A Comparison of Expected Difficulty, Actual Difficulty and Assessment of Problem Solving across GCSE Maths Sample Assessment Materials

May 2015

Ofqual/15/5679

Contents

Introduction

Following the accreditation of the reformed GCSE maths (9 to 1) specifications, concerns were expressed as to differences in the difficulty of exam boards' sample assessment materials and in their approach to the assessment of problem solving. The following programme of research was conducted to evaluate whether the concerns were valid and the differences sufficient to undermine the teaching, learning and assessment of GCSE maths.

Overview

The programme of work comprised four complementary evaluations of exam boards' sample assessments.

Study 1: A comparison of the expected difficulty of all items (questions) from exam boards' sample assessments, including comparison with items from recent GCSE maths papers and with similar qualifications from international jurisdictions.

Study 2: A comparison across exam boards of the difficulty of items from the non-calculator sample assessments, including aggregation to the level of whole question paper.

Study 3: A comparison across exam boards of the extent to which items are judged as eliciting the mathematical problem-solving construct.

Study 4: A study of the ways in which problem-solving items vary across exam boards' sample assessments.

Rationale for the four studies

The first study focused on judges' beliefs as to the likely difficulty of items. This data is relatively easy to collect. It was possible, therefore, to collect data for the sample assessments, recent question papers and for similar qualifications from international jurisdictions. This allowed a comparison between exam boards of the expected difficulty of sample assessments, a comparison of the expected difficulty of these assessments with that of question papers from the current GCSE maths specifications and a comparison with broadly similar international question papers.

Without this additional context it would have been hard to evaluate whether differences in the expected difficulty of exam boards' sample assessments were of concern. It is impossible for exam boards to perfectly control the difficulty of question papers. Slight differences in difficulty are accounted for during grade boundary setting. Grade boundaries are set commensurate with question paper difficulty – more difficult papers have lower boundaries, less difficult papers have higher boundaries. However, consistent and large differences in difficulty cannot be accounted for in this way. Large differences in difficulty can undermine confidence in the extent to which there are comparable standards across exam boards. Moreover, large differences may have a negative wash-back effect on teaching and learning with candidates for easier papers potentially experiencing a poorer mathematical education or candidates for harder papers having a less positive experience of maths.

Being able to contextualise differences in expected difficulty between exam boards' sample assessment materials with differences between recent papers from current specifications is most valuable if we assume that expected difficulty is a good predictor of actual difficulty. However, expectations of difficulty do not necessarily equate to differences in actual difficulty as experienced by candidates.[1] Candidates find challenges in items arising from context or from the specific numbers involved in a task. These challenges are often not immediately apparent to more expert mathematicians.

Thus, the second study focused on the actual difficulty of exam board sample assessments. This involved Year 11 students (15 to 16 years old) taking a non-calculator sample assessment from each exam board. This allowed interrogation of item-level performance data and comparison across exam boards. However, the challenges of having large numbers of students sit sample assessments (and the consequent marking and school feedback) meant that difficulty data was only collected for one sample assessment per exam board. The extent to which measures of expected and actual difficulty correlated across the two studies determined the extent to which expected difficulty could be used as a proxy for actual difficulty, and so the full value of the data from the first study was established.

During the development of the reformed GCSE maths qualification it was challenging to gain consensus on the parameters of the mathematical problem-solving construct and its assessment. The extent of differences across exam boards in the difficulty and functioning of problem-solving items were explored in the second study. The third and fourth studies were focused on the extent of any differences in the exam boards' approach to the assessment of problem solving and the potential implications for validity.

In the third study, GCSE maths teachers compared the extent to which items from the sample assessments elicited problem solving. This produced a scale of item

[1] Pollitt, A., Ahemd, A. and Crisp, V. (2007). *The demands of examination syllabuses and question papers*. In P. Newton, J.A. Baird, H. Goldstein, H. Patrick and P. Tymms *Techniques for monitoring the comparability of examination standards* (pp.166–206), London: QCA

validity (as perceived by the teachers). Comparing where exam boards' items fell on this scale gave an indication of which boards' sample assessments included the items that best elicited problem solving. It was also possible to examine the features of those items which fell at the top and the bottom of the scale. The fourth study provided an independent analysis of what those features might include. Five GCSE maths teachers listed the similarities and differences between problem-solving items and rated each item according to these features. These ratings were aggregated so that comparisons across exam boards could be made.

In summary, the four studies combined to provide data to compare across exam boards the expected difficulty, actual difficulty and approach to problem solving of the sample assessments.

1. Study 1 – A comparison of the expected difficulty of mathematics items

1.1 Design

Study 1 was designed to elicit experts' judgements of the expected difficulty of mathematics items in different mathematics examinations worldwide. The expected difficulty was estimated through a comparative-judgement (CJ) study (see Bramley, (2007) for a description of the use of paired comparison methods[2]). In a CJ study a series of paired comparisons are presented to judges, who are asked to decide in each case which one of the pair meets described criteria. In study 1 the judges were asked to decide:

'Which question is the more mathematically difficult to answer fully?'

CJ studies draw on Thurstone's[3] 1927 law which states that people are better at making relative judgements than absolute judgements. Online CJ systems allow judgements to be made in a distributed fashion with large numbers of judges, which has the further advantage of cancelling out individual bias. Once enough judgements have been made a scale can be created from the judgements using either the Rasch[4] or the Bradley-Terry[5] model. The construction of a scale allows properties of the model, such as the consistency of judgement and the reliability of judgement, to be evaluated.

The judgement of the difficulty of items without pre-testing data is extremely challenging. Experts are typically poor judges of the difficulties faced by novices, while even subtle aspects of question design can affect difficulty.[6] However, evidence

[2] Bramley, T. (2007). Paired comparison methods. In P. Newton, J.A. Baird, H. Goldstein, H. Patrick and P. Tymms *Techniques for monitoring the comparability of examination standards* (pp.166-206), London: QCA

[3] Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, 34, 273–286.

[4] Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut. www.rasch.org/books.htm

[5] Bradley RA, Terry ME (1952). Rank Analysis of Incomplete Block Designs I: The Method of Paired Comparisons. Biometrika, 39, 324–45.

[6] Pollitt, A., Ahemd, A. and Crisp, V. (2007). *op. cit*.

is emerging which suggests that it is possible to use CJ to create scales of expected item difficulty that are validated by external criteria.[7]

One of the threats to the generalisability of CJ studies is the extent to which judges share a homogenous view of the construct under examination. CJ studies can be used, for example, to elicit political views on a subject, where it is group differences that are of interest. In mathematics, however, studies have yet to find any substantial difference on the scales of mathematical difficulty created by groups of judges with different levels of expertise.[8]

1.2 Methods

1.2.1 Materials

The items compared in study 1 comprised: AQA, Pearson and OCR sample assessment materials (SAMs) for the reformed GCSE (9 to 1);[9] AQA, Pearson, OCR and Eduqas[10] question papers from current GCSE (2011–2012); similar assessments from ten international jurisdictions for students aged around 16 years of age (taken between 2010–2012) and Cambridge International Examinations IGCSE and O level (2011) (listed in table 1). Items from the papers of a level 3 maths qualification available in England were also included in the analysis but the findings are not included in this report as the qualification's purpose and entry differed. Where the

[7] Jones, I., Wheadon, C., Humphries, S. & Inglis, M. (2014) Was the golden age of mathematics education fifty years ago? AQA Research Report.

[8] See, for example: Raikes, N., Scorey, S. and Shiell, H. (2008). Grading examinations using expert judgements from a diverse pool of judges. Paper presented to the 34th annual conference of the International Association for Educational Assessment, Cambridge, 2008. Retrieved from www.cambridgeassessment.org.uk/Images/109766-grading-examinations-using-expert-judgements-from-a-diverse-pool-of-judges.pdf ; Jones, I. and Alcock, L. (2014). Peer assessment without assessment criteria. Studies in Higher Education, 39(10), 1774–1787; Jones, I., Swan, M. & Pollitt, A. (2014). Assessing mathematical problem solving using comparative judgement. International Journal of Science and Mathematics Education, 1–27; Bisson, M., Jones, I., Gilmore, C. & Inglis, M. (submitted). Measuring conceptual understanding using comparative judgement. International Journal of Research in Undergraduate Mathematics Education; Jones, I. & Inglis, M. (in press). The problem of assessing problem solving: Can comparative judgement help? Educational Studies in Mathematics. DOI: 10.1007/s10649-015-9607-1; Jones, I., Wheadon, C., Humphries, S. and Inglis, M. (in prep). Fifty years of A-level mathematics: Have standards changed? British Education Research Journal.

[9] At the point of conducting this study the Eduqas specification had not been accredited and so the sample assessment materials were not final. Thus, Eduqas's sample assessments were not included in this study.

[10] Eduqas is the brand of WJEC offering reformed qualifications in England

assessment materials were in a language other than English, translations were obtained through commercial translators. The purpose of each assessment and the age of the cohort taking it are summarised in appendix A.

For each assessment listed in table 1, every item was included in the study. In the case of England's tiered GCSEs (both the current and reformed versions) common items that occurred on both tiers were entered and coded as higher tier items only and not duplicated in the judging set for the foundation tier. Therefore, the item counts in table 1 for the foundation tier papers will be slightly reduced. However, when analysing the expected difficulty of the foundation tier papers, the parameters for these common items were included.

The mark schemes for the items were not presented as part of the judging to encourage judges to work through items to uncover unexpected sources of difficulty. Further, mark schemes do not exist for all international jurisdictions, which may have created bias in the judgement. Any systematic differences, therefore, in the extent to which a mark scheme modifies the expected difficulty of an item/assessment will not be identified.

No item marks were presented as part of the judging as it was considered unlikely that judges would be able to make a consistent mental adjustment for the number of marks involved in different items. Again, the number of marks per question was not available for a number of the international jurisdictions.

Table 1: List of jurisdictions, assessments and specific papers included in the study

Jurisdiction / awarding organisation	Assessment	Papers	Number of items	Paper duration (mins)
Cambridge International Examinations	IGCSE	1. Paper 2 (extended) 2. Paper 4 (extended)	35 53	90 150
	O Level	1. Paper 1 2. Paper 2	58 69	120 150
England – AQA	GCSE	1. Unit 1 Higher 2. Unit 1 Foundation 3. Unit 2 Higher 4. Unit 2 Foundation 5. Unit 3 Higher 6. Unit 3 Foundation	23 14 22 31 29 35	60 60 75 75 90 90
	GCSE (9 to 1)	1. Paper 1 Higher 2. Paper 1 Foundation 3. Paper 2 Higher 4. Paper 2 Foundation 5. Paper 3 Higher 6. Paper 3 Foundation	37 33 32 30 37 28	90 90 90 90 90 90
England – Pearson	GCSE	1. Mathematics B Unit 1 Higher 2. Mathematics B Unit 1 Foundation 3. Mathematics B Unit 2 Higher 4. Mathematics B Unit 2 Foundation 5. Mathematics B Unit 3 Higher 6. Mathematics B Unit 3 Foundation	26 23 28 35 32 34	75 75 75 75 105 90
	GCSE (9 to 1)	1. Paper 1 Higher 2. Paper 1 Foundation 3. Paper 2 Higher 4. Paper 2 Foundation 5. Paper 3 Higher 6. Paper 3 Foundation	28 26 29 27 31 29	90 90 90 90 90 90

Jurisdiction / awarding organisation	Assessment	Papers	Number of items	Paper duration (mins)
England – OCR	GCSE	1. Mathematics A Unit A Higher 2. Mathematics A Unit A Foundation 3. Mathematics A Unit B Higher 4. Mathematics A Unit B Foundation 5. Mathematics A Unit C Higher 6. Mathematics A Unit C Foundation	27 21 22 24 35 33	60 60 60 60 120 90
	GCSE (9 to 1)	1. Paper 1 (Foundation) 2. Paper 2 (Foundation) 3. Paper 3 (Foundation) 4. Paper 4 (Higher) 5. Paper 5 (Higher) 6. Paper 6 (Higher)	32 39 36 38 36 36	90 90 90 90 90 90
England – WJEC	GCSE	1. Unit 1 Higher 2. Unit 1 Foundation 3. Unit 2 Higher 4. Unit 2 Foundation 5. Unit 3 Higher 6. Unit 3 Foundation	22 27 29 29 29 26	75 75 75 75 105 90
Hong Kong (China)	Hong Kong Certificate of Education Examination (HKCEE)	1. Mathematics Paper 1 2. Mathematics Paper 2	46 54	120 90
Hungary	National Assessment of Basic Competence (NABC)	Grade 10 Booklet A (mathematics sections)	60	90
Japan	National Assessment of Academic Ability (NAAA)	1. Lower Secondary Year 3 Mathematics A 2. Lower Secondary Year 3 Mathematics B	36 15	45 45

Jurisdiction / awarding organisation	Assessment	Papers	Number of items	Paper duration (mins)
Massachusetts (USA)	Massachusetts Comprehensive Assessment System (MCAS)	1. Grade 10 Mathematics – Test Session 1 2. Grade 10 Mathematics – Test Session 2	29 30	60 60
Netherlands	VMBO TL/GL	Mathematics CSE TL and GL	24	120
New Zealand	National Certificate of Educational Achievement (NCEA) Level 1	1. Level 1 Mathematics and Statistics 91027 2. Level 1 Mathematics and Statistics 91028 3. Level 1 Mathematics and Statistics 91031 4. Level 1 Mathematics and Statistics 91037	19 18 17 19	60 60 60 60
Ontario (Canada)	Grade 9 Assessment of Mathematics	1. Academic Paper 2. Applied Paper	31 31	100 100
Scotland – SQA	Standard Grade	1. Credit Level Paper 1 2. Credit Level Paper 2 3. General Level Paper 1 4. General Level Paper 2 5. Foundation Level Paper 1 6. Foundation Level Paper 2	18 16 16 18 12 23	55 80 35 55 20 40
Shanghai (China)	Zhong Kao	Junior High School Joint Graduation and Academic Examination – Mathematics Exam	34	100
South Korea	National Assessment of Educational Achievement (NAEA)	9th Grade Mathematics	37	60

1.2.2 Transcription of items

All items were transcribed using a standard typescript. During this process, every attempt was made to eliminate any cues as to the jurisdiction/assessment from which the item was taken. Given that items drawn from England's GCSEs predominated, these were used as the style template for the other items. Modifications included using wording/phrasing that more closely matched that used in England's GCSEs (e.g. 'show your working'), and applying a standardised layout and font for the items. Words and names that may have identified countries were changed to neutral terms; this included changing non-metric units. In an attempt to ensure consistency between the old and new items, a maths expert reviewed the items where the changes were considered substantial (43 out of a total of 2,150).

