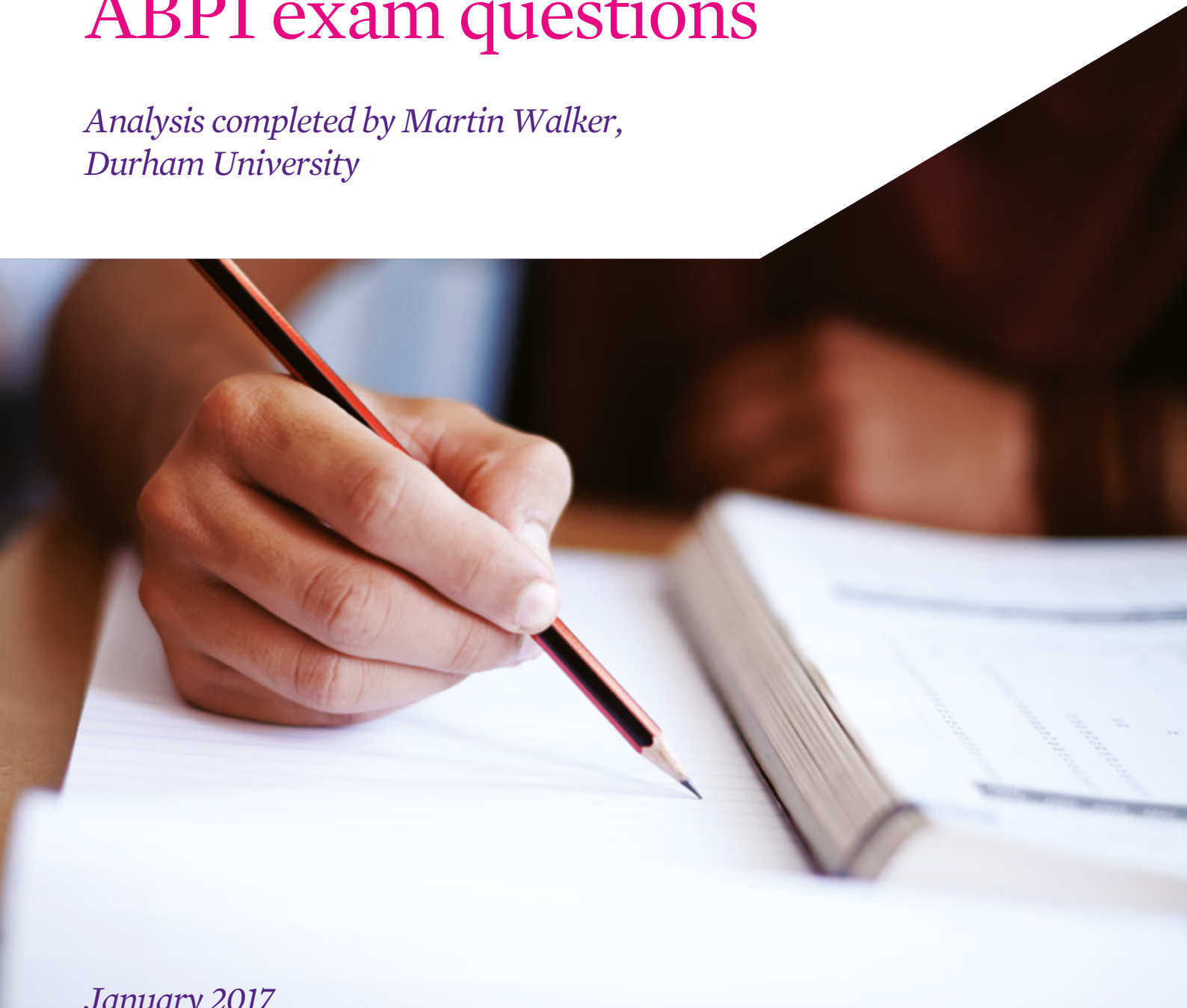
abpi

Bringing medicines to *life*

Independent analysis of ABPI exam questions

Analysis completed by Martin Walker, Durham University



January 2017

Robert Coe
Professor in the School of Education and Director of the Centre for Evaluation and Monitoring (CEM), Durham University

Foreword

Since 2015 the ABPI has taken significant steps towards producing high-quality assessments. High-stakes tests, which can change people's lives, should be of the highest quality possible and we should know what we mean by "quality".

Information regarding successful tests and test questions should probably be a combination of statistical evidence and expert judgement. Having evidence of how students of different abilities have interacted with questions of differing difficulties allows test developers to make predictions as to how future test questions will perform. Carrying out such analysis and then explaining the results to test writers is not common in the UK context and the work that the ABPI has been doing places the organisation at the forefront of professional test development in the UK.

The ABPI has at its disposal a set of test performance indicators which are well established in the academic literature from the field of test development. The focus on question difficulty, good correlations and the logic of the more able students achieving the right answers, shows a clear line of thought. Question writers can learn from questions and question types that have worked well, whilst also exploring the reasons why some questions might have led to unusual response patterns. The approach taken by the ABPI shows the power of evidence and the ABPI now has information available which can support test developers in their work and lead to continuing test improvement.

A scientific approach to test development includes reviewing previously used questions so as to learn lessons. Through two major rounds of test analysis, the ABPI has been able to consider the reliability and validity of its testing processes and to report on these areas using clear, evidence-based information. The test candidates at the ABPI are likely to understand the scientific method and should feel encouraged that the tests they will take have been developed in an evidence-based, scientific manner.

Executive summary

The ABPI introduced an accredited exam for medical representatives in 2014. The exam was first introduced in the 1960s, and since that date it has taken a variety of forms; however, before 2014 it had never been formally accredited by an external awarding body. There are two separate level 3 qualifications, a Diploma and a Certificate, in the promotion of prescription medicines.

Passing the exam is a requirement for a number of roles within pharmaceutical companies, as specified in the ABPI Code of Practice[1]. If a person is unable to pass the exam within two years of starting work in a role which requires the exam, they will not be able to continue to work in a promotional role. Hence the exam needs to provide reliable results.

Candidates sit the exam in invigilated sessions; one exam session is held every month. The papers are compiled from a secure bank of multiple choice questions which is regularly refreshed. The Diploma exam, taken by over 95% of candidates, comprises four mandatory units and a choice of disease area units adding up to at least 15 credits (two or three different disease areas must be studied).

This analysis, and a previous partial analysis of the mandatory unit questions commissioned in 2015, was carried out on questions which had been used in accredited exams since January 2014. The aim of the analysis was to:

- ensure that all questions used in exams are working as intended; and
- identify how easy or difficult each question is, to ensure that papers on different, optional topics are of a similar level of difficulty and to enable exam papers of an equal level of difficulty to be created each month.

The available data on each question was initially reviewed to ensure that the question had been answered by a sufficient number of candidates for the data to be reliable, before the question was analysed.

Questions were analysed in comparison with others in the same unit, and with questions from all units. The analysis identified the following attributes for each question:

- how easy or difficult the question is;
- how well the answer selected matches the ability of candidates answering it;
- whether high-ability candidates are more likely to select the right answer, and lower-ability candidates the wrong answer.

Of the 2,260 questions analysed, 13 questions were found to be so difficult that an average candidate has only a very low (less than 10%) chance of selecting the right answer and 206 questions were found to be so easy that the average candidate has a very high (greater than 90%) chance of selecting the correct answer. This suggests that the question bank is skewed somewhat towards easy questions. The ABPI exam team attempt to produce tests that are balanced in terms of question difficulty; this additional knowledge will help them in this aim.

