False Passes: Assessing the Impact of Guessing In Multiple Choice Tests

By Graham Barrow of GR Business Process Solutions

Over the last few issues of T & C News there have been a number of articles about the construction of Multiple Choice Questions (called 'items') and how to analyse their effectiveness. In this article I would like to turn my attention (and yours) to the impact of guesswork when assessing the effectiveness of a multiple choice test.

Over the course of this article I am going to use as my example a 100 question test, using four option items, with an overall pass mark of 70%. In my experience this represents an oft-found test of both firm specific and generic knowledge.

So let's look in the first instance at a perfectly constructed test in which each of the items has four equally attractive options. If the candidate does not know the correct answer, he is equally likely to choose any of the four options. First of all, it's worth pointing out that this test probably does not exist in reality. I have never found a test where all three distractors (wrong answers) are equally attractive for every question in the test. Nevertheless (and for the purposes of this article) let's assume that a candidate took this test and scored 70%. Fantastic: he (or she) has passed! So what does this pass tell us about the candidate's knowledge of the syllabus on which he has just been tested?

Statistically, the test result demonstrates that the candidate knew 60% of the syllabus and achieved the other 10% of his marks by guesswork. How do I know that? Well we know that he got 30 questions wrong (30%) and it is vanishingly unlikely that every time he guessed he selected the wrong answer. Across a four option item the chances are one in four that he will guess correctly so by extrapolation, three in four guesses will be wrong. Applying this to the example means that, if three in four guesses were wrong and the candidate answered thirty questions wrongly by guessing, he must also have answered 10 correctly by guesswork. So a raw score of 70% indicates a true score of 60%. But it gets worse.

As I said above, this example assumes a perfectly constructed test where all options are equally attractive. What happens if some of the options are so implausible that nobody ever picks them? This has the effect of reducing the odds of guessing correctly from one in four to one in three (or worse, if more than one option is implausible within a given item).

Let's have a look at an example which is a bit closer to real life than the one given above (although still a simplistic model).

In this test, on average, one of the distractors for each question is considered implausible and therefore, where the candidate does not know the answer he has a one in three chance of guessing correctly. If he has a raw score of 70% what, statistically, is this likely to represent in terms of his actual knowledge of the syllabus? If one in three guesses are likely to be correct it follows that two out of three will be wrong. Therefore, if our candidate gets 30 questions wrong, he must have guessed on 45 occasions which means that 15 of his correct answers are likely to be guesses, giving a true score of 55%!

Once you start doing the sums and using real data the impact of guesswork can be really frightening. The next, and final, example is taken from a real set of data provided by one of our clients (although to be fair to them, they knew there were issues with their tests which is why they spoke to us).

The data came from a 100 question test. At the point it was provided, the test had been taken by 220 candidates. For our purposes we are only interested in the results from the bottom 27% of candidates, i.e. the 60 lowest scores emerging from the test sitting, on the basis that the top performers will, self-evidently, be much better at selecting the correct answer.

Looking at these candidates we were able to determine significant information by analysing their results.

Note: a redundant distractor is here used to mean an option that NOBODY selected (a very strict application of the criteria) – in reality it is as good as redundant if fewer than 7.5% of the bottom set of candidates selected it.

8 questions were answered correctly by all the bottom 60 candidates
28 questions had two redundant distractors (a 1 in 2 chance by guessing)
33 questions had one redundant distractor (a 1 in 3 chance by guessing)
31 questions had no redundant distractors (a 1 in 4 chance by guessing)

Therefore just by guessing a candidate could achieve the following result:

Where all three distractors are implausible	8
Where two of the distractor are implausible	14
Where one of the distractors is implausible	11
Where there are no implausible distractors	8
Total by guesswork alone	**41**

This means that, statistically, the candidate only needs to know 29% of the syllabus in order to achieve an overall pass mark.

In reality of course, the marks achieved by guessing would be spread across a normal distribution centred on the relevant percentage (e.g. a large number of candidates guessing a four option test across 100 items would achieve an average mark of 25% but they would actually be spread across a range of between, say, 10% and 40%). This means that, for the above example, the scores achieved by guessing might have a range of, say, 20% to 60%.

The implications of this are clearly significant when viewed in the light of regulation and the need to carry out robust testing of both firm specific and generic knowledge. It will become ever more important for firms to be able to satisfy both the regulator and themselves that they truly understand and allow for the difference between ‘raw’ and ‘true’ scores.

So what can be done about this?

First of all, it is imperative that the testing regime does not stop with the administration and marking of the test. It is vital that post-test analysis is performed to understand (amongst other things) the level of redundant distractors (particularly amongst the bottom quartile). Once that analysis is available it then becomes important to review and revise all those questions where one or more of the distractors are never (or rarely) being selected. This can (and often does) present some difficulties.

Many times when writing items it is easy to create (as well as the correct answer) two good distractors but it can then be extraordinarily difficult to craft a plausible third option. At this point even the best of item writers can succumb to the temptation of keeping the question with a poor third distractor. Let me give you an example:

1. What happens to the yield on gilt-edged securities when interest rates go up?

    a. It increases.
    b. It decreases.
    c. It stays the same
    d. None of the above.

It only takes a moment to realise that option (d) is nonsense. What other choices are there beyond the first three? You can see what has happened, can't you? The item writer does not want to discard the question but simply cannot think of a plausible third distractor. However, in the process he has made it possible for someone of average intelligence who knows nothing about Gilts to have a one in three chance of guessing correctly rather than one in four.

At this point, the item writer needs to make a decision which is likely to be one of the following: -

1. Throw the item away and start again
2. Rewrite the stem (question element) in order to be able to create three plausible distractors
3. Come up with a different third distractor that is plausible

Of these options, (3) is nearly always the most difficult because the likelihood is that, if there was a more plausible distractor, the item writer would have come up with it first time around. The first option is heartbreaking (anyone who has ever written items will understand why). This means that the second option is the favourite.

So let's see if we can come up with a question that lets us write more plausible options. What about:

1. The yield on Gilt-edged securities has just gone up. Which **ONE** of the following options is **MOST LIKELY** to have been a direct cause:

    a. Interest rates have increased.
    b. Interest rates have decreased.
    c. Inflation has increased.
    d. Inflation has decreased.

I'm not suggesting for a minute that this is now a perfect question (apart from anything else I would want to see some statistical analysis of how the question performs in a test) but I think you would agree that it is clearly an improvement on the original.

So, in summary, it is clearly important to assess the impact of guessing within a Multiple Choice Test but that is just the start of the process. Once that assessment has been made you must have good quality analysis and very high quality item writers in order to apply all that learning to the testing regime and make it more effective.

We interpret CP157 as indicating that the regulator will increasingly be expecting firms to undertake this sort of analysis and validation of their knowledge testing regimes.

In order to help us to understand exactly how item construction impacts on test results we have constructed two tests which can be accessed through our website (http://www.grbps.com). There you will find a link to "Multiple Choice Guessing". There are two tests involved in the research, one has been constructed as far as possible in accordance with the accepted rules for MCQs and the other hasn't. Each consists of 16 questions of extremely obscure knowledge which most people will not be able to answer without guessing. Each of the 16 questions in the two tests are matched in terms of learning points, so any differential in overall scores will demonstrate how the construction of the item has influenced the result.

Anyone who likes doing general knowledge quizzes and who can spare 20 minutes of their time is invited to take part, and if you leave your email address we will send you a free copy of the report and findings. I will also endeavour to summarise the findings in a future issue of T & C news.