Multi-part items were treated as a series of individual items, given that the expected difficulty could vary across the parts. In some cases the item parts were entirely unrelated to one another; these were transcribed as separate items. Linked multi-part items (where each part related to the same basic problem) were presented in full, with the relevant section for judgement highlighted in a different colour. Where part of an item relied on the answer to a previous part, judges were instructed to assume all earlier parts had been answered correctly.

The use of a calculator or formula sheet in an assessment is likely to affect the expected difficulty of an item. If a formula sheet or a calculator was allowed for the paper and would have been helpful in answering an item (or any of the sub-parts on a multi-part item), this was indicated above the item with the phrases 'Calculator allowed' and/or 'Formula sheet provided'.

1.2.3 Participants

Forty-three PhD students studying mathematics at English universities were recruited to be judges. Judges were paid for their time. PhD students were used as they were considered to be less likely than teachers in England to be familiar with England's specifications and papers and less likely to have been exposed to any of the debate surrounding the design of the new GCSE maths.

1.2.4 Procedure

Comparisons were conducted using the online CJ platform, No More Marking.[11] Judges were given detailed instructions on how to access the platform and how to

[11] Wheadon, C. and Jones, I. (2014, June 1). Online Comparative Judgement. Retrieved April 21, 2015, from www.nomoremarking.com

make their judgements. Pairs of items were presented side by side on the screen and the judges were prompted to indicate:

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

The judging prompt was always present on the judging screen. The judges were specifically asked to judge the mathematical difficulty of the items.

It was left up to the judges how they made their judgements; the only restriction was a date by which they had to complete them. However, a combination of speed and accuracy was encouraged. For example, the instructions emphasised that there would be opportunities in the future for consistent judges. Following the judging window, judges were asked to volunteer their thoughts on the judging process, how they made their decisions and what difficulties and challenges arose. Each judge made 1,000 judgements, giving a total of 43,000 judgements (providing a minimum of 20 judgements per item). The pairs of items were distributed among judges so that the items were all seen a similar number of times.

1.3 Results

1.3.1 Analysis

The Bradley-Terry2[12] R package was used to estimate expected difficulty parameters for each item. The node package, Comparative-Judgement,[13] was used to estimate item and judge outfit, scale-separation reliability (SSR) and inter-rater reliability under the Rasch model.

1.3.2 Judge consistency and exclusion

Eight judges were excluded from the analysis on the basis of the haste and lack of consistency with which they made judgements. The eight judges had a median judgement time of less than 10 seconds per item and outfit[14] values that ranged from 1.08 to 1.48. The range of outfit values for the 35 judges included in the analysis was 0.81 to 1.10, while the range of median judgement times by judge was 11 to 35 seconds (mean = 19 seconds).

[12] Turner, H., Firth, D. (2012). Bradley-Terry Models in R: The BradleyTerry2 Package. Journal of Statistical Software, 48(9), 1–21. URL www.jstatsoft.org/v48/i09

[13] Wheadon, C. (2014, Sept). Comparative Judgement Algorithms. Retrieved April 21, 2015, from www.npmjs.com/package/comparative-judgement

[14] For an explanation of outfit, see, for example, Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19(3), 281–300.

Once the eight judges were removed, the median inter-rater reliability was assessed by repeatedly allocating judges to two groups, fitting the Rasch model independently for each group and correlating the two rank orders of the item parameters. Across 100 replications the correlation was 0.74 (sd=0.01). Reliability is quantified in CJ studies by an SSR statistic that is derived in exactly the 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 0.88. The reliability values suggest a certain degree of disagreement among the judges, but not enough to threaten the measurement properties of the expected difficulty scale created.

1.3.3 Basis of judgements

Eight judges offered their thoughts and reflections on the judging process. These comments confirm that it had not been an easy task for some judges. As one judge eloquently stated,

> More generally one is led to ask what mathematical difficulty could possibly mean. In some cases this was rather more apparent to me via some ineffable means. But once one recognises that there are a bunch of different scales of difficulty the activity becomes rather difficult.

It was apparent that the judges were aware of potential sources of bias in their judgements and sought to control them. For example, two judges commented that some items became familiar towards the end of the task but they nonetheless made a conscious effort to re-read them. There was no common pattern to the factors that the judges reported as influencing their decision making. For example, the following factors were mentioned:

- time taken to complete the item;
- the number of steps involved;
- the knowledge involved;
- the complexity of the mathematical idea;
- the format of the item (for example multiple choice);
- the need to prove a statement;
- the need for a mathematical argument or logical statement;
- whether a calculator was allowed;
- whether a formula sheet was provided.

1.3.4 Comparative-judgement analysis

Distributions of expected item difficulty parameters are shown aggregated by paper in figure 1 and by ‘qualification’ in figure 2. Figure 3 focuses on England’s current GCSE and new sample assessments. The comparisons between England’s GCSEs and similar qualifications available internationally will be the subject of future reports and are not a major focus here.

In the analyses that follow, all items are equally weighted regardless of their tariff. So, in figures 1 and 2, for example, each point represents one question. As will be shown later, the expected difficulty and item tariff are correlated, so any attempt to weight the analyses by tariff would be confounded by this relationship.

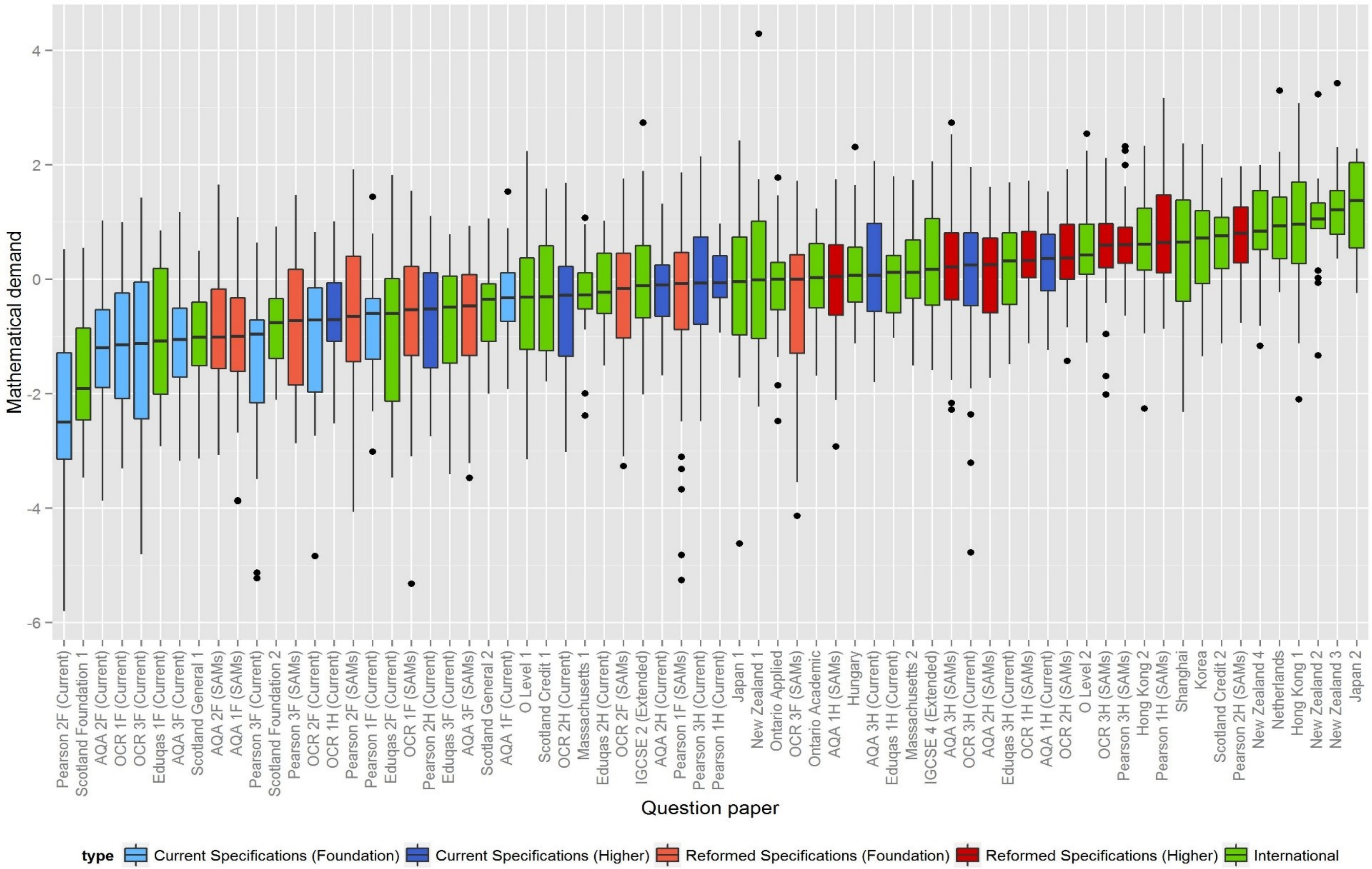


Figure 1: Box plots showing median and interquartile ranges of expected item difficulty parameters for all papers

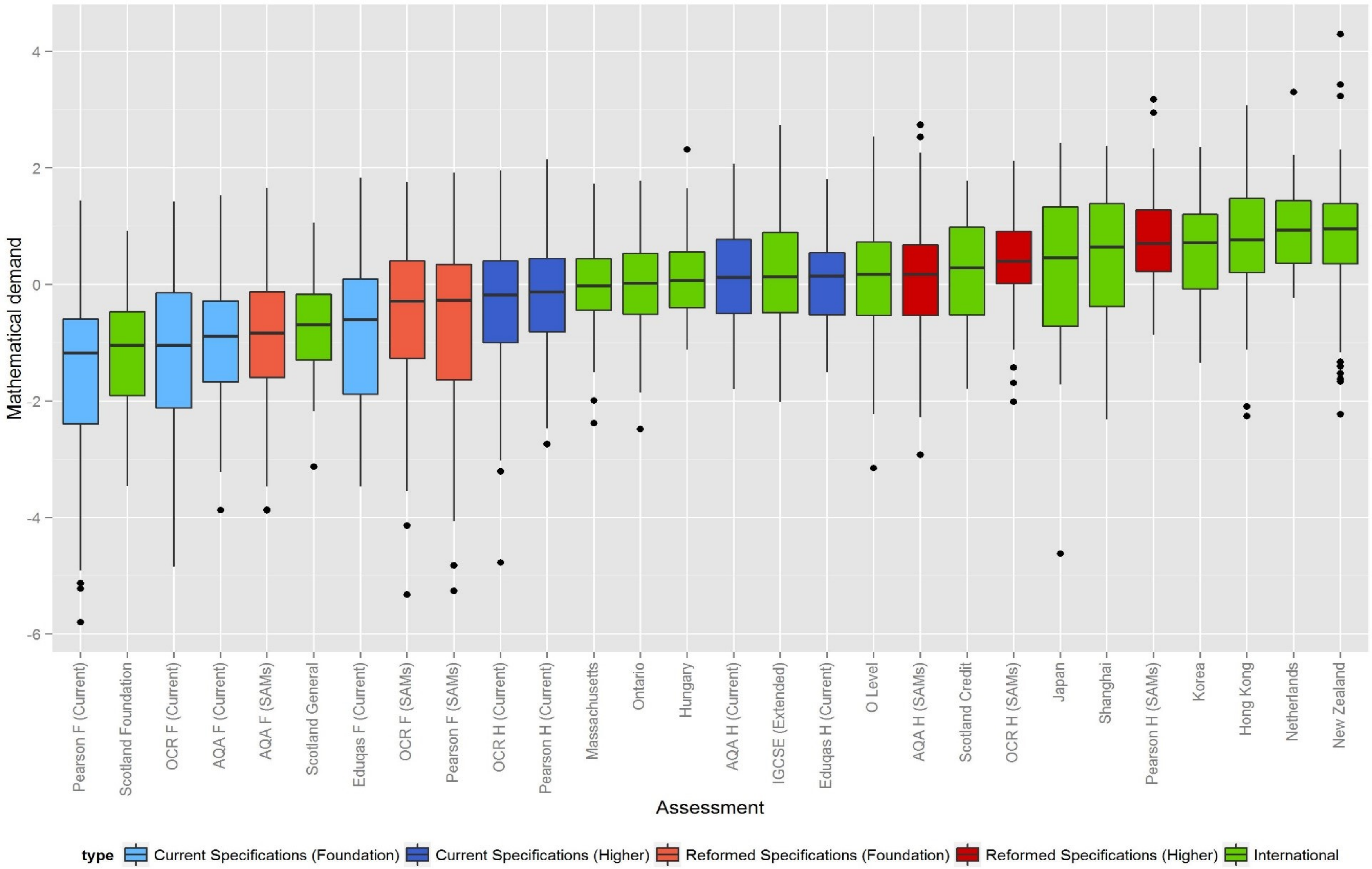


Figure 2: Box plots showing median and interquartile ranges of expected item difficulty parameters for all ‘qualifications’ (combined across papers)

