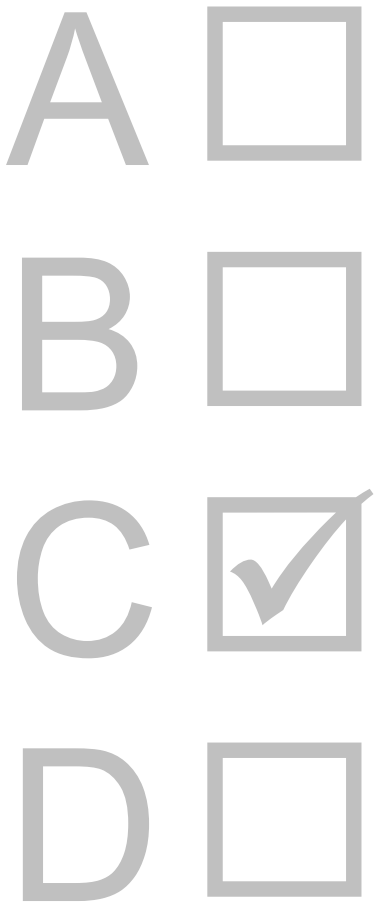


Guessing or Assessing?

Multiple Choice and the False Pass Problem

FREE PREVIEW EDITION

ORDER FORM FOR FULL VERSION INCLUDED AT END OF DOCUMENT

By Graham Barrow
and Ray Blake

GR Business Process Solutions

A ☐
B ☐
C ☑
D ☐

A ☐
B ☐
C ☑
D ☐

Guessing or Assessing?

Multiple Choice and the False Pass Problem

Graham Barrow and Ray Blake

GR Business Process Solutions

A ☐
B ☐
C ☑
D ☐

First published 2004 by GR Business Process Solutions

www.grbps.com

ISBN X XXXX XXXX X

A ☐
B ☐
C ☑
D ☐

CONTENTS

Where a page number is not shown, that section is not included in this free preview edition. Of those sections which are included, some have been abridged for inclusion here and the full report in that event contains more information

A ☐
B ☐
C ☑
D ☐

A ☐
B ☐
C ☑
D ☐

Section 1. Background to the research

Introduction to us

GR Business Process Solutions provides expert knowledge measurement services to firms, primarily within the Financial Services industry. Within these services we both advise and assist firms in the review and revision of question banks for multiple choice tests.

Why we initiated this research

We had been working in the field of multiple choice testing for some time. We had secured a number of commissions to audit question banks for organisations and businesses, and had created special software to help us analyse test data.

We knew that many of the questions we encountered in our work were more 'guessable' than others by those who did not actually know the answer. When we identified such a question, we could highlight its aberrant performance in the test bank and in most cases through a textual analysis of the question itself (the 'stem') and the offered answers (the 'options') we could repair the question to a state where it was no longer easily guessable.

Amongst all the techniques and analysis we were able to employ lay a nagging uncertainty, though. How likely was it that a poor test – one comprised of a large number of easily guessable items – might be passed by a person without the supposedly requisite knowledge for passing. In other words, how likely was a 'false pass', given either a well-constructed test or one of poor construction.

Academic research on multiple choice tests has consistently pointed to the fact that the construction of test items can (and does) influence the outcome of the result. It does this both positively (by clueing the correct answer) and negatively (by falsely clueing the wrong answer). There exists a great deal of guidance as to how MCQ items should be constructed (e.g. *Developing and Validating Multiple-Choice Test Items* by Thomas M. Haladyna) but we have been unable to find any empirical evidence to underpin this research.

This lack of empirical evidence can make it difficult for firms to agree to the expense of revising (or, in some cases, rewriting) their question banks when we have been unable to show them statistical evidence to support the need. We realised that if we wanted to cast some light on this issue, then we would have to light a torch of our own.

How we went about it

If we were to generate some sort of data for analysis, then we needed to think about people and about tasks for them to undertake. We tackled the task side of things first and decided to create two multiple choice tests. The first would represent a poorly constructed, highly 'guessable' test bank, and the second would be the control paper, constructed along best practice lines.

Because we wanted to look specifically at the effect of guessing, it was important to remove as far as we could the possibility of our people actually knowing any of the right answers. We wanted to see what sort of score they could obtain in areas they knew nothing about, purely by guessing, and how far we could create tests to both boost and limit their guessing performance. Consequently, we started gathering some of the most obscure knowledge we could locate. Each question in the guessing test bank would have a brother in the control test bank, asking about exactly the same field of knowledge, but in a differently-framed question. This would mean that, across a broad population of participants, any differences between average scores for the two tests would be due to clues provided by the construction of the items concerned.

A ☐
B ☐
C ☑
D ☐

Next we considered the people who would be our guinea pigs. Family, friends and professional colleagues could be relied upon to help, of course, but our close relationships might easily skew the results. We were also concerned to gather data from a sufficiently large experimental sample, and set ourselves a target sample of 200 test candidates. We needed a mixture of people – some who have taken such tests regularly, and some who never have, some in education and some outside, some younger, some older. In short, we needed a random sample of people.

In the old days, of course, this would have represented a problem. However, in this connected age we knew that we would simply have to make a web page available to facilitate test sittings and then publicise it. That is exactly what we did. To ensure a spread of different types of people, we publicised widely and after a small pilot exercise we made the web page available in September 2003. When we closed the tests down in November, we had exceeded our completion targets; 335 people had completed our control test and 365 had completed our guessable test.

To the extent that there is any bias in our sample of people, it is that people without internet access are excluded. We did not and do not consider this to represent a significant influence of the research outcomes given the present ubiquity of the internet both at home and in the workplace.

What we thought we'd find

When the web page was posted, we were fairly sure that we would have confirmation of our contention that a highly guessable test could be constructed and that such a test would give rise to false passes.

We also thought we would be able to identify the different types of question flaw which increase guessability and give some sort of empirical value to benchmark how far each type of flaw was likely to skew a guessed performance.

In the same way, we expected to validate how good question construction practice could limit guessability and thereby prevent widescale false passes.

A ☐
B ☐
C ☑
D ☐

Section 2. The test banks we constructed

Test bank 1 – the guessable questions

Each item and the rationale for its construction is given below. In each case, the right answer (or 'key option') is highlighted in italics.

1 Someone who is afraid of blushing is known as an…

A merinthophobe
B gephyrophobe
C ***erythrophobe***
D taphephobe

This question is clueing by grammatical construction. The use of a trailing 'an' clearly indicates that the answer should start with a vowel. Given that three of the options start with a consonant the clued-up candidate should be able to discount them.

2 The time it takes the earth to complete EXACTLY one orbit of the sun is…

A 365 days
B 365.2 days
C 365.24 days
D ***365.242199 days***

Use of the word 'exactly', particularly as it is capitalised, clues the most precise answer as correct.

3 In the USA, which was the 16th State to join the Union?

A ***Tennessee***
B Vancouver
C Alaska
D Hawaii

As it stands this question is not clueing at all. However the ability of a candidate to select the correct answer is affected by the quality of the distractors. Given that Vancouver is in Canada and the other two are widely known to have been the last to join the Union, the correct answer can be arrived at by discounting the other options. However this does not validate the candidate's knowledge of the learning point as tested. (The candidate knows what it *isn't* but not necessarily what it *is*.)