When the pattern of answers to a question was considered, and was compared to the ability of the candidates answering the question, 459 questions were found to produce an odd pattern of answers. This might be because the question is too difficult or too easy, the wording of the question, or of the possible answers, was unclear, or there might be more than one correct answer to a question.

In some instances, more able candidates tended to choose the wrong answer. This could be because the database had the wrong answer identified as correct. However, as all questions had been carefully reviewed before adding to the question bank and errors of this type had been addressed following careful checking when exam papers were set, this was not believed to be the major reason for 83 questions falling into this category.

Some questions fell into more than one of the above categories, resulting in a total of 615 questions being identified as requiring review. The report recommends that these questions be looked at by expert question writers, to try to identify why each was not performing as expected.

The ABPI is already using the data produced from this analysis to ensure that exam papers for a particular unit are set to be of a comparable level of difficulty each month, and to minimise differences in the level of difficulty between exam papers for the optional disease area units.

As a result of commissioning this analysis of the questions used in ABPI exams, the ABPI has detailed information on a large number of questions which are known to produce reliable test results.

The ABPI is now in a strong position to be able to reassure candidates and their employers that ABPI exams are valid and robust and are built on scientific principles.

Contents

1. Background to the report

1.1 Brief history of ABPI examinations

The ABPI has been running an exam for medical representatives since the 1960s. The need for people who promote medicines on behalf of pharmaceutical companies to take and pass the exam is a requirement of the ABPI Code of Practice, Clause 16. If a person is unable to pass the exam within two years of starting work in a role which requires the exam, they are likely to lose their job.

The exam has taken different forms during its history, but it remains a formal, invigilated exam. In 2013 the exam achieved accreditation and the older, unaccredited, exam was phased out. The final candidates took the unaccredited exam in December 2015.

Governance of the exam is through an independent committee comprising physicians, pharmacists and people involved in education and training, several of whom work in the NHS. Reporting to this committee is a steering group made up of training and compliance managers from ABPI member companies.

1.2 Accreditation of the exam

Accreditation of the exam did not fundamentally change the level of knowledge required, and the exam continues to assess knowledge through multiple-choice questions. However, the accredited exam has a stronger focus on demonstrating understanding of a topic and less on recall of knowledge. This was a requirement to achieve accreditation, but was also identified as a need by the exam steering group to ensure that the exam provided benefits for the individuals taking it and was not just a 'hurdle to overcome'.

Two versions of the accredited exam are offered: a Level 3 Certificate in the promotion of prescription medicines, which is appropriate for people who promote medicines only on the basis of quality, price and availability to people who do not prescribe medicines; and a Level 3 Diploma in the promotion of prescription medicines which is taken by those who promote medicines to prescribers[2].

1.3 What the exam covers

The exam is intended to ensure that all industry representatives have an appropriate background knowledge of: the industry they work in and its Code of Practice, the customers they engage with in the NHS, basic (level 3) human biology, the process for discovering and developing new medicines, and their role in monitoring patient safety.

The Certificate comprises four mandatory units:

- Unit 1 – Code of Practice and the NHS
- Unit 2 – Human body systems (circulatory, respiratory, digestive, musculoskeletal and skin systems)
- Unit 3 – Human body systems (nervous, endocrine, reproductive and urinary systems)
- Unit 4 – Development and use of medicines

The Diploma comprises the four mandatory units plus at least 15 credits from the candidate's selected disease area units.

Each unit exam draws questions from a bank to cover all assessment criteria for that unit.

1.4 Number of candidates taking the exam

The number of candidates who register and start studying for the exam is higher than the number who actually take the exam. This is because a number change jobs and no longer need to take the exam, or they leave the industry.

In the first year of the accredited exam (2014) 242 people were booked to take ABPI exams (this may be just the mandatory units, just the disease area units, or both). A further 440 were booked to take exams in 2015 and, in the first five months of 2016, 208 additional people had booked to take at least part of their ABPI exam.

Since the accredited exam was introduced in January 2014 up to and including May 2016, 890 people have booked to take these exams. Of these, 22 are studying for the Certificate and 868 the Diploma. The total number of people who actually took exams during this period was 845. This discrepancy is explained by the fact that around two people each month who are booked to take the exam fail to attend, without advising the ABPI that they no longer plan to sit the exam.

Some of the people who take the exam fail one or more units, and will have retaken these units, or are planning to do so. Others decide to leave the industry without completing their qualification, and a small number are "let go" by their company because they are unable to pass.

1.5 Exam sittings

Exams are offered once a month. Candidates take the exam in strict exam conditions at invigilated centres. Up to 80 candidates take the exam at each sitting; exams for Units 1–4 are taken in the morning and the optional unit exams are in the afternoon.

Candidates may take all unit exams on one day, or may take the mandatory units one month and the optional units on another occasion. All four mandatory units must initially be taken together. If one or more units are failed, then that unit, or units, may be retaken on another occasion. Similarly all the selected disease area units must be taken together; single units can only be taken if the remaining units have already been passed.

1.6 How well is the accredited exam working?

With around 850 people now having taken the accredited exam, sufficient information is available for a reliable analysis of the exam questions to be carried out. The analysis was carried out in July and August 2016 on all data available up to and including May 2016.

Candidates are allowed to re-sit the various units on multiple occasions. Although exams are constructed to minimise the chance of any candidate seeing a question more than once, a unit exam can be repeated several times, so there could be multiple responses from a given candidate to the same question. The data set for analysis was constructed so as to include only the first instance of the candidate answering a question.

2. Question analysis

The analysis was carried out on questions which had been used in accredited exams since January 2014. The purpose of the analysis was to:

- ensure that all questions used in exams are working as intended; and
- identify how easy or difficult each question is, to ensure that papers on different optional topics are of a similar level of difficulty and to enable exam papers of an equal level of difficulty to be created each month.

The aim of the process is to establish that ABPI examinations are based on sound scientific principles supported by the academic literature on test development.

2.1 How can we tell if a question is working?

For a question to be useful as part of a broader test, it must contribute information to the overall measurement of the persons being tested. As part of the analysis we considered the extent to which a question:

- is not so easy that the vast majority of candidates answer it correctly;
- is not so difficult that the vast majority of candidates answer it incorrectly;
- fits a pattern of performance across the rest of the test;
- is answered correctly more often by higher-ability candidates than by lower-ability candidates.

If a question satisfies all of the above then it can be considered to be a useful question that can contribute to the measurement of the candidates, ie the question:

a) is neither too easy nor too difficult;
b) correlates well with overall test performance;
c) is answered correctly by more able and incorrectly by less able candidates.

The three criteria above – a) to c) – were used to filter questions.

2.2 Method of analysis

The question response dataset was analysed using the Rasch model, one of the approaches from item response theory. The Rasch model is widely used in test development and provides information about the ability of each candidate and the difficulty of each question.

If the questions are to be used to place candidates into groups such as pass or fail then there should be an underlying model to which the data will fit satisfactorily. If the test is to be used to determine candidate ability, then the most able candidates should get most of the questions correct whereas the least able should get fewer questions correct.

For each candidate (person) the analysis produces a measure of ability. For each question (item) the analysis produces an estimate of difficulty. The interaction between the ability of a person and the difficulty of an item suggests a probability of success.

> A person of high ability should have a high chance of success on an item of low difficulty; and conversely,
>
> A person of low ability should have a low chance of success on an item of high difficulty.

If a test is well designed then it should be able to measure candidates across a reasonable range of ability. For this to be the case, the difficulty of the set of questions used should be suitably matched to the ability of the candidates.

