abpi

Bringing medicines to *life*

Question Analysis
Report 2019

Foreword

The Association of the British Pharmaceutical Industry (ABPI) represents innovative research-based biopharmaceutical companies, large, medium and small, which work to ensure the UK remains at the forefront of helping patients prevent and overcome diseases.

The ABPI examination has evolved significantly since its inception in the 1960s, when it was introduced for medical representatives, such that it is now a renowned accredited qualification with a far-reaching reputation of being outstanding for its purpose. Without doubt, the ABPI accredited exam, which is a requirement for several industry roles in order to comply with the ABPI Code of Practice for the Pharmaceutical Industry, is a cornerstone for patient trust and the reputation of the UK pharmaceutical industry.

As an industry qualification, it is paramount that the highest standards of preparation, delivery, quality assessment and independent governance are ensured and maintained. I have therefore once again commissioned the Centre for Evaluation and Monitoring (CEM) at Durham University to use scientific and evidence-based monitoring for analysing the robustness of the examination.

I know such analysis and subsequent use of results to inform planning is not common across sector-based qualifications, and as such, the ABPI is therefore, in many ways, leading professional test development in the UK. However, industry professionals make a commitment to study and pass the examination, and it is only right that the ABPI Examinations Department treats with equal commitment all aspects of the examination, from developing learning materials, revision materials and question development to examination delivery.

The thorough analysis of question data undertaken by CEM has shown the high levels to which the examination is planned and prepared, such that we are in a ‘strong position … to reassure candidates and their employers that the ABPI exams are valid and robust and are built on scientific principles’. I will not stop there, however; I intend to use the findings of this report, together with feedback sought from candidates, as the basis for continued review, development and quality control of an examination which impacts on the pharmaceutical industry, NHS and the better outcomes desired for patients.

Andrew Croydon
Director of Examinations

About the author

Martin Walker is involved in teaching and research into the effective use of assessment to improve learning. He has run postgraduate courses for teachers and examiners across the UK and has also worked with the official regulators for testing in England, Wales, Scotland and Northern Ireland.

In the academic world, Martin has spoken at Assessment Europe conferences and at the Asia Pacific Educational Assessment Conference, held in Singapore. He works regularly with Principals Academy in Singapore, delivering courses on validity in assessment and applications of test theory to schools in the region.

He has advised several companies in the City of London on aspects of professional examinations and has worked across the finance, banking and pharmaceutical sectors.

Martin has taught English and physics in a range of secondary schools and colleges in England and has written numerous books on the teaching of English and English literature. As a teacher he became involved in national examinations, eventually operating at the highest level of national test development. During this period, he noted, ‘I remember thinking that we seemed simply to be making up questions and thinking that if a question felt about right, it must be good’.

With these ideas in mind, Martin studied for an MSc in educational assessment at Durham University from 2009 to 2011, leading to a PhD study of the validity of national examinations for 16-year-old pupils in the UK. He helped to set up the UK Chartered Institute of Educational Assessors and developed many of the Institute’s teaching programmes. His current interest is in exploring national education systems to examine whether what is being taught makes sense in today’s world, and whether the assessment of what has been taught provides useful information about learning.

Executive summary

The ABPI introduced an accredited examination 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 . To remain compliant with the ABPI Code of Practice, Clause 16 states that representatives must take the examination 'within their first year of employment as a representative and must pass it within two years of starting such employment'. Hence the examination needs to provide reliable results.

Candidates sit the exam in invigilated sessions, one exam session being 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 (2019), is the latest in a series of analyses of question functioning: previous analysis of new questions was carried out in 2017 and 2016 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 each question had been answered by sufficient 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.

Table 0.1 shows the number of questions that have been analysed to date.

Year of analysis	Number of questions analysed	Number of good questions	Number of questions for review
2015 and 2016 combined	2260	1645	615
2017	427	331	96
2019	178	139	39
Total questions	**2865**	**2115**	**750**

From 2015 to 2019, of the 2,865 questions analysed, 19 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 225 questions were found to be so easy that the average candidate has a very high (greater than 90%) chance of selecting the correct answer.

When the pattern of answers to a question was considered, and was compared to the ability of the candidates answering the question, 536 questions were found to produce an unexpected pattern of answers. This might be because: the question is too difficult or too easy; the wording of the question, or that of the possible answers, might be 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 checks when exam papers were set, this was not believed to be the major reason for 101 questions falling into this category.

Some questions came into more than one of the above categories, resulting in a total of 750 questions being identified as requiring review. The 2019 analysis contributed 39 of the 750 questions and the report recommends that these questions be reviewed 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 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 the 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 examination 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. To remain compliant with the ABPI Code of Practice, representatives must take the examination 'within their first year of employment as a representative and must pass it within two years of starting such employment'.

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, with membership open to physicians, pharmacists, professionals from appropriate learned societies and people involved in education and training. Reporting to this committee is an Exam Steering Group comprising 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.

1.3 What the exam covers

The exam is intended to ensure that all industry representatives have an appropriate background knowledge of the industry in which they work, its Code of Practice, the customers in the NHS with whom they engage, 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 question 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 take the exam. This is because a number of individuals 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 have been 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 January 2014 when the accredited exam was introduced, up to and including May 2019, 2,532 people have booked to take these exams. Of these, 133 were studying for the Certificate and 2,399 the Diploma. The total number of people who took exams during this period was 2,441. This discrepancy is explained by the fact that around three people each month who are booked to take the exam fail both to attend and to advise the ABPI that they no longer plan to pursue the qualification.

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. Some decide to leave the industry without completing their qualification, whilst others (a small number) are not allowed to continue in employment in their representative role with their company because they have been unable to pass.

1.5 Exam sittings

Exams are offered once a month. Candidates take the exam in strict exam conditions at invigilated centres. Between 80 and 100 candidates can 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 another month. 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 2,440 people now having taken the accredited exam, sufficient information is available for a reliable analysis of the exam questions to be carried out. Analysis of test data was carried out in:

- July and August 2016 on all data available up to and including May 2016
- July 2017 on data from a first set of new questions
- April 2019 on data from a second set of new questions.

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 more than once, so there could be multiple responses from a given candidate to the same question. The dataset for analysis was constructed so as to include only the first instance of the candidate answering a question.

2 Question analysis

The three stages of analysis were 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, ensure that papers on different optional topics are of a similar level of difficulty and enable exam papers of an equal level of difficulty to be created each month.

The aim of the process is to establish that the 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, i.e. 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.

2.2.1 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.