4 The process known as UHT entails the product being heat treated at a temperature of not less than 135 degrees Centigrade for…

A a very short time.
B a very short time and it has then been aseptically packaged in sterile containers.
C a very short time, aseptically packaged in sterile containers, and it should have undergone minimum chemical changes in relation to the severity of the treatment required for sterilisation.
D ***a very short time, aseptically packaged in sterile containers, and it should have undergone minimum chemical, physical and organoleptic changes in relation to the severity of the treatment required for sterilisation.***

Again, this question, as it stands, does not provide any clues. However the options as given build one upon the other so it is reasonable to suppose that the most complete answer is correct.

A ☐
B ☐
C ☑
D ☐

5 What are the third and fourth volumes in Robert Rankin's "Brentford" series?

A "East of Ealing" and "The Book of Ultimate Truths"
B ***"East of Ealing" and "The Sprouts of Wrath"***
C "The Brentford Triangle" and "The Sprouts of Wrath"
D "East of Ealing" and "The Suburban Book of the Dead"

This is known as a 'counting' question. 'East of Ealing' appears three times and 'The Sprouts of Wrath' twice so it could be presumed that these two books were uppermost in the item author's mind when compiling the question. On the other hand, Option C might be falsely clued as the only answer that does not contain 'East of Ealing' in the answer.

6 Robert Banks Jenkinson's term as British Prime Minister was notable for…

A expensive foreign wars
B colonial expeditions
C ***the ruthless suppression of free speech and the freedom of the press***
D education reforms

This question has one answer that is (slightly) longer than the other three options. Traditionally, one method of guessing (in the absence of knowing the answer) is to choose the longest option.

7 Which law sought to prevent discrimination against employees or job applicants on the grounds of their sex?

A Family Law Reform Act 1969
B ***Sex Discrimination Act 1975***
C Race Relations Act 1976
D Administration of Justice Act 1982

This type of question commits the cardinal sin of lifting too many significant words from the key option into the stem. Given that one of the options contains two words that are found in the question (and the other options between them contain none), this option is bound to be strongly clued.

8 What is the atomic number of Antimony?

A 1
B 8
C 298
D ***51***

This question contains two clues. One is that where numbers are non-contiguous or follow nether any mathematical order nor logic, candidates will tend to opt for a 'middle-of-the-road' answer. Secondly, more knowledgeable candidates, whilst not knowing the answer, may be aware that Atomic Numbers 1 and 8 are common elements (Hydrogen and Oxygen) whilst there are not 298 elements in the periodic table.

A ☐
B ☐
C ☑
D ☐

9 An increase in the velocity of fluid flowing through a pipe is…

A ***likely to give rise to turbulence***
B unlikely to give rise to turbulence
C solely dependent on the length of the pipe
D impossible

Where two available options directly contradict each other the likelihood is that one of them is correct. Given that most questions are looking to test a positive effect then option A is clued.

10 Of alpha, beta and gamma rays, which is a hydrogen nucleus?

A Alpha ray
B Beta ray
C Gamma ray
D ***None of these as they are not Protons***

A 'non-sequitur' answer (as in (d) here) is likely to be strongly clued as it differs so strongly from the other three offerings. This is often indicative of a trick question.

11 The main constituents of natural gas are…

A Helium
B Hydrogen
C ***Methane and Ethane***
D Oxygen

This stem contains another grammatical clue. Use of the word 'are' in the question stem clearly indicates that the correct option will contain more than one word. As only one of the options conforms to this, it is strongly clued.

12 What are the main ingredients of the liqueur Quetsch?

A Orange
B ***Plums***
C Pear
D Apple

Similar to question 11 above this question stem is also clueing a plural answer. Unlike (11) all the options are one word; however three are singular whilst the fourth is plural providing a strong clue to the correct answer.

13 Which Italian painter produced a series of paintings for the Scuolo di San Rocco in Venice?

A Baldung
B ***Tintoretto***
C Altdorfer
D Dumoussier

This question contains strong clues as to the identity of the painter concerned. Given that only one of the available options is identifiably Italian in nature it is strongly clued.

14 In highly successful South African trials, the inoculant treatment of big round bale silage resulted in…

A ☐
B ☐
C ☑
D ☐

A ***increased milk production***
B reduced milk production
C The effects were not predictable
D There was no effect

There are two main clueing elements here. Use of the phrase 'highly successful' indicates that the answer should contain a strongly positive outcome. Additionally, two of the options (C and D) do not flow logically from the question stem (i.e. they do not complete the sentence begun in the stem.)

15 What PERCENTAGE of the atmosphere is Nitrogen?

A 1/2
B three quarters
C ***79%***
D All of it

This is another grammatically clued option. This time the question is clearly looking for an answer given as a percentage. Only one of the answers meets that criterion and is therefore clued.

16 The transition from the Triassic to the Jurassic periods was marked by…

A ***a dramatic change from high-diversity palynofloras to assemblages almost entirely composed of cheirolepidaceous conifers***
B a major upsurge in marine fossils
C an increase in ammonoid cephalopods
D the disappearance of the continental tetrapods

This is another 'weight' question where one option is of a markedly different length (and contains much greater technical detail) than the other three, and is therefore clued.

Test bank 2 – the control questions

In the companion test, each of the questions was designed to match the equivalent question in Test Bank 1.

Here are the questions with some brief notes relating to the construction of the four options.

1 Someone who is afraid of beards is known as a…

A coprastasophobe
B ***pogonophobe***
C doraphobe
D cibophobe

Another unusual phobia although this time the stem does not provide any grammatical clues and the options are of similar length.

A ☐
B ☐
C ☑
D ☐

2 The time it takes the earth to complete EXACTLY one rotation on its axis (a sidereal day) is…

A 23 hours 56 minutes 2 seconds
B 23 hours 56 minutes 3 seconds
C ***23 hours 56 minutes 4 seconds***
D 23 hours 56 minutes 5 seconds

Matched to the question relating to the time it takes the earth to orbit the sun except this time the options are of equal weight.

3 In the USA, which was the 17th State to join the Union?

A Louisiana
B Indiana
C Mississippi
D ***Ohio***

The options are the 17th, 18th, 19th and 20th states to join the Union (but not respectively!)

4 In relation to milk, homogenisation is…

A a bacterial heat treatment. The milk is heated to a temperature of 72.2 degrees centigrade for 15 seconds, effectively killing all harmful bacteria and making it safe for humans to consume for the next five days.
B ***the process where fat globules in the milk are broken down into smaller pieces, which are dispersed through the milk. This process improves both texture and taste and also extends shelf life.***
C milk where the fat content is taken out and then replaced at a consistent fat percentage to give different fat contents of milk and dairy products.
D milk that has been heat treated at a temperature of 150 degrees centrigrade for a few seconds. The milk is then transferred to airtight cartons.

Designed to match the question on UHT treatment, this contains four plausible (and real) treatments to which milk is subjected. They are Sterilisation, Homogenisation, Standardisation and UHT.

5 The second and third volumes in Stephen Donaldson's "Thomas Covenant" series are:

A "The Wounded Land" and "Lord Foul's Bane"
B ***"The Illearth War" and "The Power That Preserves"***
C "Lord Foul's Bane" and "The Illearth War"
D "The Power That Preserves" and "The Wounded Land"

A sequence of books of similar readership as Robert Rankin's "Brentford" series except this time each of the offerings appears twice, once as the first and once as the second element of the option.

6 William Petty FitzMaurice's term of office as British Prime Minister was notable for…

A reforming Britain's economic policy
B the implementation of a successful agricultural policy
C reorganising parliamentary constituency boundaries
D ***bringing an end to the American Revolution***

This is a question about another obscure British Prime Minister and four plausible outcomes of his premiership.

7 Which law introduced the requirement for remuneration for women's employment to come into line with that set for men doing the same job?