Analysis in Winsteps

The analysis was carried out using a software package called Winsteps. The various pieces of information (correlation, discrimination, difficulty vs ability etc) that appear in the report are derived from this Winsteps analysis.

Stages of analysis

The exam question bank contains over 3,500 questions. All questions which had been used in accredited exams since January 2014 were considered for analysis. This created a set of 2,692 questions.

In analysing the set of 2,692 questions the decision was taken to only consider questions which had been taken by at least 25 people (ie at least 25 'interactions'). This gave a dataset of 2,260 questions for analysis.

All candidates must take Units 1–4; thereafter the options are chosen by the candidate. There are 12 optional units, one of which (Unit 6) counts as two units. This makes the possible permutations too great to explore at the individual level.

Although no candidate will take all 16 units, it seemed reasonable to treat the various unit combinations as one homogenous test. The ABPI sets only one final boundary for each unit: candidates pass the unit or fail the unit. Any allowed combination of units can lead to a final overall pass for the qualification. It does not matter whether one candidate takes Units 1–4 plus Units 14, 15 and 16 whilst another candidate takes Units 1–4 plus Unit 6 and Unit 7. If the candidate passes each of the units then the candidate will pass the final qualification.

With this in mind, initial analysis of the questions was carried out at the level of the full dataset of 2,260 questions. Each individual candidate may have taken no more than 350 questions but there were sufficient interactions throughout the entire dataset for the test to be considered as one large test.

2,260 question set analysis

The full matrix of interactions between 845 candidates and 2,260 questions was analysed at the first stage.

Estimates of person ability and item difficulty were derived from this analysis. The full 2,260 question set analysis also produced information about questions with unusual response patterns and about the extent to which a question was good at discriminating between the more and the less able candidates.

Individual unit analysis

The analysis was re-run at the individual unit level for all of the 16 units. This provided additional information about the correlation between responses to questions and underlying ability. As each run of the test would contain different combinations of questions from the question bank it seemed sensible to explore the extent to which questions within a unit correlated with the underlying ability of the candidates. A unit might contain questions on quite different and even disparate topics, but in most cases there should be a consistent connection between candidate ability and score.

3. Quality checks for questions

The quality of each question was considered against three criteria from the analysis described above, ie that the question:

a) is neither too easy nor too difficult;

b) correlates well with overall test performance;

c) is answered correctly by more able and incorrectly by less able candidates.

3.1 Questions need to be neither too difficult nor too easy

The first quality check "hurdle" considered questions which are:

a) so easy that the average candidate will have a greater than 0.9 (90%) probability of success;

b) so difficult that the average candidate will have a less than 0.1 (10%) probability of success.

If the mean ability candidate has a less than 0.1 probability of success, then half of the candidates have a probability lower than this.

If the mean ability candidate has a greater than 0.9 probability of success, then half of the candidates have a probability higher than this.

When half of the candidates have a less than 0.1 or greater than 0.9 probability of success on a question, the question is unlikely to provide much in the way of good measurement.

3.2 Question responses should correlate with ability

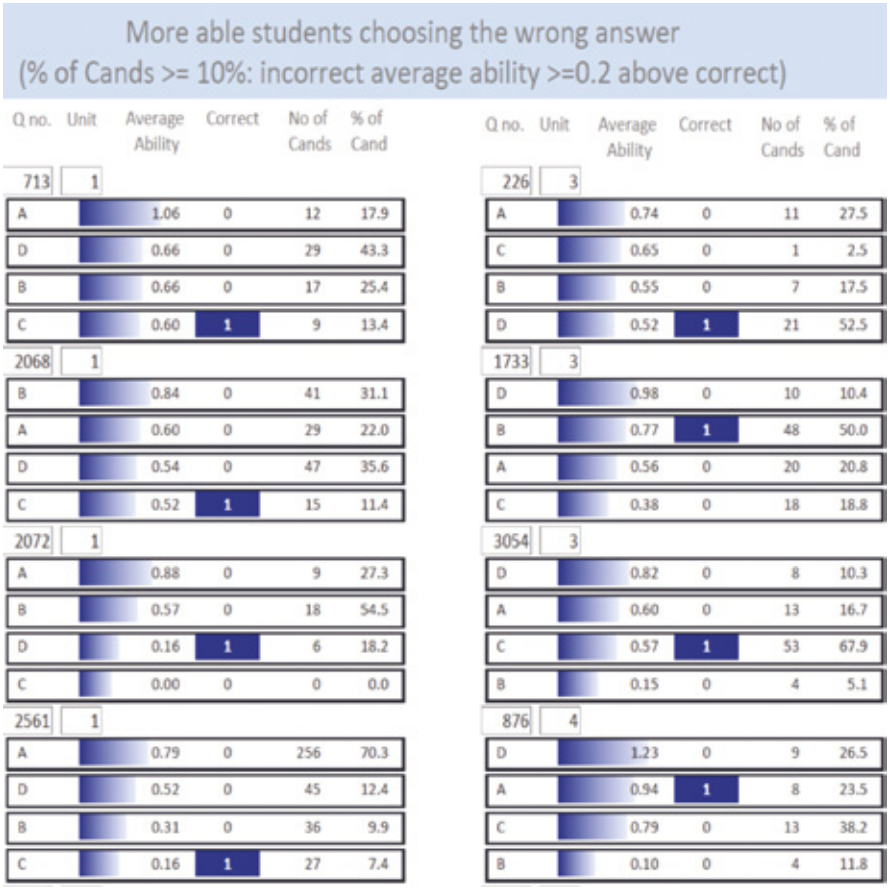
The analysis will produce estimates of ability for the candidates. A good question will produce responses which correlate well with candidate ability.

Correlation ranges between 1 and -1. A correlation of 1 between two or more variables says that the two variables fluctuate perfectly in parallel. A correlation of -1 says that as one variable increases the other variable does exactly the opposite. A correlation of zero says that there is no connection between the ways in which the variables fluctuate.

For this analysis a correlation of 0.1 or above was considered to be sufficient. Whilst this is a relatively low "hurdle" for a question to pass, the nature of the combined test – with 16 units that deliberately set out to test quite different things – helped to determine that a correlation threshold of 0.2 or 0.3 (which might be used in other test circumstances) was likely to flag up too many questions as performing erratically. The wide-ranging nature of the individual units will lead to some low correlations but this can be explained logically and is not a negative feature of the test.

3.3 Correct answers should be chosen by more able candidates (and incorrect answers by less able candidates)

Each question used in ABPI exams has four possible answers labelled A, B, C or D. As the ability of the candidates has been calculated as part of the analysis, it is possible to look at the average ability of candidates who chose each of the possible answers to a given question.



More able students choosing the wrong answer
(% of Cands >= 10%: incorrect average ability >=0.2 above correct)

Q no.	Unit		Average Ability	Correct	No of Cands	% of Cand
713	1					
		A	1.06	0	12	17.9
		D	0.66	0	29	43.3
		B	0.66	0	17	25.4
		C	0.60	1	9	13.4
2068	1					
		B	0.84	0	41	31.1
		A	0.60	0	29	22.0
		D	0.54	0	47	35.6
		C	0.52	1	15	11.4
2072	1					
		A	0.88	0	9	27.3
		B	0.57	0	18	54.5
		D	0.16	1	6	18.2
		C	0.00	0	0	0.0
2561	1					
		A	0.79	0	256	70.3
		D	0.52	0	45	12.4
		B	0.31	0	36	9.9
		C	0.16	1	27	7.4

Q no.	Unit		Average Ability	Correct	No of Cands	% of Cand
226	3					
		A	0.74	0	11	27.5
		C	0.65	0	1	2.5
		B	0.55	0	7	17.5
		D	0.52	1	21	52.5
1733	3					
		D	0.98	0	10	10.4
		B	0.77	1	48	50.0
		A	0.56	0	20	20.8
		C	0.38	0	18	18.8
3054	3					
		D	0.82	0	8	10.3
		A	0.60	0	13	16.7
		C	0.57	1	53	67.9
		B	0.15	0	4	5.1
876	4					
		D	1.23	0	9	26.5
		A	0.94	1	8	23.5
		C	0.79	0	13	38.2
		B	0.10	0	4	11.8

Figure 3:1 Ability and choice of answer

The average ability of each group of candidates who chose A, B, C or D is shown in Figure 3:1. It is logical that for each question, the more able candidates tended to choose the correct answer.

