An evaluation of the difficulty of the assessments and the characteristics of the problem-solving (AO3) items

December 2017

Ofqual/17/6329

This report was written by Stephen Holmes, Emma Howard and Tim Stratton from Ofqual's Strategy Risk and Research directorate.

Contents

1 Executive summary

We carried out two separate investigations of the summer 2017 GCSE mathematics assessments to check inter-board comparability and comparability with the sample assessments in relation to perceived item (question) difficulty. One study used a comparative judgement approach to estimate the expected difficulty of the assessments. The second study collected expert ratings of the features of problem-solving item to determine if there were any differences in approach between the exam boards.

Comparative judgement is a technique in which a number of experts independently review many pairs of items and decide each time which item is more difficult to answer. This harnesses the human ability to make accurate relative judgements rather than absolute judgements, at which we are known to be quite poor. It has several useful characteristics, including capturing a group consensus well, and avoiding individual biases (leniency or harshness) in absolute judgements.

Mathematics PhD students carried out the comparative judgement exercise on the expected difficulty of all the items from the summer papers. We included some items from the sample assessment materials, which were published when the specifications were first accredited. This allowed us to make a direct comparison between the summer and the sample assessments. We found that there were only small differences between the expected difficulty of the summer 2017 assessments for the 4 exam boards. These were slightly larger than those seen between the sample assessments, but still small in terms of the likely effect on grade boundaries. Overall the foundation tier summer assessments were marginally less difficult, and the higher tier assessments marginally more difficult than the sample assessments. In addition, for both tiers the spread of difficulty of the summer assessments was slightly larger than in the sample assessments, which would have helped to differentiate candidates of different abilities.

We separately analysed a subset of items for which marks were designated as assessing problem solving skills, Assessment Objective 3 (AO3). There was slightly more difference between the expected difficulty of these AO3 items from different exam boards than was found across all items. In our second study, we investigated AO3 items more closely, asking experienced examiners to carry out a rating exercise on these items. We found very close correspondence between the exam boards on ratings of the features of the items that relate to good quality problem solving. There was more similarity of ratings across the exam boards for the summer 2017 AO3 items than we had found previously in the sample assessments. Both studies will provide helpful data to the exam boards in setting the difficulty of their future papers and in designing their problem solving items.

2 Background

This study focuses on the difficulty of exam boards' GCSE mathematics items in 2017 and the nature of the problem solving items. This continues a programme of work to evaluate the exam boards' GCSE maths sample assessments using similar measures[1]. This earlier work included multiple strands and several phases[2] and concluded with asking the boards to adjust their sample assessments to align their difficulty and ensure they could adequately differentiate between students[3]. The focus on items with multiple marks allocated to Assessment Objective 3 (AO3), which captures factors related to mathematical problem-solving, provided evidence which could be used to inform the design of future problem-solving items.

The current study considers the first live assessments for the new reformed GCSE maths qualifications sat by candidates in summer 2017, to compare the difficulty of the live assessments and the nature of the problem solving (AO3) items contained therein to those in the sample assessments.

This study includes 3 separate strands evaluating:

- overall assessment difficulty using comparative judgement to estimate expected item difficulty including a comparison with items from the sample assessments
- difficulty of AO3 items using a subset of the comparative judgement data
- ratings of items aimed at assessing AO3 on a range of features of quality .

[1] https://www.gov.uk/government/publications/gcse-maths-final-research-report-and-regulatory-summary
[2] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/440052/2015-06-30-gcse-maths-sample-assessment-materials-post-research-review.pdf
[3] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/440053/2015-06-30-regulatory-summary-gcse-maths-sample-assessment-materials-post-research-review.pdf

3 Comparative judgement study of the expected difficulty of the 2017 assessments and sample assessments

3.1 Method

The method used closely follows that employed in the previous study to evaluate the difficulty of GCSE maths items from the sample assessments. This involves the use of mathematics PhD students to judge the mathematical difficulty of items presented in pairs and analysed using a comparative judgement framework. The distribution of difficulty within and between papers and assessments can then be analysed and visualised.

3.1.1 Materials

Items were taken from the live summer 2017 GCSE maths papers from AQA, Pearson, Eduqas and OCR. Every item was included in the study, although common items found on both the foundation and higher tier papers were included as higher tier items only for judging. The results for these items were then duplicated and included for the foundation tier papers as well for analysis. In total there were 800 unique summer 2017 items, which increased to 916 items with the common items – see Table 1. The table details the number of items in the whole assessments for each exam board and tier, in both the live 2017 papers and the sample assessments against which the live assessments are compared.

Table 1*. Number of items across the two sets of papers*

	Summer 2017		**Sample assessments**	
	Foundation	Higher	Foundation	Higher
AQA	117	104	117	106
Eduqas	127	100	103	88
OCR	146	116	134	111
Pearson	118	88	111	96

All the items were formatted to give a consistent layout and font so differences between exam boards could not be identified. Marks available for each item were not visible to judges. Multi-part items were treated as a series of individual items for judging, although judges could see the other parts of the item as in some cases this may impact on the interpretation and difficulty of the item. When a calculator was allowed for a paper this was indicated at the top of each item by stating ‘Calculator Allowed’ where relevant.

3.1.2 Anchor items

One aim of this study was to allow direct comparison of the expected difficulty of the summer 2017 assessments to that of the sample assessments. Rather than include every single item from the sample assessments in this work and duplicate judging we had already carried out, we included some items of known expected difficulty from the sample assessments in the current study which we term 'anchor items'. One hundred anchor items were added to the 800 unique summer 2017 items to be judged. In order to cover the full extent of the difficulty scale, anchor items were drawn by sampling at equal intervals along the list of items from the sample assessment research ordered by difficulty, regardless of exam board or tier.

When the statistical model was fitted to the judgement data to estimate item expected difficulties, the expected difficulty parameters of the anchor items were fixed at the value obtained in the previous work. This ensures that the modelled scale of expected difficulty was the same between the current study and the earlier work and allows the direct comparison we required.

3.1.3 Participants

33 PhD students studying mathematics at English universities were recruited to judge the difficulty of the items. This included 16 judges who had participated in a previous A level judging study[4] carried out by Ofqual and had proven to be reliable judges.

3.1.4 Procedure

Comparisons were conducted using the online platform No More Marking. Judges were given instructions on how to access the platform and how to perform the judging. Pairs of items were presented to judges side by side on the screen and judges were prompted to select:

> 'Which item is more mathematically difficult to answer fully?'

This is the same prompt as used in the previous GCSE sample assessment study. After selecting the more difficult item (a 'judgement') a new pair of items were presented. Judges were given two weeks to complete 480 judgements each. They were free to complete these judgements as and when they liked. Items were distributed among judges so each item was judged a similar number of times, with a minimum of 29 judgements per item (maximum – 39, median – 33).

[4] A level and AS mathematics: An evaluation of the expected item difficulty. Ofqual report. To be published.

3.2 Analysis

The R package, Supplementary Item Response Theory (sirt[5]), was used to estimate expected difficulty parameters for each item using the Bradley-Terry model. Additional R code was also used to estimate item and judge infit[6], scale-separation reliability (SSR) and split-half reliability.

3.2.1 Judge consistency and exclusion

One judge was excluded from the analysis. This judge had a median judgement time of 6.7 seconds and an infit value of 1.43. For these kind of judgements we normally consider a median judging time below 10 seconds to indicate a possible lack of care. The high infit value (more than 2 standard deviations above the mean) supports this conclusion. The range of median judgement times for the other judges was 9.8 seconds to 35.6 seconds with an overall median of 21.4 seconds. Infit values for the other judges ranged from 0.73 to 1.32.

Median split-half reliability was assessed by repeatedly allocating judges to two groups, fitting a Bradley-Terry model to each group and correlating the two rank orders of item difficulty. Over 100 replications the mean correlation was 0.82 (sd=0.01). Reliability is quantified in comparative judgement studies by an SSR statistic that is derived in same way as the person separation reliability index in Rasch analyses. It is interpreted as the proportion of 'true' variance in the estimated scale values. The SSR was 0.91, indicating a low degree of variance in the item expected difficulty values.

3.3 Results

We obtained facility data (the average performance of candidates on items) from the exam boards for all of the summer 2017 items. Because the facility values for items across tiers are not equivalent, the common items were used to calculate an adjustment between tiers by calculating the average difference of candidate performance between tiers. Having equated the item facilities across tiers, the correlation of facility and expected item difficulty from this comparative judgement study was 0.62. This correlation is in line with the correlation of 0.66 (unadjusted) obtained from the earlier work on the sample assessments.

[5] Alexander Robitzsch (2015). sirt: Supplementary Item Response Theory Models. R package version 1.8-9. https://sites.google.com/site/alexanderrobitzsch/software

[6] Infit is a measure of the consistency of the judgements made by a judge or for an item compared to the overall model. A high judge infit value indicates that they were either inconsistent within their own judgements, or were applying different criteria from the consensus. High item infit suggests the item is difficult to judge.

Each assessment is shown in the figures in this section as a box plot displaying the median and inter-quartile range of the expected item difficulties on a logit scale on the y-axis. This probabilistic scale describes the log odds of one item being judged more difficult than another item. The absolute value is arbitrary, in this case 0 is set equal to the mean of all the items included in the earlier work on the sample assessments. The expected item difficulties have been weighted by the item tariff (maximum mark) by duplicating each item parameter by the number of marks for that item. Each mark on the paper is therefore treated as a 1-mark item, with the same difficulty for all marks within each judged item.

3.3.1 Foundation tier

Table 2 and Figure 1 show that for the foundation tier the range of median difficulties are very similar between the sample assessments and live exams. These ranges are small and indicate highly comparable assessments. These small differences are not substantive, and can easily be accounted for in awarding with small adjustments to the grade boundaries. While AQA and Pearson have similar median difficulties across their live and sample assessments, Eduqas and OCR have slightly lower difficulty in the live assessments. Combining all of the items from the 4 exam boards, the summer 2017 assessments have a median difficulty of -0.22, compared to -0.04 for the sample assessments, showing a small reduction in difficulty for the summer tests. The spread of item difficulties (indicated by the width of the boxplots and whiskers in Figure 1) has also increased in the summer 2017 assessments, particularly with an indication of more low-difficulty items. This may have helped to make the assessments more accessible for the lowest-achieving candidates.

Table 2*. Median, mean and standard error of item difficulties for all foundation tier summer 2017 and sample assessments*

Foundation	Sample assessments			Summer 2017		
AO	Median	Mean	SE	Median	Mean	SE
AQA	-0.16	-0.29	0.07	-0.13	-0.36	0.09
Eduqas	-0.05	-0.19	0.07	-0.37	-0.32	0.09
OCR	0.15	-0.20	0.06	-0.24	-0.26	0.08
Pearson	-0.13	-0.26	0.07	-0.13	-0.31	0.09
Range	0.31	0.10		0.24	0.10	