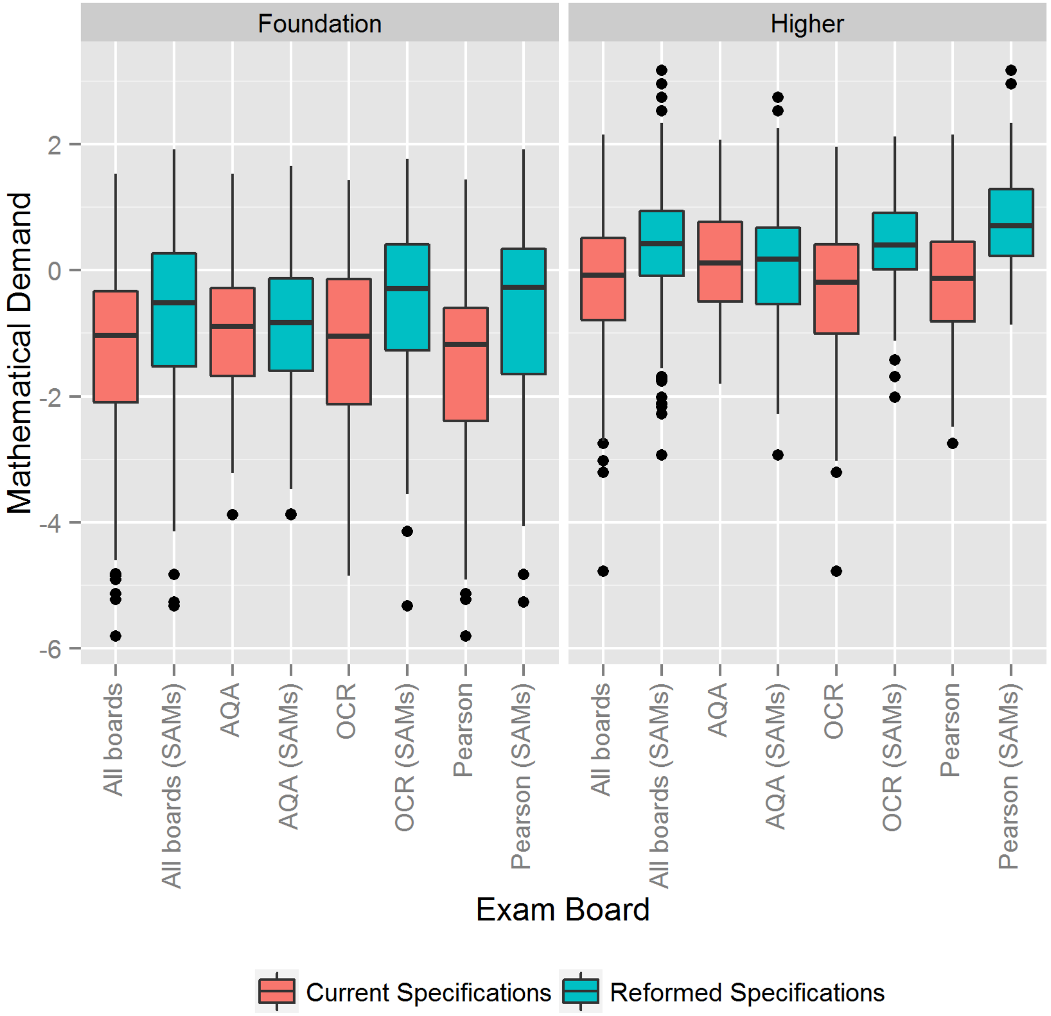


Figure 3: Box plots showing median and interquartile ranges of expected item difficulty parameters for England's GCSEs only

The following observations can be made from the analysis of the judges' expectations of relative item difficulty.

- For all exam boards the expected difficulty of the reformed GCSE sample assessments is higher than that of the current GCSE papers aggregated across boards.
- While the expected difficulty of the OCR and Pearson sample assessments is higher than that of the current papers, the expected difficulty of AQA's sample assessments is very similar to that of AQA's current papers.

- The size of the difference in expected difficulty between the sample assessments is greater than the difference in expected difficulty between current papers.
- The spread of expected item difficulty is less on the foundation tier than the higher tier sample assessments.
- The spread of expected item difficulty across AQA's foundation tier sample assessments is lower than for OCR's and Pearson's.
- Current GCSEs are judged to be of lower expected difficulty than similar international assessments.
- In general, the expected difficulty of the reformed, higher tier GCSE sample assessments is more in line with similar international assessments than the current GCSE papers.

Figures 4a to 4e show the items with the highest expected difficulty parameters. All of these seemingly challenging multi-step items fell into the domains of algebra or geometry. The most difficult item (figure 4a) may have been judged particularly difficult because an experienced mathematician would probably use calculus to solve it, rather than by analysing the roots of the quadratic and using reasoning to find the time value of the maximum point. Since calculus is usually a topic at senior secondary level, this may have influenced its high perceived difficulty.

Figures 5a to 5d show the items with the lowest expected difficulty parameters. All represent quite basic levels of arithmetic or very simple algebra. All are from England's GCSE papers (including the new SAMs), possibly reflecting the purpose of GCSE – to be accessible to students of all but the very lowest abilities.

Person A and Person B are throwing a ball to each other outside their house. Person B misses the ball and it falls to the ground.

The path of the ball can be modelled by the equation

$$h = -t^2 + 2t + 8$$

where *t* is the time in seconds since the ball is thrown, and *h* is the height in metres above the ground at any time *t*.

How much higher does the ball rise above the height of the point from which it is thrown?

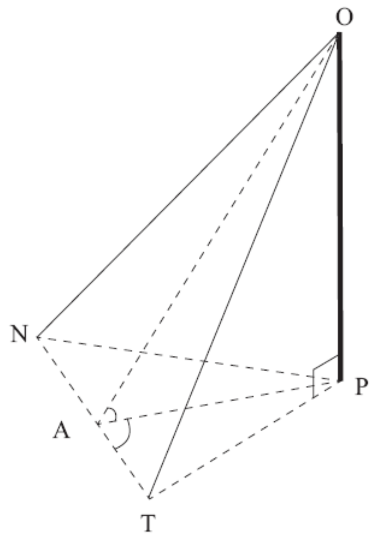
Explain what you are calculating at each step of your answer.

Figure 4a: The item with the highest item difficulty parameter, New Zealand Paper 91027 Q3d(ii) (parameter = 4.30)

Calculator allowed.

A basketball post has a set length for OP. A 3-D sketch of the goal post is given below.

OT and ON are both 90 cm long.
PT, PN and NT are all 40 cm long.
Point A is halfway along NT.



(i) Calculate the size of angle TAP.

Explain your reasoning.

(ii) Calculate the length of AP.

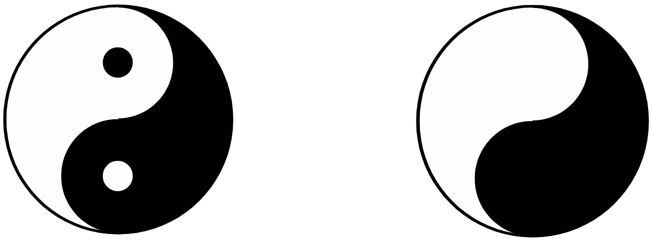
(iii) Calculate the angle OAP.

Show your working clearly.

Figure 4b: The item with the second highest item difficulty parameter, New Zealand Paper 91031 Q2b(iii) (parameter = 3.43)

Formula sheet provided.

Below on the left is the Yin-Yang symbol. This is a Chinese symbol, where the black part represents the moon (Yin) and the white part the sun (Yang).



In this question we consider a simplified version of the Yin-Yang symbol, which is depicted on the right. Here the dots are omitted.

The boundary between the black and white portions is formed by two half-circles.

(a) A circle with centre M is shown below. Draw the half-circles in this simplified Yin-Yang symbol.

(b) The below Yin-Yang symbol has a diameter of 5 cm. Show that the circumference of the black part of the symbol is as large as the circumference of the whole circle.

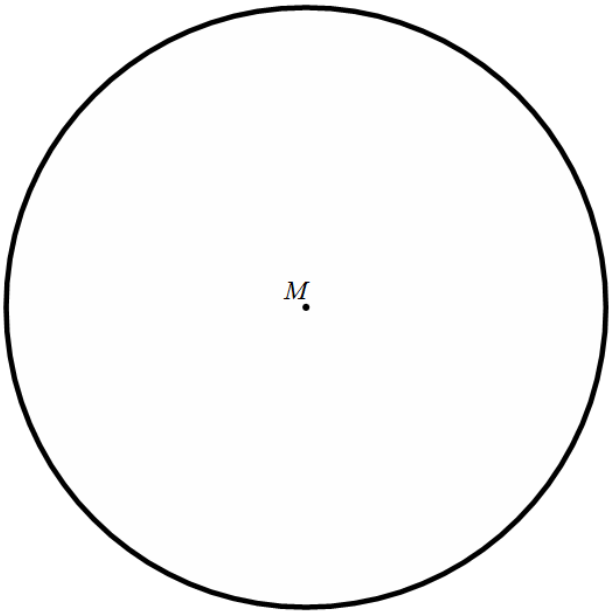


Figure 4c: The item with the third highest item difficulty parameter, Netherlands Q24 (parameter = 3.30)

Calculator allowed.

(i) In a children's play park, a ball is kicked so that its flight path can be modelled by the equation

$$h = - ax(x - 6)$$

where *h* metres is the height of the ball when it is *x* metres from the point from where it is kicked.

If the maximum height of the ball is 2 m, what is the value of *a*?

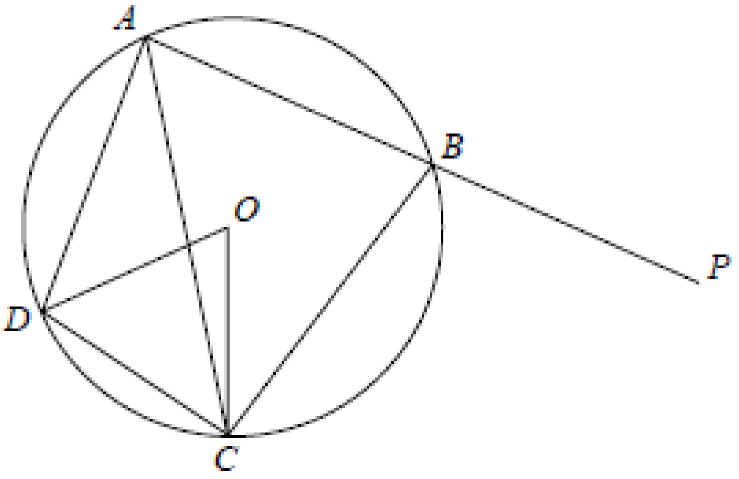
(ii) A ball is kicked from the ground and lands at a point 10 m away on the opposite side of a goalpost.

The crossbar of the goalpost is 2 m above the ground.

When the ball passes over the crossbar, it is at its maximum height of 2.5 m.

Give the equation for the height, *h* metres, of the ball above the ground at a distance, *x* metres, from where it was kicked, if the path of the ball is modelled by a parabola.

Figure 4d: The item with the fourth highest item difficulty parameter, New Zealand Paper 91028 Q3d(ii) (parameter = 3.24)



A, B, C and D are four points on a circle, centre O.
PBA is a straight line.

Angle $PBC = 100^\circ$
Angle $DAC = 23^\circ$

Show that the size of angle $OCA = 10^\circ$

You must give a reason for each stage of your working.

Figure 4e: The item with the fifth highest item difficulty parameter, Pearson SAMs Higher Tier Paper 1 Q15 (parameter = 3.17)

Simplify $d + d + d + d$

Figure 5a: The item with the lowest item difficulty parameter, Pearson GCSE Foundation Tier Paper 1 Q4a (parameter = -5.80)

Calculator allowed.

(a) Insert one of the symbols $<, >$ or $=$ to make each statement true.

(i) -5 ……………. -7

Figure 5b: The item with the second lowest item difficulty parameter, OCR SAMs Foundation Tier Paper 1 Q10a(i) (parameter = -5.33)

Solve $4x = 20$

Figure 5c: The item with the third lowest item difficulty parameter, Pearson SAMs Foundation Tier Paper 1 Q2a (parameter = -5.26)

Calculator allowed.

(a) Work out $+3 - 5$

(b) Work out $-12 - 6$

Figure 5d: The items with the fourth and fifth lowest item difficulty parameters, Pearson GCSE Foundation Tier Paper 3 Q4b (parameter = -5.23) and Pearson GCSE Foundation Tier Paper 3 Q4a (parameter = -5.13)

1.3.5 Analysis of expected difficulty by mathematical domain

The items were classified by their principal mathematical domain as defined in the latest GCSE subject content for reformed GCSEs.[15] Some items crossed more than one domain; in these cases a judgement was made as to which was the primary

[15] www.gov.uk/government/uploads/system/uploads/attachment_data/file/254441/GCSE_mathematics_subject_content_and_assessment_objectives.pdf

domain. Although there was an element of subjectivity, and no weighting by item marks, this classification allowed an analysis of the expected difficulty of the domains in the assessment materials.

Figure 6 shows that across the boards the spread of expected difficulty was greatest on the higher tier. On the higher tier, geometry was expected to be the most difficult domain and statistics and number the least difficult. On the foundation tier, probability was expected to be the most difficult domain and algebra and number the least difficult. In general, the pattern of expected difficulty across exam boards followed the trend observed for the whole paper. For example, there was no domain in which the AQA items were expected to be the most difficult on either tier. For some domains the Pearson items were expected to be the most difficult, but for others the OCR items were expected to be of highest difficulty. The expected difficulty parameters, domain and tariff for individual items on each paper can be found in appendix B.

