Quality assurance of multiple-choice tests

Martin E. Bush
London South Bank University, London, UK

Abstract

Purpose – To provide educationalists with an understanding of the key quality issues relating to multiple-choice tests, and a set of guidelines for the quality assurance of such tests.

Design/methodology/approach – The discussion of quality issues is structured to reflect the order in which those issues naturally arise. It covers the design of individual multiple-choice questions, issues relating to the question bank as a whole, choice of test format, and what can be learned through post-test analysis. The paper offers practical advice, with an emphasis on maximising test reliability.

Findings – It is recognised that considerable expertise and effort is required to undertake a thorough post-test statistical analysis, but pre-test quality assurance is relatively straightforward, if labour-intensive. The question of which is best amongst the various alternative test formats is left open.

Originality/value – The general issue of quality assurance of multiple-choice tests is surely an important one, yet the author is not aware of any other publication that deals directly with this topic.

Keywords Scoring procedures (tests), Quality assurance, Assessment, Education

Paper type General review

Introduction and terminology

Multiple-choice tests are already in widespread use within many areas of education, and their importance as a method of assessment seems likely to grow further with the advent of e-learning. They are attractive for a variety of reasons. For example, they can be used to test higher-order thinking skills in addition to knowledge recollection, there is no dependence on good writing skills, and a wide range of topics can be covered in one sitting. Also, the marking is easily automated – which makes them ideal for e-learning.

Designing good test questions is not easy, however. The strain to produce a large question bank can easily lead to poorly written and/or unintentionally overlapping questions. In this paper we review a range of quality issues relating to multiple-choice tests, and I offer practical advice regarding quality assurance both of a question bank and the tests that are derived from it, with an emphasis on maximising test reliability (defined below).

As far as quality assurance is concerned, too often the only considerations are:

(1) Do the questions look reasonable bearing in mind the subject matter?

(2) Does the distribution of marks look reasonable bearing in mind the performance of the examinees in other examinations?

The first question should be considered before any testing takes place. The second question can only be considered after the testing has taken place. Given additional effort, quality assurance can go a lot further – as this article explains.

Conventionally, an “n-choice test” is made up of a number of questions each with a question stem, one correct answer and $n - 1$ distractors (incorrect answers). We can

Quality Assurance in Education
Vol. 14 No. 4, 2006
pp. 398-404
© Emerald Group Publishing Limited
0968-4883
DOI 10.1108/09684880610703974

consider an idealised question bank to be a set of non-overlapping questions that cover the relevant subject matter fully. With respect to individual questions we can speak of examinees being "fully informed", "partly informed", "uninformed" or "misinformed". If we assume, over-simplistically, that examinees are either fully informed or uninformed with respect to every question (i.e. that they either know the right answer or they have no idea which is the right answer) then the proportion of an idealised question bank which an examinee can successfully answer without guessing would be a good indicator of the proportion of the subject matter she/he had learned.

Examinees who are misinformed with respect to particular questions will answer those questions incorrectly, but they are not guessing. Examinees who are uninformed with respect to particular questions can do no more than guess at the answers for those questions. Examinees who are partly informed with respect to those questions are able to make what is often referred to as "educated guesses". Consequently, as far as those questions are concerned, the partly informed examinees are likely to score higher marks than the uninformed examinees, while any misinformed examinees will inevitably score zero.

The key notion as far as whole tests are concerned is reliability. The reliability of a test is the degree to which it repeatably yields scores that are a true reflection of the examinees' knowledge and understanding of the subject matter. A perfectly reliable test would always yield test scores that truly reflected each examinee's level of knowledge and understanding of the subject matter, but there are a variety of unavoidable factors that make this theoretical utopia unobtainable in practice.

Pre-test quality assurance

As far as pre-test quality assurance is concerned, the quality of the question bank is an obvious issue. It is not the only issue, however. It is also important to consider the number and choice of questions to be used within individual tests, the test format, and (where relevant) the degree of overlap between distinct tests.

Quality of the question bank

The literature on multiple-choice tests is copious, but the recommendations of different authors do not always coincide. One often-repeated recommendation (e.g. Frary, 1995; Wood, 1991) is that multiple-choice tests should contain as many non-overlapping questions as possible. The most obvious reason for this is that the variability of scores due to guessing diminishes as the test length increases. Also, if examinees are presented with a set of questions that do not cover the subject fully (which is usually the case) then there is an element of luck with respect to the degree of overlap with their own knowledge and understanding; the less extensive the subject coverage, the greater the element of luck. This is another reason why multiple-choice tests should contain as many (non-overlapping) questions as possible.