2.2.2 Stages of analysis

In 2016, the exam question bank contained over 3,500 questions across the 16 units. All questions that had been used in accredited exams since January 2014 were considered for analysis. This created a set of 2,692 questions in 2016.

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

In 2017, a further 427 new questions had been taken by at least 25 candidates and these questions were analysed in 2017, using the set of good questions derived from the 2016 analysis as reference.

In 2019, a further 178 new questions were available for analysis and these questions were analysed against a set of good questions from the 2016 and 2017 analyses.

All candidates must take Units 1 to 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 to 4 plus Units 14, 15 and 16 whilst another candidate takes Units 1 to 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. By the end of the 2019 analysis, 2,865 questions had been analysed.

2.2.3 2,865 questions analysis

In each round of analysis, a matrix of interactions between candidates and questions was analysed:

2016 – 845 candidates and 2,260 questions

2017 – 1,386 candidates and 427 new questions plus 1,315 core questions

2019 – 2,100 candidates and 178 new questions plus 757 core questions.

As each round of analysis led to an increased number of questions that had very good performance statistics, a smaller set of 'core' questions was needed as a reference group for the new questions in 2019.

Estimates of person ability and item difficulty were derived from this analysis. Each round of question 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 able and the less able candidates.

2.2.4 Individual unit analysis

At each round, 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 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, i.e. 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 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 that 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. The results of the 2019 analysis are shown overleaf in Figure 3:1.

Figure 3:1 Ability and choice of answer

More able students choosing the wrong answer
(Incorrect answer group >=10% of cohort, Wrong ability – Correct ability>=0.2)

Q No.	Average Ability	Correct Ans = 1	No. of Cands	% of Cand
1032				
D	1.37	0	4	13.3
C	0.61	1	19	63.3
A	0.54	0	5	16.7
B	0.01	0	2	6.7
1033				
A	1.74	0	3	14.3
B	0.56	1	14	66.7
C	0.23	0	2	9.5
D	-0.22	0	2	9.5
1288				
B	0.90	0	8	16.7
C	0.64	0	9	18.8
A	0.60	1	28	58.3
D	-0.66	0	3	6.3
1340				
A	0.88	0	34	45.9
B	0.42	1	23	31.1
D	0.39	0	13	17.6
C	0.38	0	4	5.4
2220				
C	1.09	0	11	23.4
D	0.61	1	16	34.0
B	0.20	0	19	40.4
A	-0.23	0	1	2.1
2878				
B	0.94	0	75	28.4
D	0.75	0	22	8.3
A	0.66	1	93	35.2
C	0.58	0	74	28.0

Q No.	Average Ability	Correct Ans = 1	No. of Cands	% of Cand
3168				
A	1.02	0	7	26.9
D	0.35	0	11	42.3
C	0.26	1	7	26.9
B	0.05	0	1	3.8
3239				
B	1.26	0	3	15.0
A	0.76	1	13	65.0
C	0.45	0	3	15.0
D	0.24	0	1	5.0
3406				
D	1.65	0	4	12.5
B	1.32	1	14	43.8
A	0.96	0	9	28.1
C	0.49	0	5	15.6
567				
A	0.80	0	24	28.9
C	0.57	1	22	26.5
B	0.38	0	12	14.5
D	0.35	0	25	30.1
801				
B	1.64	0	3	10.0
D	0.65	1	23	76.7
A	-0.01	0	3	10.0
C	-0.43	0	1	3.3

22 March 2019

The average ability of each group of candidates that 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.

Only questions for which the more able group choosing the wrong answer constituted more than 10% of the total number of candidates are included in Figure 3:1. This removes the likelihood of the unusual performance of a small percentage of able candidates leading to a question being highlighted for review.

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: 2016 to 2019

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.

In 2019, four questions were identified as being ‘easy’ or ‘hard’, with:

- two questions being so difficult that the average candidate has a less than 0.1 probability of success on the question
- two questions being so easy that the average candidate has a greater than 0.9 probability of success on the question.

In 2019, the calculations took place at the whole test set level of 1,642 candidates and 935 questions. Of the 178 new questions in 2019, only four questions were extremely easy or difficult.

Figure 4:1 Questions of extremely high/low difficulty (2019)

Probability of success of avereage student

Mean Ability 0.26. SD 1.39

Question	Measure	Probability
1902	3.33	0.04
2859	2.61	0.09
2217	-2.23	0.92
1039	-3.06	0.97
4 Items		

In Figure 4:1, we can see that question number 1902 was so difficult that the average candidate had only a 0.04 probability of success on the question. This is a 4 in 100 chance of success. (Although it is a probability ranging from 0 to 1, some people are more comfortable describing this as a 4% chance of success.)

If the average candidate has a 4 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 4 in 100. There were some questions that were so easy that the average candidate had a chance of success greater than 90 in 100. The inclusion 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 marked up 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.

4.1.1 The balance of easy vs. difficult questions in the question bank

In 2016 there were:

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

In 2017 there were:

- 4 questions that were too difficult
- 17 questions that were too easy.

In 2019 there were:

- 2 questions that were too difficult
- 2 questions that were too easy.