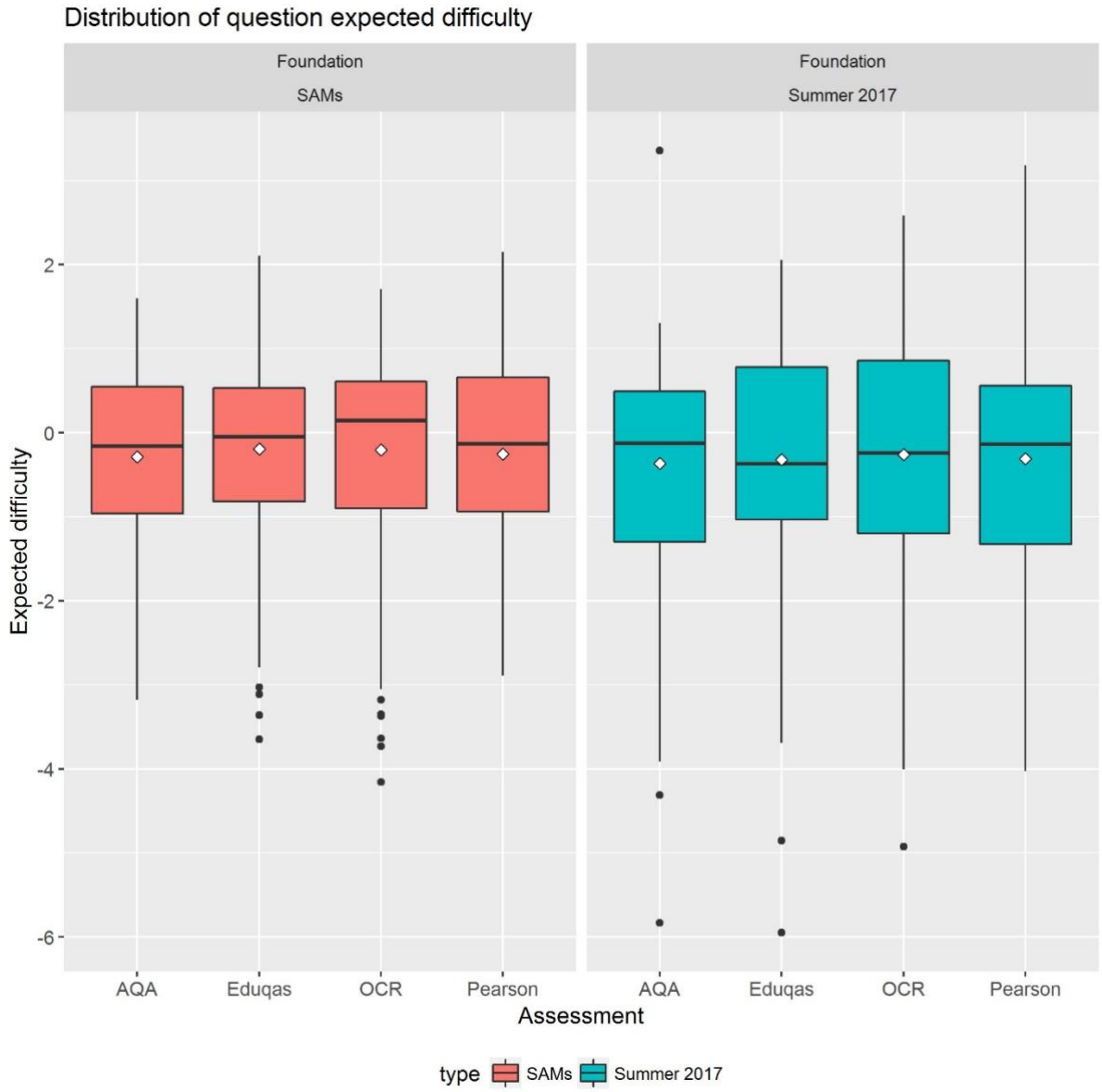


Figure 1. *Boxplots showing median and mean (white diamond) item difficulty aggregated across all exams for each exam board for foundation tier exams, weighted by item tariff.*

3.3.2 Higher tier

For the higher tier (see Table 3 and Figure 2), the range of median difficulties is higher for the summer 2017 assessments than for the sample assessments. The summer 2017 Eduqas assessment is somewhat more difficult while the OCR assessment is a little less difficult. However, the differences are not substantive, and can easily be accounted for in awarding with small adjustments to the grade boundaries. The mean values are slightly different which is consistent with the slight

skew visible in most of the summer 2017 boxplots. Greater skew and a greater range of median difficulties than is seen in the sample assessments is not surprising given that the live papers have not been through multiple rounds of adjustments.

Comparing the sample and live assessments for the higher tier, OCR's papers are closely matched in the median difficulty, while the other 3 exam boards' live assessments are slightly more difficult than their sample papers. When the items from the four exam boards are combined, the median expected difficulty is 0.96 for the summer 2017 assessments and 0.73 for the sample assessments. While the foundation tier was around 0.2 less difficult, the higher tier was around 0.2 more difficult than the sample assessments.

It is apparent from the boxplots in Figure 2 that the spread of item difficulties is larger for the live papers than the sample assessments, with more high difficulty items for most of the exam boards, but also some additional lower-difficulty items. This can only aid in differentiating between candidates of different abilities.

Table 3*. Median and standard error of item difficulties for all higher tier summer 2017 and sample assessments.*

Higher	Sample assessments			Summer 2017		
AO	Median	Mean	SE	Median	Mean	SE
AQA	0.68	0.72	0.07	0.99	0.70	0.08
Eduqas	0.77	0.73	0.06	1.21	1.10	0.08
OCR	0.69	0.61	0.05	0.71	0.57	0.07
Pearson	0.73	0.71	0.07	1.01	1.12	0.08
Range	0.09	0.12		0.50	0.55	

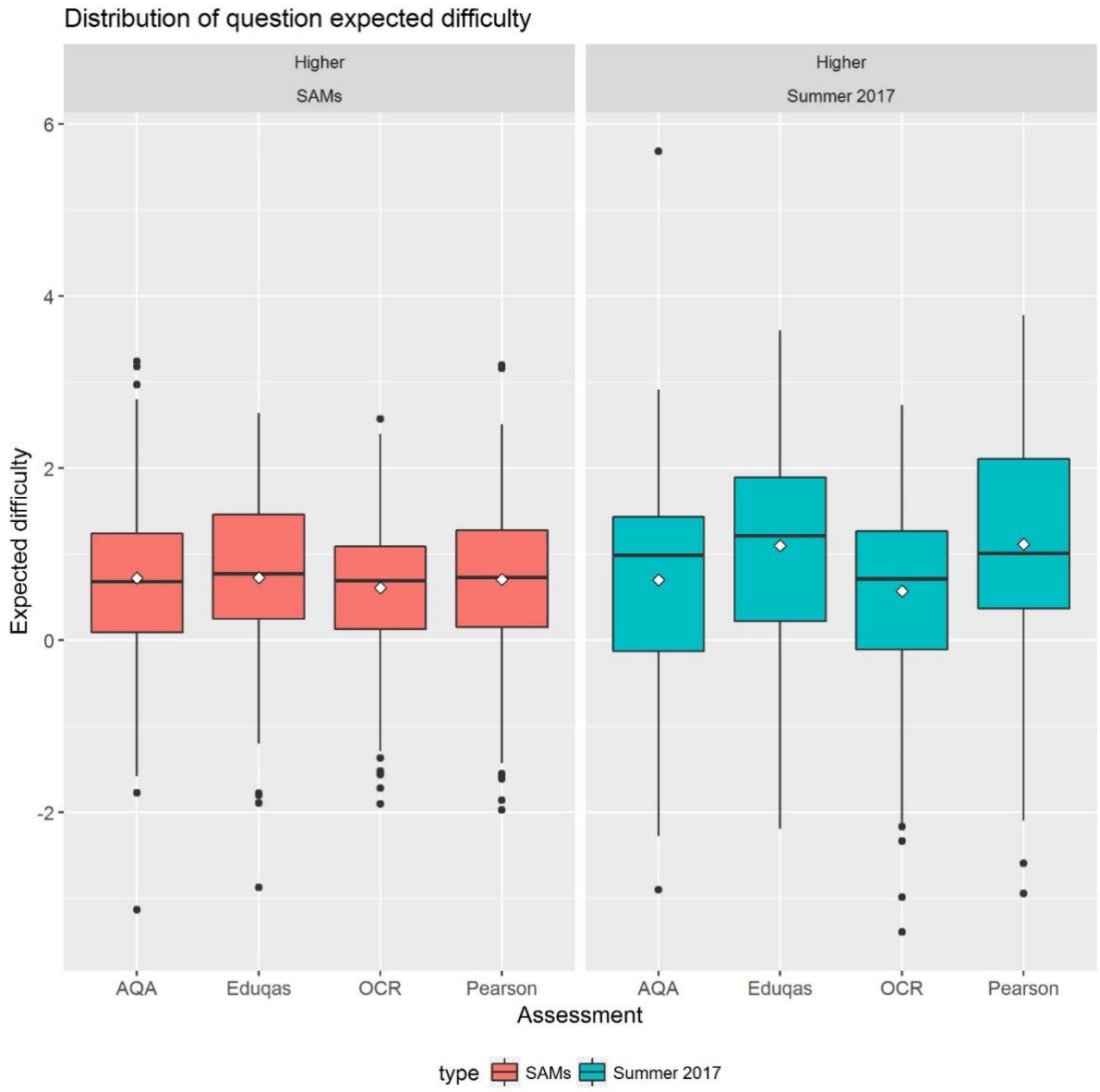


Figure 2. *Boxplots showing median and mean (white diamond) item difficulty aggregated across all exams for each exam board for foundation tier exams, weighted by item tariff.*

3.4 Example items

This section shows the five highest (Figure 3 to Figure 7) and lowest (Figure 8 to Figure 12) rated summer 2017 items, based on their expected difficulty in the study, for information. For multi-part items the item receiving the rating is the part highlighted in yellow in the figure. The assessment objective mark assignments are given in the figure captions.

3.4.1 Items with the highest expected difficulty

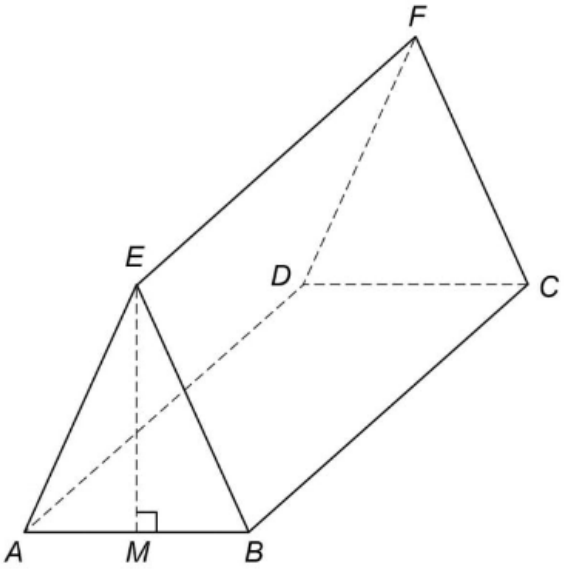
Calculator allowed

Rectangle *ABCD* is the horizontal base of a triangular prism *ABCDEF*.

$AE = BE$

E is vertically above *M*, the midpoint of *AB*.

$AB = 16$ cm $\quad AE = 17$ cm $\quad BC = 30$ cm



a) Show that $EM = 15$ cm

b) Work out the size of angle *ECM*.

Figure 3. *The item with the highest difficulty (AQA – Paper 3 – Higher, Item 25b, item score = 5.68, targeting AO1 - 1 mark and AO3 - 3 marks)*

Calculator allowed

L is the circle with equation $x^2 + y^2 = 4$

$P\left(\frac{3}{2}, \frac{\sqrt{7}}{2}\right)$ is a point on **L**.

Find an equation of the tangent to **L** at the point *P*.

Figure 4. *The item with the second highest difficulty (Pearson – Paper 2 – Higher, Item 23, item score = 3.78, targeting AO1 - 1 mark and AO2 - 2 marks)*

a) The diagram shows a large shipping container at rest on horizontal ground.

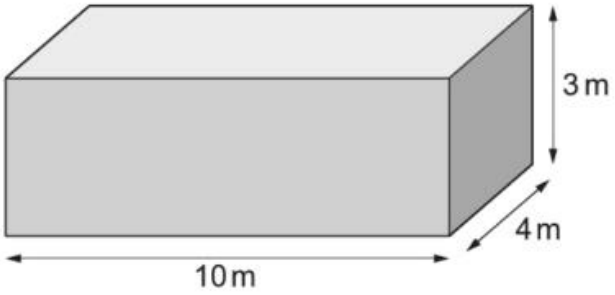


Diagram not drawn to scale

The weight of the container is 32 000 N.

Work out the pressure exerted on the ground by the shipping container.

Give your answer in **N/m²**.

b) A table is at rest on horizontal ground.

The table has 4 legs.

Each leg has a height of 50 cm.

The volume of material in one leg is 450 cm^3.

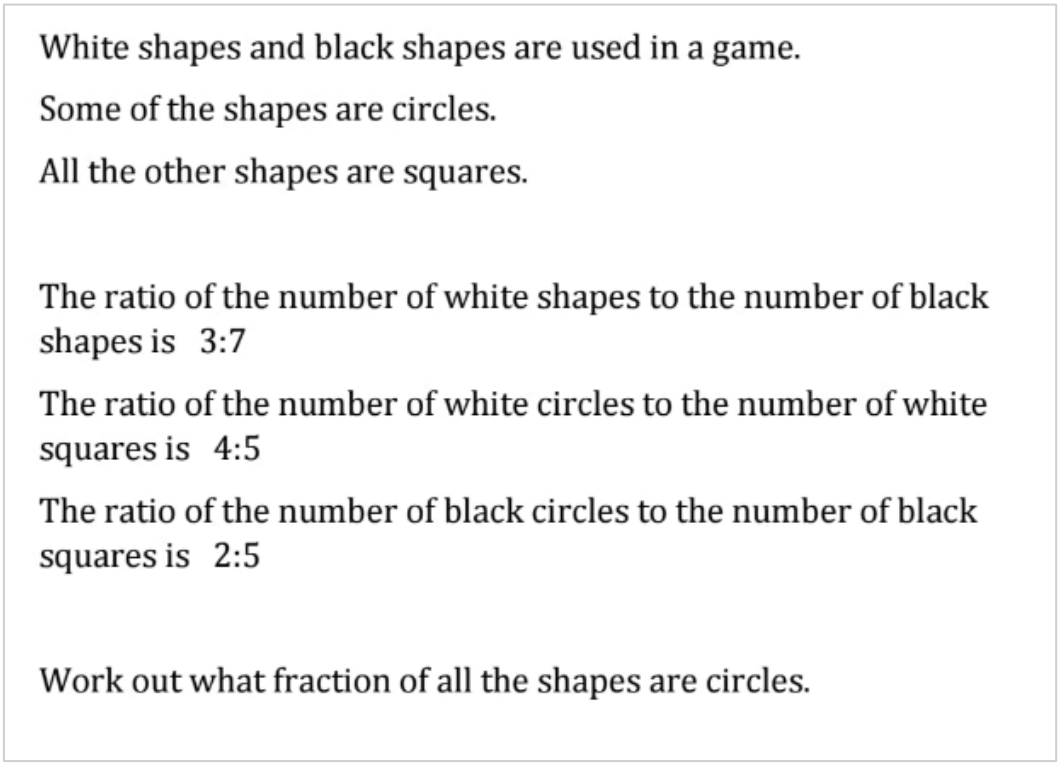
The table weighs 54 N.