The need for a large question bank is exacerbated whenever there is a requirement to generate two or more distinct tests for different groups of examinees. Having access to a previously written question bank can be extremely helpful of course, but this does not alleviate the problem entirely because each question still needs to be considered very carefully to ensure validity with respect to the knowledge and understanding that is expected of the current examinees.

It can be very difficult to avoid overlap between questions. Rather than trying to eliminate overlaps, a more pragmatic approach is to mark overlapping questions as such. Indeed, the inclusion of overlapping questions within a question bank may be the only way to enable the generation of multiple tests in which no question is re-used (and no test contains overlapping questions). Culwin (1998, p. 56) goes so far as to propose a "question template based system" in which 32 different variations of each question can be generated by using both the positively-phrased and the negatively-phrased versions of the question stem and supplying four alternative but valid answers to each, in such a way that if an answer is valid for the stem in its positive (negative) form then it is invalid for the stem in its negative (positive) form.

For example, here are positively- and negatively-phrased versions of a question stem with four valid answers in each case concerning the differences between mammals and reptiles:

(1) *(Positive form). Which of the following is true with respect to mammals versus reptiles?*
   (a) Mammals have three ear bones, while reptiles have just one.
   (b) Mammals are warm-blooded, while reptiles are cold-blooded.
   (c) Mammals have a thoracic cavity with a diaphragm, but reptiles do not.
   (d) Mammals have thin skin with hair, while reptiles have thick and scaly skin.

(2) *(Negative form). Which of the following is untrue with respect to mammals versus reptiles?*
   (e) Reptiles have heart ventricles separated by septa, but mammals do not.
   (f) Mammals are cold-blooded, while reptiles are warm-blooded.
   (g) Reptiles have a thoracic cavity with a diaphragm but mammals do not.
   (h) Mammals have thick and scaly skin while reptiles have thin skin with hair.

So, for example, one version of this question would present the positive form of the stem accompanied by answer (a) and distractors (e), (f) and (g). Another version of this question would present the negative form of the stem accompanied by answer (e) and distractors (a), (b) and (c). These two derived questions are obviously overlapping, but we might also have other questions that are (perhaps inadvertently) overlapping in less obvious ways. Consider the following question for example: "How many ear bones do mammals have? (a) one; (b) two; (c) three; (d) four." This overlaps with some variants of the former question, but not all of the variants.

Optimum test reliability demands more than just lengthy tests with non-overlapping questions. It also demands moderately difficult questions containing equally plausible distractors, plus (nevertheless) a high average score. These requirements can perhaps be most easily understood by considering the following pathological cases:

- If there were any questions that all the examinees got right, and/or any that they all got wrong, then the spread of marks would be entirely due to differing performance on the remaining questions. In this case the effective test length – and consequently the reliability – would be reduced.
- If uninformed examinees can identify implausible distractors then they are able to act as, and hence they become indistinguishable from, partly informed examinees – and therefore reliability suffers.
- If scores are generally low it could be that many of the answers were guessed – in which case, again, the test reliability will have been reduced.

There are numerous readily available publications – see Frary (1995) for example – that give excellent advice about how to write effective multiple-choice questions. As far as pre-test quality assurance of the question bank is concerned, peer inspection is the natural way to tackle this. Reviewers may find the following checklist helpful as a starting point:

- Is each question clearly and unambiguously expressed, with correct spelling and grammar?
- Have any uncommon words been used that could be replaced by more familiar words having the same meaning? (This is an especially important consideration whenever the language of the test is foreign to any of the examinees.)
- Are there any questions in which one or more of the distractors are too obvious, or for which it could be argued that one (or more) of the supposed distractors is in fact an alternative correct answer to the question?
- Have all of the overlapping questions within the question bank (if any) been marked as such?
- Do the questions within the question bank collectively cover the subject matter, or is there an area in which additional questions could be added?
- Are there enough questions in the question bank to enable the generation of sufficiently long tests without overlapping questions within a test, and – assuming that more than one test is required – without excessive overlap or (worse) re-use of questions within different tests?