A ☐
B ☐
C ☑
D ☐

A Employment Rights Act 1982
B Equal Opportunities Act 1979
C Sex Discrimination Act 1975
D ***Equal Pay Act 1970***

This is similar to the item in Test Bank 1 relating to Sex Discrimination except that each of the four options could plausibly have enacted the relevant legislation. There is not the same language clue of words in the stem directly echoed in the key option.

8 What is the atomic number of Osmium?

A 73
B 74
C 75
D ***76***

Unlike the question relating to the atomic number for Antimony, the options here are sequential and in order.

9 The volume of a fixed amount of gas is…

A directly proportionate to the effects of the gravity acting upon it
B ***inversely proportionate to the total amount of pressure applied***
C equal irrespective of the gas
D equal to the equvalent mass of the gas when in its liquid state

A highly technical question with four, seemingly, technical options.

10 Which of these is NOT a 'flavour' of quark?

A ***Left***
B Up
C Top
D Strange

A very obscure question from Quantum physics with four equally unlikely options.

11 The main constituents of town gas are hydrogen and…

A ***carbon monoxide***
B carbon dioxide
C nitrogen
D methane

Unlike the matching question in Test Bank 1 there are no grammatical clues in the question stem and all the options are constituents of town gas (although some are only in small quantities).

12 The liqueur cachaca is made from what primary constituent?

A plums
B barley
C ***sugarcane***
D oranges

A ☐
B ☐
C ☑
D ☐

This item has four choices with no grammatical clues.

13 Which painter created the Cubist picture 'De Meidagen' in 1940?

A ***Hulsbergens***
B Hamdorff
C Heijnckes
D Nibbrig

In its way, this is almost the perfect question to test people's ability to guess. With all due deference, this painting is virtually unheard of, as are the artists. This time, though, there is a commonality in terms of the apparent nationality of the names.

14 Which of the following food additives when combined with molasses is most likely to lead to an increase in cows' milk production?

A palm oil
B ***halibut oil***
C groundnut oil
D whale oil

This question was designed to match the question relating to the successful trials of the 'inoculant treatment of big round bale silage' in question one. The authors were both surprised (and somewhat relieved) to be able to find an equivalent question to match it with.

15 What percentage of the earth's crust is made up of Iron?

A ***6%***
B 7%
C 8%
D 9%

As for question 2, this is a fairly obscure piece of knowledge accompanied by four options in ascending order.

16 Sauropods reached the zenith of their evolution during the…

A permian-triassic transition
B triassic-jurassic transition
C ***jurassic-cretaceous transition***
D cretaceous-tertiary transition

This is another obscure piece of knowledge with four plausible options. It would be very hard to find any clues within the stem or options to help with a guess.

These then were the two sets of questions with which we embarked on the research. We firmly believe that, if a representative audience were to be asked these questions without the benefit of a multiple choice format (i.e. they had to provide their own answer) the resulting percentage of correct scores would be less than 1%. This means that, within normal statistical bounds, any differences in the overall scores across the two tests can be attributed to the manner of the test's construction and not variance in candidate knowledge.

As far as possible, we have sought to ensure that if a candidate KNOWS the answer to any question in Test Bank 1 (as opposed to being able to work it out from the stem and options) he or she is equally likely to KNOW the answer to the matched question in Test Bank 2.

We distributed the key option fairly evenly between positions A, B, C and D. In Test Bank 1, there are 4 in each position, but in the control test, Test Bank 2, we distributed them differently so that position C actually has only 3 key options, balancing this by allocating 5 to position B. This adjustment was to subject to analysis the commonly-held belief that a clueless candidate is best advised to choose option C in any case. There is much folklore surrounding this principle, and we were keen to see whether our sample group was aware of and acted upon this folk wisdom.

A ☐
B ☐
C ☑
D ☐

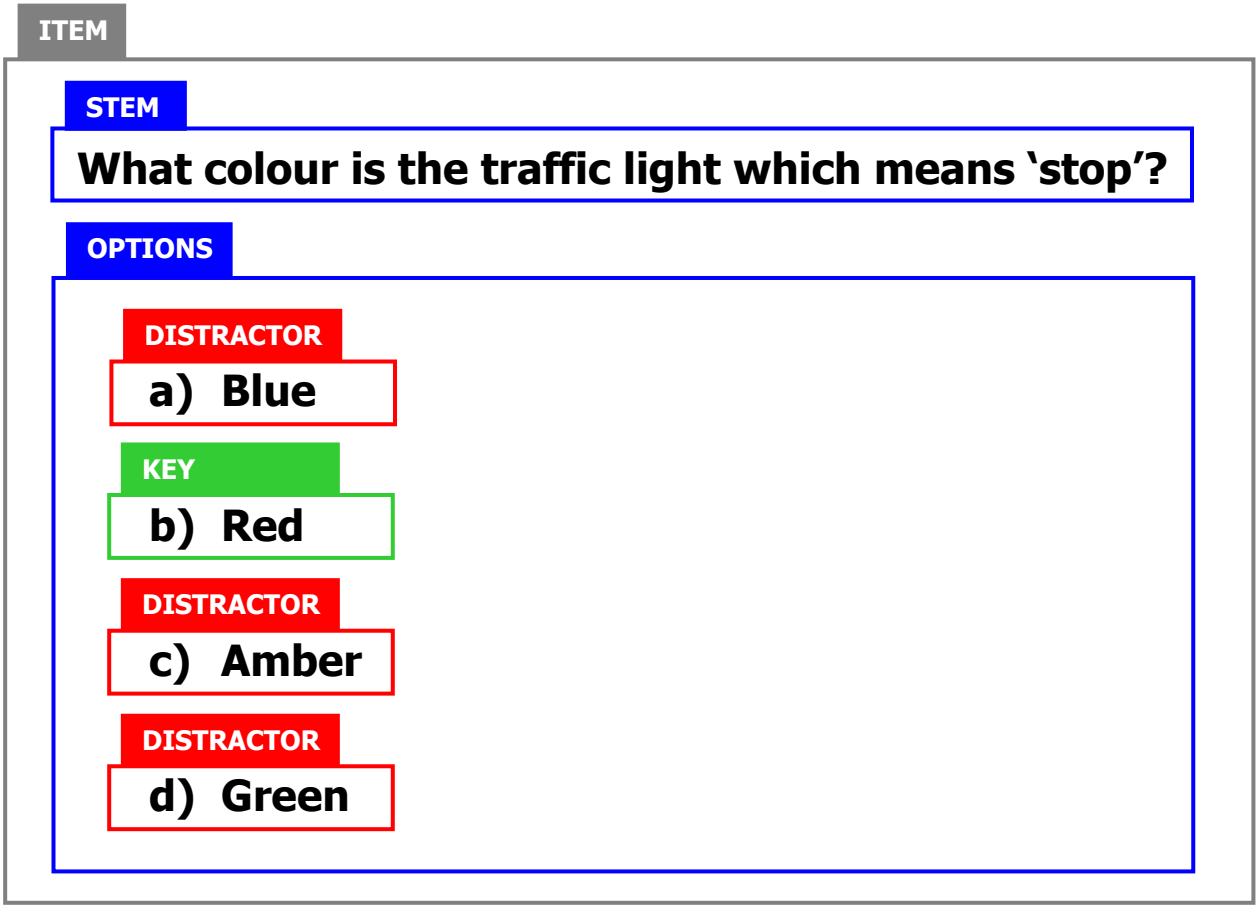
Section 3. Gross analysis

Defining terms

To ensure full understanding of the analysis which follows, it will be worthwhile to define some of the technical terms commonly used in the field.