By considering the base of the table legs, work out the pressure exerted on the ground by the table.

Give your answer in **N/cm²**.

You must show all your working.

Figure 5. *The item with the third highest difficulty (Eduqas – Paper 1 – Higher, Item 9b, item score = 3.6, targeting AO1 - 2 marks and AO3 - 3 marks)*


White shapes and black shapes are used in a game.

Some of the shapes are circles.

All the other shapes are squares.

The ratio of the number of white shapes to the number of black shapes is 3:7

The ratio of the number of white circles to the number of white squares is 4:5

The ratio of the number of black circles to the number of black squares is 2:5

Work out what fraction of all the shapes are circles.

Figure 6*. The item with the fourth highest difficulty (Pearson – Paper 1 – Higher, Item 14, item score = 3.4, targeting AO1 - 1 mark and AO3 - 3 marks)*

Calculator allowed

Ellen works for a company that sells cars.

Her **monthly** pay is

- a salary of £1470
- 28% of the total **profit** the company makes from her sales
- a £250 bonus **if** she sells at least 15 cars.

The table shows information about the cars she sold last year.

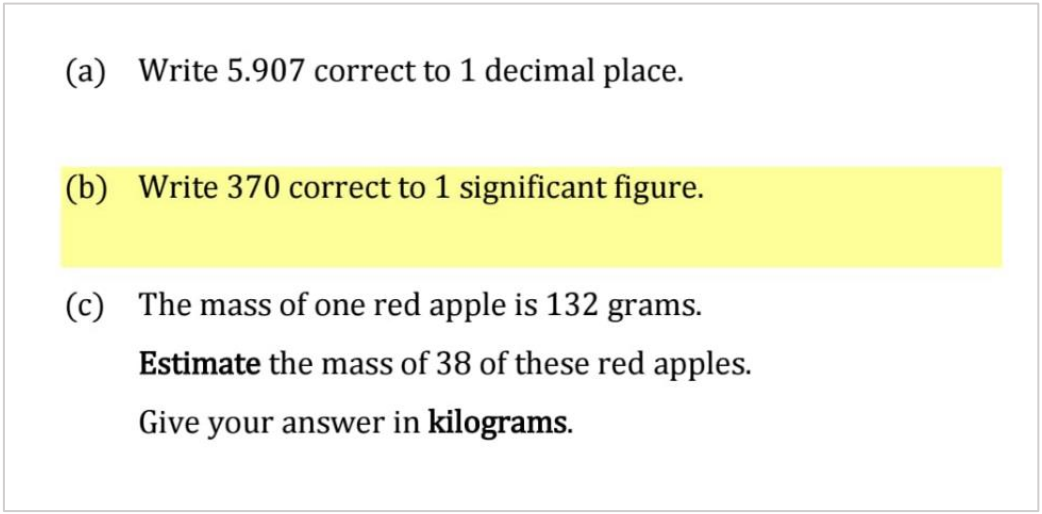
Total cost to the company	Total income for the company	Number of months when she sold at least 15 cars
£464 500	£538 000	3

Was Ellen's total pay for the **year** more than £40 000?

You **must** show your working.

Figure 7. *The item with the fifth highest difficulty (AQA – Paper 2 – Foundation, Item 18, item score = 3.36, targeting AO1 - 1 mark and AO3 - 5 marks)*

3.4.2 Items with the lowest expected difficulty


(a) Write 5.907 correct to 1 decimal place.

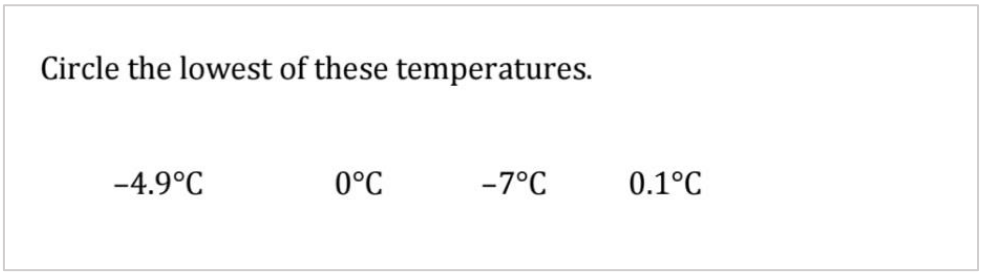
(b) Write 370 correct to 1 significant figure.

(c) The mass of one red apple is 132 grams.

Estimate the mass of 38 of these red apples.

Give your answer in **kilograms**.

Figure 8*. The item with the lowest difficulty (Eduqas – Paper 1 – Foundation, Item 6b, item score = -5.95, targeting AO1 - 1 mark).*


Circle the lowest of these temperatures.

−4.9°C 0°C −7°C 0.1°C

Figure 9*. The item with the second lowest difficulty (AQA – Paper 3 – Foundation, Item 1, item score = -5.83, targeting AO1 - 1 mark).*

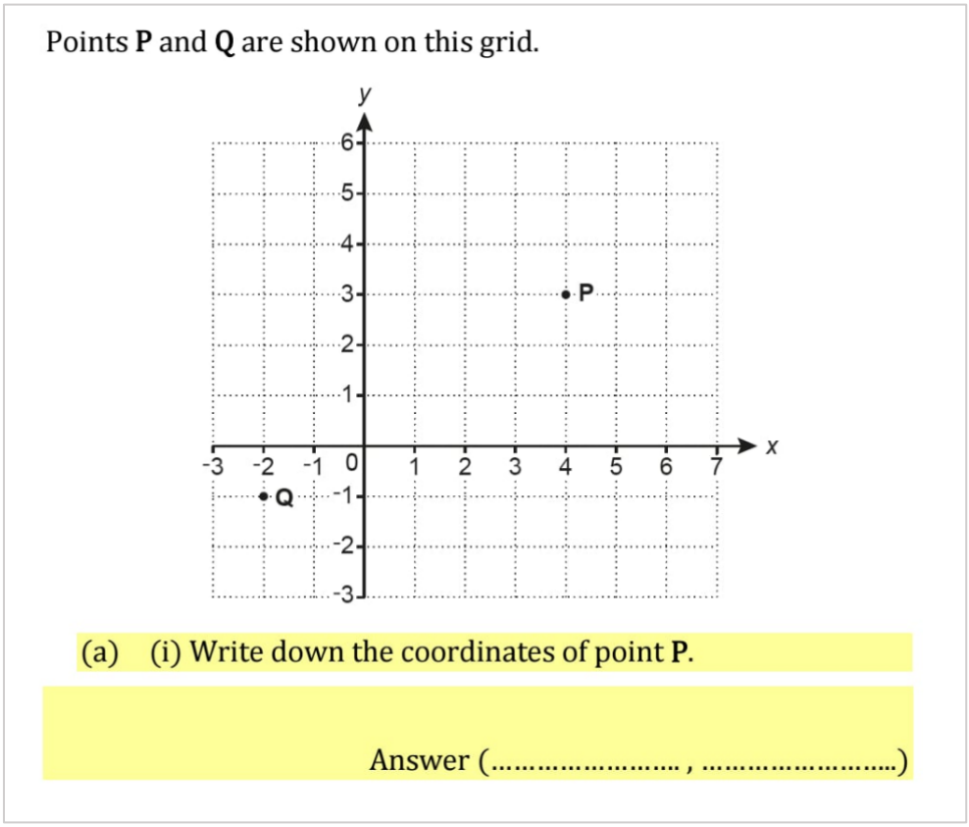


Figure 10. *The item with the third lowest difficulty (OCR – Paper 1 – Foundation, Item 5ai, item score = -4.93, targeting AO2 - 1 mark)*

Figure 11. *The item with the fourth lowest difficulty (Eduqas – Paper 2 – Foundation, Item 7a, item score = -4.86, targeting AO1 - 1 mark)*

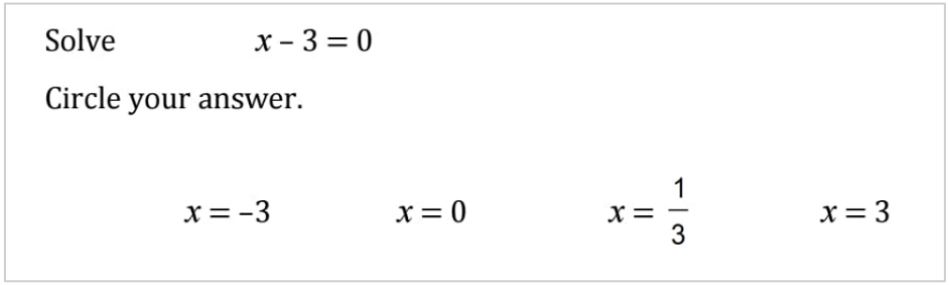


Figure 12*. The item with the fifth lowest difficulty (AQA – Paper 1 – Foundation, Item 4, item score = -4.31, targeting AO1 - 1 mark)*

3.5 Discussion

Differences between the assessments from the four exam boards were not large. For the foundation tier, the range of assessment median difficulties was the same for summer 2017 as the sample assessments. For higher tier, the difference between assessments was larger, but a difference of around 0.5 is not very large. Although we cannot assume that the cohorts taking these assessments with each exam board are of equivalent ability, in actual awarding the grade boundaries are located fairly centrally in the mark distributions (grade 3-5 for foundation tier and grade 6 and 7 for higher tier) varied by no more than 10 percent across exam boards. This provides some support that the differences seen in the median expected difficulty here do not represent substantive differences.

Within each exam board, the summer 2017 assessments were fairly close in median difficulty to their respective sample assessments, with the largest difference seen for the Eduqas higher tier summer 2017 assessment which was almost 0.5 more difficult than the sample assessment. Consistent with the grade boundary differences noted above, in the previous sample assessment research, a difference of 0.5 on the expected difficulty scale corresponded with a mean mark difference of around 10 percent. Averaged across all exam boards the foundation tier summer 2017 assessments were slightly less difficult than the SAMs, while the higher tier summer 2017 assessments were slightly more difficult. The distributions of item difficulty were also generally wider in the summer 2017 assessments, which would have helped in differentiating between candidates of different abilities. Given that the sample assessments passed through several rounds of modification in order to closely align their difficulties, the small inter-board differences in difficulty between the summer 2017 assessments is acceptable.

4 Problem-solving item difficulty

One of the changes to the reformed GCSE mathematics was more of an emphasis on problem solving. This was implemented through changes to the wording of Assessment Objective 3 (AO3), which captures problem-solving features (see Figure 13), and an increase in the proportion of AO3 marks in the whole assessment.

Because of the importance of this change, in our previous investigation of the sample assessments we looked closely at the characteristics of a set of items with several AO3 marks allocated. The first investigation on problem-solving items from the summer 2017 papers, was to evaluate their difficulty using the comparative judgement data described in section 3.

AO3: Solve problems within mathematics and in other contexts

Students should be able to:

- translate problems in mathematical or non-mathematical contexts into a process or a series of mathematical processes
- make and use connections between different parts of mathematics
- interpret results in the context of the given problem
- evaluate methods used and results obtained
- evaluate solutions to identify how they may have been affected by assumptions made

Figure 13*. Assessment Objective 3*

4.1 Methods

Items with marks allocated to AO3 were identified from mark allocation data provided by each exam board[7]. These were not necessarily all predominantly "problem-solving" type items, since only a single AO3 mark was required for selection, but at a minimum they contained elements of problem-solving according to the exam boards' own classification. AO3 marks make up 25% of the foundation tier assessments, and 30% of the higher tier assessments.

These items were then extracted from the full set of comparative judgement data. Table 4 shows the count of items. Common items are counted in both tiers.

[7] Comparable data was not available for the final set of sample assessments and so the analysis in this section is restricted to the summer 2017 items.