Alternative test formats

Given a (quality assured) question bank, the next consideration should be choice of test format. There is a wide variety of novel and worthy test formats that can be used in preference to the conventional one; see Bush (1999) for an overview. However, this seems to be the area in which the recommendations of different authors coincide least of all. There are many contenders for the "best practice" test format, and it may be that there is no one best format for every different situation. Several of these alternative test formats are designed to counter the artificial inflation of marks (see below) and also the variability of marks due to guessing (again, see below), which are both detrimental from the point of view of test reliability. In the remainder of this paper we focus on the use of negative marking and "liberal" multiple-choice tests, but readers should be aware that there are other alternatives.

It is easy to calculate how guessing is likely to result in the artificial inflation of marks. With a conventional test we can consider each examinee's score as being made up of two components; one due to known answers and the other due to correctly guessed answers. Consider, for example, a conventional four-choice test (or a series of tests) consisting of 100 questions, and a particular examinee who is fully informed with respect to 40 per cent of the questions in the test. If that examinee was completely uninformed with respect to the other 60 per cent of the questions, and therefore resorted to blind guesses for those questions, then his or her most likely score would be $(40 + 0.25*60)\% = 55\%$. In reality it is quite unlikely that an examinee will be either fully informed or completely uninformed in relation to every question; it is much more likely that they will be partly informed for at least some of the questions. Hence a score of 55 per cent would most probably consist of more than 15 per cent (0.25*60%) gained through guesswork (assuming some educated guesswork), and therefore less than 40 per cent due to known answers.

In an attempt to quantify the effects of variability due to guessing, Burton (2004) calculated that for a 30-question, four-choice test, assuming a random choice of questions on the part of the test setter and that examinees guess blindly whenever they do not know the correct answer to a question, there is a 25.8 per cent chance of a student with 50 per cent knowledge scoring higher than a student with 60 per cent knowledge. However, if every examinee refrained from guessing then the figure of 25.8 per cent would be reduced to 18.1 per cent; the remaining variability being due to luck with respect to inclusion of questions that the examinees either could or could not answer.

It is impossible to eliminate guessing entirely, but it can – and arguably should – be discouraged. One approach is to require examinees to associate a level of confidence with each of their selected answers, and to add or subtract marks for each question according to a formula that takes this into account as well as whether or not they have selected the correct answer.

An alternative, simpler approach is to use a conventional test format but to score it using negative marking. Given an n-choice test this means awarding $n - 1$ marks for every correct selection and deducting one mark for every incorrect selection, so that examinees are likely to lose as many marks as they gain overall through guessing. Negative marking is very well known and quite widely used. Furthermore, it has been shown that negative marking does indeed reduce guessing, although not consistently. For example, examinees with lower grade expectations seem more likely to guess (Bereby-Meyer *et al.*, 2002).

If negative marking is applied as described above, why not allow examinees to select as many answers as they wish. For example, a particular examinee might want to select two answers to a certain question in a four-choice test in the belief that one of them is correct. If she/he is right (either by virtue of being partly informed or due to a lucky guess) then she/he will score 3 marks for their correct selection minus 1 mark for their incorrect selection, making a total of two marks for that question. If she/he is wrong (either by virtue of being misinformed or due to an unlucky guess) then she/he will score a total of minus two marks for that question. Such tests are referred to as “liberal” tests in Bush (2001). It is worth noting that examinees may choose to treat such tests as though they are conventional tests – thereby selecting one answer for each and every question – and in practice they often do. This novel test format is equivalent to the so-called “elimination procedure” proposed by Coombs *et al.* (1956), in which examinees are asked to eliminate as many incorrect answers as they can.

A potential problem with many unfamiliar test formats is that the examinees can become perplexed when thinking about alternative tactics. For example, in a four-choice liberal test there are 16 (2^4) possible responses per question – or rather 15, since selecting every answer is equivalent to selecting none. By contrast, in a four-choice conventional test there are only four possible (reasonable) responses per question – or, if negative marking is used, five possible responses, since selecting nothing is a reasonable response.

Post-test quality assurance

There are two complementary ways of tackling post-test quality assurance; one is through statistical analysis, the other is through student feedback.

Some obvious statistics relating to tests include number of examinees, highest and lowest scores, median and mean averages, standard deviation and skew. These are relatively easy to calculate, and these are often the only statistics that are calculated. The distribution of marks for one test could also be correlated with the distribution of marks for other tests, and this may or may not yield some useful insights.