Anatomy of an item

Item	*- a multiple choice question (MCQ) stem together with four options, three of which are distractors and one of which is the key*
Distractor	*- an incorrect option within an MCQ item*
Key	*- the correct option within an MCQ item*
Stem	*- the question element of an MCQ item*



Facilty

Facility factor	*- a decimal between 0 and 1 which expresses the proportion of candidates who correctly answered the question (e.g. a Facility factor of 0.6 means that 60% of candidates answered the question correctly)*
Facility Index	*- the questions forming a test ordered by Facility Factor with the question answered correctly by the fewest candidates being first*

Facility is the accepted measure of how easy or hard a question is, based on how often it has been answered correctly by candidates. Clearly, the higher the figure, the easier the question was found to be by the sample group of candidates. The bigger this group is, the more reliable the figure derived, as with all statistical methods. Our general view is that item statistics based on a sample of fewer than 50 candidates should be regarded with caution.

A ☐
B ☐
C ☑
D ☐

Facility as defined here related to individual items, but it is possible to find the mean facility of all questions in a test or a test bank.

Discrimination

Discrimination factor	*- a decimal between 0 and 1 which identifies the difference between how many of the top 27% scorers in the test overall answered a question correctly compared to the bottom 27%.*

Discrimination is a major reason for why we conduct tests in the first place. It allows us to distinguish between more and less able candidates and this distinction should apply whether the test is measuring absolute or relative values. The reason for this is that however well matched a group of people being examined is, there must be some variance in ability and the test should demonstrate this even if all the candidates (in an absolute test) achieve a pass mark.

In order to understand why we look at the top and bottom 27% we need to explain a little more about what discrimination is trying to do.

We take the results achieved on any question by a group of people who have performed best in the test overall (i.e. the candidates with the best scores) and compare them to the results achieved by the group who have done least well (the candidates with the lowest scores). It is a *sine qua non* that each question should be answered correctly more often by the top group than the bottom as it should never be the case that less able candidates demonstrate more knowledge in the subject than more able candidates.

Clearly there is a minimum size that should apply to this measurement as very small numbers are more likely to skew the result than a larger group would.

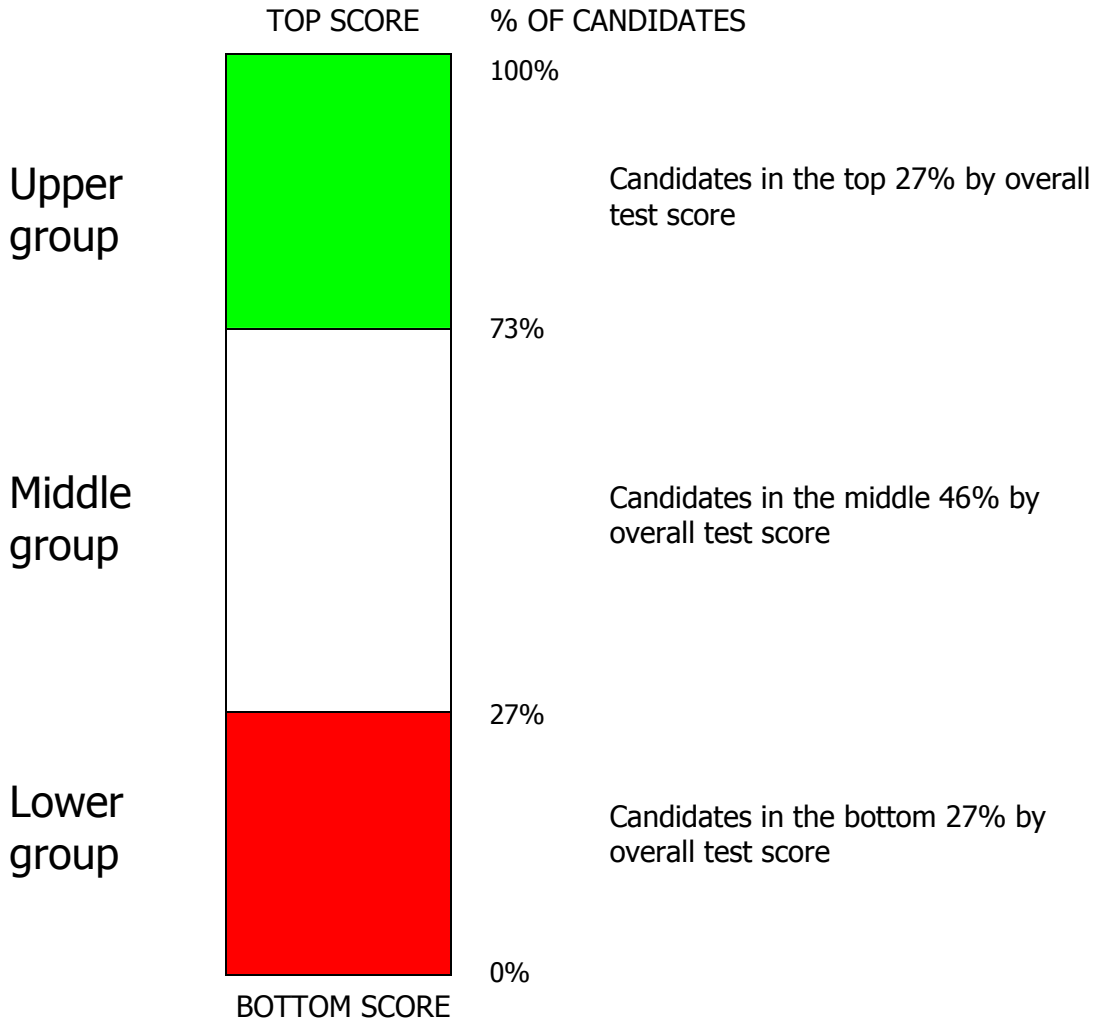
An example of this effect might be the outcome of repeated tosses of a coin. Whilst over time the ratio of heads to tails will always tend to 50:50 over a small number of coin tosses it could easily be, say, 100:0 or 75:25.

For this reason we believe it unlikely that any great reliance can be placed on discrimination for numbers less than 100, particularly as this group needs to be subdivided in order to have an upper and lower group. And this is where the 27% figure comes in.

For discrimination to have significance it is important that each group is composed of a statistically significant number of the overall candidates but it is also important that there is as wide a separation between the groups as possible. Over time this figure has settled down to a range of 25% - 30% and the most commonly used percentage amongst these is 27% and this is the one we have chosen, both to use in our software system (iOTA) and to apply to this research.

The diagram overleaf illustrates this principle graphically.

A ☐
B ☐
C ☑
D ☐



We can compare how well the Upper group and the Lower group performed in any given question, and conclude from this how well the questions discriminates between more able and less able candidates.

For instance, if 80% of the Upper group answered a question correctly compared to 50% of the Lower group, the item would have a discrimination factor of 0.3. Questions should, self evidently, always have a positive discrimination factor and, ideally, this should be more than 0.25

A test composed entirely of high discriminators will be very effective at favouring more able candidates. As we shall show, however, 'more able' may not always equate with 'more knowledgeable'.

Distractor analysis

Distractor Analysis	- *a measure of how many candidates selected each of the four options within an item*

Distractors, of course, exist to distract, to present credible alternatives so that candidates cannot select the correct answer other than by possessing the requisite knowledge or by pure chance. Where there are four options, there is a one in four chance of a guesser selecting the correct answer.

We would expect most guessing to be taking place in the Lower group, of course. If the distractors are working well, then wrong answers in the Lower group should be spread roughly evenly between all of them. Where a distractor is selected by fewer than 7.5% of candidates in the Lower group, we classify this distractor as redundant, since it is not proving sufficiently credible.