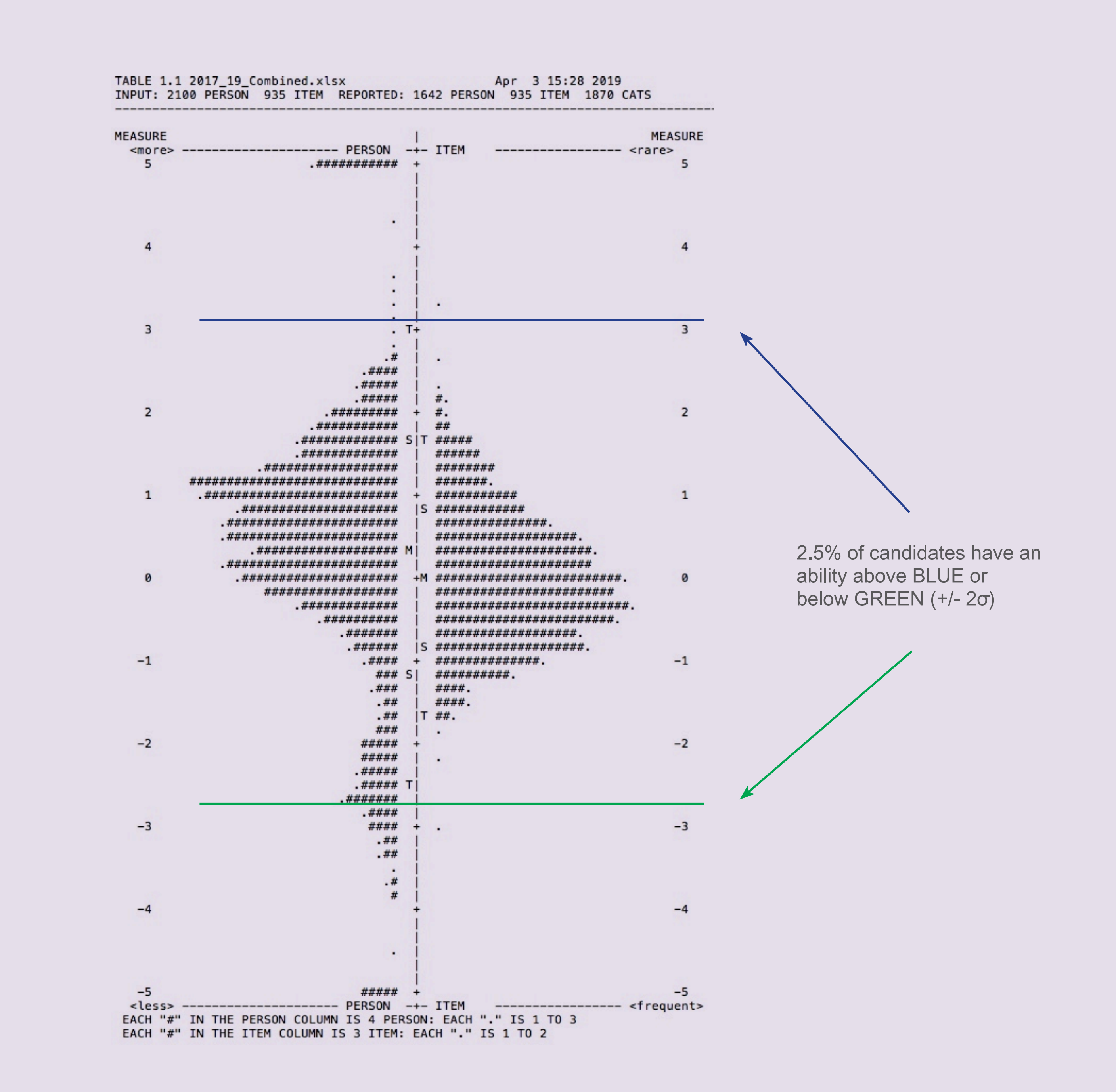
The first round of analysis (2016) suggested that the question bank was skewed somewhat towards easy questions. The question development team had been aware of this from the previous round of analysis and had attempted to create individual unit tests which took equal numbers of easy and difficult questions from the bank so as to produce unit tests that were balanced in terms of question difficulty.

Between 2016 and 2019, the number of extremely difficult or extremely easy questions has fallen considerably. This suggests that the question writing team has taken account of previous research findings and been able to adjust the demand of new questions as they are being produced. As the ABPI has measures of difficulty for all questions in the bank, it is a relatively simple task to create tests with an average difficulty of zero; meaning that neither easy nor difficult questions are over-represented in any single instance of a test.

4.1.2 Visualising candidate ability and question difficulty – the Wright Map

It is possible to represent all 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 from the 2019 analysis shows the alignment between 935 questions and 1,642 in Figure 4:2.

Figure 4:2 Wright Map



```
TABLE 1.1 2017_19_Combined.xlsx                    Apr  3 15:28 2019
INPUT: 2100 PERSON  935 ITEM  REPORTED: 1642 PERSON  935 ITEM  1870 CATS
---------------------------------------------------------------------------------

MEASURE                                 |                               MEASURE
  <more> --------------------- PERSON  -+- ITEM    ---------------- <rare>
    5                     .###########  +                                   5
                                        |
                                        |
                                        |
                                     .  |
                                        |
    4                                   +                                   4
                                        |
                                     .  |
                                     .  |
                                     .  |  .
                                     .  |
    3                                . T+                                   3
                                     .  |
                                    .#  |  .
                                 .####  |
                                .#####  |  .
                                .#####  |  #.
    2                       .#########  +  #.                               2
                          .###########  |  ##
                        .############# S|T #####
                        .#############  |  ######
                   .##################  |  ########
          ############################  |  #######.
    1      .###########################  +  ###########                      1
                .#####################  |S ############
              .#######################  |  ###############.
              .#######################  |  ###################.
                  .################### M|  #####################.
              .#######################  |  #####################
    0           .#####################  +M #########################.       0
                    ##################  |  ########################
                        .#############  |  ##########################.
                           .##########  |  ########################.
                              .#######  |  ###################.
                               .######  |S ####################.
   -1                            .####  +  ##############.                 -1
                                   ### S|  ##########.
                                  .###  |  ####.
                                   .##  |  ####.
                                   .##  |T ##.
                                   ###  |  .
   -2                            #####  +                                  -2
                                 #####  |  .
                                .#####  |
                                .##### T|
                              .#######  |
                                 .####  |
   -3                             ####  +  .                               -3
                                   .##  |
                                   .##  |
                                     .  |
                                    .#  |
                                     #  |
   -4                                   +                                  -4
                                        |
                                        |
                                     .  |
                                        |
                                        |
   -5                            #####  +                                  -5
  <less> --------------------- PERSON  -+- ITEM    ---------------- <frequent>
 EACH "#" IN THE PERSON COLUMN IS 4 PERSON: EACH "." IS 1 TO 3
 EACH "#" IN THE ITEM COLUMN IS 3 ITEM: EACH "." IS 1 TO 2
```

2.5% of candidates have an ability above BLUE or below GREEN (+/- 2σ)

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 correctly more times than wrongly, whilst the negative numbers relate to questions which were answered wrongly more often than correctly.

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 is even alignment between candidates: wherever there is a candidate ability between two standard deviations above and below the mean ability, there are questions.

There is a slightly unusual pattern to the distribution of candidates. Whereas the questions are normally distributed, the candidates appear to be in two distinct groups. The distribution of candidate ability is bi-modal, i.e. there are two distinct distributions (seen as two groups on the left-hand side of the map). Below a candidate ability measure of approximately -1.5 logits, there is a distinct group of low-ability candidates and this group appears to be normally distributed. This is a slightly unexpected feature and could possibly be due to the sampling that has taken place to construct the reference set of 757 questions. There is information here for the ABPI in that there is a group of around 80 candidates who were of such low ability as to have been unlikely candidates for the tests.

The only questions that are of very high or very low difficulty are shown by the two dots on the Item side of the Wright Map at approximately 3.2 and -3 logits. These are the questions identified in Section 4.1 above.