However, even though the distribution of marks for a particular test might look perfectly reasonable there may be significant problems associated with certain questions. Two key attributes of individual questions are their "difficulty" and their "discrimination". The difficulty of a question is the proportion of examinees found to have answered it correctly. The term discrimination in this context is the degree to which the performance of the examinees on that question correlates with performance on the test as a whole. For example, given a question of difficulty 0.5 that discriminates perfectly, those who answered the question successfully will have scores in the top half of the group while those who did not will have scores in the bottom half.

Clearly, if a question is found to have been too hard, or too easy, or it does not discriminate well between high and low scorers, then it should be reviewed before being re-used in future tests. If a particular distractor is very rarely selected then it may be too obviously wrong. Alternatively, if a distractor is selected often then it might arguably be a valid answer to the question. In either case, the reliability of the test will have been harmed to some extent. There are a number of software packages designed specifically for this kind of analysis that, in some cases, have been available for many years (Kehoe, 1995); they calculate question difficulty and discrimination and many other statistics.

From the point of view of post-test quality assurance, statistical analysis along the lines described above is recommended as being the most thorough way to identify potentially problematic questions and hence to achieve process improvement. It can be done automatically within an e-learning environment, but undertaking such an exercise independently clearly demands considerable expertise and effort. An easier, complementary option is to elicit either written or verbal feedback from the examinees; they may have some very worthwhile points to make about a multiple-choice test they have just taken.

Conclusions

We have seen how three fundamental aspects of quality assurance can be applied to multiple-choice tests:

(1) peer inspection of the question bank;

(2) adoption of best practice regarding test format; and

(3) process improvement through post-test analysis and feedback.

The first two aspects should receive attention before any tests are set. As far as peer inspection of the question bank is concerned, we saw that it is important to check for question validity, absence of implausible distractors, avoidance of overlapping questions within a test (although overlapping questions within a question bank is OK – provided that they are marked as such), etc. Tests should be lengthy, and should cover as much of the subject matter as possible.

As far as the test format is concerned, we saw that it is important to discourage guessing. Negative marking is one way to achieve this. The adoption of negative marking also opens the door to allowing examinees to select multiple answers per question; this is the notion of liberal testing. There are various other alternative test formats, but there is no consensus as to which one(s) can be considered to be "best practice".

Post-test statistical analysis – expertise and resources permitting – is required to reveal the truth about whether the examinees found all of the questions to be moderately difficult and all of the distractors within each question to be equally plausible. Feedback from students can be helpful too.

Finally, this article has focused on quality assurance from the point of view of optimizing test reliability, but that should not be the only consideration. For example, including a few easy questions may help to prevent weak students from becoming overly disillusioned when they receive their marks, which can be a very real problem when negative marking is used.

References

Bereby-Meyer, Y., Meyer, J. and Flascher, O. (2002), "Prospect theory analysis of guessing in multiple choice tests", *Journal of Behavioral Decision Making*, Vol. 15 No. 4, pp. 313-27.

Burton, R. (2004), "Multiple choice and true/false tests: reliability measures and some implications of negative marking", *Assessment and Evaluation in Higher Education*, Vol. 29 No. 5, pp. 585-95.

Bush, M. (1999), "Alternative marking schemes for on-line multiple choice tests", *Proceedings 7th Annual Conference on the Teaching of Computing, Jordanstown*, available at: www.caacentre.ac.uk/resources/bibliography/biblio2.shtml (accessed 7 January 2005).

Bush, M. (2001), "A multiple choice test that rewards partial knowledge", *Journal of Further and Higher Education*, Vol. 25 No. 2, pp. 157-63.

Coombs, C., Milholland, J. and Womer, F. (1956), "The assessment of partial knowledge", *Education and Psychological Measurement*, Vol. 16, pp. 13-37.

Culwin, F. (1998), "Web hosted assessment – possibilities and policy", *ACM SIGCSE Bulletin*, Vol. 30 No. 3, pp. 55-8.

Frary, R. (1995), "More multiple-choice item writing do's and don'ts", *Practical Assessment, Research & Evaluation*, Vol. 4 No. 11, available at: http://PAREonline.net/ (accessed 7 January 2005).

Kehoe, J. (1995), "Basic item analysis for multiple-choice tests", *Practical Assessment, Research & Evaluation*, Vol. 4 No. 10, available at: http://PAREonline.net/ (accessed 7 January 2005).

Wood, R. (1991), *Assessment and Testing: a Survey of Research*, Cambridge University Press, Cambridge.

Corresponding author

Martin E. Bush can be contacted at: martin.bush@lsbu.ac.uk