Table 4. *Number of identified AO3 items by exam board and tier.*

Board	Foundation Tier	Higher Tier	Total
AQA	25	29	54
Eduqas	30	30	60
OCR	43	43	86
Pearson	23	28	51
	121	*130*	*251*

4.2 Results

In the following analysis the AO3 items were weighted by the number of AO3 marks assigned to them (rather than the total number of marks per item in the overall comparative judgement analysis in section 3), so items with more AO3 marks contributed more to the overall AO3 difficulty distribution. The expected item difficulties weighted by AO3 marks are shown for the two tiers in Figure 14 and Figure 15.

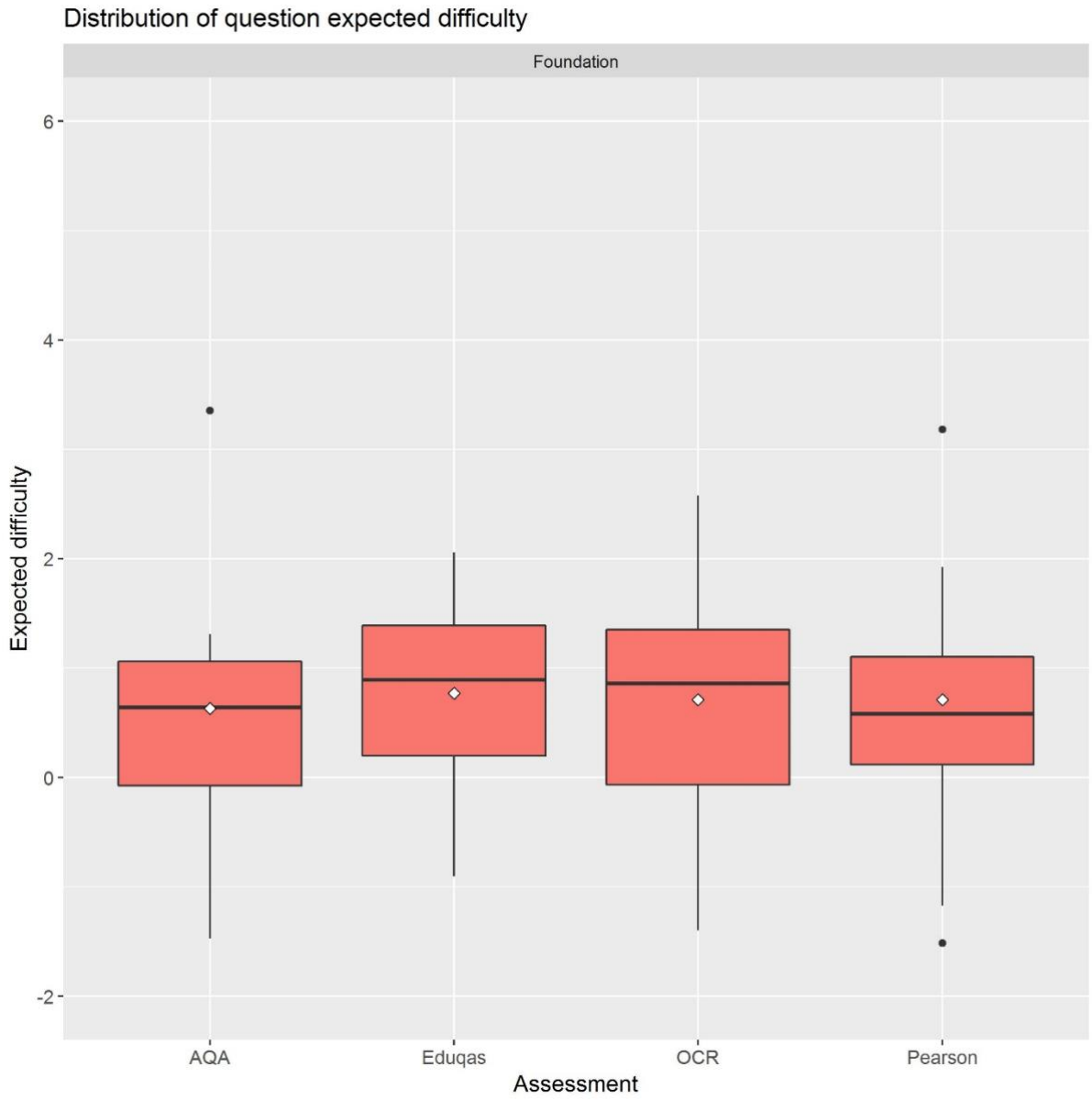


Figure 14: *Box plots showing median, mean and interquartile ranges of expected item difficulties for AO3 items from the foundation tier summer 2017 assessments by exam board.*

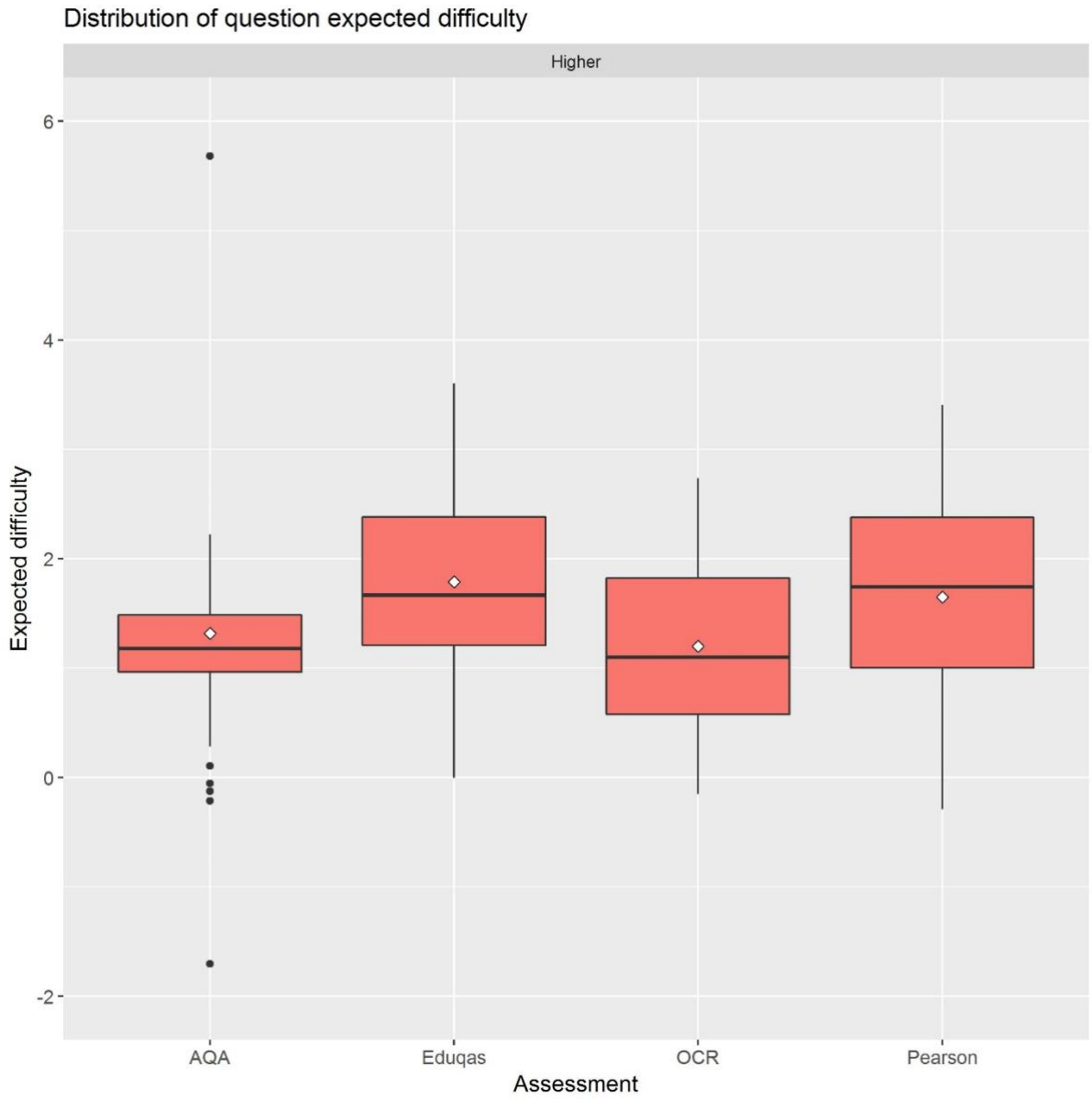


Figure 15: *Box plots showing median, mean and interquartile ranges of expected item difficulties for AO3 items from the higher tier summer 2017 assessments by exam board.*

The median expected difficulty values for both the AO3 items and the overall assessments are listed in Table 5. This also lists the difference between these two median measures to show how AO3 items differ from the whole assessment.

Table 5. *Median difficulty of items containing AO3 marks, all items ('overall'), and difference between these two values, by tier and exam board.*

board	AO3 median	Overall median	Difference
Foundation tier			
AQA	0.64	-0.13	0.77
Eduqas	0.89	-0.37	1.26
OCR	0.86	-0.24	1.10
Pearson	0.58	-0.13	0.72
Higher tier			
AQA	1.18	0.99	0.19
Eduqas	1.67	1.21	0.46
OCR	1.10	0.71	0.41
Pearson	1.74	1.01	0.73

AO3 items are perceived to be harder than the average item difficulty. For foundation tier, the estimated difficulty of AO3 items is comparable between exam boards (range of 0.58 to 0.89) with Eduqas and OCR having the hardest AO3 items by a small margin. The AO3 items are clearly estimated to be amongst the very hardest items on the foundation tier, with a judged median difficulty around 1.0 higher than the overall median difficulty.

For higher tier, there is a bigger difference between exam boards. Eduqas and Pearson papers are estimated to be slightly more difficult judged on overall median difficulty, and this is partly explained by the AO3 items which are judged to be more difficult than those of AQA and OCR (by around 0.5-0.6). This suggests that much of the difference in estimated overall difficulty could arise from AO3 items, particularly for Pearson.

For the higher tier Pearson assessment, the difference between the median of the AO3 items and the overall assessment is of a similar magnitude to the foundation tier assessments. However the higher tier AO3 items of AQA, Eduqas and OCR are only around 0.2-0.5 more difficult on average than the overall median. Figure 15 suggests that AQA may lack stretching AO3 items, while Eduqas and Pearson include some

high estimated difficulty AO3 items which are contributing to their slightly higher overall difficulty.

5 Problem-solving item feature analysis

In conjunction with the analysis of problem solving item difficulty described in section 4, we also analysed the characteristics of a subset of these items. Forty-five whole items with the largest number of marks assigned to AO3 were rated by a group of examiners and subject experts against a set of dimensions representing features of the items taken from our previous research into the 2015 GCSE maths sample assessments[8]. This rating exercise took place at a day-long meeting, which allowed discussion of the items and how they functioned as problems. The aim of this study was to see if there are any board-specific patterns in the way problem-solving items are designed.

5.1 Methods

5.1.1 Rating dimensions used

The problem-solving items were rated against 11 separate dimensions, together with an overall rating of how good they were at eliciting problem solving. Our previous research into sample assessments used the Kelly's Repertory Grid[9] method to generate a set of dimensions on which 33 items with 4 or more AO3 marks varied. We obtained 23 dimensions using this approach.

Following publication of the main report, our follow-up analysis[10] showed that a subset of these dimensions were significantly correlated with ratings of the quality of problem solving elicited by the items. These 8 dimensions (see Table 6) were included in the current study, together with 3 additional dimensions that although not significantly related to problem-solving quality in the previous study, were either explicitly mentioned in AO3, or are generally considered to capture a desirable quality for problem-solving items.

We asked our participants to give an integer rating from 1 to 5 for each item on each dimension. We also collected an overall rating of problem solving elicited by the item. This rating was made after all the other ratings, giving participants the greatest opportunity to consider all aspects of the item. For this final rating we used the same scale from 1 to 5 but allowed participants to use decimals if they wanted. This rating

[8] https://www.gov.uk/government/publications/gcse-maths-final-research-report-and-regulatory-summary

[9] Kelly, G.A. (1955). The Psychology of Personal Constructs: Vols 1 and 2. (New York: WW Norton).

[10] Holmes, S.D., He, Q. and Meadows, M. (2017) An investigation of construct relevant and irrelevant features of mathematics problem-solving questions using comparative judgement and Kelly's Repertory Grid, Research in Mathematics Education, 19:2, 112-129, DOI: 10.1080/14794802.2017.1334576

took the place of the problem-solving quality generated in Phase 3 (the comparative judgement exercise) of the previous sample assessment research.