The only area of the Wright Map that shows a small number of questions of appropriate difficulty is the very top of the candidate ability range. This should be a concern only if a category such as 'distinction' is to be awarded to some candidates. The small number of questions targeted at the most able will probably lead to bunching of the candidates' total marks in this region. As the test bank as a whole has numerous high difficulty questions available to the test setters, this should not be a problem over the long term.

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.

As there are multiple available routes through the qualification and no two candidates need take the same set of units in order to qualify, correlation was explored at the level of the interaction between the 178 new questions and the 757 core reference questions. Although it is likely that some units will not correlate well with other units in terms of content, candidate performance in each individual unit should be stable. Candidates who chose to take Unit 8 should be as competent in that unit as the candidate who chooses to take Unit 9 will be on Unit 9 etc.

As each candidate must take Units 1 to 4, there is common ground here regardless of the route a candidate chooses. This suggests that correlation between candidate ability and score on an individual question can be considered to be a useful indicator of the performance of the question. Because of the many possible routes that can be taken, the cut-off point for correlation was taken as 0.1. This is relatively generous (0.2 or 0.3 could have been used in a more restricted test setting), yet still allows unusual patterns to be observed. The candidate ability measure is derived from the total score achieved but is not simply the total score itself. There should be a clear correlation between candidate ability and score on a question, i.e. a large proportion of able candidates should answer a question correctly whilst candidates of lower ability should answer incorrectly. Ability-score correlations below 0.1 are shown opposite in Figure 4:3.

Figure 4:3 Questions with low correlation between ability and score

Correlation <0.1

NAME	Correlation <0.1	Expected Correlation	Actual Expected
3236	0.0933	0.3376	-0.24
1325	0.0924	0.3053	-0.21
2376	0.0908	0.3595	-0.27
3199	0.09	0.2261	-0.14
3391	0.0874	0.291	-0.20
2621	0.0828	0.2493	-0.17
3392	0.0724	0.2957	-0.22
3341	0.063	0.33	-0.27
2220	0.0567	0.388	-0.33
2595	0.0495	0.3089	-0.26
1540	0.0482	0.2902	-0.24
1299	0.0429	0.3661	-0.32
589	0.0362	0.3289	-0.29
1288	0.0335	0.3427	-0.31
567	0.0183	0.3266	-0.31
801	0.0051	0.268	-0.26
2637	0.0045	0.3204	-0.32
2872	-0.0004	0.3299	-0.33
3239	-0.0061	0.3196	-0.33
49	-0.0109	0.2277	-0.24
1041	-0.0239	0.3489	-0.37
2495	-0.0398	0.2758	-0.32
2201	-0.0398	0.3071	-0.35
2878	-0.0582	0.3549	-0.41
1032	-0.0777	0.3499	-0.43
1033	-0.0986	0.3522	-0.45
1340	-0.1791	0.3046	-0.48
3168	-0.1816	0.3267	-0.51

In the 2019 analysis, there were 27 questions that had low correlations (<0.1). These questions were included in the set of questions to be discussed with the ABPI question writing team. Two questions had a strongly negative ability-score correlation – questions 1340 and 3168. On these two questions, the less able are more likely to do well whilst the more able are less likely to do well. Such correlations suggest that there is an error in the question that needs to be identified.

Some reasons for the low correlations are 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. Typical findings 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 178 new questions for 2019, there were 27 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. This occurs at an ability difference of 0.2 logits (see Appendix A for explanation).

Figure 4:4 More able candidate choosing the wrong answer

More able students choosing the wrong answer
(Incorrect answer group >=10% of cohort, Wrong ability – Correct ability>=0.2)

Q No.	Average Ability	Correct Ans = 1	No. of Cands	% of Cand
1032				
D	1.37	0	4	13.3
C	0.61	1	19	63.3
A	0.54	0	5	16.7
B	0.01	0	2	6.7
1033				
A	1.74	0	3	14.3
B	0.56	1	14	66.7
C	0.23	0	2	9.5
D	-0.22	0	2	9.5
1288				
B	0.90	0	8	16.7
C	0.64	0	9	18.8
A	0.60	1	28	58.3
D	-0.66	0	3	6.3
1340				
A	0.88	0	34	45.9
B	0.42	1	23	31.1
D	0.39	0	13	17.6
C	0.38	0	4	5.4
2220				
C	1.09	0	11	23.4
D	0.61	1	16	34.0
B	0.20	0	19	40.4
A	-0.23	0	1	2.1
2878				
B	0.94	0	75	28.4
D	0.75	0	22	8.3
A	0.66	1	93	35.2
C	0.58	0	74	28.0
3168				
A	1.02	0	7	26.9
D	0.35	0	11	42.3
C	0.26	1	7	26.9
B	0.05	0	1	3.8
3239				
B	1.26	0	3	15.0
A	0.76	1	13	65.0
C	0.45	0	3	15.0
D	0.24	0	1	5.0
3406				
D	1.65	0	4	12.5
B	1.32	1	14	43.8
A	0.96	0	9	28.1
C	0.49	0	5	15.6
567				
A	0.80	0	24	28.9
C	0.57	1	22	26.5
B	0.38	0	12	14.5
D	0.35	0	25	30.1
801				
B	1.64	0	3	10.0
D	0.65	1	23	76.7
A	-0.01	0	3	10.0
C	-0.43	0	1	3.3

22 March 2019

In most cases, the total number of candidates taking a question was quite small and the unexpected behaviour of five or ten candidates could skew the data for the question. It is still worth investigating why more able candidates may have chosen the wrong answer, if only to add to the knowledge of the question writers.

Question 2878, however, had been taken by 264 candidates. Altogether 93 candidates chose the correct answer, which was A, but these candidates were of lower ability than those who chose B or D as the answer. Altogether 75 candidates chose B as the answer and this group had the highest average ability of any of the groups who answered the question. There is something in option B that suggested that this was the correct choice for the most able candidates. The question bank had already been checked for data entry errors, so the ABPI team was confident that A was in fact the correct answer in the marking software.

It still remains the case, however, that there could be more than one correct answer or that more able candidates have found something that the questions setters had not anticipated. For example, in question 3168 (shown in Figure 3:1) the group with the highest ability (1.02) chose answer A. The correct answer was in fact answer C. (The white number 1 on the blue background denotes the correct answer for the question.) In the case of questions 1288, 2678 and 3168, two groups of more able candidates chose answers other than the correct one. These candidate performances are not logical and suggest that the question would benefit from being reviewed. There is something in the question that has caused those with the highest ability in the rest of the test to choose the wrong answer, whilst those with a lower ability in the rest of the test chose the correct answer.

