Repairing Items – How Post-Test Analysis Can Help

By Graham Barrow of GR Business Process Solutions

In my last article I wrote about the need to train item writers in order to ensure that items were not only written correctly but also tested the designated audience at the right level.

At the end of that article I mentioned that even with the best item writers and an effective edit committee it was still not possible to be certain how well an item would perform until it had been unleashed upon its unsuspecting audience. Of course the only way of telling if it has worked post-test is to carry out some form of analysis.

I have talked extensively in previous articles about both the need for, and how to perform, post-test analysis so I'm not going go over that again except to emphasise that even if you only have the percentages for how many candidate's answered each question correctly, that is a helpful start. Of course if you have the full panoply of facility, discrimination, distractor redundancy etc., so much the better.

Once this is to hand it is also helpful to agree the parameters outside of which items would be automatically earmarked for review. In my own company we use the following yardsticks: -

> Facility – any question achieving > 95% or < 50%
> Discrimination – any question achieving a negative discrimination score
> Redundant distractors – any question with 1 or more distractors chosen by fewer than 7.5% of candidates from the lowest scoring group (bottom quartile)

Of course, you can apply your own set of norms but these are a reasonable starting point and if you don't have this data, use whatever data you do have. The one thing you can be sure of, if items are achieving these types of scores in analysis, something is wrong. And if that is the case, the offending item needs to be either repaired or discarded before being let lose on the candidates again.

So once an errant item has been identified where do you go from there? My focus in this article is to help those responsible repair items wherever possible, as this is almost always cheaper and more efficient than throwing it away and starting again from scratch. However that will not always be the case, as we shall see.

The first place to start is with the syllabus. Often a question fails to register an analysis within the accepted parameters because it is testing a learning point which is outside of the syllabus to which the candidates have been trained. In these circumstances it would not be surprising if the item produced an anomalous analysis. If the question is testing a point that candidates are simply not being taught then there is very little that can be done to repair it and it should be discarded and replaced with an item that correlates to the syllabus.

If the item clearly does relate to a defined learning point within the syllabus then we need to look elsewhere for the issues causing the poor analysis. And the next place to look is at the item itself.

Very often, items are constructed that are “time critical” by which I mean that the answer can change over time because of changes to legislation, tax limits, minimum and maximum premiums etc.

Let me give you an example. Suppose you had a question in your bank along the lines of:

> Mrs McGillicuddy left an estate valued at £487,500. She has made two partially exempt transfers (PETs), one of £25,000 four years ago and a second amounting to £50,000 three years ago. Which ONE of the following amounts correctly represents the amount of inheritance tax payable on her death?

I’m not going to try and work up four examples but it only takes a minute to realise that a change to either the nil rate band or the full inheritance tax rate or, indeed, to the rules relating to PETs would immediately cause an answer that had been right to become wrong. Even worse, one of the distractors might have, unintentionally become the right answer which will really mess up the results.

If it is the case that legislative, budget or other changes have caused the item to become incorrect, this is itself a signal to check the item bank for other questions (that might not have been used since the changes came into effect) to be checked.

Once you have checked the item bank for currency (and amended where necessary), or if this is not the reason for the item to have performed poorly a bit more digging will be required. Let’s look at a few scenarios and see if we can work out what might be going on.

Scenario 1 – an item is answered correctly more often by candidates from the bottom quartile than the top.

Provided you are sure that the correct answer is being marked this is a certain indication that there is a deep flaw in the question. After all, it should never be the case that less able candidates answer a question correctly more often than more able candidates. So what might be causing the problem?

In our experience, the most common cause of this issue is a question that is worded in such a way as to make the more able candidate think that it is a “trick” question or certainly one that the answer is not what they consider to be the “obvious” one. This is almost invariably brought about by obscure or poorly worded items that have served to confuse (or it is a “double bluff” question, designed to look like a trick but not really – and I’ve seen those before!) The best way to remedy this item is to make the wording in the stem clearer and easier to understand.

Consider the following example:

> When is it not possible to have Waiver of Premium added to our XYZ policy?
>
> Option a) It is always possible to have Waiver of Premium added.

Now this situation may not exist in real life but bear with me for the sake of the article. This is the sort of question that those who are less able will take at face value and select option “a” but the very best people may well be able to conjure up a scenario where it was not possible to have WoP and therefore reject option "a" as being to *precise* an answer.

Experience suggests that the very brightest candidates tend to reject answers containing absolutes (always, never, must, must not, etc.) in favour of less specific options (usually, rarely etc.).

Scenario 2 – an item has been answered poorly (say less than >50% correct answers) and there are equal numbers of candidates form both the top and bottom quartile answering correctly.

This almost always indicates that candidates are, by and large, guessing the answer, hence the even spread of scores. This is mainly caused by items that have not been included in the training which candidates have received, is a piece of knowledge used so rarely that candidates have not retained it or the item is written so confusingly that no-one is quite sure what is being asked of them. Let me give you an example of the latter:

> When would it be false to say that you cannot have Waiver of Premium when taking out one of our XTZ policies?
>
> Option a) When a client has not been previously declined for

You see my point? There are three negatives being employed here, two in the stem and one in the option which makes it extraordinarily difficult to establish what is being asked of the candidate let alone the right answer.

Much simpler to ask:

> When can Waiver of Premium be added to one of our XYZ policies?

The other common reason that we see this type of analysis is where the question has been based on a particularly obscure piece of information about some product particulars which the vast majority of people being tested have never seen or had to use before.

The author well remembers seeing an item some years ago relating to Income Protection asking under what circumstances a policyholder travelling abroad outside of certain countries (all of which were listed in the stem) and unable to work for a period of more than nine months, would be unable to claim on their policy.

This was a scenario that was unlikely to crop up in real life and the item served no great value in the test (if it ever happened, you could not reasonably expect the Financial Adviser to know, by heart, the consequences of such an event). It's a bit like expecting a doctor to know every adverse reaction to every drug he ever prescribes off by heart. In fact I'd rather he looked it up – just to be sure.

Again if this is the case it may be that you have to discard the item in favour of one related to a more easily accessible or more generally understood learning point.

Sometimes, of course, an item will perform poorly and it will be difficult to see why. In this case we recommend setting up a small focus group to work through the rogue item(s) with you and talk through their reasons for responding the way they do.

This can be a very illuminating way of seeing how your items actually work in practice as it takes away your suppositions and replaces them with at least a degree of reality. Once you have gained some insight into how your audience is reading your items you can begin the re-engineering process.

For this we recommend pretty much the same process as authoring items (although it should be much quicker). Ideally the item writer should collaborate with a subject matter expert (except where both are combined) in order to ensure both the construction and technical content of the item are correct and, working with the results of the analysis or focus group, repair the item as necessary (e.g. reword the stem to make it clearer, replace obvious or misleading distractors etc.).

From there the item should go back to the edit committee for review prior to being signed off and inserted back into the item bank where its predecessor should be archived (preferably with a brief note to explain why it was withdrawn from the item bank, in case there is any future query).

And once all this has happened? Next time it is used in an exam you go through it all again!