The information in Figure 3:1 is arranged by question, and then within the question it is arranged in descending order of the average ability of the group choosing A, B, C or D. The expected outcome would be that the group that chose the correct answer would be the group with the highest average ability of the four groups who chose A, B, C or D.

However, this is not always the case. Take question 226 shown in Figure 3:1 as an example. The group with the highest ability (0.74) chose answer A. The correct answer was in fact answer D. (The white number 1 on the blue background denotes the correct answer for the question.) In the case of question 713, it was the least able candidates who chose the correct answer. This is not logical and suggests that the question would benefit from being reviewed. There is something in the question that has caused those with the highest ability on the rest of the test to choose the wrong answer whilst those with the lowest ability on the rest of the test chose the correct answer.

A question was flagged as needing further investigation if its responses did not follow this logical pattern of the most able candidates choosing the correct answer. Such questions tend to randomise the overall data, as they break the connection between ability and score.

4. Results of the analysis

4.1 Questions which were too easy or too difficult

The analysis produced information about questions which were "too easy" or "too difficult". The determining factor was the probability of success of the average candidate on the question.

Altogether 219 questions were identified as being "easy" or "hard", with:

13 questions being so difficult that the average candidate has a less than 0.1 probability of success on the question;

206 questions being so easy that the average candidate has a greater than 0.9 probability of success on the question.

The analysis resulted in two lists of questions. One list contained the questions sorted by increasing probability of success; the other list contained the same information sorted by unit. This would allow the ABPI to review the information based on the placing of a question by overall difficulty or by difficulty within a unit. In all cases the calculations took place at the whole test level of 845 candidates and 2,260 questions.

Unit	Q no.	Measure	Prob of Avg Cand
1	2561	3.31	0.07
			1 questions
1	1704	-1.56	0.90
1	703	-1.67	0.91
1	2077	-1.75	0.92
1	2569	-1.86	0.93
1	1707	-2.06	0.94
1	704	-2.09	0.94
1	2088	-2.15	0.94
1	2052	-2.24	0.95
1	711	-2.27	0.95

Figure 4:1 Examples of easy and hard questions

In Figure 4:1 we can see that Unit 1 contained one question (number 2561) which was so difficult that the average candidate had only a 0.07 probability of success on the question. This is a 7 in 100 chance of success. (Although it is a probability ranging from 0 to 1, some people are more comfortable describing this as a 7% chance of success.)

If the average candidate has a 7 in 100 chance of success on the question and half of the candidates are less able than the average candidate, then it follows that half of the candidates have a chance of success that is lower than 7 in 100.

There were many more questions which were so easy that the average candidate had a chance of success greater than 90 in 100. The inclusion of large numbers of very easy items in a test is likely to:

a) compress the mark range;

b) make it more difficult to distinguish between different levels of performance;

c) waste a limited number of measurement opportunities.

Questions that were very easy or very difficult were collated by the ABPI for further review by the question-writing team. Working out why a question that had been written with the best intentions proved to be very easy or very difficult is a useful part of the process of question development.

The balance of easy vs difficult questions in the question bank

There were:

- 13 questions that were too difficult;
- 206 questions that were too easy.

This suggested that the question bank was skewed somewhat towards easy questions. The question development team had been aware of this from the previous round of analysis and had attempted to create individual unit tests which took equal numbers of easy and difficult questions from the bank so as to produce unit tests that were balanced in terms of question difficulty.

Visualising candidate ability and question difficulty – the Wright map

It is possible to represent all of the candidate abilities and question difficulties diagrammatically. A commonly used diagram is the Wright map (named after Ben Wright, an early proponent of Rasch analysis). The Wright map for the 2,260 questions taken by 845 candidates is shown below in Figure 4:2.

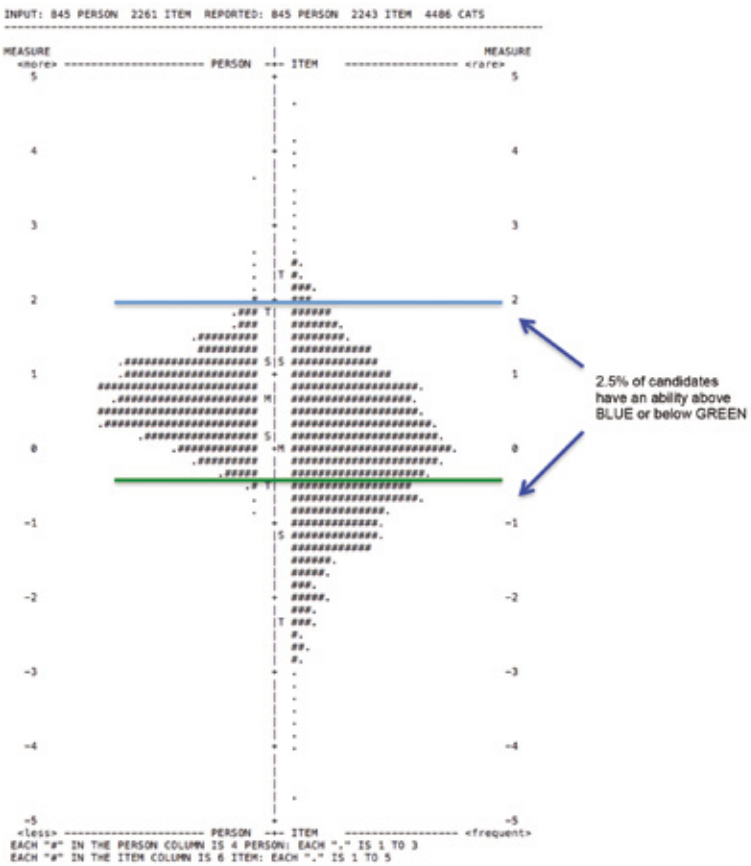


Figure 4:2 Wright map

The scale is displayed on both the left- and right-hand sides of the Wright map. The scale is a logarithmically derived interval scale on which:

- the positive numbers for the person (candidate) ability measure on the left-hand side of Figure 4:2 relate to candidates who scored more questions right than wrong; the negative numbers relate to candidates who scored more questions wrong than right;
- for the item (question) difficulty on the right-hand side of Figure 4:2, the positive numbers relate to questions that were answered wrongly more times than correctly, whilst the negative numbers relate to questions which were answered correctly more often than wrongly.

An important feature of a test is the alignment between the ability of the candidates and difficulty of the questions. If the test above was perfectly aligned then there would be questions to measure each level of ability and the questions would be distributed evenly in relation to candidate ability. We can see in Figure 4:2 that there are more easy questions aligned with (and beyond) the bottom of the candidate ability range than there are difficult questions at the top.

In fact this is simply a visualisation of the 16 difficult questions versus the 206 easy questions that were in the question bank.