In 2019, 11 of the 178 questions 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 above list of 11 will also appear in the list of questions that have low correlations.

4.4 New questions with unusual characteristics (2019)

Some questions showed unexpected performance against more than one of the stated criteria. This is to be expected, as the criteria that indicate unusual question performance are linked. If more able candidates choose the wrong answer, this will affect the correlation between ability and score and the question will also appear in the list of unusual ABCD responses seen in Section 4.3. The 39 questions that had unexpected response patterns are shown below in Figure 4:5.

39 “ Unusual” New Questions by Unit 2019

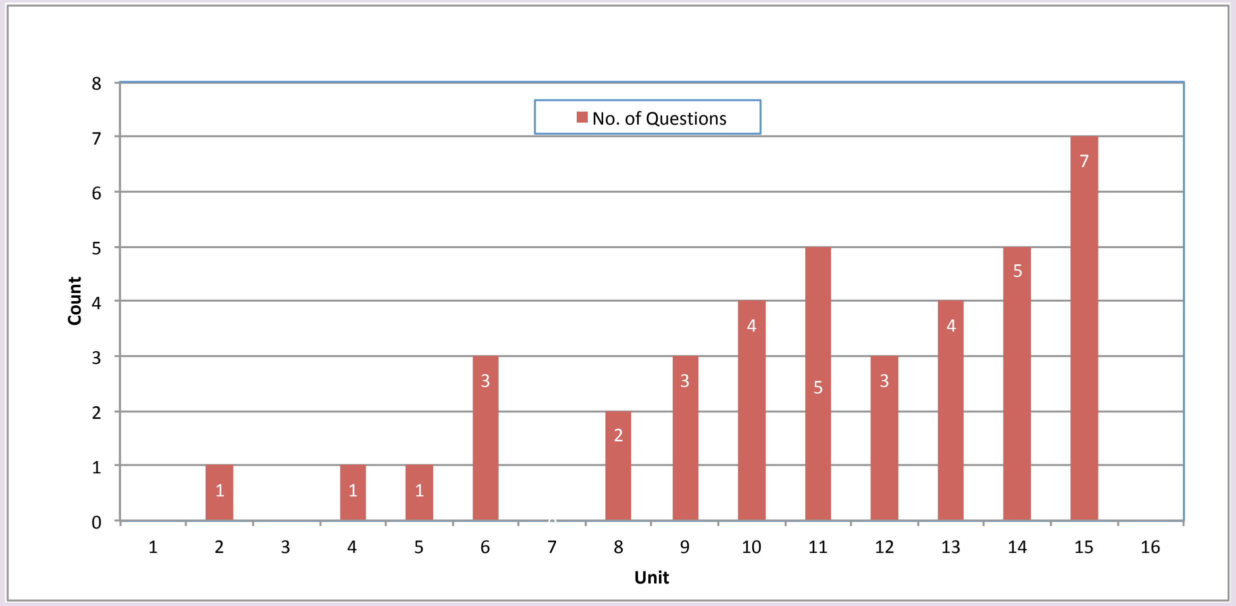


Figure 4:5 Unusual questions by unit, 2019

The criteria are shown by ‘reason’ in Figures 4:6 and 4:7 below.