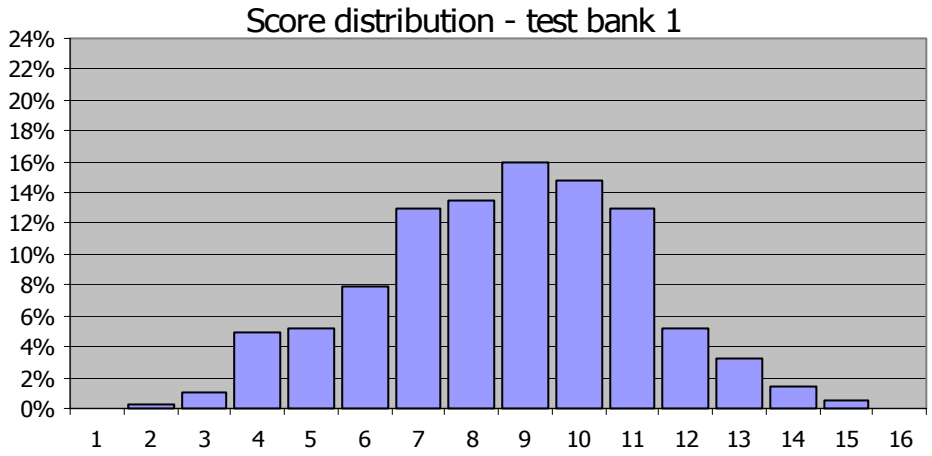
A ☐
B ☐
C ☑
D ☐

As we shall see, a significant incidence of redundant distractors in a test can seriously compromise its effectiveness.

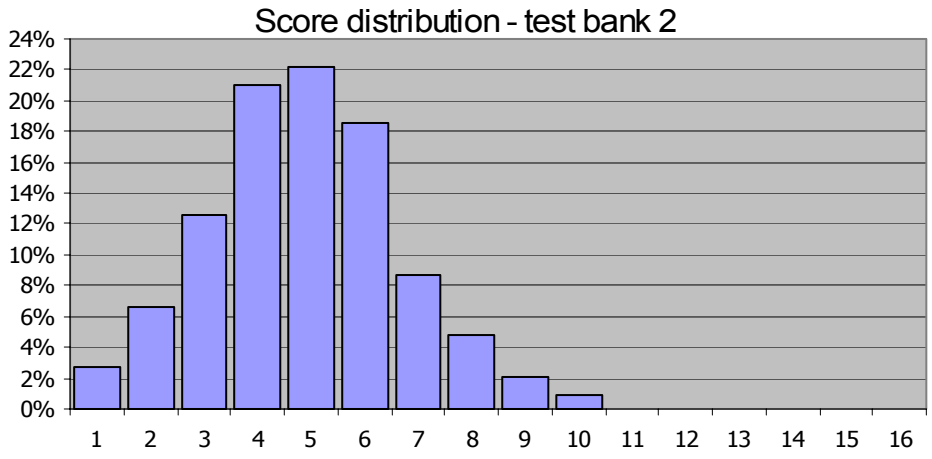
Headline statistical findings

Before considering the item bank in detail we want to present the overall performance figures as we believe this will help to give context to the detailed analysis.

Let's start off with the overall score distribution charts. For Test Bank 1 it looked like this:



And for Test Bank 2:



Some of the differences will be immediately obvious but, for the sake of clarity, we would like to talk through them in some detail.

Firstly, and most encouragingly, both sets of results look like they follow a normal distribution curve or, as it is more commonly known, a bell curve.

A ☐
B ☐
C ☑
D ☐

False Passes

One of the central points of conducting this research was to try and measure the effects of guessing on pass rates. For example, if we required candidates to get half of the questions right, how many would have achieved this standard?

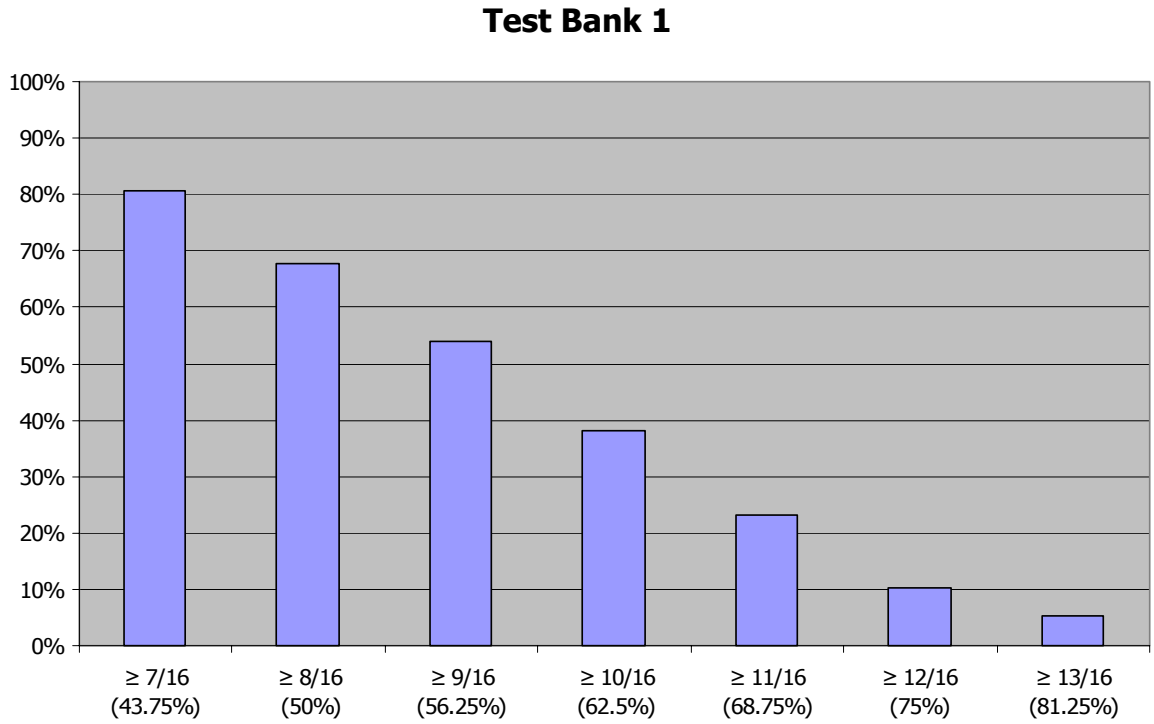
Bearing in mind that the tests did not gauge knowledge, but purely guesswork (albeit aided in one case by a variety of structural clues in the questions) any "pass" awarded would have to count as a false pass. That is to say, we would be ascribing to a passing candidate a level of knowledge or technical competence which that candidate did not in actuality possess.