This way of thinking might be unfamiliar to some test developers. It is quite common to encounter the "let's give candidates some easy ones at the start to settle them in" when setting an exam paper. It might be reasonable to place some easier questions at the start of an individual unit test but these "easier" questions should not be so easy that everyone answers them correctly. If everyone gets a question correct (or wrong) then the question provides no information about where candidates are on a scale of ability and a limited number of measurement opportunities is then reduced further. There are numerous relatively easy questions which will still provide some measurement information and these questions lie close to but above the green line in Figure 4:2.

4.2 Question responses correlate with ability

Each time a question is answered by a group of examination candidates it produces a pattern of 1s (correct answers) and 0s (incorrect answers). This pattern can be compared to the information about candidate ability that has been generated by the analysis.

Because each of the 16 individual units can test quite different things, it is unreasonable to expect high correlations at the whole test level. For this reason the correlation between candidate ability and question success was considered at the level of the individual unit.

For the ABPI qualification, each unit must be passed individually.

With this in mind, it seemed appropriate to consider correlation at the individual unit level. The analysis was run on each individual unit, as though the unit were a complete test. The 16 runs of the analysis produced correlations for each question within a unit.

Unit	Question ID	Answer	% Wrong Ans less able >10%	Correlation	Avg Person Prob rel to Unit Avg	Easy or Hard	MEASURE	Prob relative to ALL avg	No of Qns	16 Unit avg ability
1	702	D		-0.0342	0.98	Easy	-3.13	0.98	21	0.68
1	703	D			0.91	Easy	-1.46	0.89		
1	704	D			0.94	Easy	-1.85	0.93		
1	706	D		-0.055			-0.64	0.79		
1	710	B			0.96	Easy	-2.27	0.95		
1	711	D			0.95	Easy	-1.98	0.93		
1	712	B		0.0768			0.5	0.54		
1	713	C	17.9104	-0.1341	0.09	Hard	3.14	0.08		
1	717	D			0.96	Easy	-2.38	0.96		
1	1704	C			0.91	Easy	-1.4	0.89		
1	1707	B			0.94	Easy	-1.9	0.93		
1	1709	D			0.96	Easy	-2.34	0.95		
1	1710	A			0.96	Easy	-2.21	0.95		
1	2052	D			0.95	Easy	-2.04	0.94		
1	2068	C	31.0606	0.0748	0.09	Hard	3.13	0.08		
1	2077	B		0.0729	0.92	Easy	-1.64	0.91		
1	2088	B			0.95	Easy	-1.98	0.93		
1	2096	B		0			0	0.66		
1	2561	C	70.3297	-0.1493	0.06	Hard	3.66	0.05		
1	2565	D		-0.0087			1.96	0.22		
1	2569	C		-0.0476	0.93	Easy	-1.74	0.92		

Figure 4:3 Example information for question writers

Figure 4:3 shows one worksheet from the 16 separate unit worksheets that were provided. The correlation of the question performance to underlying ability is shown in the column headed "Correlation". The information was provided in the format of an Excel workbook to enable the ABPI team to extract information that is relevant to the particular task at hand. Here we are considering correlation.

In the ABPI question bank question set for Unit 1 there were 10 questions that had low correlations (<0.1). A number of these questions were discussed at a meeting with the ABPI question-writing team.

Some reasons for the low correlations were easy to identify:

- questions that are very easy or very hard will tend to have poor correlations;
- some correct answers had been chosen by large numbers of otherwise low-ability candidates.

Beyond these "self-generating" low correlations, the question-writers were asked to explore possible reasons why questions had unusually low correlations. Some findings during the meeting included instances where there was:

- more than one possible correct answer to a question;
- potentially confusing language in the question stem;
- potentially confusing language in one or more of the possible answers.

For a question to be working well and providing good information about candidates, the responses to the question should follow a pattern that is driven by candidate ability rather than extraneous factors which are not being measured.

Across the whole question set there were 459 questions for which the correlation between ability and score was less than or equal to 0.1.

4.3 Correct answers chosen by more able candidates (and incorrect answers by less able candidates)

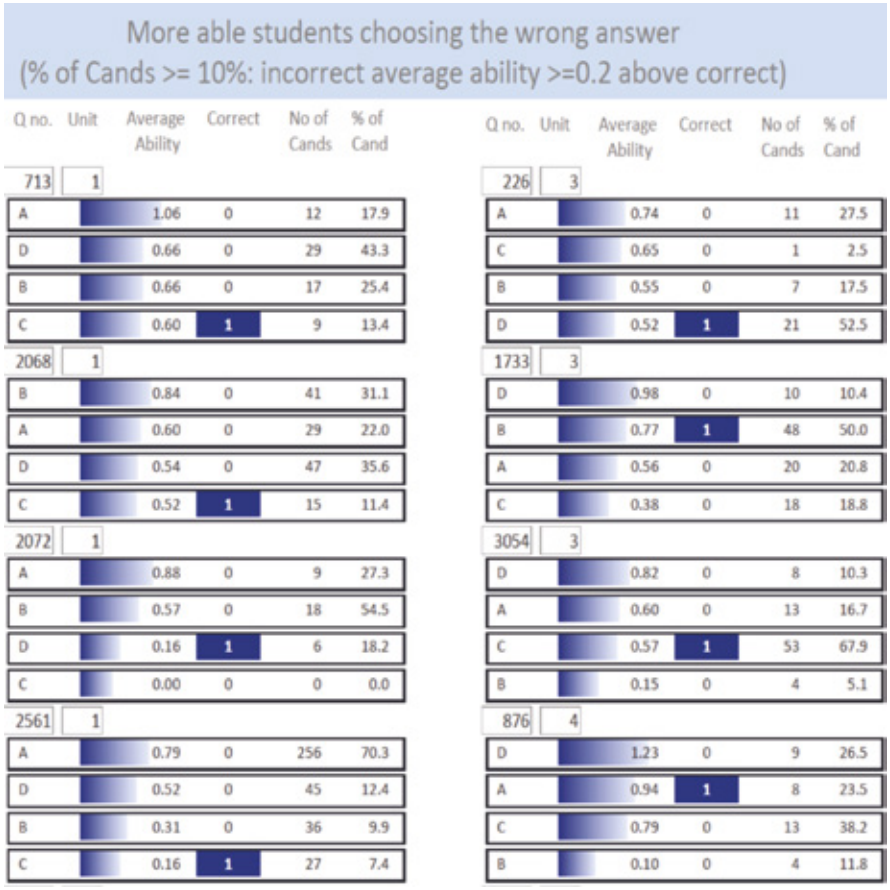
The third of the "hurdles" for question performance was an analysis of the extent to which the more able candidate tended to pick the correct answer.

As stated in Section 3.3:

> The expected outcome would be that the group that chose the correct answer would be the group with the highest average ability of the four groups who chose A, B, C or D.

For some questions there were only small numbers of candidates who chose a particular answer. This could mean that one or two candidates who made unusual choices could distort the data for a question. For this reason two additional filters were applied to the information about ability and choice of answer:

1) the proportion of lower-ability candidates who chose the correct answer must be more than 10% of the total number of candidates who answered the question;
2) the gap in ability between "more" and "less" able must be great enough for candidates to have a 5% increase or decrease in probability of success.