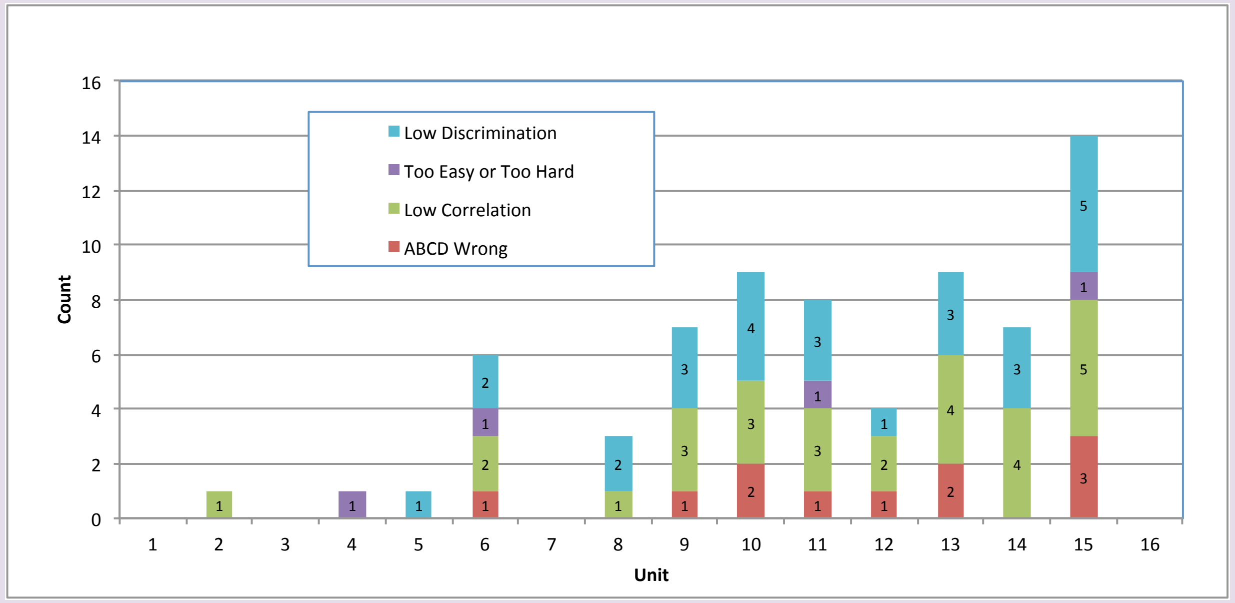


Figure 4:6 Unusual new questions by reason; bar chart, 2019

39 “ Unusual” New Questions by reason 2019

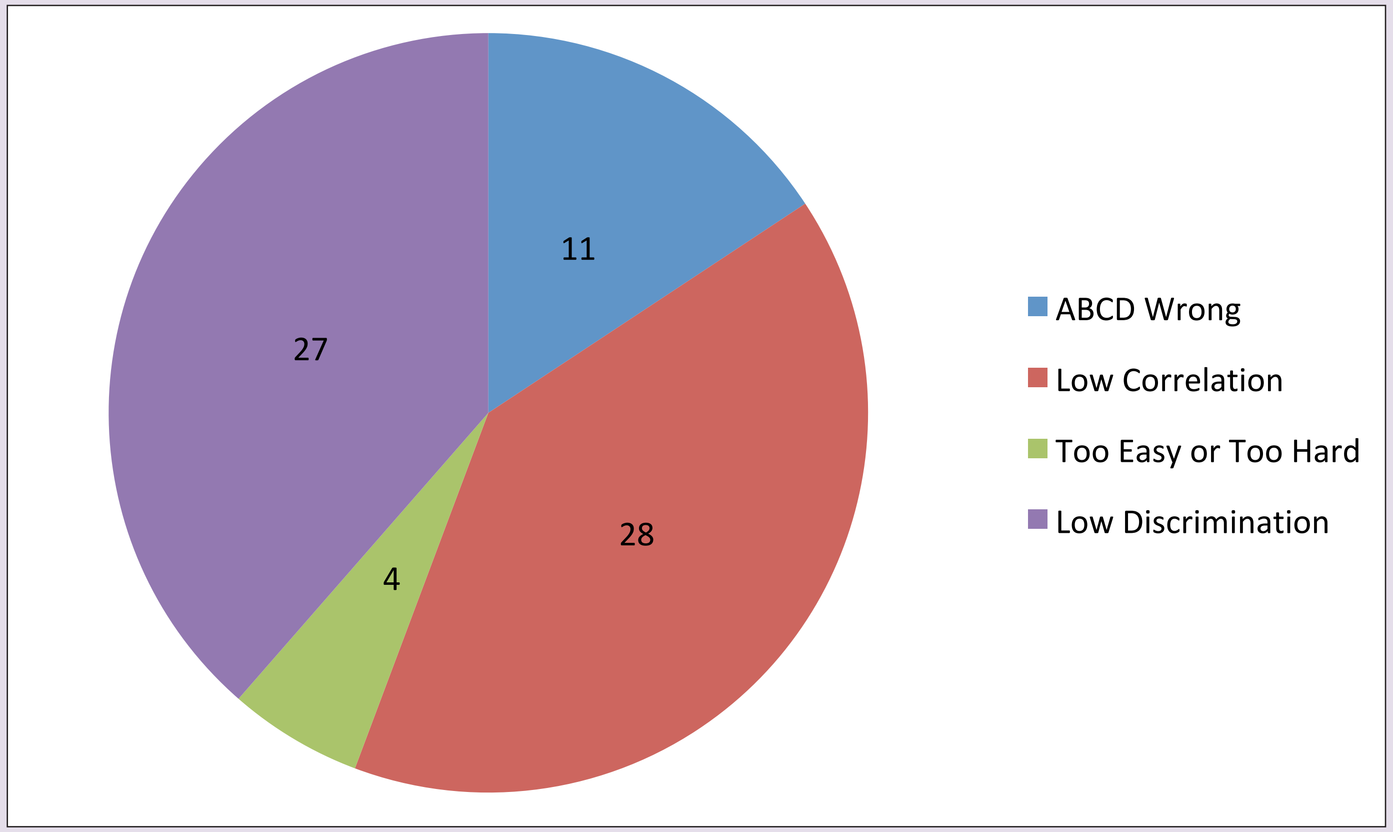


Figure 4:7 Unusual new questions by reason; pie chart, 2019

Figure 4:5 shows the total number of questions with unusual performance characteristics in each unit. Figure 4:6 shows the breakdown of reasons for questions considered as having unusual performance. For example, Figure 4:5 shows that three questions in Unit 6 demonstrated unusual candidate responses and Figure 4:6 shows that there were six reasons in total for these three questions to have been included in the list of unusually performing questions.

The proportion of questions considered unusual was the same in 2019 as in 2017:

- 2019: 139 questions from a set of 178 were considered good questions: 78% of the questions were good, 22% required further investigation
- 2017: 331 questions from a set of 427 were considered good questions: 78% of the questions were good, 22% required further investigation

From 2017 to 2019, four out of five new questions proved to be good questions, whereas in the 2016 round of analysis, results were as follows:

- 2016: 73% of the questions were good, 27% required further investigation.

This suggests that the ABPI could expect approximately four new questions out of five to be accepted for use in the question bank and that for 80 new good questions, 100 questions would have to be commissioned.

4.5 Good new questions (2019)

Of the 178 new questions analysed in 2019, 139 could be considered to be good enough to be added to the question bank. Figure 4:8 shows the breakdown of good new questions by unit for 2019.

Of these 139 questions, 126 questions had performed very well with performance statistics, well above the minimum thresholds explained in Section 3. Ability–score correlation was greater than or equal to 0.2 for these questions, suggesting that there was a strong link between general ability in the test and the likelihood of choosing the correct answer in the question.

A total of 23 questions exceeded the minimum performance requirements but were close to the minimum in the case of one or more of the performance metrics used. ABPI staff were given this detailed question performance information in a spreadsheet.