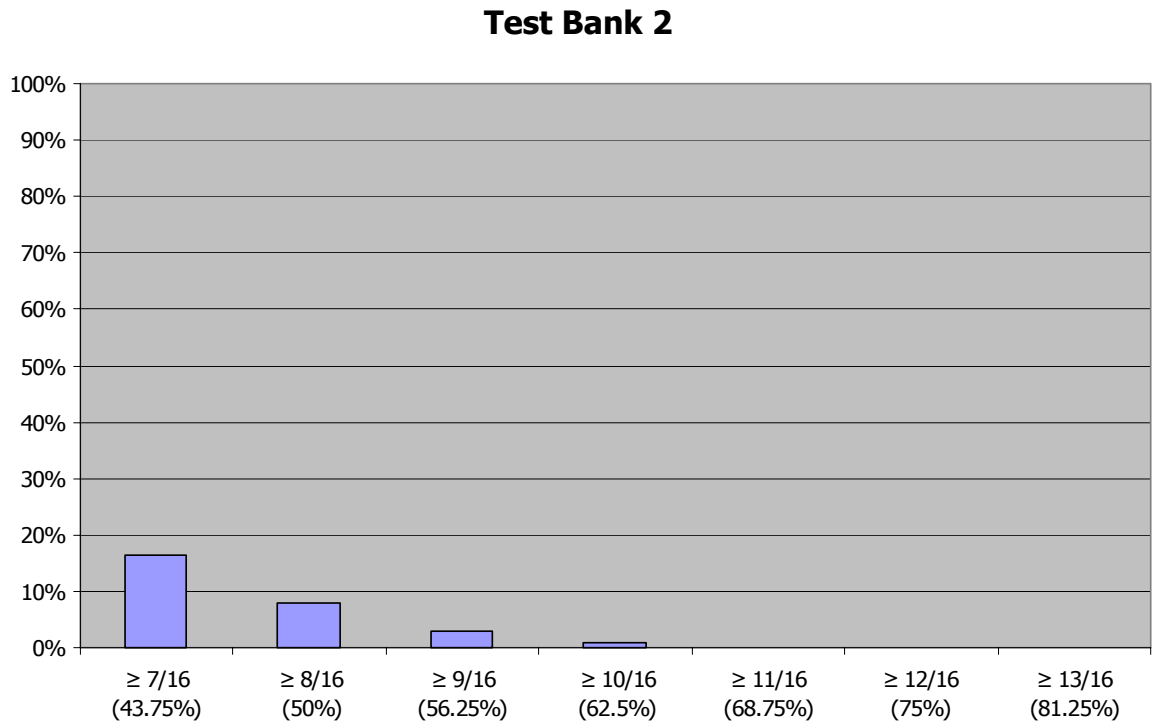
We have tabulated the results against a series of different pass marks to see what percentage of our respondents would have achieved a "pass" if this had been a real test.

The result for each of the Item Banks was as follows:

Pass mark		Test Bank 1		Test Bank 2	
		Number achieving	% of candidates	Number achieving	% of candidates
≥ 7/16	43.75%	294	80.55%	55	16.47%
≥ 8/16	50.00%	247	67.67%	26	7.78%
≥ 9/16	56.25%	197	53.97%	10	2.99%
≥ 10/16	62.50%	139	38.08%	3	0.90%
≥ 11/16	68.75%	85	23.29%	0	nil
≥ 12/16	75.00%	38	10.41%	0	nil
≥ 13/16	81.25%	19	5.21%	0	nil

When converted to bar charts the same data looks like this:





It is fair to say that we were shocked at these results.

At a pass mark of around 45% (not uncommon in Higher Education) 80% of our respondents to Test Bank 1 would have passed – by guessing, whereas in Test Bank 2, this figure drops to just over 16%. Even at a pass mark of around 70% (a figure commonly applied in Financial Services tests) almost a quarter of our participants would have succeeded, whereas in Test Bank 2 nobody managed to score more than 62.5% and only a tiny fraction managed to score above 50%.

This is a really startling difference. Given that many of our respondents gave us feedback to the effect that they found both tests extremely difficult, the fact that a massive difference appears in the results between the two, even though the majority of our respondents were unaware of any difference, gives us a great deal of concern.

We can only surmise how many people have been passed as competent, across a wide range of disciplines, when in fact their score has been a manifestation of the manner in which the test was constructed and not a measure of their intrinsic knowledge of the subject under test.

Overall Facility and Discrimination

Let's now turn our attention to the overall performance of the items in the two banks in terms of their facility and discrimination factors.

The data for each of the tests is shown overleaf.

A ☐
B ☐
C ☑
D ☐

Test Bank 1

Question number	Facility factor	Discrimination factor	Index no.
6	0.14	0.26	1
5	0.19	0.22	2
16	0.26	0.37	3
4	0.29	0.22	4
8	0.38	0.49	5
1	0.49	0.39	6
2	0.52	0.57	7
12	0.52	0.41	8
15	0.58	0.63	9
9	0.59	0.38	10
10	0.65	0.32	11
14	0.66	0.53	12
3	0.75	0.41	13
11	0.82	0.32	14
13	0.84	0.32	15
7	0.94	0.09	16

Test Bank 2

Question number	Facility factor	Discrimination factor	Index no.
6	0.10	0.08	1
14	0.12	0.17	2
13	0.15	0.11	3
10	0.16	0.19	4
5	0.19	0.19	5
8	0.21	0.30	6
3	0.26	0.37	7
15	0.26	0.22	8
11	0.31	0.31	9
2	0.32	0.37	10
1	0.33	0.29	11
16	0.39	0.18	12
4	0.43	0.41	13
9	0.48	0.49	14
7	0.50	0.30	15
12	0.63	0.34	16

Again, it is quite obvious from the data that the two test banks performed quite differently. The range of facility in Test Bank 1 was 0.14 to 0.94 (80%), whilst in Test Bank 2 it was 0.10 to 0.63 (53%). The lowest and highest discrimination factors recorded for Test Bank 1 were 0.09 and 0.63 respectively (a range of 0.54) whilst for Test Bank 2 these figures were 0.08 to 0.49 (a range of 0.41).

Overall the average facility for Test Bank 1 was 0.54 and for Test Bank 2 it was 0.30. The average discrimination for each test was 0.37 and 0.27 respectively. Test Bank 1 was easier and discriminated more effectively between those candidates who were able to detect (either consciously or subconsciously) the clues embedded within the items.

Internal test bank consistency

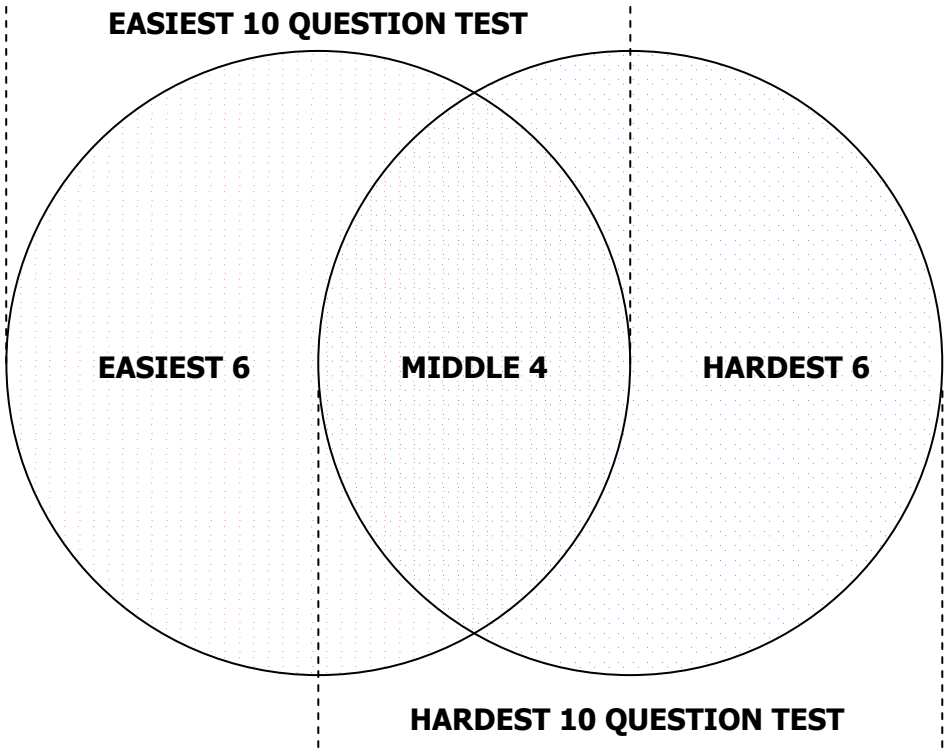
A ☐
B ☐
C ☑
D ☐

Each of our two test banks contains 16 questions, and we did not arrive at this number by chance. Ideally, we would have been able to construct much larger test banks in a wider range of subjects and with far more of each type of fault. However, each would need a control question and we were ever mindful that we could not intrude too far into the time of our volunteer candidates, who might abandon our tests if they took too long to complete. This did turn out to be a valid concern, since although 365 people completed the first test, only 335 of those people went on to complete the second test. It seems our instinct to keep the bank as small as possible was right.