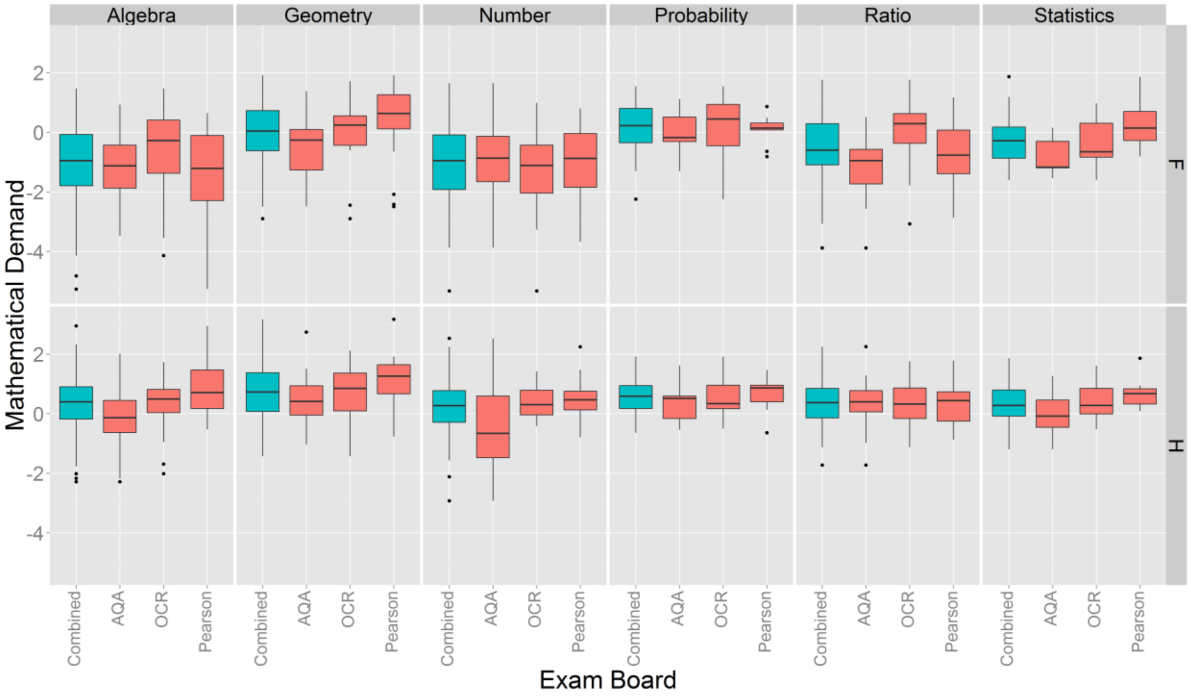


Figure 6: Box plots showing median and interquartile ranges of expected item difficulty parameters for the reformed specifications by mathematical domain

1.3.6 Analysis of expected difficulty by assessment objective

The items were classified by assessment objective.[16] The assessment objectives can be summarised as:

- Use and apply standard techniques (AO1);
- Reason, interpret and communicate mathematically (AO2); and
- Solve problems within mathematics and in other contexts (AO3).

They are fully described in appendix C. Some items cross more than one assessment objective; in these cases the item was allocated to the primary assessment objective. Where the marks associated with an assessment objective were equal, the item was allocated to the assessment objective with the higher labelling (e.g. an item equally split between AO1 and AO2 would be allocated to AO2). In other words, allocation was systematic but arbitrary in nature. When the allocation was reversed (i.e. the item was allocated to the assessment objective with the lower labelling) the findings were broadly comparable.

Figure 7 shows that, in general, on the foundation tier, AO1 items were expected to be less difficult than AO2 and AO3 items. This was also the case for the higher tier items although the effect was less pronounced. The AO2 and AO3 items were of similar levels of expected difficulty. In general, the pattern of expected difficulty across exam boards followed the trend observed for the whole paper. In general, AQA items were expected to be the least difficult whatever the assessment objective they were intended to measure. There were, however, exceptions. For example, OCR's AO3 items were expected to be the least difficult AO3 items on the higher tier. The differences in expected difficulty between boards were greatest for AO1 and AO2.

[16] The classification conducted by exam boards was used. This classification was scrutinised and challenged during accreditation so systematic differences in the allocation of items to AOs ought not to exist.

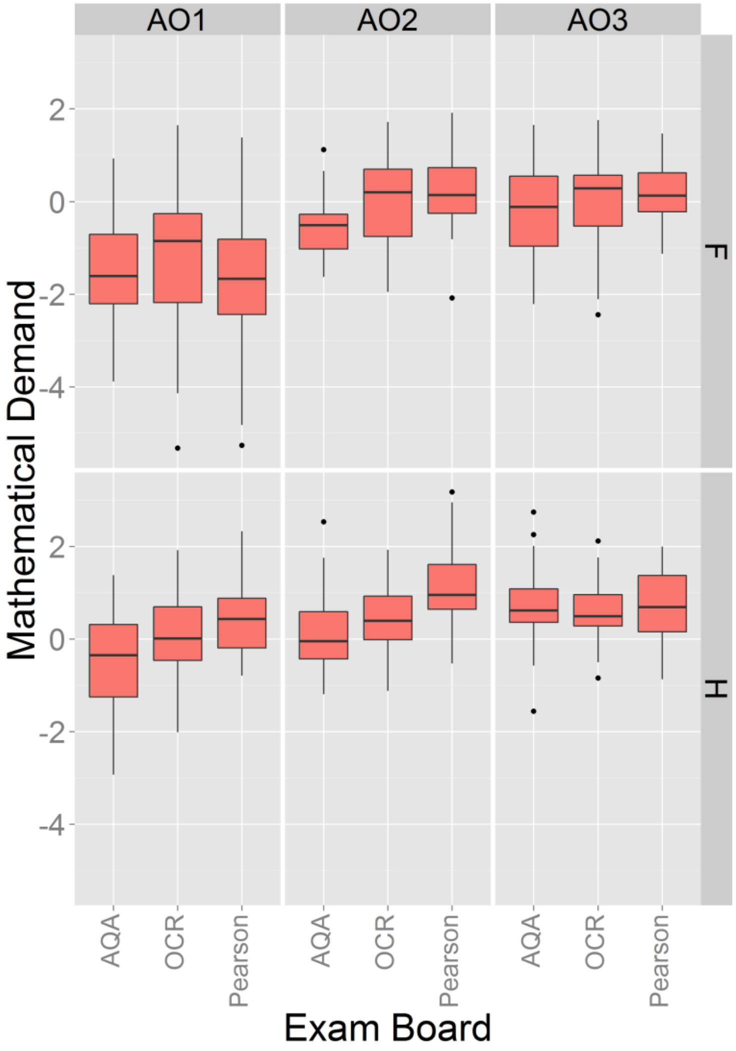


Figure 7: Box plots showing median and interquartile ranges of expected item difficulty parameters for the reformed specifications by assessment objective

1.3.7 Analysis of expected item difficulty by tariff

The relationship between the maximum mark and the expected difficulty of the items was explored. Figure 8 shows there was a tendency for items with low mark tariffs to be perceived as easier than items worth more marks (r=0.47). However, there are many examples of items with low mark tariffs which were expected to be relatively hard and items with high mark tariffs which were expected to be easy. It would be wrong to assume that assessments with large numbers of small tariff items are automatically easier than papers comprising higher tariff items.

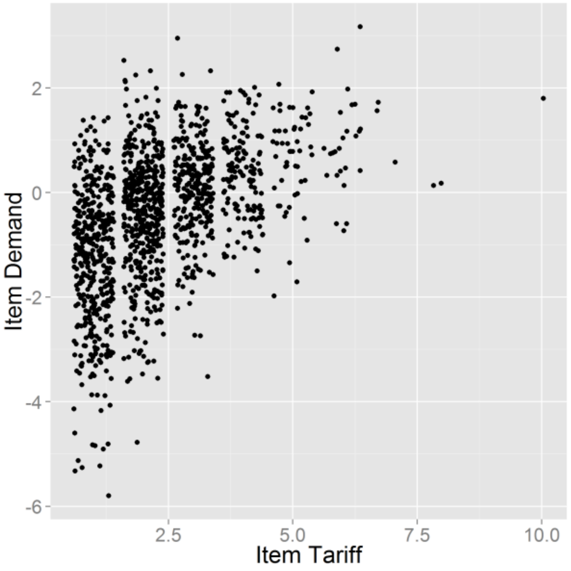


Figure 8: Expected item difficulty by item tariff

1.4 Simulations of student performance using model expected difficulty parameters

The item parameters estimated by fitting the Rasch model have units of logits and are, in absolute terms, arbitrary. Given the arbitrary nature of this scale it is difficult to understand the consequences of any differences. To provide a more interpretable representation of the differences in item parameters shown in figures 2 and 3, marks on these items of students of different abilities were simulated. These were then used to construct simulated mark distributions for each exam paper, allowing transformation from the logit scale to a more meaningful mark scale.

In addition to providing a more practically meaningful scale, notional grade boundaries can be set on these simulated distributions. This allows the variation in notional grade boundary position to be compared across exam boards to spot any potential problems with awarding. Performing this analysis for the current papers provided a useful basis for comparison. It is important to remember that this would represent differences in grade boundary position only if the difficulty of the exam papers were to mirror the expected mathematical difficulty exactly.

1.4.1 The simulation process

To produce the mark distributions for each paper, the item level response patterns for 10,000 students were simulated. Based on the Rasch model for dichotomous data, the probability of each student (given his or her ability) responding correctly to each item (given its expected difficulty) was calculated.

Given the question asked of judges, the item parameters represent the expected difficulty of responding fully to an item. It is therefore extremely challenging to simulate partial credit in this model and the calculated probabilities represent the chances of a fully correct response. The calculated probabilities of a correct response were used to simulate the attribution of either zero marks for an incorrect response or the maximum item tariff for a correct response to each student for each item to allow the intended weight of items to be reflected. These simulated item response patterns were then summed to form a student level score for the paper, allowing construction of an overall mark distribution for the simulated cohort.

1.4.2 Item parameters

To reflect the uncertainty in the expected item difficulty parameters, the parameter values used for the simulation were drawn randomly from distributions for each item. These were normal distributions, centred on the parameter estimate with a standard deviation equal to the standard error of the estimation. To reflect the uncertainty in the item level parameters in the overall mark distribution, ten sets of item parameters (and therefore ten mark distributions) were produced for each exam paper with different random item parameters being drawn for each run.

1.4.3 Setting grade boundaries

To evaluate the impact that differences in expected difficulty could have on the grade boundary position for the different question papers, notional grade boundaries were located on each mark distribution. The notional boundaries provide a common basis for comparison and allow estimation of the impact of differences in expected difficulty between papers. To establish this common basis for comparison it was necessary to define some plausible grade outcomes for each tier. For the purposes of the comparisons made here, all points are referenced to the grade scale for the current specifications (that is grades A* to G rather than 9 to 1 to be used for the reformed specifications).[17] This means that grades A, C and F boundaries are defined here for the current papers and the reformed versions despite the impending change to the grading scale.

The typical tier level cumulative percentage outcomes selected here are as shown in table 2. It should be noted that the grade boundaries set in these simulations will not necessarily relate strongly to those set operationally. There are three reasons for this. First, the grade outcomes quoted in table 2 are indicative only. Second,

[17] Note that given the use of comparable outcomes for the setting of grade boundaries in the reformed GCSE specifications, grades A and 7 are equivalent and grades C and 4 are equivalent and, therefore, the differences here are largely only in notation.

expected difficulty, which is the subject of this study, does not have a one-to-one relationship with actual difficulty. Therefore the mark distributions which are observed on the papers from the current specifications will differ from those simulated. Further, differences between simulated and operational mark distributions will exist due to differences between the actual ability profile of students sitting the question papers and between the actual and simulated ability distributions. Nevertheless, identification of the grade boundaries through simulation here is instructive as it indicates the consequences of differences in expected mathematical difficulty.

Table 2: Notional grade outcomes used to set grade boundaries

	Cumulative percentage outcome		
Tier	A	C	F
Foundation	–	30	90
Higher	30	90	–

1.4.4 Person parameters

The value of the item parameters clearly impacts on the shape of the mark distributions; so too, however, will the assumed distribution of student ability parameters used for the simulation. There are three factors that should be considered when selecting the person parameters, which may affect the findings:

1. relative inter-tier student ability
2. overall ability of the simulated students relative to the expected item difficulty
3. spread of ability of students relative to the expected item difficulty.

Under the assumption that the ability of students achieving a grade C on the foundation tier and those achieving a grade C on the higher tier are the same, the information in table 2 can be used to define the first of these conditions. When defining the distributions from which to select person ability parameters, irrespective of the overall cohort ability (point 2) or the spread of student abilities (point 3), the foundation and higher tier ability distributions overlapped such that, at a certain point on the ability scale, 30 per cent of the foundation tier and 90 per cent of the higher tier students had that level of ability or higher. By definition this point would also be coincident with the grade C boundary position on both tiers. Figure 9 shows an example set of ability distributions that realise this condition with the zero point on the ability scale being the point at which these conditions are met.

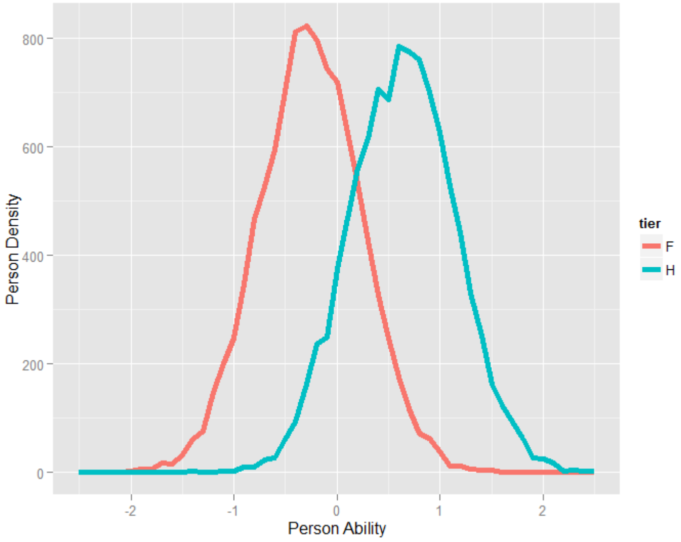


Figure 9: Example inter-tier ability distributions

The definition of the absolute student ability and spread of abilities is more challenging to define. One approach would be to reference the distributions back to those that occurred operationally in the current papers. However, due to the modular nature of the specifications used and the currently available data, this modelling is not trivial and would only act as a very loose approximation as a single set of ability distributions across all exam boards and question papers would be unlikely to approximate well. For this reason, the simulations were performed for all combinations of a range of relative cohort abilities (between -0.4 and +0.4 on the latent scale) and a range of spreads of tier level ability distributions (between 0.1 and 0.9 on the latent scale – equivalent to a variation of 0.1 to 1.2 across the combined ability distributions).