Table 6. *Dimensions used for the problem-solving item rating exercise. The dimensions in bold text were significantly related to problem-solving quality in our previous work. For each dimension a rating of 1 was associated with poorer problem-solving quality while 5 indicated high quality.*

Pole with rating of 1	Pole with rating of 5
Numerical / mathematical answer	**Open-ended written answer**
Low level of language demand	**High level of language demand (unusual words used)**
General knowledge not needed	**General knowledge needed**
Little or no text to be read	**High quantity of text to be read**
No selection of parameters to do the calculation	**Requires selection of parameters to do the calculation**
Requires using obvious standard method	**No obvious standard method**
Obvious first step	**Non-obvious first step**
Single approach	**Multiple possible approaches**
Does not require evaluation of assumptions	Requires student to evaluate assumptions made
Does not require connections between different parts of maths	Requires connections between different parts of maths
Intermediate steps given or implied	Intermediate steps not obvious
	Overall rating of problem solving elicited by the item (1 = low, 5 = high)

5.1.2 Selection of items

Initially, whole items with 4 or more AO3 marks allocated from the summer 2017 mathematics papers, were selected, based on the allocation of marks to assessment objectives provided by the exam boards. As this led to an unbalanced number of items per exam board, some additional items with 3 AO3 marks were also randomly selected, predominantly from papers for which we had an examiner attending our meeting (see Participants below). In total, there were 45 items included in the study. These items included all parts of the numbered question, although we informed our participants which parts contained no AO3 marks and they were instructed to ignore these parts when making their ratings. A summary of the items included in this study, their tier and mark tariffs is shown in Table 7.

Table 7. *Summary of items included in this evaluation. For tier, F = Foundation tier, C = Common item on both tiers, H = Higher tier.*

Board	Number of items	Tier (F/C/H)	Minimum AO3 mark	Maximum AO3 mark	Minimum total mark	Maximum total mark
AQA	11	5/2/4	3	5	4	8
Eduqas	12	3/2/7	4	8	4	10
OCR	11	2/3/6	3	5	4	13
Pearson	11	4/4/3	3	4	4	6

5.1.3 Participants

In total, 13 participants were recruited for the study.

With assistance from each of the four boards, eleven examiners were recruited (3 from AQA, OCR and Pearson, and 2 from Eduqas). We asked the exam boards to put forward experienced markers who had examined on at least one of the new specification GCSE mathematics papers sat in summer 2017, but had not been involved in item writing for these papers. Eight were Assistant Principal Examiners and two were experienced assistant examiners. The examiners recruited covered both foundation and higher tier papers and marked papers from which 58% of the study items had been drawn.

In addition to the ten exam board examiners, two Ofqual subject experts were recruited from Ofqual's subject expert pool[11]. The subject experts had relevant subject qualifications and experience such as teaching, item writing and/or examining in the subject. The subject experts had no involvement with the GCSE mathematics papers sat in summer 2017 i.e. they had not taught the specifications, examined, or written items for any of the exam boards.

5.1.4 Materials

As well as providing question papers and mark schemes sat in summer 2017, the exam boards also provided assessment grids (showing the mark allocations and domain) and Principal Examiner reports (or equivalent). These reports were used as

[11] Ofqual looks for certain types of experience, qualities and characteristics in our external experts. Eligibility criteria to become a subject expert can be found here: https://www.gov.uk/guidance/apply-to-become-an-external-advisor-to-ofqual

a source of additional information about how the items had functioned in measuring problem-solving item and how candidates had tackled them.

5.1.5 Procedure

The examiners and subject experts were invited to a one day group meeting in which the aim was to rate the AO3 items against the 11 dimensions and to rate the overall problem solving elicited by the item. There were three parts to the meeting: familiarisation with the task, evaluation of items with commentary, and evaluation of items without commentary (where no examiner who had marked that paper was present).

Familiarisation. Participants familiarised themselves with the dimensions. Each dimension was presented in turn, with an explanation and discussion of the meaning and application of the dimension. For each dimension, participants practiced evaluating three items from the sample assessments, and were then shown the average scores awarded in the previous work and given an opportunity to discuss any differences or issues with applying the dimension. Participants reported confidence in their ability to evaluate items along these dimensions after this task.

Evaluation of study items with commentary. For items from papers where an examiner was present, the examiner provided a brief verbal commentary on how the item functioned, drawing on their experience of marking many responses in the summer. This commentary included points such as candidate performance, aspects of the item that encouraged problem solving, the number of approaches taken, and any notable/novel approaches to answering the item.

For each item, the group followed the following procedure:

1. independent initial ratings by all participants for each dimension were recorded;
2. a brief commentary was given by the relevant examiner and supplemented with details from the Principal Examiner's Report. Participants were able to ask items and discuss the functioning of the item;
3. independent adjustments were made to ratings in light of the commentary and discussion.

Evaluation of items without commentary. For items from papers where no examiner attended, participants worked through items on their own giving independent ratings. Participants were allowed to discuss with their neighbours, but did not do so in the majority of cases. Full discussion round the table was not encouraged as everyone was working through the items at different rates.

Participants undertook familiarisation at the beginning of the meeting day. The item rating was carried out in 4 sessions with both the morning and afternoon involving one session where items were discussed (with commentary) followed by a session

without commentary. The order of items presented in each session was random. The dimensions were scored in the order as presented in Table 6.

5.2 Results

5.2.1 Reliability of rating scores

All ratings were made on a scale from 1 to 5. Mean ratings and mean standard deviations were calculated across all items for each dimension (Table 8). The mean standard deviation indicates the consistency of ratings between participants, with large deviations for dimensions that were more problematic to rate, and small deviations for dimensions that were less problematic to rate. All mean standard deviations were between 0.5 and 1. As in Holmes et al. (2017)[12], the dimensions with low variability were those which captured surface features. For instance, in the current study, determining whether the response required numerical/mathematical answers, or an open-ended written answer (mean SD = 0.55), and the amount of text to be read (mean SD = 0.65). The dimension with greatest variability between judges involved the determination of parameters required to do the calculation (mean SD = 0.97).

The data was further analysed for inter-rater reliability using a two-way random effects intra correlation coefficient (ICC) model with the consistency agreement measure. Inter-rater reliability was good to excellent, with ICC estimates falling between .80 and .96 (see Table 8). ICC estimates generally mirrored the variability in ratings, i.e., where mean SD was lower, ICC estimates were higher.

Generally, the mean ratings were below the mid-point (i.e. 3) of the scale for each dimension. In particular, the dimension 'requires student to evaluate assumptions made' has the lowest mean rating (mean = 1.75). This may be surprising since this is explicit in the wording of AO3: one of the bullet points is *'evaluate solutions to identify how they may have been affected by assumptions made'.* It may therefore be expected that this feature scored more highly. However, this dimension also received a score of 1.92 for the SAMs and so it appears that this may be a feature that is assessed more sparingly on AO3 items.

The dimension with the highest mean rating was 'requires selection of parameters to do the calculation' (mean = 3.10), followed by 'intermediate steps not obvious' (mean = 2.90). Both of these elements are central to what makes a problem – the lack of an easily identifiable and easily applied standard approach. Interestingly the next highest dimension was the 'overall rating of problem solving elicited by the question'

[12] Holmes, S.D., He, Q. and Meadows, M. (2017) An investigation of construct relevant and irrelevant features of mathematics problem-solving questions using comparative judgement and Kelly's Repertory Grid, Research in Mathematics Education, 19:2, 112-129, DOI: 10.1080/14794802.2017.1334576

(mean = 2.88). Our participants thought that on average there was a reasonable level of problem-solving quality demonstrated, and the fact this was rated higher than most of the dimensions, probably indicates that no one feature is required to make a good problem-solving item, and each problem contains a different subset of features.

Table 8. *Mean ratings, standard deviations of ratings and ICC estimates and ICC 95% confident intervals for the rating of problem-solving and 11 dimensions.*

Dimension		Mean rating	Mean SD	Inter-rater reliability - ICC estimate consistency
Pole with rating of 1	Pole with rating of 5			
Overall rating of problem solving elicited by the question: Low	Overall rating of problem solving elicited by the question: High	2.88	0.68	0.85
Numerical / mathematical answer	Open-ended written answer	1.88	0.55	0.96
Low level of language demand	High level of language demand (unusual words used)	2.20	0.82	0.89
General knowledge not needed	General knowledge needed	1.80	0.66	0.91
Little or no text to be read	High quantity of text to be read	2.77	0.65	0.96
No selection of parameters to do the calculation	Requires selection of parameters to do the calculation	3.10	0.97	0.82
Requires using obvious standard method	No obvious standard method	2.52	0.83	0.81
Obvious first step	Non-obvious first step	2.31	0.89	0.85
Single approach	Multiple possible approaches	2.42	0.83	0.83
Does not require evaluation of assumptions	Requires student to evaluate assumptions made	1.75	0.56	0.94
Does not require connections between different parts of maths	Requires connections between different parts of maths	2.23	0.76	0.80
Intermediate steps given or implied	Intermediate steps not obvious	2.90	0.88	0.82

5.2.2 Difference in AO3 items across Boards

Previous work showed that in the sample assessments the ratings of some dimensions capturing features of problem-solving differed between the exam boards. One aim of the current study was to determine if this was the case for the summer 2017 papers. The participants' mean ratings for the study items for each dimension were analysed between the boards (see Table 9).

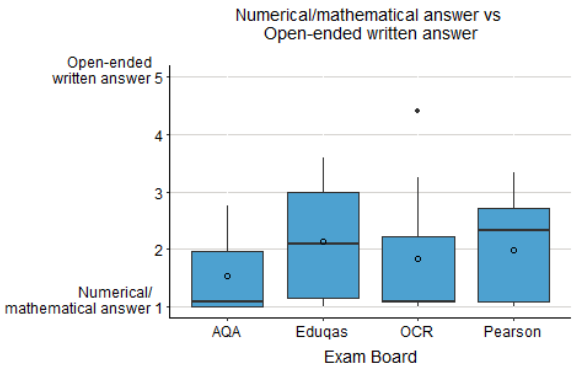
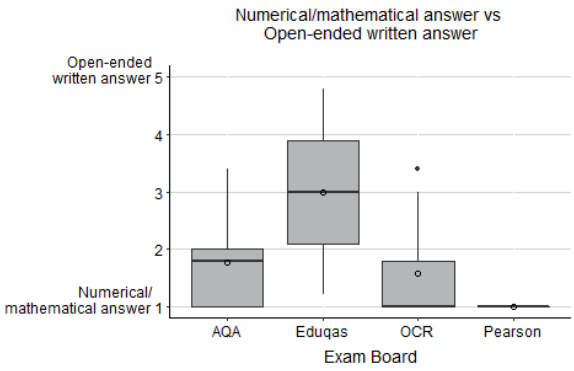
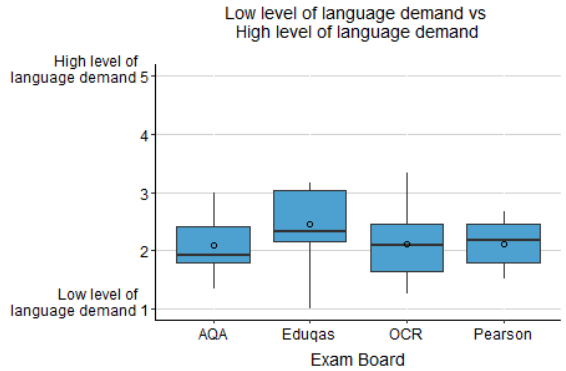
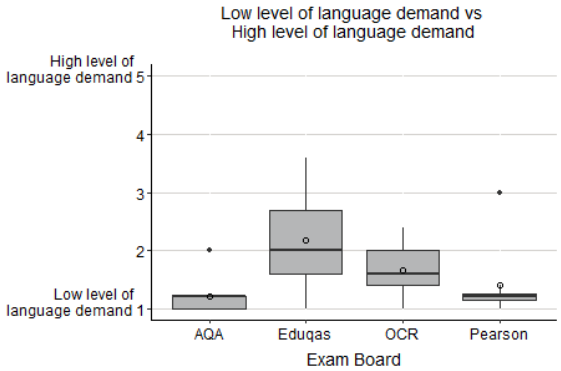
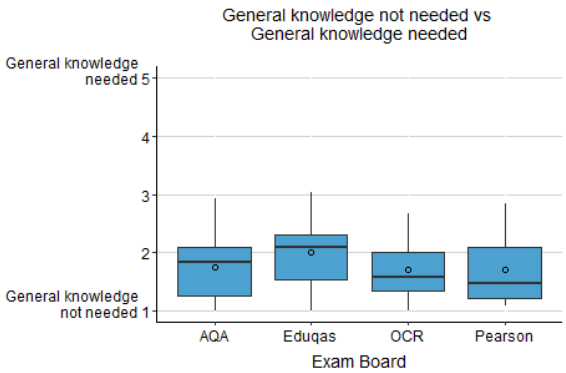
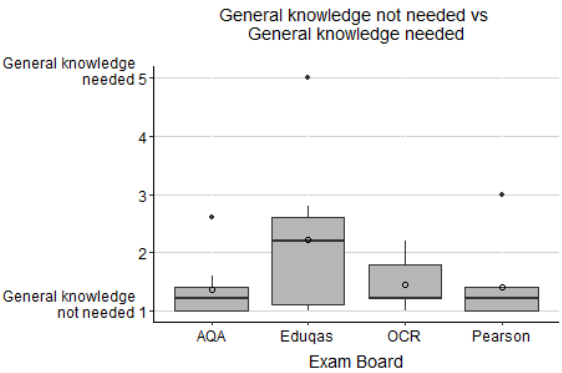
To determine if there were differences between the dimensional features across the boards, 12 one-way between groups analyses of variance (ANOVAs) were run. The ANOVAs indicated that there were no significant differences in any of the dimension ratings between the boards (*Fs* < 2.53*, p*s, >.07; see final column of Table 9). For two of the dimensions there was a marginally significant difference ($p = 0.071$). Overall though, this analysis indicates that there are no substantive differences across the boards in the features captured by the 11 dimensions as well as the overall problem-solving quality of the items for the summer 2017 papers.