More able students choosing the wrong answer
(% of Cands >= 10%: incorrect average ability >=0.2 above correct)

Q no.	Unit		Average Ability	Correct	No of Cands	% of Cand
713	1					
		A	1.06	0	12	17.9
		D	0.66	0	29	43.3
		B	0.66	0	17	25.4
		C	0.60	1	9	13.4
2068	1					
		B	0.84	0	41	31.1
		A	0.60	0	29	22.0
		D	0.54	0	47	35.6
		C	0.52	1	15	11.4
2072	1					
		A	0.88	0	9	27.3
		B	0.57	0	18	54.5
		D	0.16	1	6	18.2
		C	0.00	0	0	0.0
2561	1					
		A	0.79	0	256	70.3
		D	0.52	0	45	12.4
		B	0.31	0	36	9.9
		C	0.16	1	27	7.4

Q no.	Unit		Average Ability	Correct	No of Cands	% of Cand
226	3					
		A	0.74	0	11	27.5
		C	0.65	0	1	2.5
		B	0.55	0	7	17.5
		D	0.52	1	21	52.5
1733	3					
		D	0.98	0	10	10.4
		B	0.77	1	48	50.0
		A	0.56	0	20	20.8
		C	0.38	0	18	18.8
3054	3					
		D	0.82	0	8	10.3
		A	0.60	0	13	16.7
		C	0.57	1	53	67.9
		B	0.15	0	4	5.1
876	4					
		D	1.23	0	9	26.5
		A	0.94	1	8	23.5
		C	0.79	0	13	38.2
		B	0.10	0	4	11.8

Figure 4:4 More able candidate choosing the wrong answer

The data on more able candidates choosing the wrong answer were provided in the form shown in Figure 4:4.

The first question in Figure 4: 4 is number 713 and was from Unit 1. A total of 67 candidates had answered this question. The group that chose A as the answer had a higher-than-average ability than the three groups who chose B, C or D; however, A was not the correct answer. The correct answer is shown in Figure 4:4 as a number 1 in the column headed “Correct”; the correct answer to question 713 is C.

The question bank had already been checked for data entry errors so the ABPI team was confident that C was in fact the correct answer. This means that there is something about answer A that caused 12 candidates, who are on average the most able on this question, to pick A as the answer.

Question 3054 in Unit 3 was taken by 78 candidates. The most able of the four groups chose D and this represented 10.3% of the total number who answered the question. Answer A was chosen by the next most able group, representing 16.7% of the entry for the question. The correct answer – C –was chosen by 67.9% of candidates but these were not the highest-ability candidates.

There were 83 questions which exhibited the pattern discussed above.

It is natural that questions on which the more able candidates did not choose the correct answer would also generate low correlations between candidate ability and score. Many of the questions that appear in the list of 83 will also appear in the list of questions that have low correlations.

5. Conclusions and recommendations

5.1 The ABPI question bank

The ABPI has a question bank in which 2,260 questions have been taken by sufficient numbers of candidates to provide meaningful data.

The three criteria, a) to c), were used to filter questions:

a) is neither too easy nor too difficult;
b) correlates well with overall test performance;
c) is answered correctly by more able and incorrectly by less able candidates.

2260 Qns	
Unit	Number of Qns
1	105
2	230
3	216
4	225
5	99
6	294
7	100
8	136
9	85
10	146
11	118
12	95
13	96
14	125
15	48
16	142
Total	2260

Figure 5:1 Number of questions per unit

Of the 2,260 questions available for analysis, a total of 615 questions were suggested for review before being used again.

Criteria for review	Number of questions
Too easy or too difficult	219
Poor correlation	459
More able choosing wrong answer	83

Figure 5:2 Question numbers for review

Some questions were flagged for review because of more than one criterion. This explains why the total number of questions shown in Figure 5:2 is greater than 615.

The recommendation to the ABPI is that all 615 questions should be reviewed. It was also suggested that information should be gathered centrally relating to the typical causes of such unusual question performance. Although there are 615 questions which should be looked at, it is likely that there will be only a handful of reasons as to why questions have not worked as expected.

Possible reasons include:

- questions are so easy or difficult that candidate performance on the question does not correlate to candidate ability (most get it right or wrong irrespective of ability as measured by the other questions);
- misleading wording in the question stem or answers leads the more able candidates to give an answer other than the one that was the intended correct answer;
- there is more than one possible answer to the question;
- there is no correct answer to the question;
- a question has been set on information that has changed in the supporting study materials;
- the focus of a question has changed over time but the question has not been updated.

Creating tests of comparable difficulty

As ABPI staff now have the difficulty measures for every one of 2,260 questions, it is possible to combine questions of known difficulty into a unit exam with the correct number of questions of known overall difficulty.

5.2 Reflections and next steps

The ABPI has now commissioned two rounds of detailed test analysis on all of the questions in the ABPI question bank. This 2016 analysis shows that of the 2,260 questions available for analysis,

- 615 questions should benefit from being reviewed;
- 1,645 questions are performing well.

This is a real achievement for any test developer. The ABPI now has detailed information suggesting that a large bank of items can provide reliable test results and that the measurement properties of these items are known in detail.

The additional ability to create tests of comparable difficulty is also a significant achievement.

Any questions from external organisations regarding test validity could be addressed by a combination of the question level analysis and the professional judgement of ABPI staff and expert question-writers.

At a time when it is not usual for UK professional organisations to provide validity evidence about their tests, or to be able to provide such evidence were it requested, the ABPI has moved in the direction of gathering as much evidence as possible about its own tests. Being able to include test questions which are known to work, from a large bank of such questions, puts the ABPI in a strong professional position. Building a good validity argument would require:

- professional judgement about the overall testing structure;
- statistical evidence that the tests are measuring real attributes in candidates.

The ABPI is in a strong position to be able to provide such evidence and to reassure candidates and their employers that the ABPI tests are built on scientific principles.

As far as the author is aware, it has not been common practice in the UK for bodies who set professional examinations to carry out analysis of the questions they use, as the ABPI has done. In this respect the ABPI is leading the way in ensuring that its exams are reliable and give valid results that can be trusted by exam candidates and their employers.

6. References

[1] http://www.pmcpa.org.uk/thecode/Documents/Code%20of%20Practice%202016%20.pdf

[2] ABPI Code of Practice 2016 Clause 16.3 http://www.pmcpa.org.uk/thecode/Documents/Code%20of%20Practice%202016%20.pdf

7. Appendix

A Technical information

A.1 Sufficient question response data

Although there were candidate responses to over 2,500 questions, some questions were either:

a) new questions;
b) part of a unit that few candidates have opted to take;

and had only small numbers of responses. As the intention was to use item response theory and the Rasch model to analyse the interactions between questions and candidates, a question would need to have been answered enough times for there to be reasonable data. The Institute for Objective Measurement recommends that:

"a sample of 50 well-targeted examinees is conservative for obtaining useful, stable estimates. 30 examinees is enough for well-designed pilot studies." (Rasch.org, 2016)

The numbers of candidates taking each question were considered at four levels of interaction:

Questions that had been taken by:

a) over 100 candidates;
b) over 75 candidates;
c) over 50 candidates;
d) over 25 candidates.

The number of questions that exceeded each of these four respective thresholds is shown in Figure A:1.

Threshold	No. of questions	Comments
Over 100 cands	1085 questions	Very good number of interactions between items and persons BUT from 2694 to 1085 = 60% loss
Over 75 cands	1428 questions	
Over 50 cands	1730 questions	
Over 25 cands	2261 questions	Acceptable number of interactions between items and persons 16% loss of items from 2694

Figure A:1 Numbers of candidates taking questions

If the threshold for the number of interactions between candidate and question were to have been set at 100 interactions, then 1,085 of the 2,694 questions would have been removed from the analysis. As the main purpose of the analysis was to find out which questions seemed to work well and which questions might be giving the test team less good information, the loss of this large number of questions, representing 60% of the available question bank, was considered to be too great.

