking these questions in a multiple choice format can be very difficult. If you asked a student "What effect does the above code have on our DataSet?" the distracters may give themselves away – the student may easily be able to see that the code isn't doing what the distracter claims.

There are a several alternatives to this approach. For example, asking the student whether the code will have the desired effect may allow the writing of more plausible distracters, or alternatively, asking them to analyse some code and then make a comparison with some known code. Another example of testing analysis, by identifying patterns, is shown below:

SD2SetI methods in the array of Boolean implementation follow one of 4 patterns of processing:

(i) access or change one value
(ii) traverse all of the set, doing something
(iii) traverse the set[s] until the answer is clear
(iv) process two sets in parallel to produce a new set

Consider the code below which could be used to produce the complement of the set. This code would fit one of the patterns above. Which other methods also follow this pattern?

(a) makeEmpty, isEmpty, size
(b) isEmpty, isSubsetOf
(c) union, intersection
(d) forAll, select
(e) include, exclude

Another method of testing a student's higher cognitive skills is through the use of linked sequential questions which allow the examiner to build on a concept. An example of this method would be to ask a number of questions each of which makes a small change to a piece of code, and to ask what effect that change would have on the functioning of a program. The student could be required to use the outcome of each question to answer the subsequent question. Using this technique, however, care needs to be taken to avoid unfairly penalising the student through accumulated or sequential errors.

5 A checklist for writing effective multiple choice questions

In this paper, we have discussed ways of improving the quality of multiple choice questions and provided guidance in writing questions to test the higher levels of cognition. To assist teachers in implementing these recommendations we have created a checklist, attached as Appendix A, that provides a handy summary of the issues

associated with writing effective multiple choice questions.

The issues discussed in section 3 can be categorised as those that relate to individual questions, and those that relate to an examination containing multiple choice questions.

In relation to individual questions we have identified:

- Correct grammar and wording;
- "Not" and the use of double-negatives;
- Multiple correct answers;
- Order of options;
- "All of the above" and "none of the above"; and
- Writing plausible distracters.

In relation to the construction of an examination:

- the order in which the questions appear, and
- the number of options that should be used.

Whilst the number of options may appear to relate to individual questions, it is really a matter that should be considered when constructing examinations, as to have different numbers of options in questions may cause confusion to the students.

The checklist covers the major issues presented in the paper and we believe that it will assist teachers of Information Technology subjects, and indeed any discipline, to set more effective questions. Being able to set good questions which test higher cognition allows teachers to use multiple choice questions in end of semester tests with confidence, not just as a convenience for low-valued mid-semester tests and formative assessment.

There are many other issues relating to multiple choice questions that we have not discussed (such as whether it is appropriate to have a guessing correction) however our checklist is a god starting point to help novices avoid common pitfalls.

6 Conclusion and future work

We have attempted to advise teachers of Information Technology of the vast amount of research that has been undertaken into writing multiple choice questions. We have discussed the terminology used to describe multiple choice questions and their limitations, as well as a range of factors that should be considered when composing questions.

Further, we have described how multiple choice questions can be used to test more than straight recall of facts. We gave specific examples which test students' comprehension of knowledge and their ability to apply and analyse that knowledge and suggest that sequentially dependent questions also facilitate testing of higher cognition.

The most important contribution is the creation of a checklist which draws together these ideas into a concise form that will be beneficial to lecturers setting multiple choice exams.

In other related work, the authors are implementing a web-based multiple choice management system. A stand-alone prototype of this system (Rhodes, Bower and Bancroft 2004) is currently in use, while the web-based system will allow further features, including concurrent access and automatic generation of paper-based examinations.

7 References

Bloom, B.S. (ed.) (1956): *Taxonomy of educational objectives. Handbook 1. The cognitive domain* New York, David McKay.

Bruno, J. and Dirkzwager, A. (1995): Determining The Optimal Number Of Alternatives To A Multiple-Choice Test Item: An Information Theoretic Perspective *Educational and psychological measure,* **55**(6):959-966.