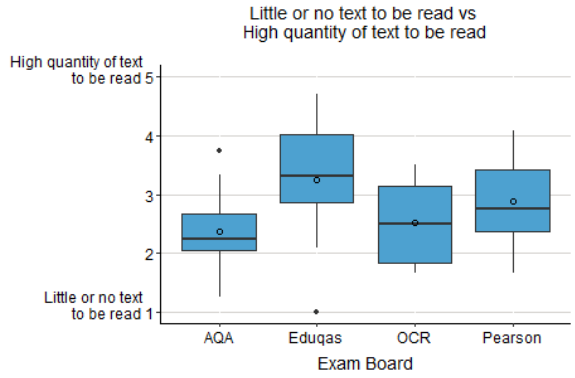
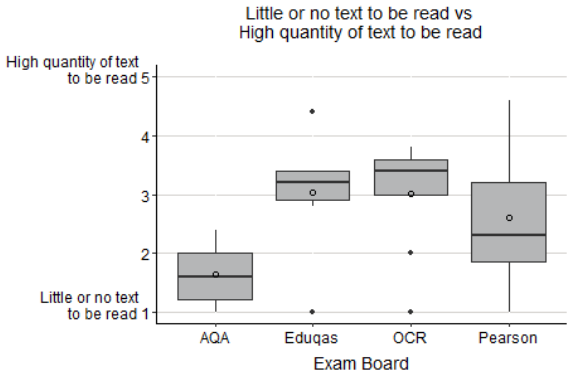
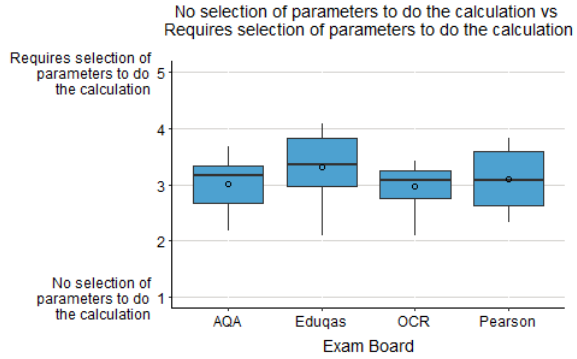
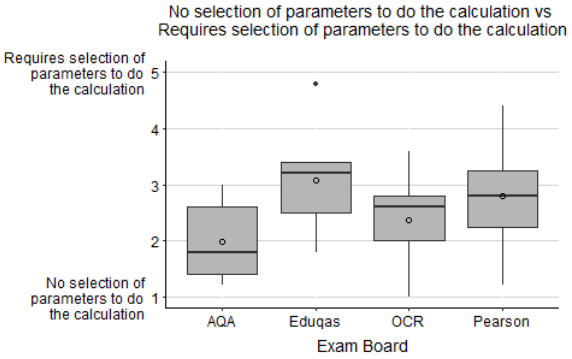
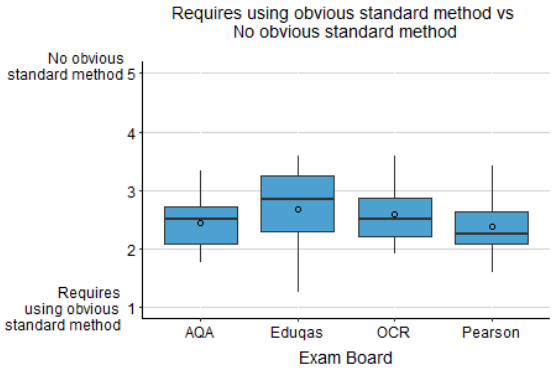
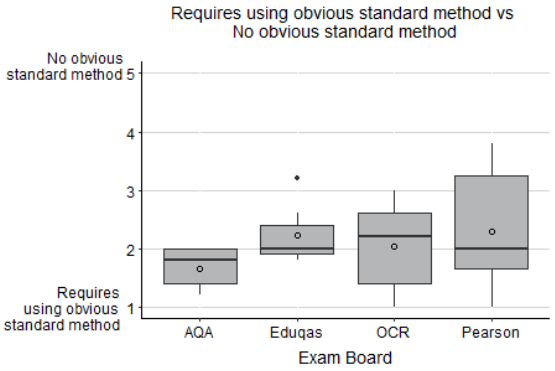
Because different raters were involved in this study than in the previous one, and slightly different procedures were used, it would be inappropriate to statistically compare these ratings. However, Figure 16 shows that the ratings for the dimensions are more similar across the boards for the summer 2017 papers than for the sample assessments.

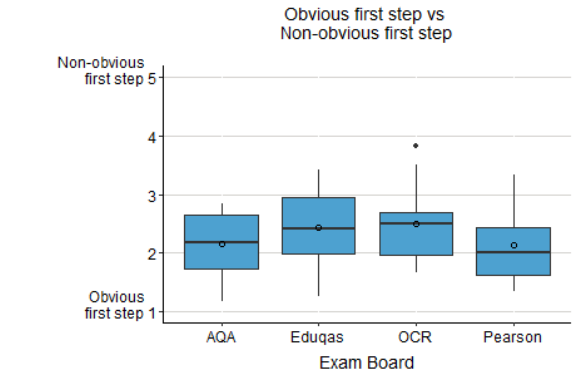
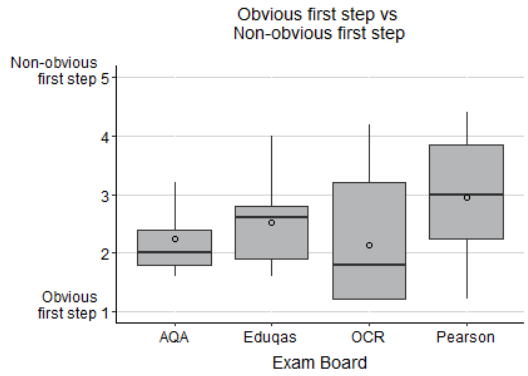
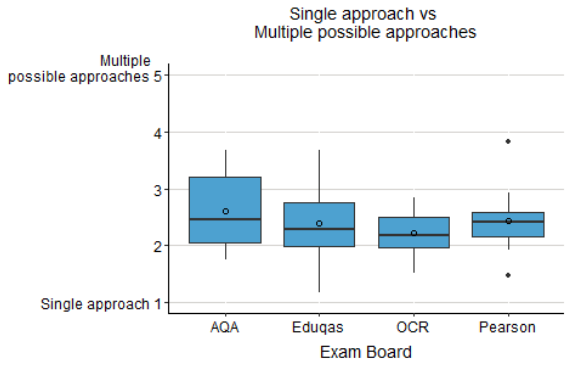
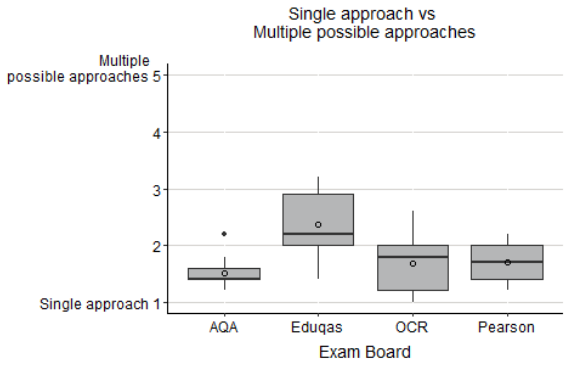
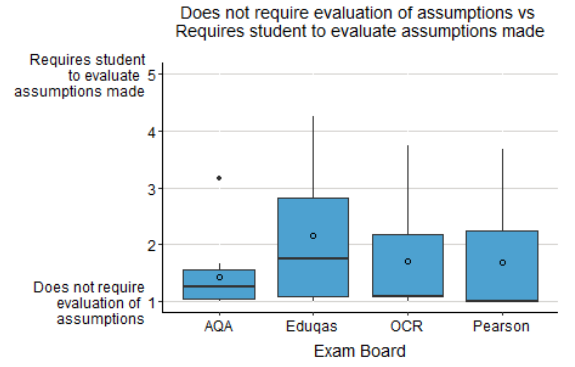
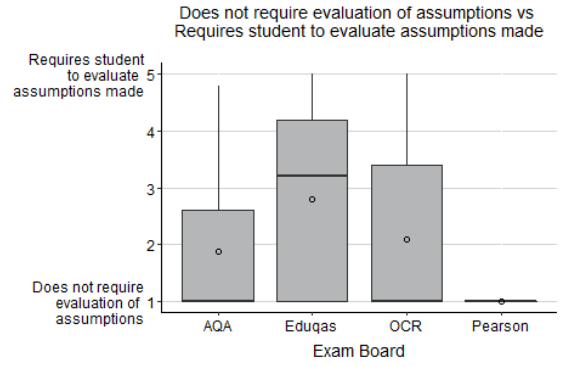
Table 9. *Mean ratings and mean standard deviations of ratings, by dimension, across the exam boards, and summary of one-way ANOVA.*

Dimension		AQA		Eduqas		OCR		Pearson		One-way ANOVA between boards
Pole with rating of 1	Pole with rating of 5	Mean	Mean SD	Mean	Mean SD	Mean	Mean SD	Mean	Mean SD	
Numerical / mathematical answer	Open-ended written answer	1.52	0.43	2.14	0.67	1.83	0.48	1.99	0.63	$F(3, 41) = 0.903$, $p = 0.448$, $\eta^2 = 0.066$
Low level of language demand	High level of language demand (unusual words used)	2.09	0.79	2.45	0.89	2.12	0.81	2.12	0.78	$F(3, 41) = 1.850$, $p = 0.366$, $\eta^2 = 0.079$
General knowledge not needed	General knowledge needed	1.75	0.56	2.00	0.79	1.71	0.63	1.70	0.66	$F(3, 41) = 0.652$, $p = 0.586$, $\eta^2 = 0.048$
Little or no text to be read	High quantity of text to be read	2.37	0.61	3.25	0.67	2.53	0.66	2.88	0.67	$F(3, 41) = 2.525$, $p = 0.071$, $\eta^2 = 0.185$
No selection of parameters to do the calculation	Requires selection of parameters to do the calculation	3.02	0.97	3.30	0.96	2.97	1.07	3.09	0.90	$F(3, 41) = 0.923$, $p = 0.438$, $\eta^2 = 0.067$
Requires using obvious standard method	No obvious standard method	2.43	0.84	2.67	0.78	2.59	0.95	2.38	0.74	$F(3, 41) = 0.694$, $p = 0.561$, $\eta^2 = 0.051$
Obvious first step	Non-obvious first step	2.16	0.91	2.43	0.89	2.49	0.94	2.13	0.80	$F(3, 41) = 0.915$, $p = 0.442$, $\eta^2 = 0.067$
Single approach	Multiple possible approaches	2.61	0.84	2.40	0.83	2.23	0.89	2.43	0.74	$F(3, 41) = 0.755$, $p = 0.526$, $\eta^2 = 0.055$

Dimension		AQA		Eduqas		OCR		Pearson		One-way ANOVA between boards
Pole with rating of 1	Pole with rating of 5	Mean	Mean SD	Mean	Mean SD	Mean	Mean SD	Mean	Mean SD	
Does not require evaluation of assumptions	Requires student to evaluate assumptions made	1.43	0.52	2.16	0.65	1.70	0.60	1.69	0.48	$F(3, 41) = 1.088$, $p = 0.365$, $\eta^2 = 0.080$
Does not require connections between different parts of maths	Requires connections between different parts of maths	2.14	0.71	2.17	0.77	2.35	0.86	2.26	0.70	$F(3, 41) = 0.417$, $p = 0.742$, $\eta^2 = 0.031$
Intermediate steps given or implied	Intermediate steps not obvious	2.85	0.82	2.91	0.92	3.22	0.93	2.61	0.87	$F(3, 41) = 2.525$, $p = 0.071$, $\eta^2 = 0.185$
Overall rating of problem solving elicited by the question: Low	Overall rating of problem solving elicited by the question: High	2.76	0.69	3.02	0.66	2.91	0.69	2.83	0.66	$F(3, 41) = 0.622$, $p = 0.605$, $\eta^2 = 0.046$

2017 mathematics papers	Sample Assessment Materials.
	