1.4.5 Simulated mark distributions

Figures 10 and 11 show, as an example, the simulated mark distributions for the current Pearson question papers and Pearson sample assessments respectively. The offset of the ability of the cohort relative to the items was set to zero and the spread of tier level ability parameters was 0.5. Shown on these plots are the locations of the notional grade boundaries identified as outlined above. Although the boundaries shift up or down depending on the expected difficulty of the paper, the gap between the boundaries varies little. A full set of mark distributions for all current exam papers and sample assessments using these ability parameters is provided in appendix D. It is clear for the example Pearson plots in figures 10 and 11 that the higher expected mathematical difficulty of the sample assessments relative to the current question papers has led to a less negatively skewed distribution with lower grade boundaries.

The main purpose of these simulations is to identify whether or not the level of variability in grade boundaries for the sample assessments would differ from the variability in the current exam papers. Shown in figures 12 and 13 are the simulated grade boundaries for the current papers and sample assessments (ability offset = 0, sd of tier level ability = 0.5). Table 3 shows the standard deviations for, and range of, grade boundaries, expressed as a percentage of the maximum mark, across the different papers.

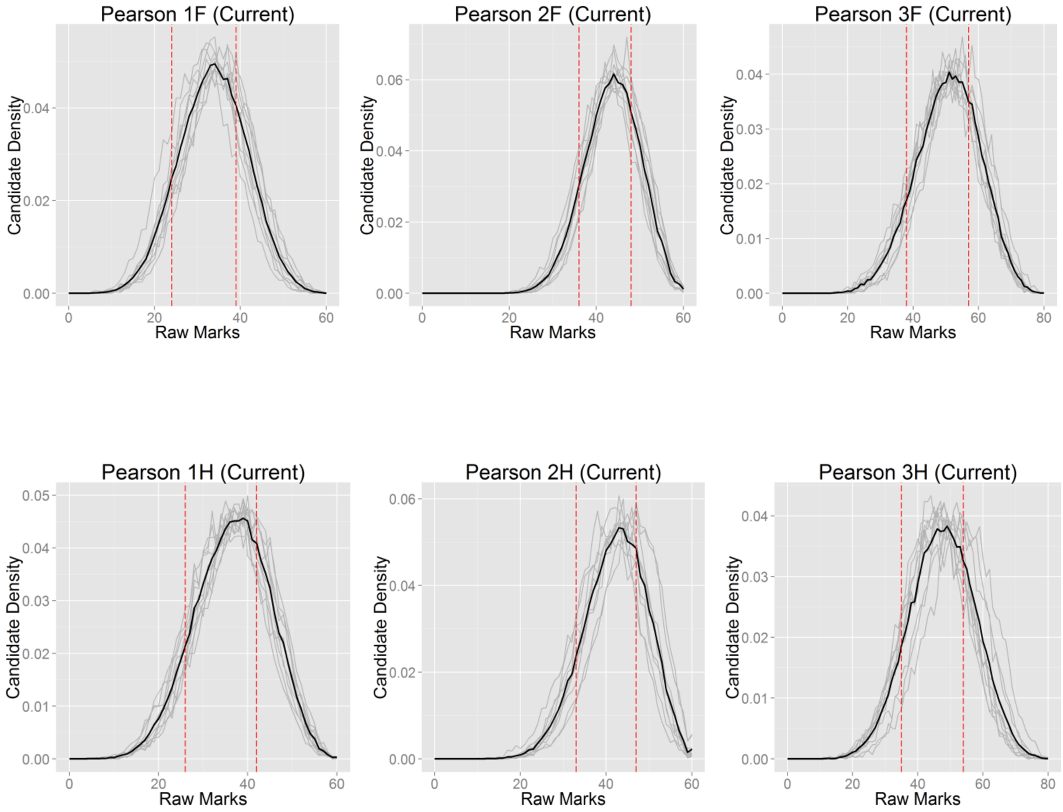


Figure 10: Simulated mark distributions for the current Pearson question papers. Red lines indicate the notional grade boundary positions

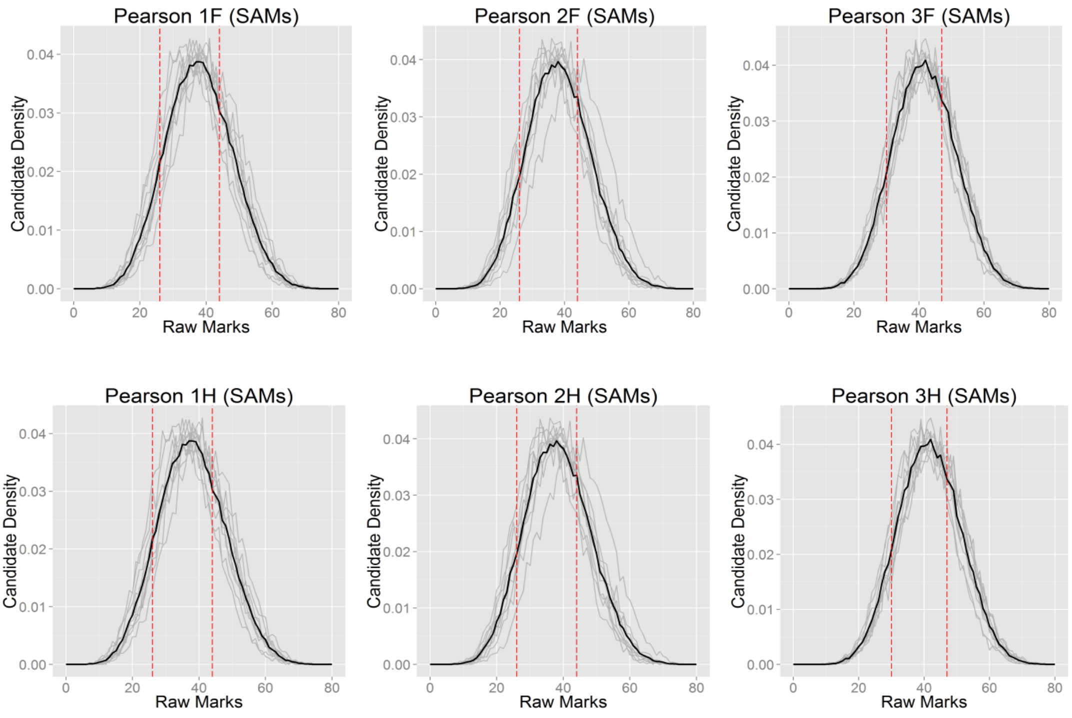


Figure 11: Simulated mark distributions for the Pearson sample assessments. Red lines indicate the notional grade boundary positions

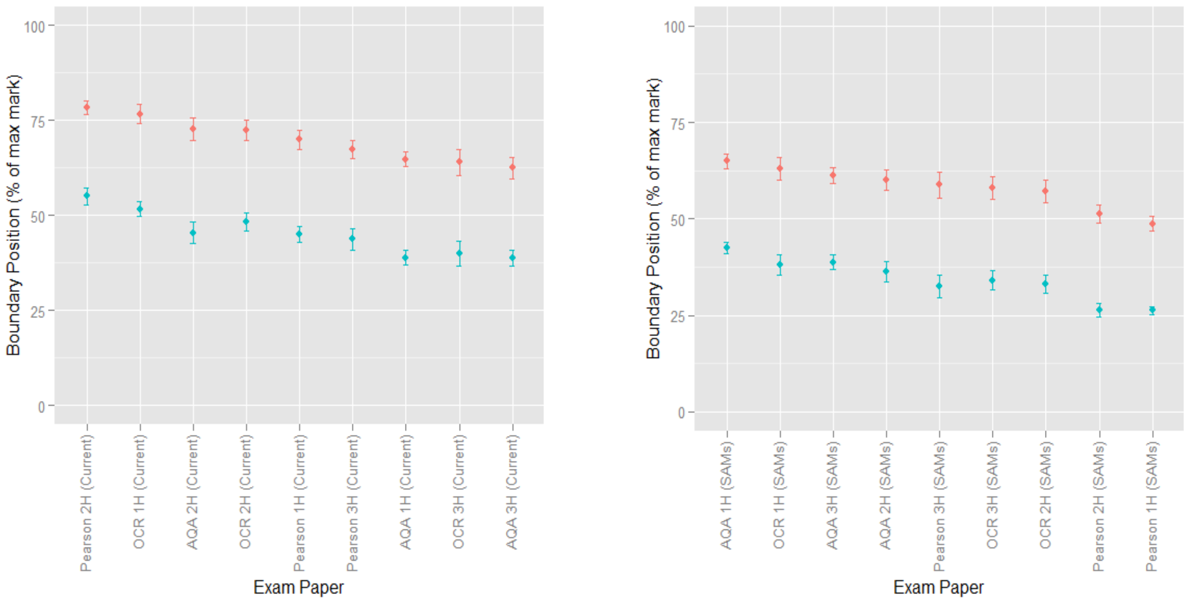


Figure 12: Simulated grade A (pink) and grade C (blue) boundary positions for the higher tier current exam papers and sample assessments

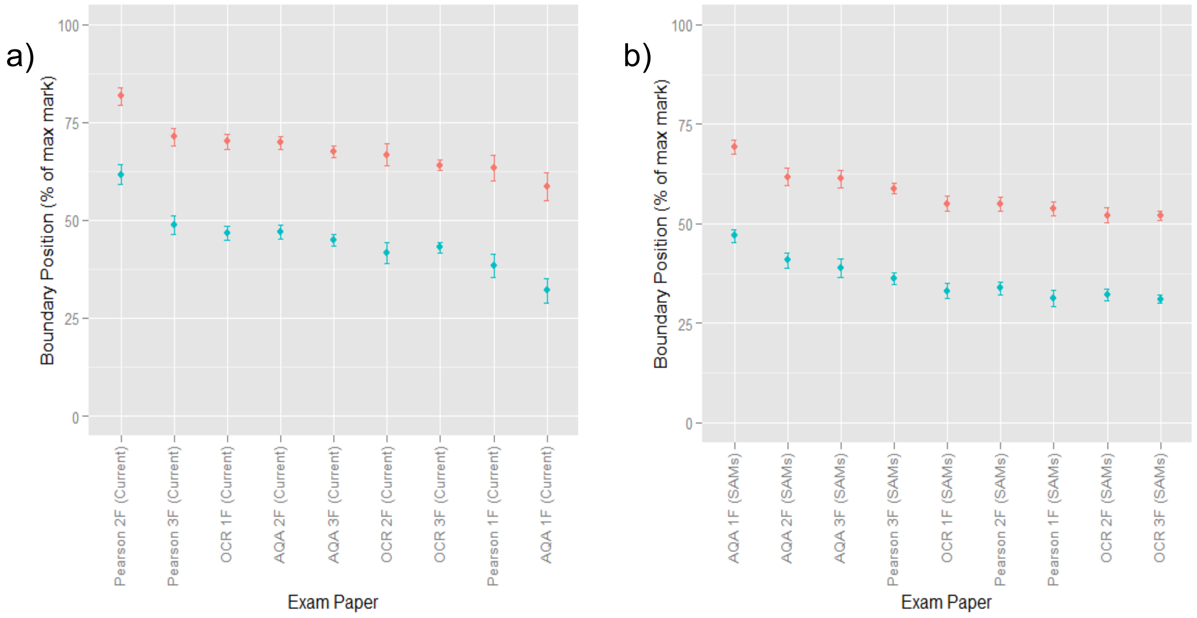


Figure 13: Simulated grade C (pink) and grade F (blue) boundary positions for the foundation tier a) current exam papers and b) sample assessments

Table 3: Standard deviation of grade boundary positions for simulated mark distributions (ability offset = 0 and ability spread = 0.5)

		SD of boundary position (%age of max mark)			Difference between highest and lowest boundary (% of max mark)		
		A	C	F	A	C	F
Current	Foundation	-	5.52	7.50	-	19.84	28.14
	Higher	6.37	6.16	-	17.08	18.15	-
Sample assessments	Foundation	-	6.13	5.76	-	18.14	16.91
	Higher	4.82	5.08	-	16.25	16.25	-

To establish whether or not the choice of student ability distribution has a significant impact, similar analyses were performed for a range of student ability. Figure 14 shows the variation in boundary spread and range across a range of overall cohort abilities and spreads of ability.

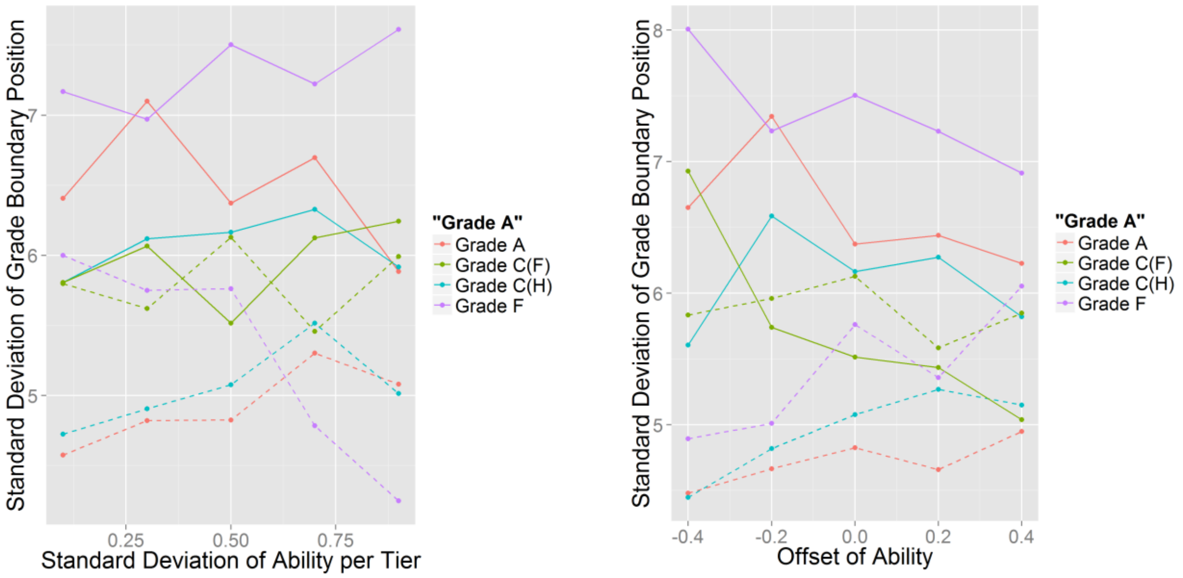


Figure 14: Variation in grade boundary spread for the current papers (solid) and sample papers (dotted)

The results illustrated in figure 14 suggest that, if the expected mathematical difficulty of the items translates into actual item difficulty, the variation in grade boundary position for the sample papers will be lower than for the current papers.