139 Good New Questions by Unit 2019

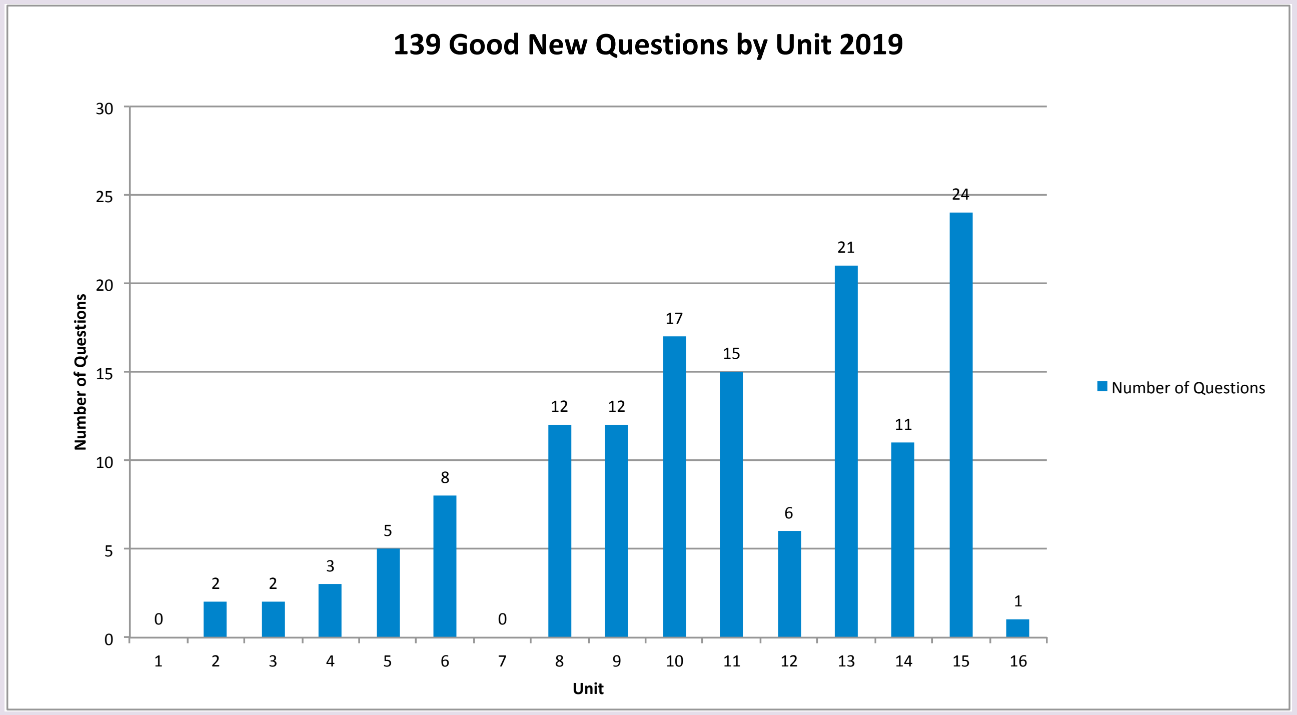


Figure 4:8 Good new questions by unit, 2019

4.5.1 Average difficulty of new unit questions (2019)

Fourteen units had good new questions following the 2019 analysis. Unit 1 and Unit 7 had no new questions in this round of analysis. The difficulty measure of each question is known and so the average difficulty of the new question in a unit can be calculated. Figure 4:9 shows the average difficulty of new questions in each of the 14 units. The scale used is the logit scale that also appears on the Wright Map.

Average difficulty of new questions 2019

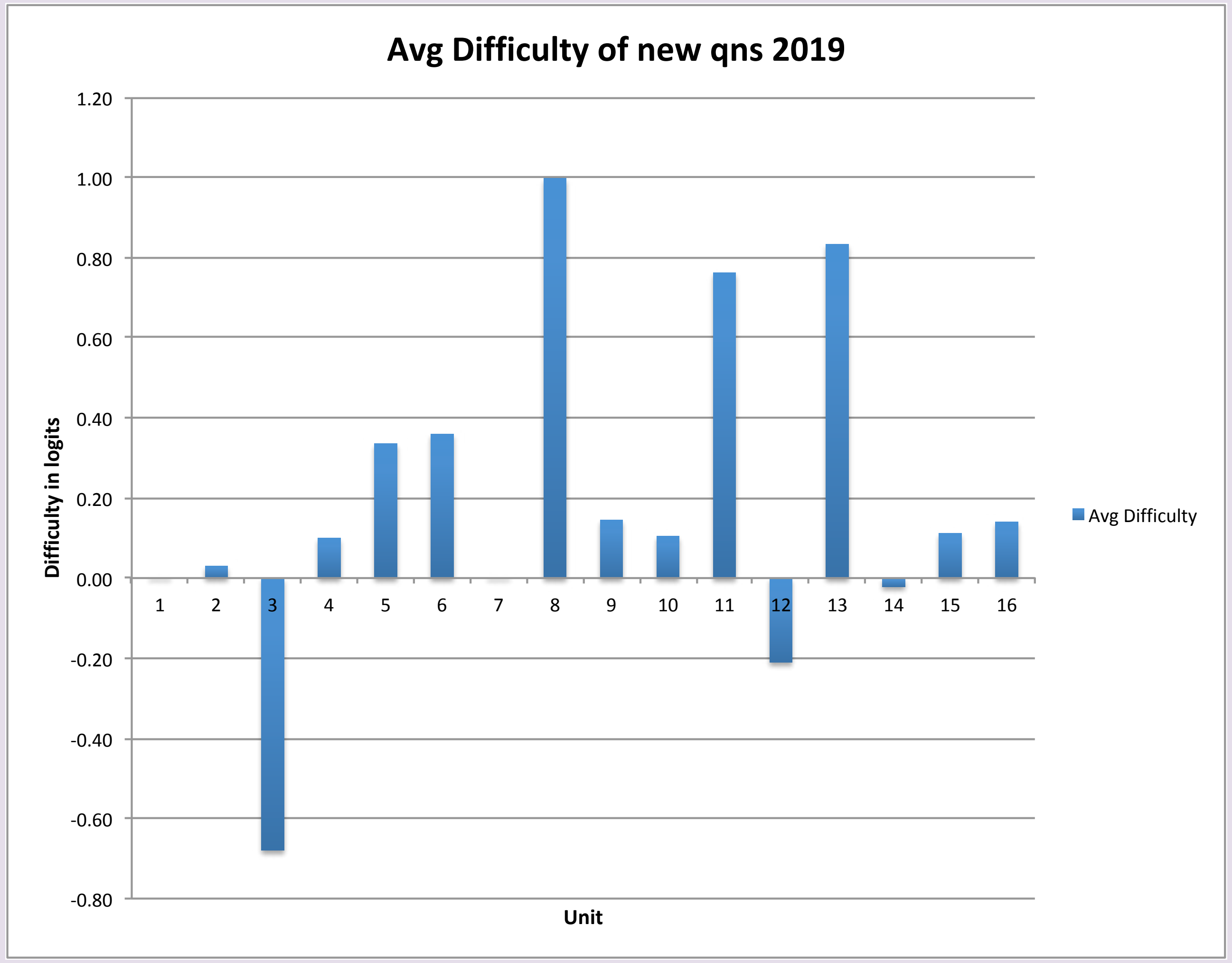


Figure 4:9 Average difficulty of new questions by unit, 2019

The number of new questions per unit varied from one question for Unit 16 to 24 questions for Unit 15. The number of new questions and the average difficulty of the new set of questions in the unit appear in Figures 4:8 and 4:9 and are summarised below in Figure 4:10.

Figure 4:10 Number of questions per unit and average difficulty, 2019

Unit	No of Items	Average Difficulty
Unit 1	0	
Unit 2	2	-0.03
Unit 3	2	-0.68
Unit 4	3	0.10
Unit 5	5	0.34
Unit 6	8	0.36
Unit 7	0	
Unit 8	12	1.00
Unit 9	12	0.15
Unit 10	17	0.10
Unit 11	15	0.76
Unit 12	6	-0.21
Unit 13	21	0.83
Unit 14	11	-0.02
Unit 15	24	0.11
Unit 16	1	0.14

The new questions will be added to the previously ratified questions in the question bank so the fact that some units have slightly more or less difficult new questions should not be an issue. As long as question setters continue to assemble tests with an overall difficulty average of (or close to) zero, the addition of some more or less difficult questions to the question bank will not affect future ABPI tests. If, however, all 12 of the new Unit 8 questions were to be used in one single test then this would have the effect of raising the difficulty of that particular test.