	
	

2017 mathematics papers	Sample Assessment Materials.
	
	
	

2017 mathematics papers	Sample Assessment Materials.
	
	
	

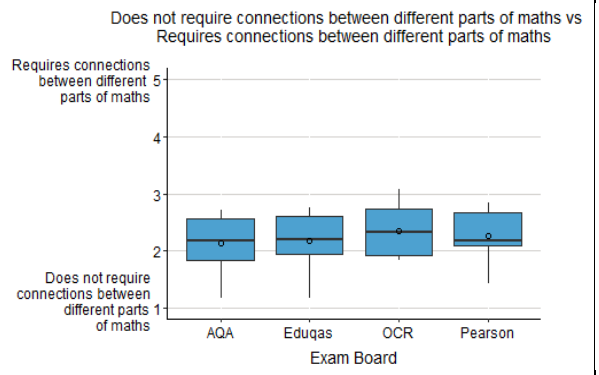
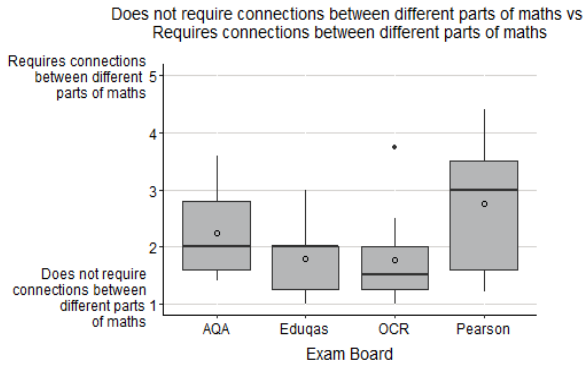
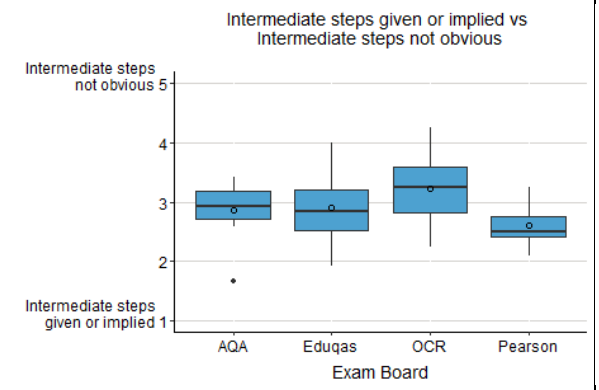
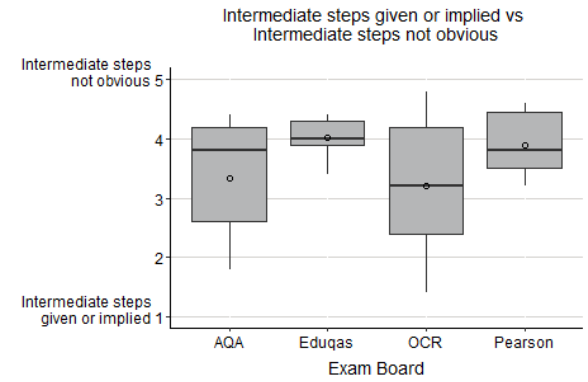
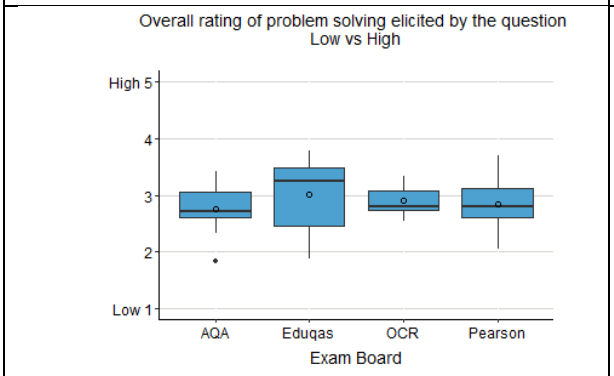
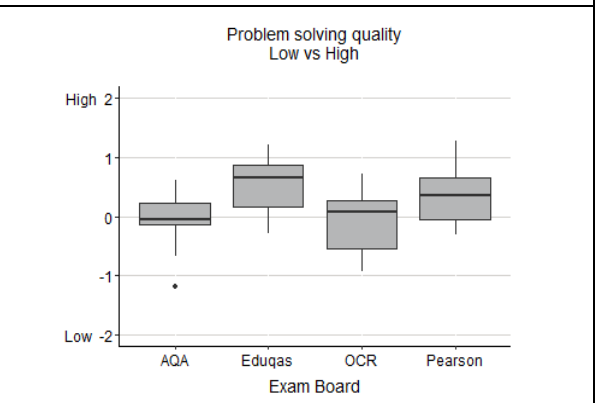
2017 mathematics papers	Sample Assessment Materials.
	
	
	

Figure 16. *Box plots showing median, mean and interquartile ranges of dimension ratings for items from the summer 2017 papers, and the sample assessments. Ratings were scored from 1 to 5 (except for 'Problem solving quality' for the sample assessments which is derived from a comparative judgement study and ranges from approximately -2 to +2). Some dimensions across the summer 2017 papers and sample assessments are positively skewed, with the median score being in line with the lowest score.*

Looking in more detail at just the rating of problem-solving quality of the summer 2017 items, we can split this data by tier (with common items included in both tiers). The overall pattern across exam boards is repeated within both tiers (see Figure 17), and the rated problem-solving quality was slightly higher for items from the higher tier (mean = 2.99) than the foundation tier (mean = 2.75). It is worth noting here that there was just a weak relationship between problem solving quality and item difficulty. Using the whole-item facility from the exam board data as the measure of difficulty (whole items were not judged in the comparative judgement study), we obtained correlations of 0.345 for the foundation tier items and 0.209 for the higher tier items (common items were included in both tiers). Like the previous work on sample assessments, this indicates to a large extent that it is possible to write good problem-solving items independent of difficulty.

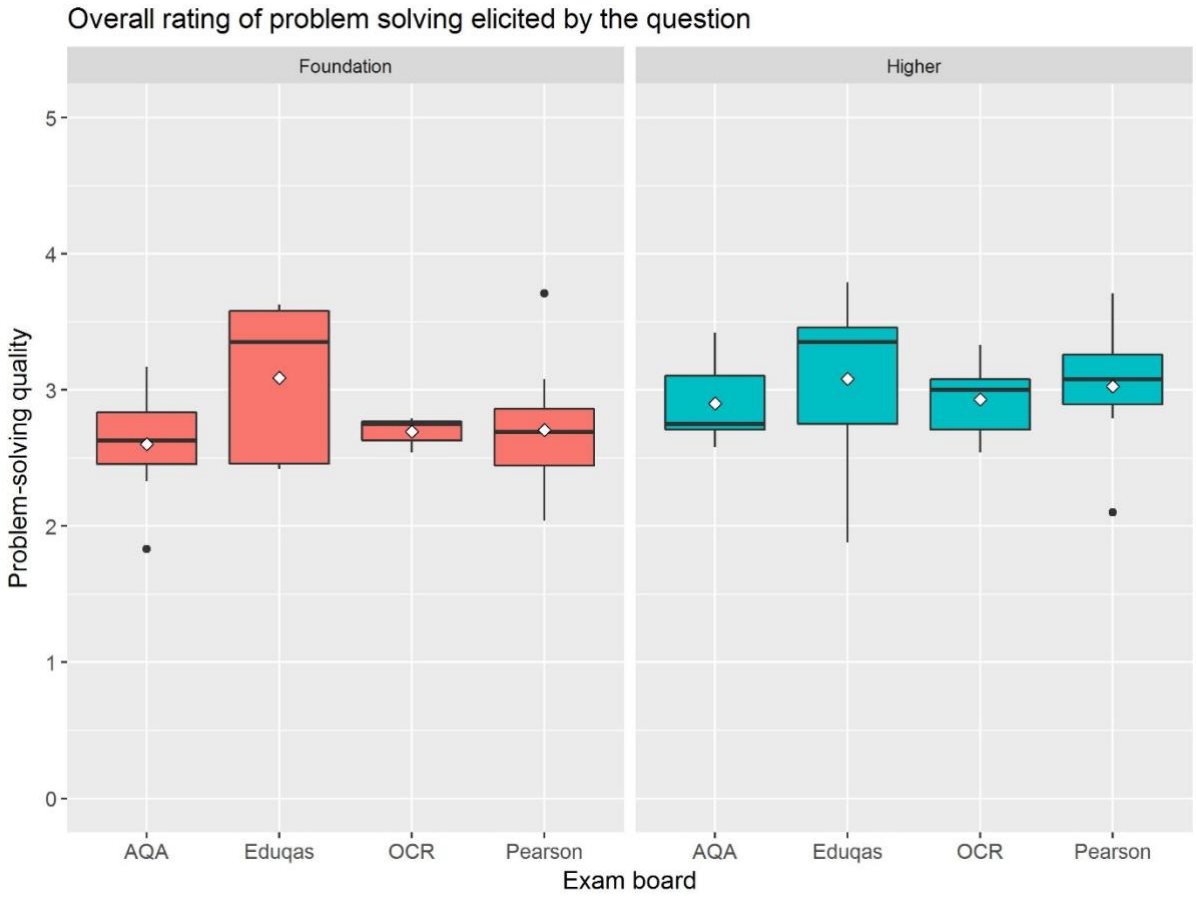


Figure 17. *Box plots showing median, mean and interquartile ranges of the overall problem-solving quality dimension for items from the summer 2017 papers split by paper tier.*

Overall, the data indicates that the AO3 problem solving items present in the summer 2017 papers show no significant differences across the exam boards in terms of the features measured here. This is in contrast to the previous work on the sample assessments, where four of the dimensions included here showed significant differences. Therefore there appears to be greater similarity in problem-solving items between the boards in the 2017 live assessments than in the sample assessment materials.

5.2.3 Examples of items rated highly on problem-solving quality

Items that were rated highly for 'Overall rating of problem-solving elicited by the question' were also rated highly on a number of other dimensions. We include the three items that had the highest overall rating of problem-solving here as examples of good practice (Figure 18 to Figure 20).

a) The diagram shows a large shipping container at rest on horizontal ground.
[2 marks]

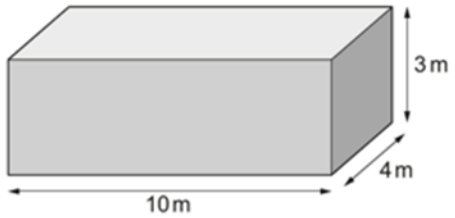


Diagram not drawn to scale

The weight of the container is 32 000 N.
Work out the pressure exerted on the ground by the shipping container.
Give your answer in **N/m²**.

b) A table is at rest on horizontal ground.
The table has 4 legs.
Each leg has a height of 50 cm.
The volume of material in one leg is 450 cm³.
The table weighs 54 N.
By considering the base of the table legs, work out the pressure exerted on the ground by the table.
Give your answer in **N/cm²**.
You must show all your working. [5 marks]

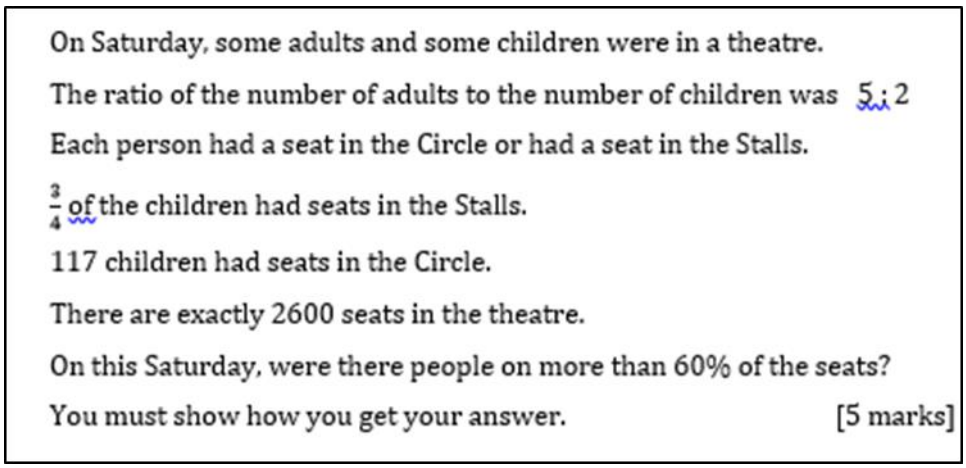
c) (i) State one assumption you have made in your answer to part (b).
[1 mark]

(ii) How would your answer to part (b) change if you had not made this assumption? [1 mark]

Figure 18*. The item that was rated highest on overall problem solving elicited by the question (mean = 3.79). Adapted from Eduqas GCSE Mathematics paper 1 (higher tier) 2017.*

Figure 18 illustrates the highest rated of all items, which was also rated highly on five feature dimensions:

- the level of language demand
- the quantity of text to be read
- the requirement to select parameters to do the calculation
- non-obvious first step
- the requirement to evaluate assumptions made.


