The Role of Analysis in Multiple Choice Question Tests

By Graham Barrow of GR Business Process Solutions

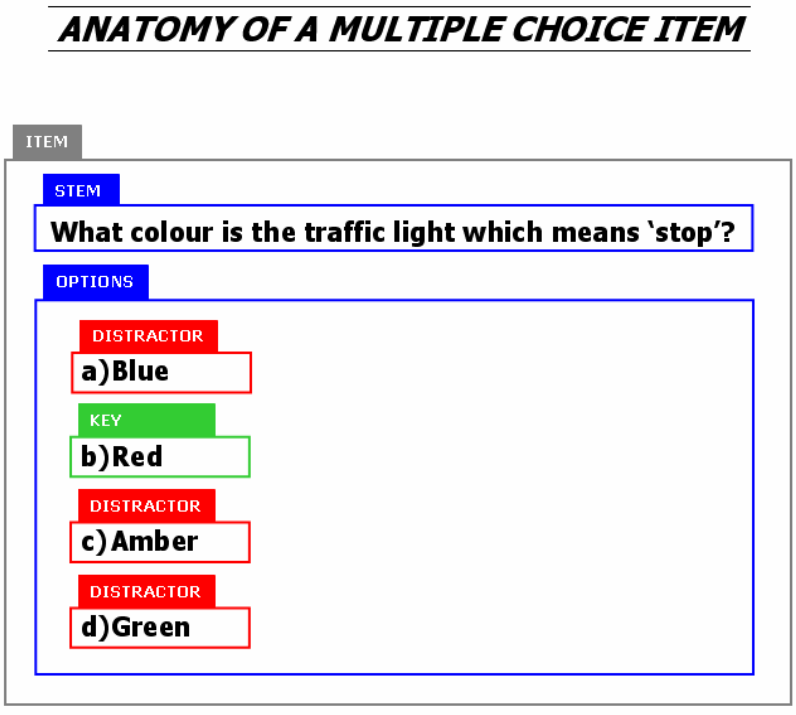
Many articles recently have referred to the role of Multiple Choice Tests in assessing competence and some of them have made mention of the role of post-test analysis. I would like to take that one stage further and look in rather more depth at some of the forms of analysis that can be carried out and at how they can make a positive impact on a firm's knowledge training and assessment programmes.

However before we do, let's have a quick review of the steps that need to have taken place before post-test analysis becomes a practical or worthwhile endeavour.

First and foremost is the creation of a clear, comprehensive and relevant syllabus on which to base the test (and, of course, the training material which precedes it). Essentially, without the syllabus, a pass in any test simply means that the candidate knew enough to pass the test but not, necessarily, enough to perform his or her job role competently. The syllabus should represent the knowledge and skills necessary to achieve competence, the training material should deliver the syllabus to the candidates effectively and the test should then provide the means to validate that the candidate has achieved the requisite level of competence. No problems there then!

For the purposes of this article I am going to make the assumption that the syllabus and training material have been written and are in excellent shape. So what then could go wrong?

First, as Margaret Spong pointed out in her article in the autumn 2002 edition of T & C News, it is necessary to construct clear and effective items which test specific learning points and which have suitably constructed stems and distractors. Now we need to make sure they are working correctly. With the best will in the world (and an excellent edit committee) it is not always possible to know in advance how well an item is going to perform. This is where post-test analysis is invaluable.



So what sort of analysis can we perform? Well, the list is pretty long, ranging from the very

straightforward to the decidedly esoteric but, for the purposes of this article, I am going to concentrate on the areas that make the biggest impact on refining the effectiveness of objective testing.

The first area is known as the 'facility factor'. This is simply a decimal between 0 and 1 which expresses how easy (or difficult) a particular question was, where 0 means that nobody answered the question correctly and 1 means that everybody did. By extrapolation a facility factor of 0.5 means that half of all people answering this question got it right. The facility factor can be measured both historically and for a particular test and making this comparison can itself be a useful guide for highlighting a potential issue.

Suppose, for example, that a particular question has historically scored a facility factor of 0.75, meaning that three out of four candidates have, in the past, answered this question correctly. Suddenly, on one particular test, the facility factor drops to 0.15. What could have happened? Firstly, maybe the answer is no longer correct (a question about income tax which has not been corrected post budget). Or maybe something was said in the training course that has misled the candidates. It's difficult to say exactly, but the sudden change in facility tells you that you need to investigate.