1.5 Summary of findings

While the current GCSEs were judged to be of lower expected difficulty than similar international assessments, in general, the expected difficulty of the reformed higher tier GCSE sample assessments was more in line with international assessments. The comparability of the GCSE with similar international assessments will be the subject of a future in-depth report. However, it is worth guarding against superficial comparisons. A wide range of assessments, for different ages, abilities and purposes, was included. Moreover, the actual difficulty of the assessments will vary according to how they are operationalised. For example, a seemingly challenging but predictable assessment can be very easy for students to complete. To some extent the international assessments are best conceptualised as representing the curriculum aspirations of that jurisdiction.

For all exam boards, the expected difficulty of the reformed GCSE sample assessments was found to be higher than that of the current GCSE papers aggregated across boards. However, while the expected difficulty of the OCR and Pearson sample assessments was higher than that of their current papers, the expected difficulty of AQA's sample assessments was very similar to that of their current papers. It is worth noting though, that the expected difficulty of the current AQA GCSE was higher than that of OCR or Pearson.

The size of the difference in expected difficulty between the sample assessments was greater than the size of the difference in expected difficulty between current papers. However, simulations suggested that the grade boundary setting process[18] would compensate for differences in difficulty of this magnitude, with higher boundaries set on easier papers and vice versa. Nonetheless, the potential wash back on teaching and learning needs careful consideration.

The spread of expected item difficulty was lower on the foundation tier than the higher tier. This was the case for all exam boards' sample assessments. This may be of concern as the foundation tier assessment in the new GCSE covers grades 1 to 5, whereas the higher tier assessment supports a wider range of grades from 4 to 9. It may be that the higher tier assessments will fail to discriminate sufficiently between students to allow reliable grading.

In general, the pattern across exam boards of expected difficulty by mathematical domain and assessment objective followed the trend observed for the whole paper. For example, there was no domain in which the AQA items were expected to be the most difficult on either tier, and AQA items tended to be expected to be easiest whatever the assessment objective they were intended to measure. The differences in expected difficulty between boards were judged to be greatest for AO1 and AO2 rather than AO3.

[18] http://webarchive.nationalarchives.gov.uk/20141031163546/http:/ofqual.gov.uk/standards/summer-2014-exams/#our-approach-to-summer-2014-awarding

2. Study 2 – A comparison across exam boards of the difficulty of the non-calculator sample assessments

2.1 Design

Study 2 involved a sample of Year 11 students preparing for their Maths GCSE, taking one non-calculator paper from all four exam boards as a mock examination. Time pressure meant it was not possible to test all the sample assessments in this way. The exam board was randomised within each class within each school to ensure that the groups taking each paper were randomly equivalent. Responses were marked online by experienced markers using standard procedures. The marks were then analysed to obtain item difficulty parameters that were aggregated to whole assessment level.

2.2 Method

2.2.1 Materials

The non-calculator papers from each exam board's higher and foundation tier sample assessments were included in this study. AQA, Pearson and OCR split their assessment into three 90-minute papers for each tier, with one of the three papers not allowing the use of calculators. Eduqas took a different approach, with two larger papers at each tier (one non-calculator) each of two hours and 15 minutes duration.

The length of the papers had to be standardised across exam boards to allow a fair comparison of difficulty. Eduqas's non-calculator papers (135 minutes) contained 120 marks. To reduce the Eduqas papers to make them equivalent to a 90-minute paper, 40 marks were removed from the higher tier paper and 36 from the foundation tier paper. To do so without introducing bias, whole items were selected for removal following these principles, across a paper:

1. the proportion of marks within each mathematical domain should remain the same

2. the proportion of marks assigned to each assessment objective should remain the same

3. the proportion of items falling into the top, middle and bottom thirds of expected difficulty (as identified by Eduqas) should remain the same.

In addition, an attempt was made to retain the same proportion of common items and a similar balance of items assessing new versus existing topics.[19] Proposed item deletions were agreed with Eduqas and did not lead to any fundamental distortion of the content of the papers.

Papers were transcribed into a neutral format with no identifying marks and using a common font to avoid potential bias. Each item started on a new page. The layout of the response space was reproduced as per the original paper, including lines, spaces and the prompt for the final answer. A generic rubric sheet for the front of all eight papers was used. A unique, anonymous code was used to identify the papers throughout the study, including during standardisation and marking. Hence, there was no way to identify the papers, beyond teachers' and markers' previous familiarity with the sample assessments.

2.2.2 Participants

Schools with Year 11 students preparing for their maths GCSE in June 2015 were recruited for the study. Motivation for participation was stimulated by the promise of student and item level analyses that would support preparation for the GCSE in June 2015 and for teaching for the reformed qualifications, which begins in September 2015.

It was hoped that Year 11 students approaching their live examination date would be motivated to perform at their best on the sample papers. As the main purpose of the study was to consider relative performance on the papers, any motivation and preparedness effects would have to affect one paper more than another for the results to be confounded.

The number of participants required was estimated by the precision of relative item facility that could be achieved. It was calculated that, based on bootstrap simulations of dichotomous and polytomous items, 500 participants per paper would be needed to achieve an estimate of item facility within +/-0.04 of the true item facility (with 95 per cent confidence).

2.2.3 School recruitment

A range of approaches to school recruitment were taken in parallel. First, a random selection of 600 English schools was drawn from the Department for Education's

[19] The reformed mathematics 2015 GCSE contains several topics that were not included in the existing GCSE specifications, and the foundation tier includes some topics that previously were covered only in the higher tier.

Edubase,[20] and the heads of maths were emailed and asked whether they would be interested in participating.

Second, a number of organisations (e.g. teacher associations) advertised the study in newsletters or email posts. Third, direct pre-existing links with schools were used: these schools were contacted directly and asked whether they would like to participate. Fourth, a letter was sent to approximately 600 secondary schools, informing them of the study. Finally, digital media was used to inform people of the research programme, providing research summary details[21] and contact details should they wish to take part.

There was no need for a perfectly nationally representative sample of students or schools as the relative item/paper difficulties were more important than the absolute values. However, to ensure a balanced representation of schools, the final selection was informed by the number of students at the school at the end of key stage 4, average key stage 2 point score of those students and the percentage of students achieving five or more GCSE grades A* to C, including in English and maths.

The schools on this shortlist were then contacted again and asked to sign a consent form and to provide a student list, including information regarding preferred tier. Replacement schools were selected when schools were not able to provide all of the information needed at any stage or decided that they no longer wished to take part.

2.2.4 Sitting the exam

The administration of the exam was the schools' responsibility. They received question papers that included instructions to candidates to be read out beforehand (these were largely reproduced on the paper rubric). The students completed the tests as mocks under exam conditions, as determined by their teachers. Completed exam papers were then returned and scanned into the marking system.

2.2.5 Markers

Fifty experienced maths markers were recruited from lists provided by the exam boards. Table 4 shows the number of markers with experience of marking for each board. Most of the markers had prior experience of marking for Pearson (due to

[20] register of educational establishments in England and Wales

www.education.gov.uk/edubase/home.xhtml;jsessionid=7A36E335E7ACC2ACBFD82C82F5023596

[21] www.gov.uk/government/publications/gcse-maths-summary-of-research-programme

Pearson's current demand for markers). Any bias in marking was monitored through the marking of seeds (see later).

Table 4: Number of recruited markers with experience of marking for the four exam boards[22]

Exam board	Number of markers
AQA	10
Pearson	41
OCR	11
Eduqas	8

2.2.6 Pre-standardisation, standardisation and marking quality control

Pre-standardisation involved making amendments to the mark schemes based on the expert input of four experienced maths principal examiners (PEs). A PE was recruited from each exam board. They had not been involved in the development of the sample assessment materials.

The four PEs carried out an independent review of several completed scripts for each paper against the mark scheme, noting any need for clarification of the mark scheme. The PEs then met for a day-long meeting to discuss these potential clarifications and amendments, and to finalise the mark scheme. The amendments were strictly to deal with ambiguous responses and the detailing of additional alternative methods, not changes to the way marks were assigned. These amended and annotated final mark schemes were used for the marking.

The PEs also identified items for which there may be some subjectivity in the awarding of marks. The quality control of the marking of these items was crucial and took the form of seed items interspersed during each marker's marking allocation (seed items are items pre-marked by the PE against which markers' marking was compared). Twenty-three items were identified for seeding (three per paper, including one that was a common item across the tiers; some of the harder items were not attempted by many students). Approximately 30 responses to each item were marked by all four PEs. For each item approximately ten responses were selected as seeds. These were responses spread across the mark range, and upon which there was unanimous (or very close to) agreement over the marks to award. Seeds were

[22] Numbers sum to more than 50 due to markers with experience of marking for more than one board.

mostly responses with intermediate marks, although some full or zero mark responses were included so as to detect any tendency on the part of the markers to avoid awarding marks at the extremes. The median mark awarded to a seed by the PEs is referred to as the ‘true mark’ from here on.

Each marker marked four papers (two foundation and two higher). Papers were allocated to markers using a matrix, so where possible no two markers would mark the same combination of four papers. They marked batches of items in order (for example 20 item 1 responses, 20 item 2 responses and so on). Every marker marked all the seeds for their allocated papers. As they marked items, the relevant section of the annotated mark scheme was displayed on-screen, along with the marking guidance notes (which explain the allocation of process and method marks, and so on).

2.3 Results

2.3.1 Analysis

Classical test theory and Rasch analysis were used to analyse the item, paper and student performance.

2.3.2 Number of students and representativeness of the school sample

While the intention was to recruit a balanced number of students by paper and tier, in the event, the numbers were somewhat uneven (see table 5). As the schools chose the tier of entry for their students it was challenging to balance numbers by tier. The disparity by exam board was largely due to the use of random allocation of papers to boards rather than a strict spiralling of the papers. Absentees on the days of testing also caused some of the disparity.

Table 5: Number of students per paper

	OCR		AQA		Pearson		Eduqas	
	F	H	F	H	F	H	F	H
Number of students	362	648	325	618	353	627	341	591

The percentage of students achieving GCSE grades A* to C in maths in 2014 for the sample of schools was 69 per cent, which is very similar to the national average in 2014 of 68 per cent.

2.3.3 Marking reliability

Table 6 shows the accuracy of marking of the seeds across the mark range for that item. The mean absolute mark difference is within a mark, suggesting high marking reliability.

Table 7 shows the accuracy of seed marking by question paper. The mean absolute mark difference and mean bias are weighted by number of responses to each seed of each item. The mean absolute mark difference for each question paper is within a mark, suggesting high marking reliability. The mean bias is less than half a mark, suggesting that markers were not systematically severe or lenient in their marking for any of the papers. The greatest level of bias was a quarter of a mark severity for the Pearson higher tier paper. As the three seed items on the Pearson higher tier paper each showed a different pattern (generous – accurate – severe), however, it cannot be concluded that the entire question paper was marked severely.

Table 6: Seed marking by item

Tier	Exam board paper	Item	Mean abs mark diff	Maximum seed mark	Mean bias (avg mark – true mark)	Number marked
Foundation	OCR	6(b)	0.24	4	0.04	208
	OCR	12(c)	0.05	5	0.04	197
	AQA	9	0.08	3	0.04	250
	AQA	15	0.39	5	-0.03	220
	Pearson	6	0.20	5	-0.02	264
	Pearson	8	0.33	3	-0.09	240
	Pearson	9	0.33	4	-0.31	264
	Eduqas	7	0.22	4	-0.16	248
	Eduqas	8	0.63	4	-0.32	221
	Eduqas	11	0.63	2	0.63	156
Higher	OCR	8	0.17	3	0.08	251
	OCR	9	0.17	4	0.00	274
	OCR	14	0.23	4	-0.16	198
	AQA	9	0.08	4	-0.02	217
	AQA	13(a)	0.54	3	0.28	230
	AQA	18	0.10	4	-0.08	220
	Pearson	2	0.75	6	-0.56	274
	Pearson	9	0.08	4	-0.02	200
	Pearson	15	0.52	2	0.15	99[23]
	Eduqas	7	0.20	6	-0.08	276
	Eduqas	8(a)(b)	0.16	4	0.11	230
	Eduqas	11	0.29	5	-0.25	200

[23] Most markers had finished their allocation by the time question 15 was due to be marked, but the items for this seed were still marked by 11 or more examiners each.

Table 7: Seed marking by question paper

Tier	Exam board	Mean abs mark diff	Mean bias (avg mark – true mark)	Number marked
Foundation	OCR	0.15	0.04	405
	AQA	0.22	0.01	470
	Pearson	0.29	-0.14	768
	Eduqas	0.47	-0.02	625
Higher	OCR	0.19	-0.02	723
	AQA	0.24	0.06	667
	Pearson	0.48	-0.25	573
	Eduqas	0.21	-0.06	706

2.3.4 Student performance on the sample assessments

The performance of students is summarised in table 8. While non-responses were recorded as distinct from zero scores, they were treated as zero in the following analyses. For the purpose of comparison, all papers were scaled to have a maximum available score of 100. All papers had good internal consistency, with the lowest omega values at 0.85. The scaled mean scores of students on the AQA papers were higher than those of students sitting the other boards' papers. This is illustrated graphically below (see figures 15 and 16). The differences in difficulty between exam boards were statistically significant. Indeed, due to the relatively large samples of students taking the papers even, differences of one mark (which could readily be corrected in the setting of grade boundaries) would be statistically significant. For this reason inferential tests are not reported.