So why 16 rather than 15 or 20, for instance? To answer that question, it is necessary to consider the way many multiple choice tests are deployed and used. For instance, in the UK financial services industry, where the majority of our clients operate, such tests are used widely in helping firms meet their regulatory obligation to prove their employees are competent in a range of knowledge areas. Typically, this operates on a '3 strikes and out' principle. If an employee fails the first sitting, he is allowed another. If he fails that sitting, he is allowed a third and final chance. Accepted practice is that no more than 70% of the questions are common between sittings. Therefore, to fully support a 10 question test, it is necessary to have 16 questions – the original 10, plus 6 extras to rotate out 30% of the test questions for each of the two subsequent tests.

Of course, once people begin taking different versions of the test, there is bound to be a suspicion that some people get easier tests than others, and that the luck of the draw can give unfair advantages. Consequently it is important that a bank provides as far as possible a uniform facility factor.

To analyse how our two test banks performed, we decided to calculate the overall average facility of both the hardest and easiest 10 question papers which could be constructed from each of them. To create the hardest test, we used the 10 questions with the highest facility factors. Choosing the 10 with the lowest facility factors created the easiest test. The middle 4 questions in the bank in terms of facility were common to both papers, as represented in the diagram below.



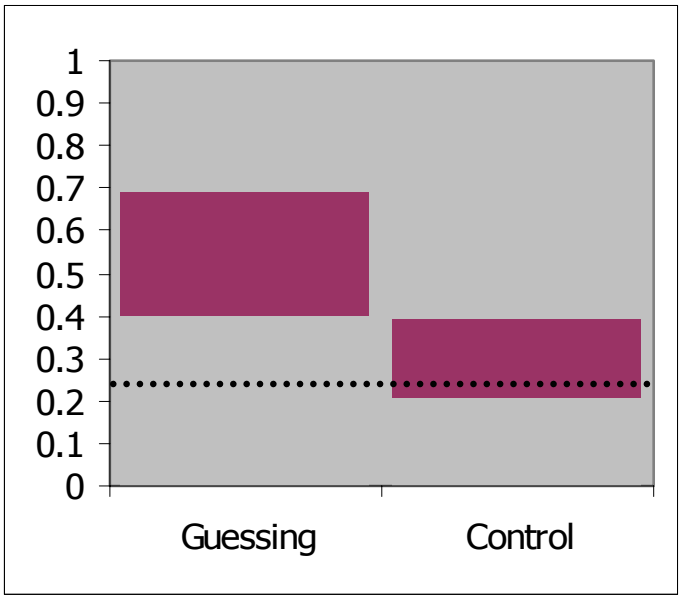
A ☐
B ☐
C ☑
D ☐

The closer together the easiest and hardest tests performed, the more fair and reliable they would be. People who deserved to pass and people who deserved to fail would be most consistently graded accordingly by a test bank with low variance.

Here is how our two item banks performed in this analysis:

	Average facilities	
	Guessing test bank	Control test bank
Easiest 10	0.687	0.391
Hardest 10	0.396	0.208
Variance	**0.291**	**0.183**

Charting this performance gives some interesting perspectives on the test banks.



The two bands represent the variance for the two tests, whilst the dotted line represents the mean facility which would be attributable purely to unclued guessing (0.25 given that there are 4 options offered for each question.)

Three main observations arise from this analysis:

1. The variance for the guessing test bank is nearly twice that for the control test bank, indicating that the control test bank will produce a fairer and more reliable test with a more predictable facility.
2. The lowest facility for the guessing test bank is greater than the **highest** facility for the control bank, indicating that for all possible tests which can be created from the banks, the guessing test will be easier to pass.
3. The variance for the control test bank is around the expected unclued mean facility of 0.25, whereas because the variance for the guessing test bank begins well above this level, no test can be constructed from that bank which meets the expected unclued mean facility.

A ☐
B ☐
C ☑
D ☐

Section 4. Question by question analysis of the test

Item analysis methods

Using our own iOTA (intelligent Objective Test Analysis) software, we subjected the individual items to a battery of tests. We have included the raw iOTA analysis for each question here, together with our commentary on the results. An iOTA analysis is laid out as the example below:

Question number:	12	**A**	Facility index position:	8
Correct answer:	B		Facility factor:	0.52

Number of people selecting each option:

B

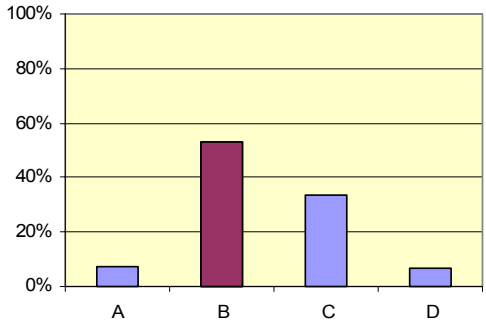
	A	**B**	C	D
Upper group	3	**73**	19	4
Lower group	11	**32**	47	9

Discrimination index scoring:

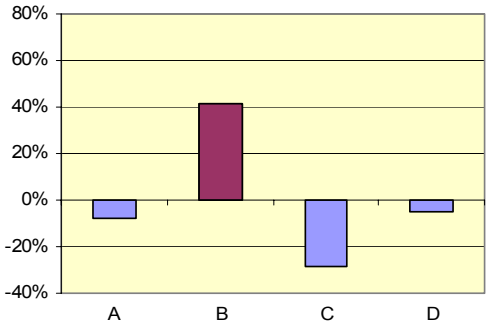
C

	A	**B**	C	D
Upper group	0.03	**0.74**	0.19	0.04
Lower group	0.11	**0.32**	0.47	0.09
Discrimination index	-0.08	**0.41**	-0.28	-0.05

D **Option distribution**



E **Option discrimination**



Analysis of question: **F**

Strength of discrimination	*Strongly positive*	*Positive*	*Nil or negative*
	●		
Appropriateness of distractors	*All OK*	*One redundant*	*2+ redundant*
			●

A ☐
B ☐
C ☑
D ☐

Key to analysis

A Basic details of the question, identifying the key option and facility data

B A breakdown of numbers of candidates selecting each option from the Upper Group and from the Lower Group. This forms the basis of the discrimination calculations

C The discrimination indices for each option of the question

D Option distribution chart – graphically depicts the figures at area B

E Option discrimination chart – graphically depicts the figures at area C

F Dashboard analysis of the item in question (see details of dashboard drivers below

Dashboard Drivers

Strength of discrimination:	A DI of 0.25 or higher is considered strongly positive, indicating that this question was significantly more likely to be answered correctly by generally more able people. Of course, in the context of a guessing test, discrimination is not particularly desirable, since it indicates a skill at 'beating the system' by spotting clues.
Appropriateness of distractors:	A distractor is deemed inappropriate if selected by fewer than 7.5% of the Lower Group. Redundant distractors are a problem because if any answers are easily discountable by most candidates, then a proportionately higher number of candidates will select the right answer without actually knowing the underlying learning objective. In terms or probability, they have increased their odds from one in four to one in two, if two distractors are patently implausible.