The second area is known as the 'discrimination factor'. This measures how effectively each question discriminates between able and less able candidates. It is calculated by looking at the results obtained by the top and bottom 27% of candidates (when listed in order of result achieved). Why 27%? It's a long story but essentially this figure ensures both a wide separation and a statistically significant group size.

The Discrimination factor is calculated by working out the percentages of the upper and lower groups who answered the question correctly and then subtracting the percentage for the lower group from the percentage for the upper. Let's look at an example.

> 200 people sat a test.
> Of the top 27% (54 people), 45 answered question 1 correctly (83%)
> Of the bottom 54 people only 18 (33%) answered question 1 correctly.
> This question has a discrimination factor of 0.5 (i.e. 50% more of the upper group answered the question correctly).

Unlike the facility factor, discrimination can have a range of between -1 and +1, where -1 means that everyone in the upper group answered the question incorrectly whilst everyone in the bottom group got it right and, of course *vice versa* where the factor is +1.

Clearly, if more people in the lower group answer a question correctly it is an indication that something is wrong. Options here range between the fact that everyone is guessing and it just happened that more people guessed correctly from the lower group, or that the question has been poorly worded and the upper group believe there is some merit to one of the other answers, or, of course, the answer may just be wrong! Either way it needs a closer look.

Now let's look at the last one of the analysis factors which I am going to include in the present article. This relates to the distribution of answers selected and is known as 'distractor analysis'. Distractors by their very nature need to 'distract' the candidate from the correct answer, in which case they need to be both plausible and attractive if they are going to do the job properly. After all, the whole point of a Multiple Choice Test is to

identify which of the candidates really understands the syllabus in question and which need to have more development before they can be said to be competent.

In order for this to happen, with the minimum possible impact of guesswork, each of the distractors must 'earn their keep'. In order to measure this we look at the number of times each distractor is selected by the lower group. We are not so interested in the upper group here because, by the very nature of the fact that they are stronger candidates they should be harder to distract from the correct answer. However if a distractor is not being selected by at least 7 ½% of the lower group it can be reasonably said to be 'redundant' and should either be discarded or at least reworded in order to make it more attractive and plausible.

So what is the problem if one of the distractors is redundant?

One of the weaknesses of MCQs is the potential for guesswork to impact on the final result. If a question has four potential options there is a statistical likelihood that a candidate would score, on average, 25% just by guessing. However if a group of candidates all take a test and guess each question, whilst the average will come out at around 25% the actual scores will range in a fairly standard bell curve distribution between about 10% and 40%. This is still well below the usual level of pass mark. However, if some of the questions have distractors which are easily discountable it can move the top end of the bell curve upwards to quite a degree and could, ultimately, threaten the integrity of the test.

Finally, it would be helpful to have some measure of the reliability of the answer and, in this case, it is useful to look at the upper group and analyse how many chose the correct answer. If it was more than 70% it is reasonable to assume that the answer, as chosen, was correct. If it was between 40% and 70% then it would be worthwhile checking the answer (particularly if the question discriminated poorly between the upper and lower groups). If less than 40% of the upper group selected the correct answer it is highly questionable whether the question as framed is acceptable even if it is correct.

Let's try and bring all this together with an example.

Question number: 1 **Correct answer:** D
Facility factor: 0.27

Number of people selecting each option:

	A	B	C	**D**
Upper group	0	0	16	**9**
Lower group	0	0	18	**7**

Discrimination index scoring:

	A	B	C	**D**
Upper group	0.00	0.00	0.64	**0.36**
Lower group	0.00	0.00	0.72	**0.28**
Discrimination index	0.00	0.00	-0.08	**0.08**

So, what's going on here? According to the answer file the correct answer is 'D'. However, only 16 people chose that answer whereas 34 chose 'C'. Although the correct answer is discriminating positively it is very weak and leaves one to believe that no-one was entirely sure whether 'C' or 'D' was correct and they were therefore guessing (and 'C' could in fact

be correct given that it was chosen by 73% of candidates as opposed to 27% for ‘D;). On the other hand nobody in either the upper of lower group chose either ‘A’ or ‘B’ and they need to be urgently reviewed and rewritten or replaced. If neither of these options is possible the item itself may need to be rewritten or replaced.

The ability to perform and interrogate this type of analysis is essential if firms are going to be able to demonstrate to the regulator that their testing regimes are robust and fit for purpose. Should you decide that it is time you overhauled your question bank, analysis of your historic results is the best possible way to prioritise your review.