The decision was taken to run the analysis for all questions for which there were at least 25 interactions. This reduced the final dataset for analysis to 2,261 questions. If it was found that this run of the analysis provided useful information about the 2,261 questions, then filtering the data at the threshold of at least 25 interactions for each question might be sufficient.

There were 845 candidates in the final dataset.

A.2 The nature of the analysis

The question response dataset was analysed using the Rasch model, one of the approaches from item response theory.

If the questions are to be used to place candidates into groups such as Pass or Fail, then there should be an underlying model to which the data will fit satisfactorily. If the test is to be used to determine candidate ability, then the most able candidates should get most of the questions correct whereas the least able should get fewer questions correct. The idea that the only two parameters which can be allowed to influence the outcome of a candidate's test score are the candidate's ability and the difficulty of the questions is central to the Rasch model.

Estimates of candidate ability and question (item) difficulty are produced by the analysis. The raw score question data is transformed into an interval scale with the units of the scale being derived logarithmically and known as logits. (Rasch, 1960/1980)

For each candidate (person) the analysis produces a measure of ability, measured in logits. For each question (item) the analysis produces an estimate of difficulty. The interaction between the ability of a person and the difficulty of an item suggests a probability of success. A person of high ability should have a high chance of success on an item of low difficulty; conversely, a person of low ability should have a low chance of success on an item of high difficulty.

A.2.1 2,261 question analysis

The full matrix of interactions between 845 candidates and 2,261 questions was analysed at the first stage.

Estimates of person ability and item difficulty were derived from this analysis. The full 2,261 question set analysis also produced information about questions with unusual response patterns and about the extent to which a question was good at discriminating between the more and the less able candidates.

A.2.2 Individual unit analysis

The analysis was re-run at the individual unit level for all of the 16 units. This provided additional information about the correlations between responses to questions and underlying ability. As each run of the test would contain different combinations of questions from the question bank it seemed sensible to explore the extent to which questions within a unit correlated with the underlying ability of the candidates. A unit might contain questions on quite different and even disparate topics but there should be a consistent connection between candidate ability and score.

B Quality checks for questions

B.1 Quality criteria for questions

The quality of each question was considered against the following three criteria.

The question:

a) is neither too easy nor too difficult;
b) correlates well with overall test performance;
c) is answered correctly by more able and incorrectly by less able candidates.

B.2 Questions are too difficult or too easy

Each question has a measure of difficulty derived from the analysis. The question difficulty is measured in units called logits. The scale used is an interval scale, ie the increase in difficulty from 1 logit to 2 logits is the same increase as from 2 logits to 3 logits etc. The scale operates in the same way as a metre ruler for measuring length.

Each person has a measure of ability represented on the same logit scale used to represent question difficulty. The gap between a candidate's ability and the difficulty of a question allows a probability of success to be calculated.

As the ability of every candidate is estimated during the analysis it is possible to report the average (arithmetic mean) candidate ability on the logit scale. The mean candidate ability for 845 candidates on 2,261 questions was 0.68 logits. For the purpose of this analysis we will use the term "average candidate ability" to refer to the mean candidate ability.

The probability of success for a candidate on a question is given by

$$P(X=1|B_n,D_i)=\frac{e^{(B_n-D_i)}}{1+e^{(B_n-D_i)}}$$

(Rasch, 1960/1980)

Figure B:1

where:

β_n is the ability of candidate n

D_i is the difficulty of question i

This first quality check "hurdle" considered questions which are:

a) so easy that the average candidate will have a greater than 0.9 probability of success;
b) so difficult that the average candidate will have a less than 0.1 probability of success.

If the mean ability candidate has a less than 0.1 probability of success, then half of the candidates have a probability lower than this. In this case the question is probably so difficult that it will not contribute good information to measurement.

If the mean ability candidate has a greater than 0.9 probability of success then half of the candidates have a probability higher than this. In this case the question is probably so easy that it will not contribute good information to measurement.

As we want all questions to be contributing to good measurement, this "too hard / too easy" filter is a useful initial way of highlighting questions that are not performing well.

This way of thinking might be unfamiliar to some question writers. It is quite common to encounter the "let's give them some easy ones at the start to settle them in" mentality amongst question writers. By all means put some easier questions at the start of an individual unit test but these "easier" questions should not be so easy as to produce no variance between candidates. Variance contains information. Some people should get a question correct whilst other get it wrong. This variance should be linked to ability, ie the more able tend to choose the correct answer.

If everyone (or almost everyone) gets a question correct (or wrong) then the question provides no information about where candidates are on a scale of ability.

This suggests that the question bank is skewed somewhat towards easy questions. The question development team had been aware of this from the previous round of analysis and had attempted to create individual unit tests which took equal numbers of easy and difficult questions from the bank so as to produce unit tests that were balanced in terms of question difficulty.

B.2.1 Visualising candidate ability and question difficulty – the Wright map

Looking at lists of questions can be helpful but does not suit everyone. It is possible to represent all of the candidate abilities and question difficulties diagrammatically. A commonly used diagram is the Wright map (named after Ben Wright, an early proponent of Rasch analysis). The Wright map for the 2,261 questions taken by 845 candidates is shown below in Figure B:2.

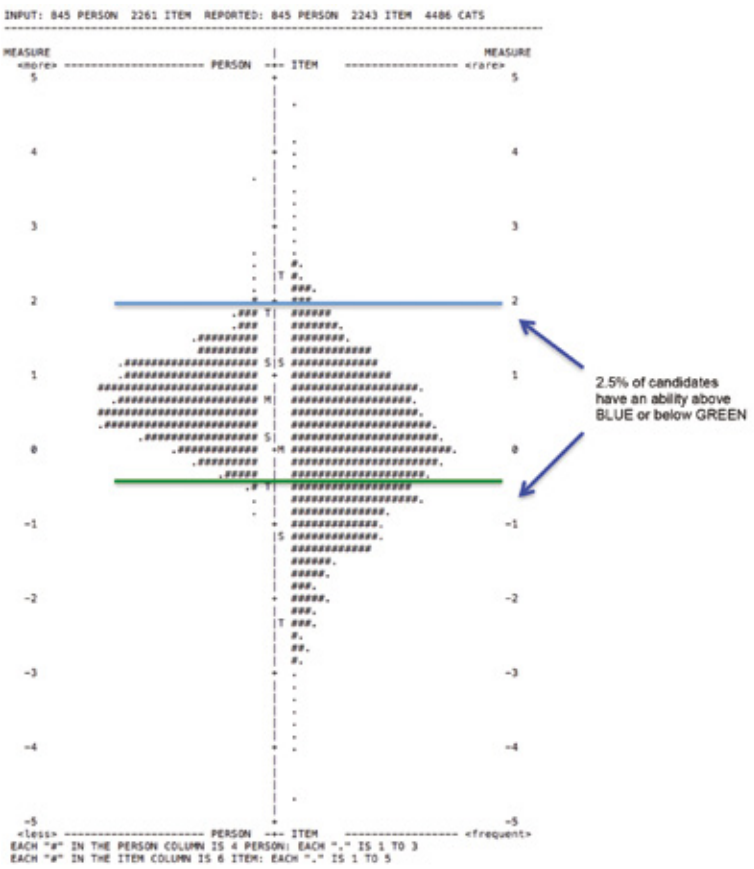


Figure B:2 Wright map 1

The logit scale is displayed on both the left- and right-hand sides of the Wright map. The scale is a logarithmically derived interval scale on which:

- the positive numbers for the person (candidate) ability measure relate to candidates who scored more questions right than wrong; the negative numbers relate to candidates who scored more questions wrong than right.
- for the item (question) difficulty, the positive numbers relate to questions that were answered wrongly more times than correctly, whilst the negative numbers relate to questions which were answered correctly more often than wrongly.