For each of the sixteen question pairs, we will show the iOTA analysis for the version from Test Bank 1 and the version from Test Bank 2. We will show the question again in each case for ease of reference and will then give a commentary looking at the analysis for the two alternative questions.

A ☐
B ☐
C ☑
D ☐

Question 1 – Phobias

Test bank 1 version – guessable question

1 Someone who is afraid of blushing is known as an…

A merinthophobe
B gephyrophobe
C ***erythrophobe***
D taphephobe

Question number:	1	Facility index position:	6
Correct answer:	C	Facility factor:	0.49

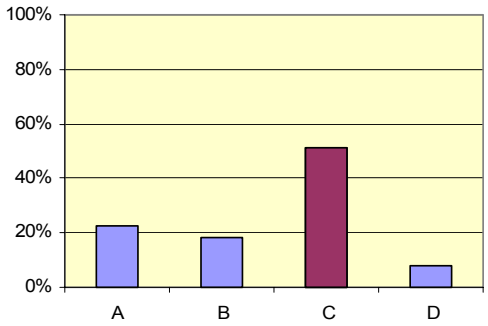
Number of people selecting each option:

	A	B	**C**	D
Upper group	16	11	**70**	2
Lower group	29	25	**31**	14

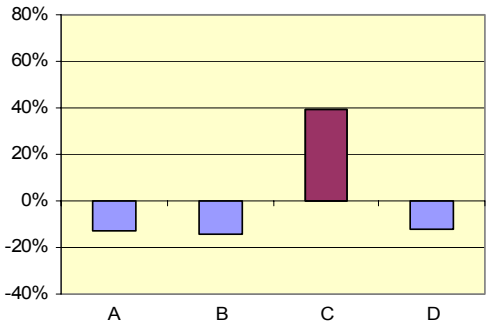
Discrimination index scoring:

	A	B	**C**	D
Upper group	0.16	0.11	**0.71**	0.02
Lower group	0.29	0.25	**0.31**	0.14
Discrimination index	-0.13	-0.14	**0.39**	-0.12

Option distribution



Option discrimination



Analysis of question:

Strength of discrimination	*Strongly positive*	*Positive*	*Nil or negative*
	●		
Appropriateness of distractors	*All OK*	*One redundant*	*2+ redundant*
	●		

A ☐
B ☐
C ☑
D ☐

Test bank 2 version – control question

1 Someone who is afraid of beards is known as a…

A coprastasophobe
B ***pogonophobe***
C doraphobe
D cibophobe

Question number:	1	Facility index position:	11
Correct answer:	B	Facility factor:	0.33

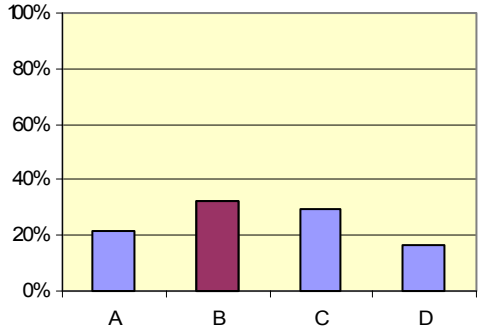
Number of people selecting each option:

	A	**B**	C	D
Upper group	14	**42**	18	16
Lower group	25	**16**	35	14

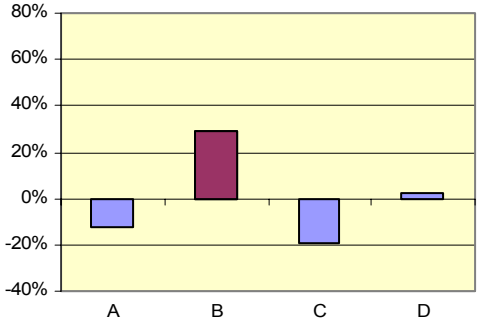
Discrimination index scoring:

	A	**B**	C	D
Upper group	0.16	**0.47**	0.20	0.18
Lower group	0.28	**0.18**	0.39	0.16
Discrimination index	-0.12	**0.29**	-0.19	0.02

Option distribution



Option discrimination



Analysis of question:

Strength of discrimination	*Strongly positive*	*Positive*	*Nil or negative*
	●		
Appropriateness of distractors	*All OK*	*One redundant*	*2+ redundant*
	●		

Commentary

A ☐
B ☐
C ☑
D ☐

The attractiveness and plausibility of the options was identical for both the upper and lower groups, although clearly the proportion selecting each option varied. In order of popularity the choices were:

C) ***Erythrophobe*** – 101 (51%)
A) Merinthophobe – 45 (23%)
B) Gephyrophobe – 36 (18%)
D) Taphephobe – 16 (8%)

Approximately 40% more of the upper group chose the correct option than the lower group.

This is one of those questions that, once the clue has been pointed out (the trailing 'an…' in the stem indicating that the correct option starts with a vowel) candidates find the answer blindingly obvious. However it is clear from the results that the ability of candidates to spot this clue varies widely and anecdotal evidence from some of our candidates would seem to indicate that, even where the correct answer has been chosen, it is not always because the candidate has *consciously* identified the clue.

Many of our candidates, when asked why they had chosen option (C) replied that it 'looked right' or that they just 'knew' it was the right answer. We believe this probably indicates that they had registered the verbal clue subconsciously.

Option (D) was clearly the least attractive although it is hard to formulate a good reason why this should be so. The only identifiable difference that we can establish is that the first three options all have four syllables whilst the fourth has only three. Whether this mismatch had any influence on choice selection however, it is difficult to say.

Nevertheless looking at the distribution in the control test we see the following:

B) ***Pogonophobe*** – 58 (32%)
C) Doraphobe – 53 (29%)
A) Coprastasophobe – 39 (22%)
D) Cibophobe – 30 (17%)

Unlike for Test Bank 1, the lower group selected in a different order, which was C, A, B, D.

It is readily apparent that this question has achieved a much more even distribution, although again option (D) is the least favoured. We would expect most, if not all, of the questions in the control test to show some element of discrimination because, as explained in the gross analysis, there was a much lower standard deviation recorded for this test which meant that only two marks separated the lower and upper groups.

Clearly, given that the average score for the control test was 5/16 and a score of 6/16 would probably place the candidate in the upper group answering any question correctly has a significant impact on the overall analysis.

For this reason, we believe that the overall distribution varies little from chance and that there is no obvious clueing taking place.

A ☐
B ☐
C ☑
D ☐

A ☐
B ☐
C ☑
D ☐

ORDER FORM

Send to:

GR Business Process Solutions
Braedon
Newell Road
Hemel Hempstead
Herts HP3 9PD
United Kingdom

Guessing or Assessing?

Multiple Choice and the False Pass Problem
by Graham Barrow and Ray Blake

Number required:	Cost each:	Total cost:
	£39.00	

Make cheques or Postal Orders payable to ***GR Business Process Solutions***. Only sterling cheques can be accepted, but you may pay in most currencies by credit card via our website – www.grbps.com.

Send to:

Name: _____

Address: _____

_____

_____

_____

Country: _____

Email address: _____

(only for use in case of a problem with your order – we will not use for marketing nor pass to any third party.)