Carter, J., Ala-Mutka,K., Fuller,U., Dick,M., English,J., Fone,W., Sheard,J. (2003): How Shall We Assess This? *ACM SIGCSE Bulletin, Working group reports from ITiCSE on Innovation and technology in computer science education,* **35**(4): 107-123.

Cizek, G. and O'Day, D. (1994): Further investigation of nonfunctioning options in multiple-choice test items. *Educational & Psychological Measurement,* **54**(4):861-872.

Geiger, M. and Simons, K. (1994): Intertopical sequencing of multiple-choice questions: Effect on exam performance and testing time. *Journal of Education for Business*, **70**(2):87-90.

Haladyna, T. and Downing, S. (1993): How many options is enough for a multiple choice test item? *Educational & Psychological Measurement journal* **53**(4):999-1010.

Hansen, J. and Dexter, L. (1997): Quality multiple-choice test questions: item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business*; **73**(2):94-97.

Isaacs, G. (1994): *HERDSA Green Guide No 16. Multiple Choice Testing*. Campletown Australia, HERDSA

Knowles, S. and Welch, C. (1992): A meta-analytic review of item discrimination and difficulty in multiple-choice items using "None of the Above" *Educational and psychological measurement*, **52**(3): 571-577.

Kolstad, R.K. and Briggs, L. (1990): Multiple choice rules prevent the selection of wrong options in examinations. *Education,* **110**(3): 360-363.

Kolstad, R.K., Goaz, P. and Kolstad, R.A. (1982): Nonrestricted multiple-choice examination items. *Journal of Dental Education*, **46**(8): 485-488.

Kolstad, R.K., and Kolstad, R.A. (1994): Applications of conventional and non-restrictive multiple-choice examination items. *Clearing House*, **67**(6): 317-319.

Lister, R. (2000): On Blooming First Year Programming, and its Blooming Assessment. *Proceedings of the Fourth Australasian Computing Education Conference*, Melbourne, Australia, 158-162, ACM Press.

Lister, R (2001) Objectives and objective assessment in CS1. *ACM SIGCSE Bulletin, ACM Special Interest Group on Computer Science Education* **33**(1): 292-296.

Rhodes, A., Bower, K. and Bancroft, P. (2004): Managing Large Class Assessment *Proceedings of the sixth conference on Australian computing education*, Dunedin, New Zealand, **30**:285-289, Australian Computer Society, Inc.

Sidick, J. and Barrett, G. (1994): Three-alternative multiple choice tests: An attractive option *Personnel Psychology*, **47**(4):829-835.

Wilson, T. and Coyle, L. (1991): Improving multiple-choice questioning: Preparing students for standardized tests. *Clearing House* **64(**6):422-424.

APPENDIX A

A checklist for setting multiple choice questions

Individual Questions:

Grammar and Wording

a/an

- [ ] Does the option start with a vowel? If so, put the a/an in the option, not the stem.

is/are & plural

- [ ] Do the options contain singular and plural? Put the is/are in the option, not the stem.

wording

- [ ] Are your options roughly same length and written in a similar style?

Not and double negatives

- [ ] If you have used “not”, have you emphasized it’s use - e.g. with bold and italics?
- [ ] Have you avoided the use of double negatives?

Multiple correct answers

- [ ] Have you avoided the use of multiple correct options?

Order of options

- [ ] Are your options ordered in some logical manner (e.g .alphabetically)?

None of the above/All of the above

- [ ] Have you carefully considered their use?
- [ ] Have you ensured that the key isn't easily identified by other correct/incorrect options?

Constructing the Examination:

Patterns in answers

- [ ] Have you avoided patterns in the position of the key (e.g. ABCD-ABCD)?

Number of options

- [ ] Do you really need more than 3 plausible options?
- [ ] Have you got the same number of options for each question?