An important feature of a test is the alignment between the ability of the candidates and difficulty of the questions. If the test above was perfectly aligned then there would be questions to measure each level of ability and the questions would be distributed evenly in relation to candidate ability. The mean candidate ability would also be at the same point on the scale as the mean question difficulty.

We can see in Figure B:2 that there are more questions aligned with (and beyond) the bottom of the candidate ability range than there are at the top.

The letters M, S and T refer to:

M = Mean
S = 1 standard deviation
T = 2 standard deviations

There are far more easy questions with difficulty measures below two standard deviations of candidate ability than difficult ones above.

B.3 Score on the question correlates with ability

An ability measure is calculated for each candidate. The score (1 or 0) is known for each question. If a question is answered incorrectly by candidates with high ability and/or correctly by candidates of low ability, then the question will provide potentially false information.

There should be a clear correlation between ability on the test and score on a question. This is explored by looking at the correlation between the scores and the person ability measures. The Pearson point measure correlation coefficients are calculated for each question. (Winsteps.com, 2016)

The level of "cut-off" regarding what is considered to be a good correlation vs a poor correlation is arbitrary. The nature of the ABPI test dictated that a particular approach to correlation be adopted. Whereas other statistics were computed at the whole test level, correlation was computed at the individual unit level. There is no reason to suppose that performance on one particular unit will correlate well with every other unit and different candidates may take different units in the test.

A correlation below 0.1 (between ability and score) was used as the indicator for a question that should be looked at again.

B.4 The question is answered correctly by more able and incorrectly by less able candidates

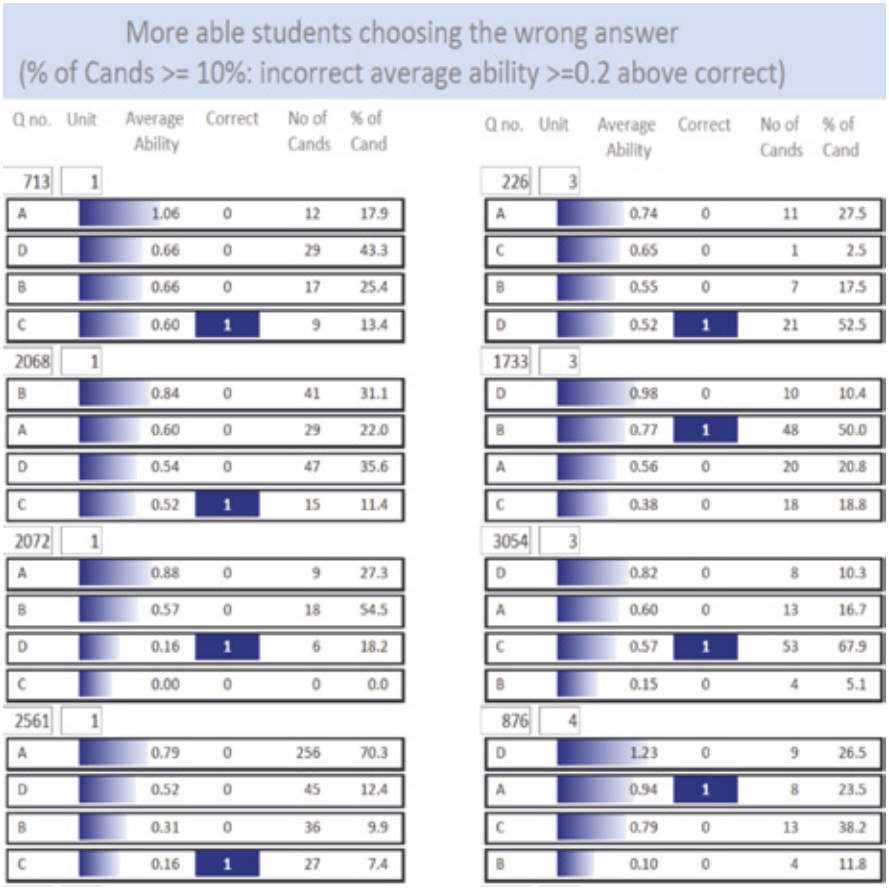
The analysis produced information about the four options, A, B, C and D for each question. The average ability of the candidates who chose each of the possible answers was calculated. This produced four average ability figures for each question. Logically, the group of candidates with the highest average ability should be the group who chose the correct answer.

For some questions there were only small numbers of candidates who chose a particular answer. This could mean that one or two candidate who made unusual choices could distort the data for a question. For this reason two additional filters were applied to the information about ability and choice of answer:

1) The proportion of lower-ability candidates who chose the correct answer was more than 10% of the total number of candidates who answered the question.
2) The gap in ability between "more" and "less" able was great enough for candidates to have a 5% increase or decrease in probability of success. This occurs when the average ability versus question difficulty is 0.2. If the value 0.2 is the difference between ability and difficulty (see Figure B:1) then the candidate's chance of success shifts from 0.5 to either 0.55 or 0.45, depending which way round the ability versus difficulty difference is expressed.

Stages 1) and 2) above are arbitrary and could be changed if necessary.

An example of the ABCD ability vs score information is shown below in Figure B:3.



More able students choosing the wrong answer
(% of Cands >= 10%: incorrect average ability >=0.2 above correct)

Q no.	Unit	Average Ability	Correct	No of Cands	% of Cand
713	1				
A		1.06	0	12	17.9
D		0.66	0	29	43.3
B		0.66	0	17	25.4
C		0.60	1	9	13.4
2068	1				
B		0.84	0	41	31.1
A		0.60	0	29	22.0
D		0.54	0	47	35.6
C		0.52	1	15	11.4
2072	1				
A		0.88	0	9	27.3
B		0.57	0	18	54.5
D		0.16	1	6	18.2
C		0.00	0	0	0.0
2561	1				
A		0.79	0	256	70.3
D		0.52	0	45	12.4
B		0.31	0	36	9.9
C		0.16	1	27	7.4

Q no.	Unit	Average Ability	Correct	No of Cands	% of Cand
226	3				
A		0.74	0	11	27.5
C		0.65	0	1	2.5
B		0.55	0	7	17.5
D		0.52	1	21	52.5
1733	3				
D		0.98	0	10	10.4
B		0.77	1	48	50.0
A		0.56	0	20	20.8
C		0.38	0	18	18.8
3054	3				
D		0.82	0	8	10.3
A		0.60	0	13	16.7
C		0.57	1	53	67.9
B		0.15	0	4	5.1
876	4				
D		1.23	0	9	26.5
A		0.94	1	8	23.5
C		0.79	0	13	38.2
B		0.10	0	4	11.8

Figure B:3 More able candidate choosing the wrong answer

C References

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Expanded edition (1980) with foreword and afterword by B.D. Wright (ed.) Vol. Expanded edition (1980) with foreword and afterword by B.D. Wright). Copenhagen / Chicago: Danish Institute for Educational Research / The University of Chicago Press.

Rasch.org. (2016). Sample Size and Item Calibration [or Person Measure] Stability. Retrieved from http://www.rasch.org/rmt/rmt74m.htm

Winsteps.com. (2016). Correlations: point-biserial, point-measure, residual. Retrieved from http://www.winsteps.com/winman/correlations.htm

The Association of the British Pharmaceutical Industry
A company limited by guarantee registered in England & Wales number 09826787
Registered office: 7th Floor, Southside, 105 Victoria Street, London SW1E 6QT
T: +44 (0)207 930 3477 exams@abpi.org.uk

abpi.org.uk

RMI-0070-1216