5 Conclusions and recommendations

5.1 ABPI question bank

The ABPI has now carried out analysis of 2,865 questions that have been taken by sufficient numbers of candidates to provide meaningful data.

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.

Figure 5:1 shows the number of questions analysed in each round from 2016 to 2019.

Year of analysis	Number of questions analysed	Number of good questions	Number of questions for review
2015 and 2016 combined	2260	1645	615
2017	427	331	96
2019	178	139	39
Total questions	**2865**	**2115**	**750**

5.1.1 Questions referred for review

Of the 2,865 questions available for analysis, a total of 750 questions were suggested for review before being used again, and 2,115 questions could be considered to have performed well enough to be added to the question bank.

Figure 5:2 Number of questions to be reviewed

Criteria for review	Number of questions		
	2015 & 2016	2017	2019
Too easy or too difficult	219	21	4
Poor correlation	459	50	28
More able choosing wrong answer	83	7	11

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 750.

The recommendation to the ABPI is that 39 questions from the 2019 analysis should be reviewed. It is also suggested that information should be gathered centrally regarding the typical causes of such unusual question performance. Although there are 750 questions which will have been reviewed by mid-2019, 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.

5.1.2 Good questions 2016–2019

The 2,115 good questions that were considered suitable for inclusion in the question bank are spread across the 16 units, as shown in Figure 5:3.

Unit 6 is a double-weighted unit and has the largest number of questions (228) available in the question bank. Unit 7 has the smallest number of available questions (78). Within these totals, however, there may be sub-sections of units that must be represented in each test, which would further restrict the choice of questions from the bank. Some units may require additional questions in the near future. It is also likely that some questions will be removed from the bank periodically as unit content changes, in order to keep up with new developments and with changes in regulations etc.

5.1.3 Creating tests of comparable difficulty

As ABPI staff now have the difficulty measures for every one of 2,115 good questions, it is possible to combine questions of known difficulty into a unit exam with the correct number of questions of known overall difficulty.

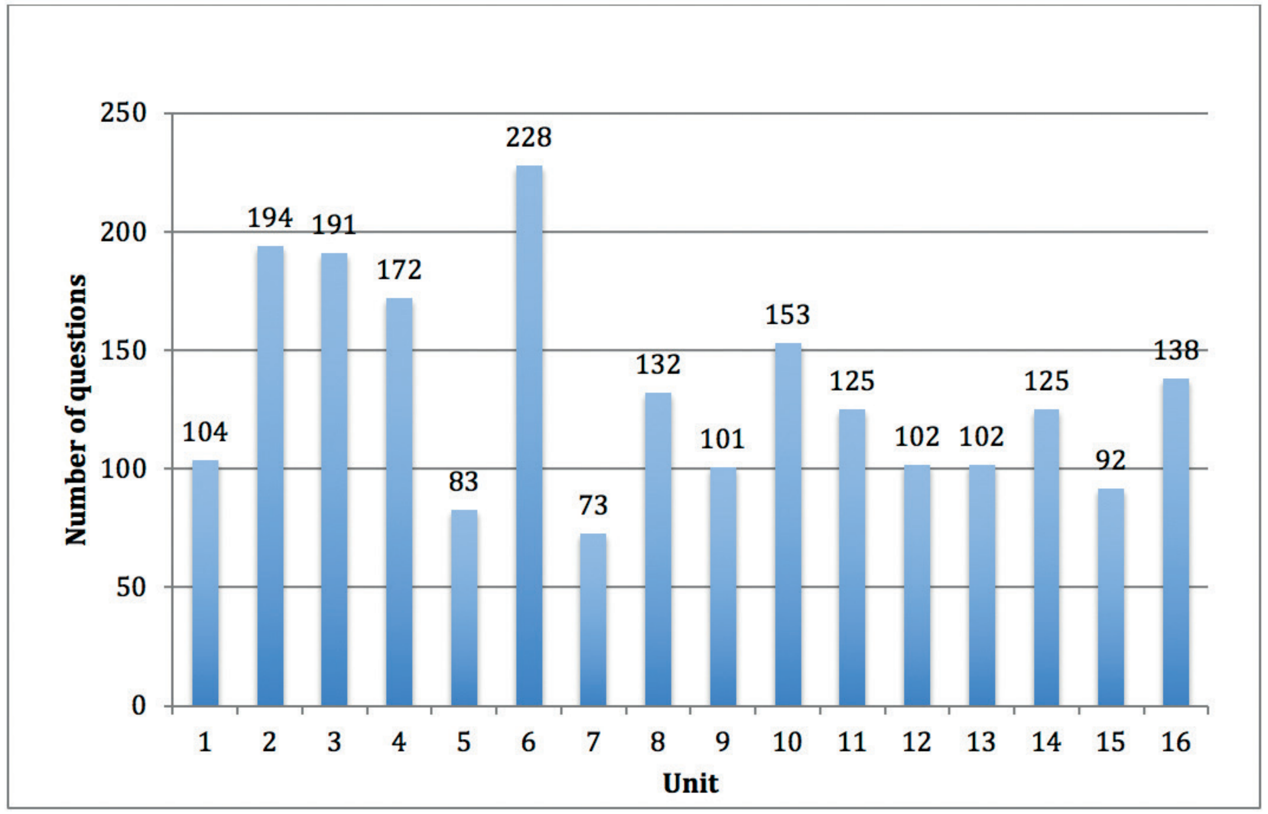


Figure 5:3 All good questions by unit, 2016–2019

5.2 Reflections and next steps

The ABPI has now commissioned three rounds of detailed test analysis on all the questions in the ABPI question bank. The 2016, 2017 and 2019 analyses show that of the 2,865 questions available for analysis,

- 750 questions should benefit from being reviewed
- 2,115 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 permanent staff and expert question writers.

At a time when it is not usual for UK professional organisations to provide validity evidence for 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 are leading the way in ensuring that their exams are reliable and give valid results that can be trusted by exam candidates and their employers.

6 References

i ABPI Code of Practice 2019 Clause 16.3
http://www.pmcpa.org.uk/thecode/Documents/ABPI%20Code%20of%20Practice%202019.pdf

ii ABPI Code of Practice 2019 Clause 16.3
http://www.pmcpa.org.uk/thecode/Documents/ABPI%20Code%20of%20Practice%202019.pdf

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)20 7930 3477   getintouch@abpi.org.uk   www.abpi.org.uk

abpi.org.uk

RMI-0124-0519