Table 8: Question-paper analysis

	OCR		AQA		Pearson		Eduqas	
	F	H	F	H	F	H	F	H
Number of students	362	648	325	618	353	627	341	591
Max available mark	100	100	80	80	80	80	84	80
Mean score	24.44	27.98	32.13	27.13	18.90	14.44	21.56	17.81
Standard deviation	15.82	16.96	14.82	13.51	10.51	10.40	10.81	14.97
Scaled max mark	100	100	100	100	100	100	100	100
Scaled mean score	24.44	27.98	40.16	33.91	23.62	18.04	25.67	22.27
Scaled sd	15.82	16.96	18.53	16.89	13.14	13.01	12.87	18.71
Cronbach's alpha	0.87	0.88	0.90	0.88	0.84	0.83	0.81	0.88
McDonald's omega_t	0.90	0.90	0.91	0.90	0.88	0.85	0.86	0.91

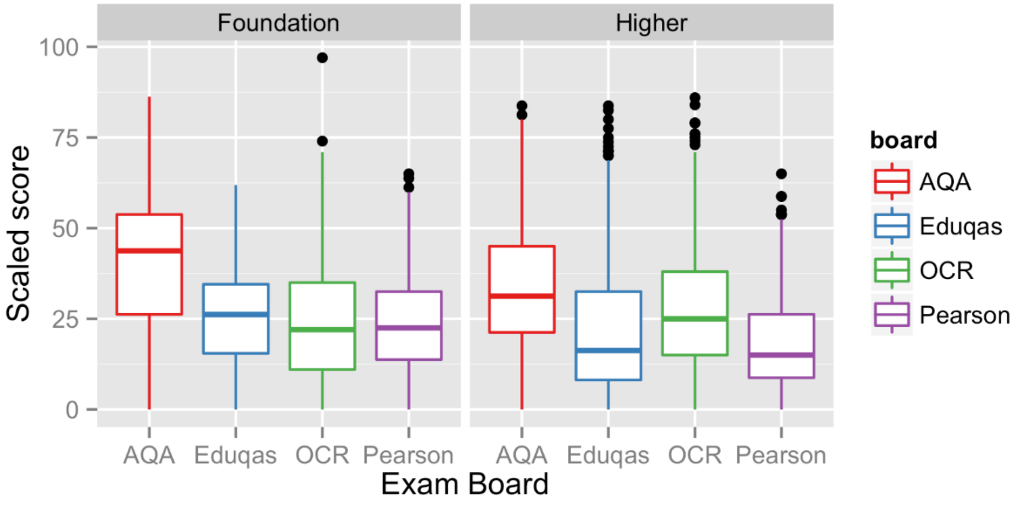


Figure 15: Box plots showing median and interquartile ranges of student scores

The distributions of raw scores are presented in figure 16. Live examinations rarely produce ideal mark distributions, but distributions as skewed as these are rare. With the exception of AQA's foundation tier assessment, the distributions are highly positively skewed. This is more so on the higher than the foundation tier. This suggests that the assessments were too difficult for the students and, as a consequence, the assessments have failed to satisfactorily discriminate between students of differing ability. This is particularly extreme for the Pearson higher tier assessment. The extent to which this might be due to a lack of motivation on the part of the students or to a lack of preparation for the new subject matter incorporated into the reformed GCSE will be explored.

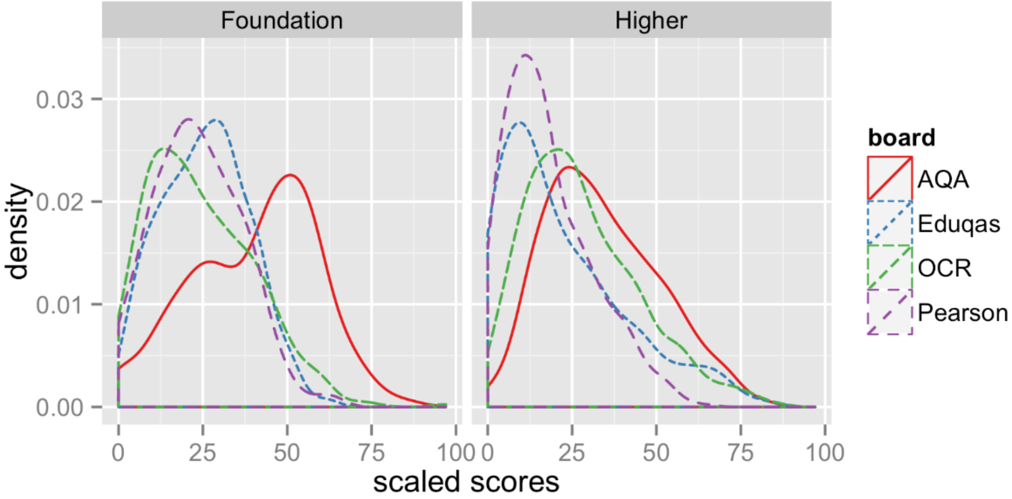


Figure 16: Raw score distributions

Exam boards identified items that included content that was new (not included in the current qualification). As well as identifying entirely new material, exam boards also highlighted content previously restricted to higher tier but now also included on foundation tier. The number of marks associated with these items is described below in table 9. It is noticeable that AQA identified fewer marks addressing new content than the other exam boards. However the two non-calculator papers from AQA happen to have the least new content of all six AQA SAM papers, and so this apparent difference is caused only by the papers selected.

Table 9: Marks for new content

		Marks for new content	
Board	Tier	Raw	Scaled
OCR	Foundation	30	30.00
AQA	Foundation	9	11.25
Pearson	Foundation	19	23.75
Eduqas	Foundation	20	23.81
OCR	Higher	27	27.00
AQA	Higher	10	12.50
Pearson	Higher	22	27.50
Eduqas	Higher	25	31.25

The difference in difficulty between the items testing new and old content is compared across exam boards in figure 17. As one would expect, students scored better on the items assessing old content than on the items assessing new content (for which they had not been prepared). On the higher tier items measuring new content, students scored higher on AQA’s items than they did on other boards’ items. On the foundation tier items measuring old content, students scored higher on AQA’s items than they did on other boards’ items. On the higher tier items measuring old content, students scored lower on Pearson’s items than they did on other boards’ items. The difference between students’ scores on the new and old content gives an estimate of the impact of the lack of preparedness on students’ scores. Even on the old content the average scores were lower than one would expect on a real examination. An additional factor is that the new content may have impacted on student performance on the items testing old content by demotivating them. Hence, the lower proportion of new content in AQA’s paper compared to the other boards may account for some of the difference in difficulty across boards. It seems unlikely, however, that this could explain the substantial differences observed.

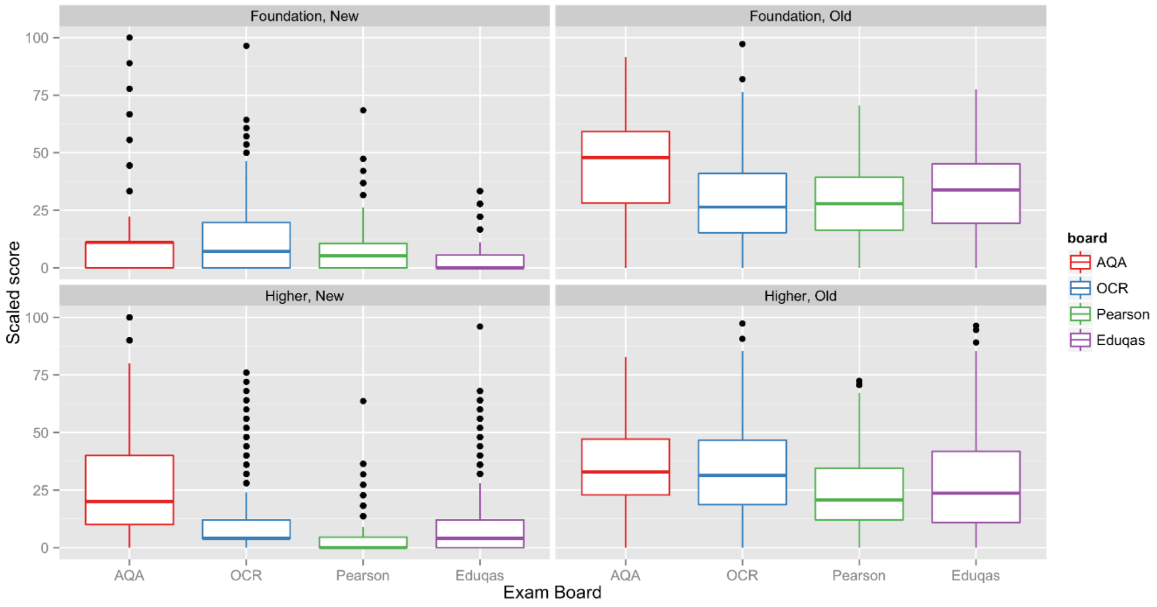


Figure 17: Box plots showing median and interquartile ranges of student scores for old and new content

2.3.5 The pre-test effect

Research has suggested that test-taking motivation affects test performance. In a low-stakes testing environment, where there are little or no consequences associated with test performance or perceived benefits for the test-takers, the performance on a test can be considerably lower than the performance under high-stakes conditions. In a study on the effect of motivation on the performance of the Key Stage 2 National Curriculum Science tests that were being pre-tested under a low-stakes condition, the overall pre-test effect represented an increase in average test facility of 14 per cent in 2006 and 13 per cent in 2007 when comparing live testing with pre-testing.[24]

If scores on the tests are boosted by 13 per cent (to emulate the pre-test effect) and the old content isolated from the new content, it is possible to consider how hard the new papers might be, without the confounding factors of motivation and new content. Adding 13 per cent to all scores assumes that the pre-test effect is constant across the score range. This would be highly unlikely in reality. The alternative was to add 13 per cent of each student's score to produce a new score. This assumes that the

[24] Pyle, K., Jones, E., Williams, C. and Morrison, J. (2009) Investigation of the factors affecting the pre-test effect in national curriculum science assessment development in England, Educational Research 51, 269–282.

pre-test effect is linear, that those students with relatively high scores (and who are probably motivated) are just as affected as those students with low scores. This is also highly unlikely in reality. For this purpose it was assumed that the pre-test effect was constant. This means that even the scores of students with zero per cent were boosted to 13 per cent. Clearly, this analysis provides only a rough estimate of the likely functioning of the assessments in 2017.

Figure 18 illustrates the placement of the grade boundaries at cumulative percentages of 30 per cent and 90 per cent to represent grade A and C boundaries on the higher tier, and grade C and F boundaries on the foundation tier. Note that the odd shape of the mark distribution at the bottom end is a product of all students' scores being boosted. The higher tier grade A boundaries would be placed at just above half marks for AQA and OCR, and just below half marks for Pearson and Eduqas. The higher tier grade C boundaries range from 16 per cent of the maximum mark for Eduqas to 28 per cent of the maximum mark for AQA. The higher tier grade boundaries are low in comparison with the papers from the current specifications (table 10). The boundaries on the foundation tier papers for grade C range from 70 per cent of the maximum mark for AQA to 50 per cent for Pearson. In comparison with the papers from the current specifications, a grade boundary set at 70 per cent of the maximum mark is not unusually high, but a grade boundary at 50 per cent is low. A percentage mark difference of 20 per cent between grade boundaries between different boards is not unusual.

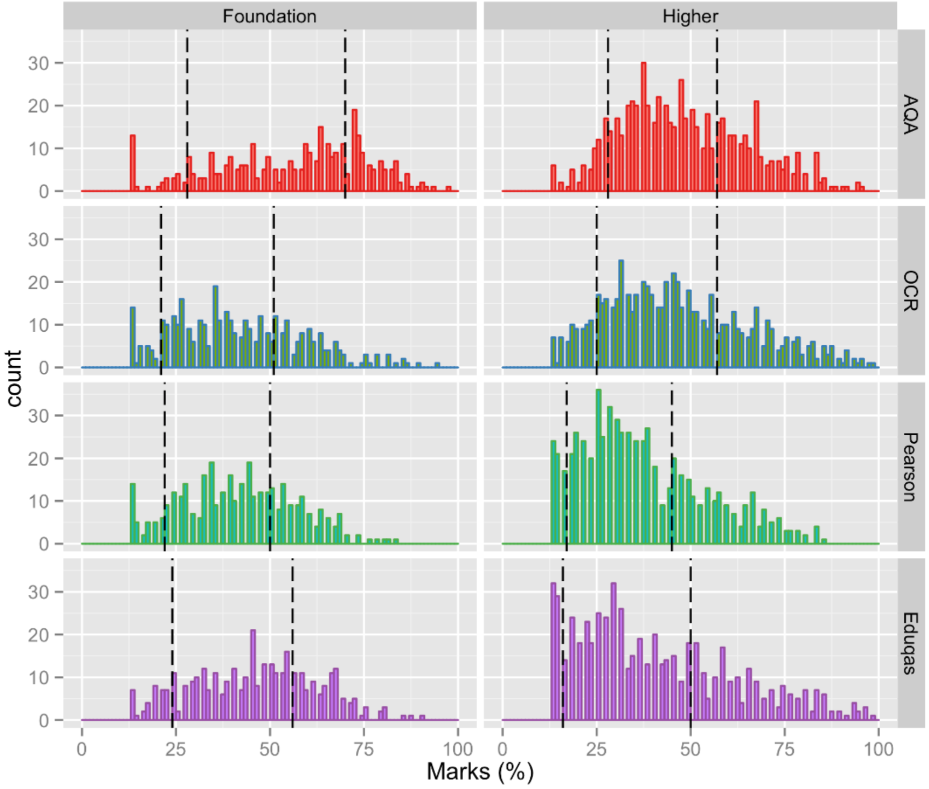


Figure 18: Histograms for existing content adjusted for the pre-test effect with notional grade boundaries overlaid to achieve cumulative percentage outcomes of 30 and 90 per cent (higher tier: grade A and C; foundation tier: C and F)

Table 10: Grade boundaries on the current papers in study 1

	Grade A		Grade C		Grade F	
Question paper	Mark	as percentage	Mark	as percentage	Mark	as percentage
AQA 1H	34	63.0	18	33.3	-	
AQA 2H	35	53.0	16	24.2	-	
AQA 3H	59	73.8	33	41.3	-	
AQA 1F	-		34	63.0	16	29.6
AQA 2F	-		37	56.1	15	22.7
AQA 3F	-		58	72.5	24	30.0
OCR 1H	34	56.7	14	23.3	-	
OCR 2H	34	56.7	17	28.3	-	
OCR 3H	83	83.0	56	56.0	-	
OCR 1F	-		35	58.3	16	26.7
OCR 2F	-		35	58.3	14	23.3
OCR 3F	-		75	75.0	36	36.0

Pearson 1H	44	73.3	21	35.0	-	
Pearson 2H	45	75.0	22	36.7	-	
Pearson 3H	58	72.5	35	43.8	-	
Pearson 1F	-		44	73.3	21	35.0
Pearson 2F	-		48	80.0	26	43.3
Pearson 3F	-		69	86.3	29	36.3

2.3.6 Rasch analysis

The partial credit Rasch model[25] (PCM) was used to provide additional information about the relative difficulty and the functioning of the sample assessments. The random equivalent group design used made it possible to compare the exam boards' assessments directly. In the following analysis, the average of the ability measures for all students taking a particular paper was set to zero such that the category step thresholds of the items from different papers (for the same tier) can be compared directly because the ability distribution of the students taking each paper was assumed to be similar.