On Saturday, some adults and some children were in a theatre.

The ratio of the number of adults to the number of children was 5 : 2

Each person had a seat in the Circle or had a seat in the Stalls.

$\frac{3}{4}$ of the children had seats in the Stalls.

117 children had seats in the Circle.

There are exactly 2600 seats in the theatre.

On this Saturday, were there people on more than 60% of the seats?

You must show how you get your answer. [5 marks]

Figure 19*. The item that was rated second highest on overall problem solving elicited by the question (mean = 3.71). Adapted from Pearson GCSE Mathematics paper 2 (foundation and higher tier) 2017.*

Figure 19 illustrates the item rated second highest for 'overall rating of problem-solving elicited by the question'. This item also scored highly on seven dimensions:

- the requirement for general knowledge
- the quantity of text to read
- the requirement to select parameters to do the calculation
- a non-obvious standard method
- non-obvious first step
- multiple possible approaches
- connections between different parts of maths.

A vending machine sells drinks.
Each drink costs 50 pence.
A sign on the machine shows the coins that can be used to buy the drinks.

Drinks: 50p

This machine accepts
50p, 20p, 10p and 5p coins only

NO CHANGE IS GIVEN

(a) Complete the table to show the 13 different ways of paying the exact amount for a drink. [2 marks]

	50p	20p	10p	5p
Number of each coin	1			
		2	1	
		2		2
		1	3	
		1	2	2
		1		
			5	
			4	
			1	8
				10

(b) The machine has a display that shows how much cash has been put in.
The machine resets the display to £0.00 after each drink is taken.
The cash container in the vending machine is emptied every night.
When it was emptied, the cash container contained the following coins:

50p	20p	10p	5p
10 coins	15 coins	31 coins	20 coins

(i) Work out the greatest possible number of drinks that could have been sold.
You must state any assumption that you make. [5 marks]

Number of drinks sold:
Assumption made:

(ii) Comment on the effect that your assumption has had on your solution. [1 mark]

Figure 20. *The item that was rated third highest on overall problem solving elicited by the question (mean = 3.63). Adapted from Eduqas GCSE Mathematics paper (foundation tier) 2017.*

Figure 20 illustrates the item which was rated third highest for ‘overall rating of problem-solving elicited by the question’. This item was also rated high on five dimensions:

- Open-ended written response
- the level of language demand
- the quantity of text to be read
- multiple possible approaches
- the requirement to evaluate assumptions made.

5.3 Item features related to good problem solving.

Although the primary aim of this study was not to determine features of items that relate to good problem solving, this analysis will be fruitful in highlighting the types of features that elicit valid problem-solving items, which, in turn, can inform AO3 item writing. As our measure of problem-solving quality is not a direct measure, but more an expert view, it is worth noting that any relationships found here do not necessarily mean that the feature is key to making a good problem-solving item, but rather that it is a feature used by experts to make their judgements of quality.

We looked at the correlations between ‘overall rating of problem solving elicited by the question’ with each of the other dimensions. The Pearson correlation coefficients are listed in Table 10, with the pole associated with higher problem-solving quality listed in the second column, and this pole being associated with the positive correlation.

All results were significant at the $p < .05$ level when uncorrected for multiple comparisons. Applying a correction for multiple comparisons (either a conservative Bonferroni correction with an alpha level of 0.0045 or the less strict Holm-Bonferonni correction[13]) leaves the nine correlations in bold at the top of Table 10 significant. Therefore, conclusions drawn about the two dimensions at the bottom of Table 10 are tentative and their related findings should be interpreted with caution.

Five of the 11 dimensions strongly correlated with the overall rating of problem solving ($r > 0.5$). These correlations were higher than any found in the previous work on the sample assessments. These five dimensions capture the strategy, methods and steps taken in problem solving which are clearly important for an AO3 problem-solving item.

[13] Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70

Four dimensions moderately correlated with the rating of problem-solving. Dimensions which capture the linguistic demands of the item text 'high level of language demand (unusual words used)', and 'high quantity of text to be read' were significant predictors. Despite the potential to add construct-irrelevant difficulty, increasing language quantity (and apparently also difficulty) is associated with a richer problem-solving context. Features of making connections between different parts of maths and the need for general knowledge, were also moderately correlated with the rating of problem-solving. These features capture the need to think creatively and use knowledge in order to generate innovative solutions to problems.

Table 10. *Pearson correlation coefficients and uncorrected statistical significance for the relationship between the dimensions and the rating of overall problem solving elicited by the item. After correcting for multiple comparisons, dimensions in bold text remain statistically significant.*

Dimension		Pearson's correlation with overall rating of problem solving	
Pole with rating of 1	Pole with rating of 5	*r(43)* =	*p* = (one-tailed)
Requires using obvious standard method	**No obvious standard method**	**0.833**	**<.001**
No selection of parameters to do the calculation	**Requires selection of parameters to do the calculation**	**0.807**	**<.001**
Obvious first step	**Non-obvious first step**	**0.776**	**<.001**
Intermediate steps given or implied	**Intermediate steps not obvious**	**0.608**	**<.001**
Single approach	**Multiple possible approaches**	**0.565**	**<.001**
Does not require connections between different parts of maths	**Requires connections between different parts of maths**	**0.447**	**.001**
Low level of language demand	**High level of language demand (unusual words used)**	**0.425**	**.002**
Little or no text to be read	**High quantity of text to be read**	**0.419**	**.002**
General knowledge not needed	**General knowledge needed**	**0.412**	**.002**
Numerical / mathematical answer	Open-ended written answer	0.258	.043
Does not require evaluation of assumptions	Requires student to evaluate assumptions made	0.250	.049

Note. Dimensions are listed in order of correlation coefficient size.

The feature with the highest correlation with the overall problem-solving rating was 'no obvious standard method' to calculate the answer ($r(43) = 0.833$, $p < .001$). Figure 21 shows an item which was highly rated overall, and with the joint highest mean rating for the dimension 'no obvious standard method' (mean = 3.58) as well as

the highest rating on ‘intermediate steps not obvious’ (mean = 4.25) and ‘non-obvious first step’ (mean = 3.83).

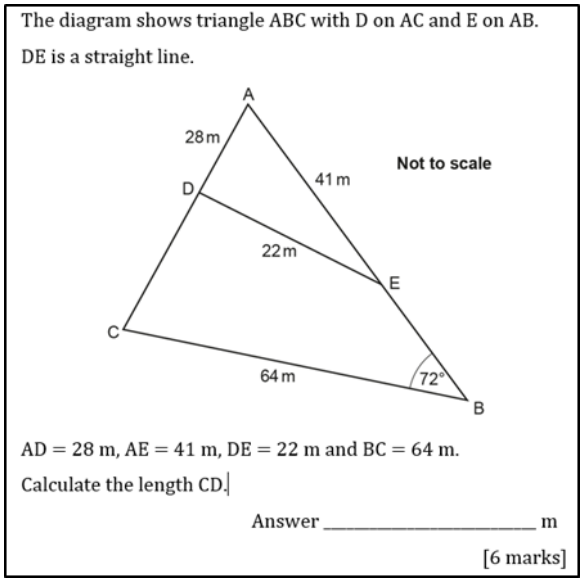


Figure 21. *An item that was rated high on overall problem solving elicited by the question (mean = 3.29) and was also rated high on several dimensions related to its non-obvious nature (see text). Adapted from OCR GCSE Mathematics paper 1 (higher tier) 2017.*

Two dimensions had weak correlations with the rating of overall problem-solving, and do not reach significance when corrected for multiple comparisons. It is likely that good problem solving items can have either numerical or written response formats, and do not always require evaluations of assumptions.

5.4 Qualitative feedback

Discussion between participants during the meeting and post-meeting feedback from our subject experts raised a couple of interesting points regarding the accessibility of items. The first issue related to how obvious the first step to solving the problem was. Participants anecdotally expressed that where candidates are presented with a problem they are unable to start to solve, or where they get stuck at an early stage, the item is not further attempted. This is particularly so for problems that are not broken into sub-parts (not scaffolded). This results in the candidate failing to acquire marks for the remainder of the problem solving, when they may in fact know how to calculate later parts.

Of course, the non-obvious way into a problem is an important and desirable feature of problem items, so there is an inherent tension here. Two ways of reducing this

impact are suggested. First, given that mark schemes frequently award method marks where wrong numbers are used, it was felt that where candidates might not know how to start a problem but knew how to carry out the later calculations required in a problem, they could be encouraged to start the problem part-way through. They could do this by making an estimate of the outcome from the earlier stages of the problem, stating their estimate clearly and basing later calculations on this value. Second, care should be taken by item writers not to make the way into a problem too hard for items not intended to be high demand.

As identified in previous work, there were also concerns that the language demand of the items may introduce construct-irrelevant difficulty which stops students accessing the item for reasons beyond mathematical ability. This may be particularly true for candidates sitting the foundation tier. The findings in this study clearly identify aspects of linguistic demand as positively related to problem-solving quality, such as the type and quantity of text. This text provides an opportunity to create an enhanced problem-solving context with which to meet the AO3 demands. Item writers should therefore be mindful of linguistic demand with relation to accessibility to the item, using text judiciously to set the problem context whilst not stopping candidates from demonstrating mathematical ability.

5.5 Discussion

The study aimed to determine if there were differences across the boards in the features of the AO3 items sat in mathematics papers in summer 2017. The data indicates that unlike in the sample assessments, there are no statistically significant differences in ratings on the dimensions between the boards. There is therefore greater consistency across exam boards in the features of the AO3 items in the summer papers, compared to those in the sample assessments.

The study provided data which allowed further examination of the features which promote good problem-solving. Several features positively correlated with rated problem-solving quality. Fairly strong correlations were obtained for a group of dimensions related to the strategy and approach required to solve the problem, such as non-obvious and multiple methods, non-obvious first and intermediate steps, the need to select parameters for the calculation and the application of different areas of maths. A second group of dimensions with slightly lower correlations related to language involved in setting the item and background knowledge required.

Overall the average ratings we obtained on the dimensions were not that high – they were mostly below the mid-point of the scale. The way individuals use rating scales is never fully transparent, so it is hard to interpret absolute values. However, the moderate ratings do suggest that no individual features appeared that widely across all items. This is not surprising when you consider the many different approaches possible for setting problems. No one feature is a firm requirement of a good

problem, but conversely it is to be expected that a good problem might clearly exhibit one or more of the features described in this study.

Therefore, the dimensions identified here as related to good problem-solving should be incorporated as much as possible when constructing problem items, whilst balancing appropriate difficulty and the need to avoid raising difficulty unduly through excessively difficult ways into starting to solve the problem, or non-mathematical elements, for instance, from unreasonable language demand or unreasonable expectations of general knowledge.

6 Overall Conclusions

Differences in overall difficulty between the exam boards' papers are moderate and easily dealt with through adjustment of grade boundaries in awarding. The level of difficulty has been set well, especially when it is considered that the sample assessment materials went through several rounds of adjustment to align their difficulty as close as they are. The broader distributions of item difficulty in the summer assessments will be helpful in differentiating effectively between candidates.

AO3 items are amongst the most difficult items on the tests. Differences in difficulty were slightly larger for the AO3 items than for the overall assessment difficulty, and there was more variation in the higher tier items between exam boards than the foundation tier items. However, qualitative analysis of a subset of items with the most allocated AO3 marks showed no significant differences in the features of items across the exam boards.

We have extended our previous analysis of the features of items related to ratings of good problem solving. Identification of these features should help item setters and Principal Examiners when writing items and designing their complete assessments to optimise the quality of items and balance of items across the tests.

We wish to make our publications widely accessible. Please contact us at publications@ofqual.gov.uk if you have any specific accessibility requirements.

This publication is available at www.gov.uk/ofqual.

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346