The Rasch analysis software WINSTEPS[26] which implements the PCM was used to conduct the analysis. Wright Maps were plotted using the R package wrightmap.[27]

2.3.7 Model assumptions and model fit

Two important assumptions are required for the PCM: unidimensionality and local independence. Unidimensionality requires that one ability or a single latent variable is measured by the items in the test. Local independence requires that test-takers' responses to any items in the test are statistically independent when their underlying ability influencing their performance on the whole test is held constant. A low correlation was found between item residuals (-0.02 to -0.04), which suggests that local independence was broadly maintained. The percentage of variance explained by the Rasch model under exploratory factor analysis varied from 48.8 per cent to

[25] Wright, B. and Masters, G. (1982) Rating scale analysis, Rasch Measurement. Chicago, IL: MESA Press.

[26] Linacre, J.M. (2014) Winsteps® (Version 3.81.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved January 1, 2014. Available from www.winsteps.com

[27] Torres Irribarra, D. and Freund, R. (2014) Wright Map: IRT item-person map with ConQuest integration. Available at http://github.com/david-ti/wrightmap

63.1 per cent, which suggests that the tests broadly loaded on a single dimension. Further details relating to model fit are included in appendix E.

2.3.8 Test characteristic curve (expected test score distribution)

Once the Rasch model has been fitted, the expected scores on the papers can be compared. Figure 19 compares the test characteristic curves (TCCs) of the four papers from each of the two tiers. Again, the available marks on the papers were scaled to have a maximum score of 100. The test characteristic curve shows the relationship between the expected score on the test and person ability. When the curves for different tests are placed on the same ability scale and have the same shape, a test on the left will be easier than tests on the right, since for the same ability the expected score on the test will be higher than those on the other tests.

The average ability for this study was set to 0, which produces expected scores of less than half marks for every paper. The difference plot (figure 20) shows the difference in expected scores compared with the AQA papers at the same level of ability across the different question papers.

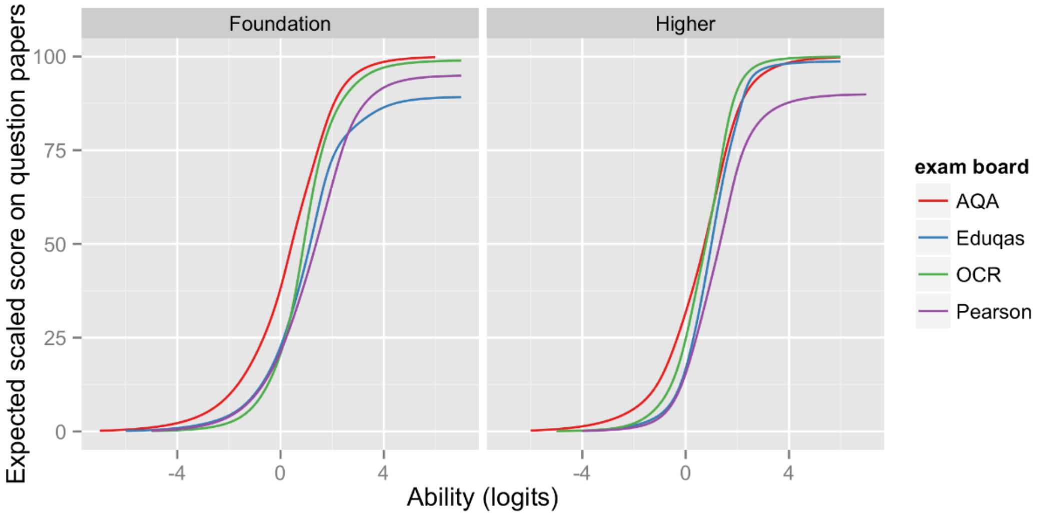


Figure 19: Expected scaled scores on the question papers

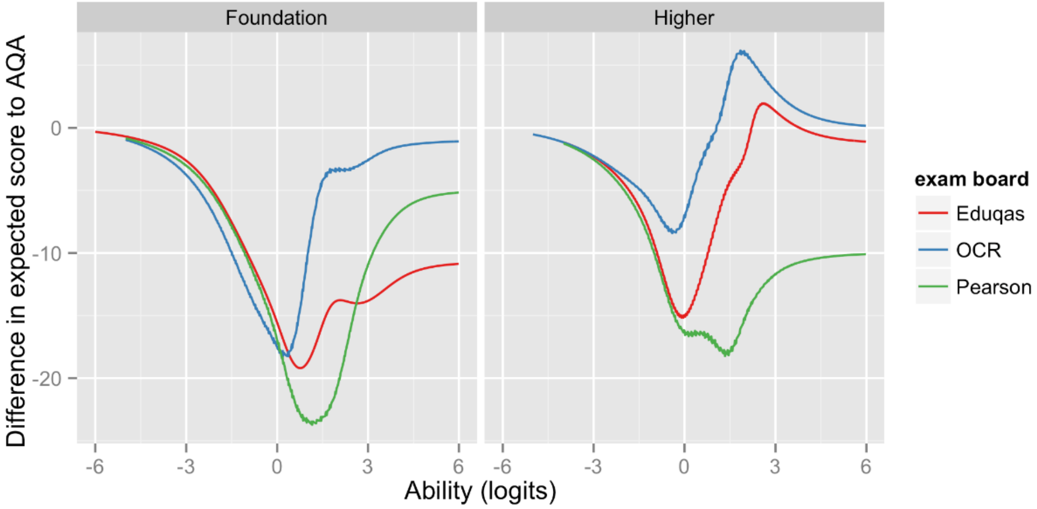


Figure 20: Difference in expected scaled scores compared with AQA

For the foundation tier papers, the AQA paper was considerably easier than the other papers across the full range of ability. For the other three papers, difficulty varied with ability. For students above average ability in the study cohort, the Pearson paper proved more difficult than the other two papers. For the higher tier, the Pearson paper was more difficult than the other papers. The AQA paper was slightly easier than the other papers below the ability of about 1.0 logits but slightly harder than the OCR paper above 1.0.

2.3.9 Test information functions

Figure 21 compares the test information functions between the four papers from each of the two tiers. The test information function provides information about how well the test produces estimates of person abilities over the full range of the ability scale. Large test information would suggest smaller measurement error at specific points on the ability continuum. For all of these papers, the test information is maximised (and the error of measurement minimised) at ability levels higher than the cohort who sat the tests.

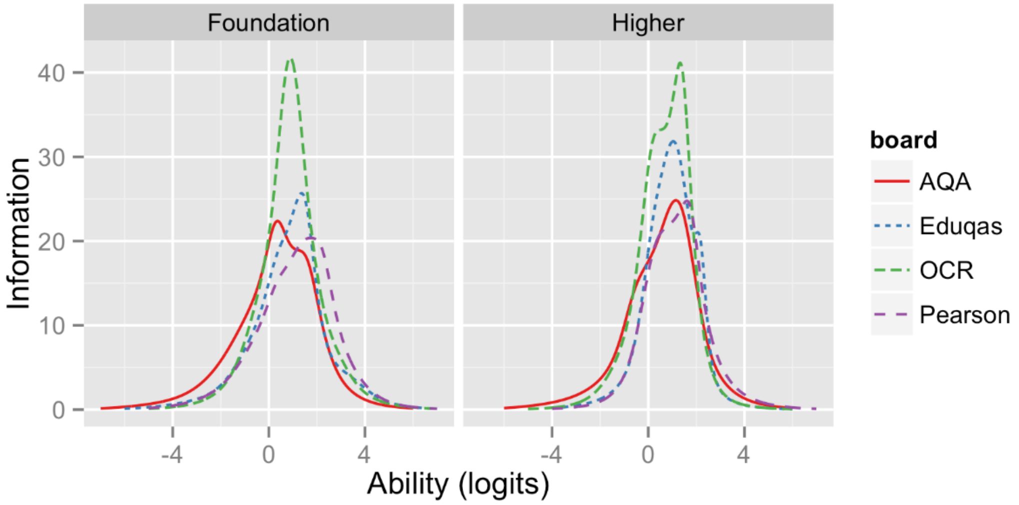


Figure 21: Test information functions

2.3.10 Person ability and item (category) difficulty distributions – Wright Maps

Figures 22 and 23 compare the distribution of category step thresholds of the items and the distribution of person ability for the eight papers (Wright Maps). For each of the tiers, the distribution of person abilities for the four papers was similar, which suggested that the assumption of equivalence in ability between the four groups held reasonably well.

Compared with the raw score distributions (figure 16), the distribution of pupils on the Rasch ability scale is more symmetric as the Rasch scale removes the floor and ceiling effects associated with raw scores. As would be expected, the category thresholds generally increase with item order. For all the papers, the items were generally too difficult relative to the abilities of the students in the study cohort from a measurement perspective.

Figures 24 and 25 compare the distribution of person ability and item difficulty (which was calculated as the mean of the category threshold values) for the four papers from each of the tiers. Since, for each tier, the groups taking the four papers are equivalent in terms of ability distribution and the calibration was centred on persons, the difficulty of the items from different papers can be compared directly. For both tiers, the AQA items tended to be easier than the other three papers.

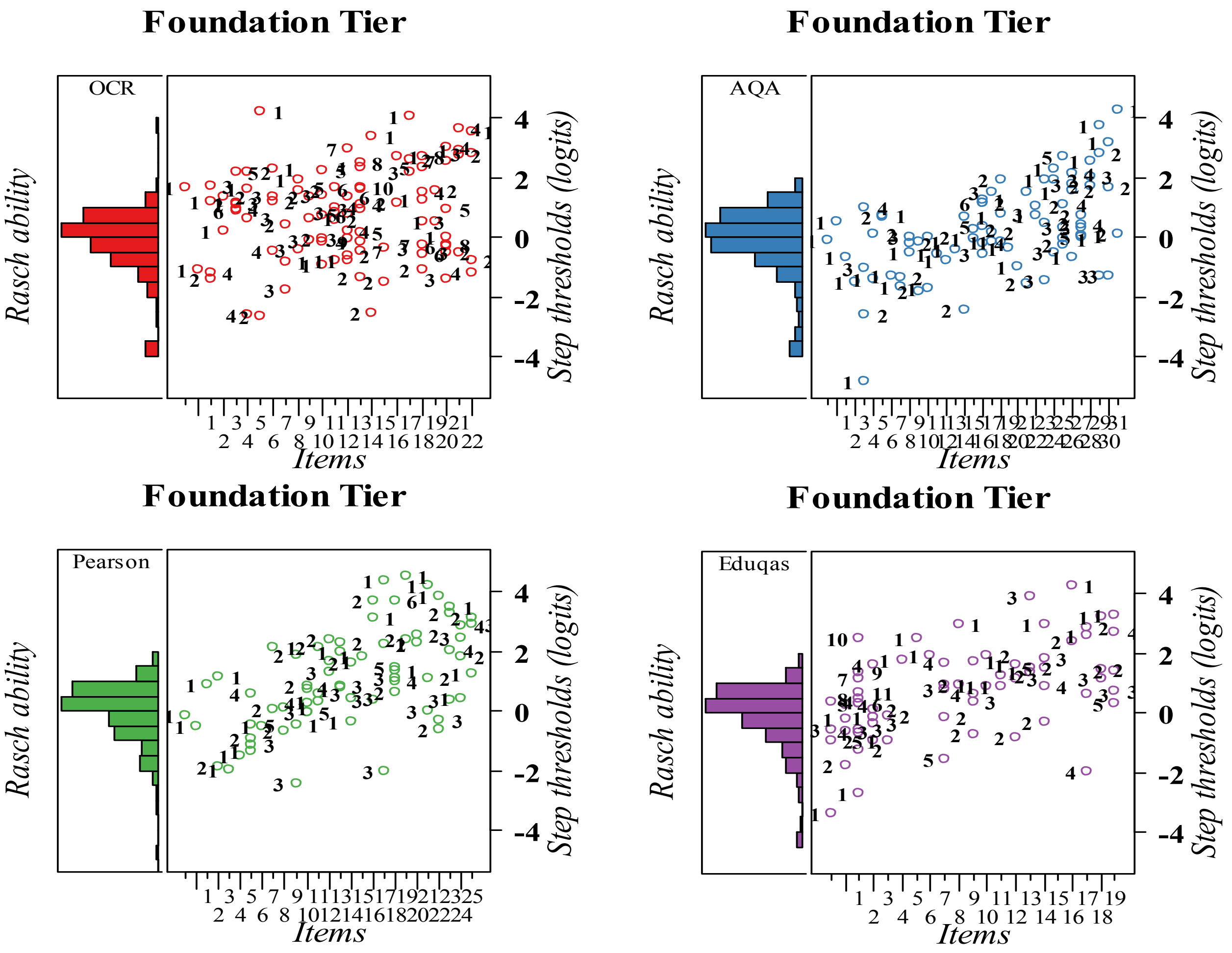


Figure 22: Wright Maps for the foundation tier

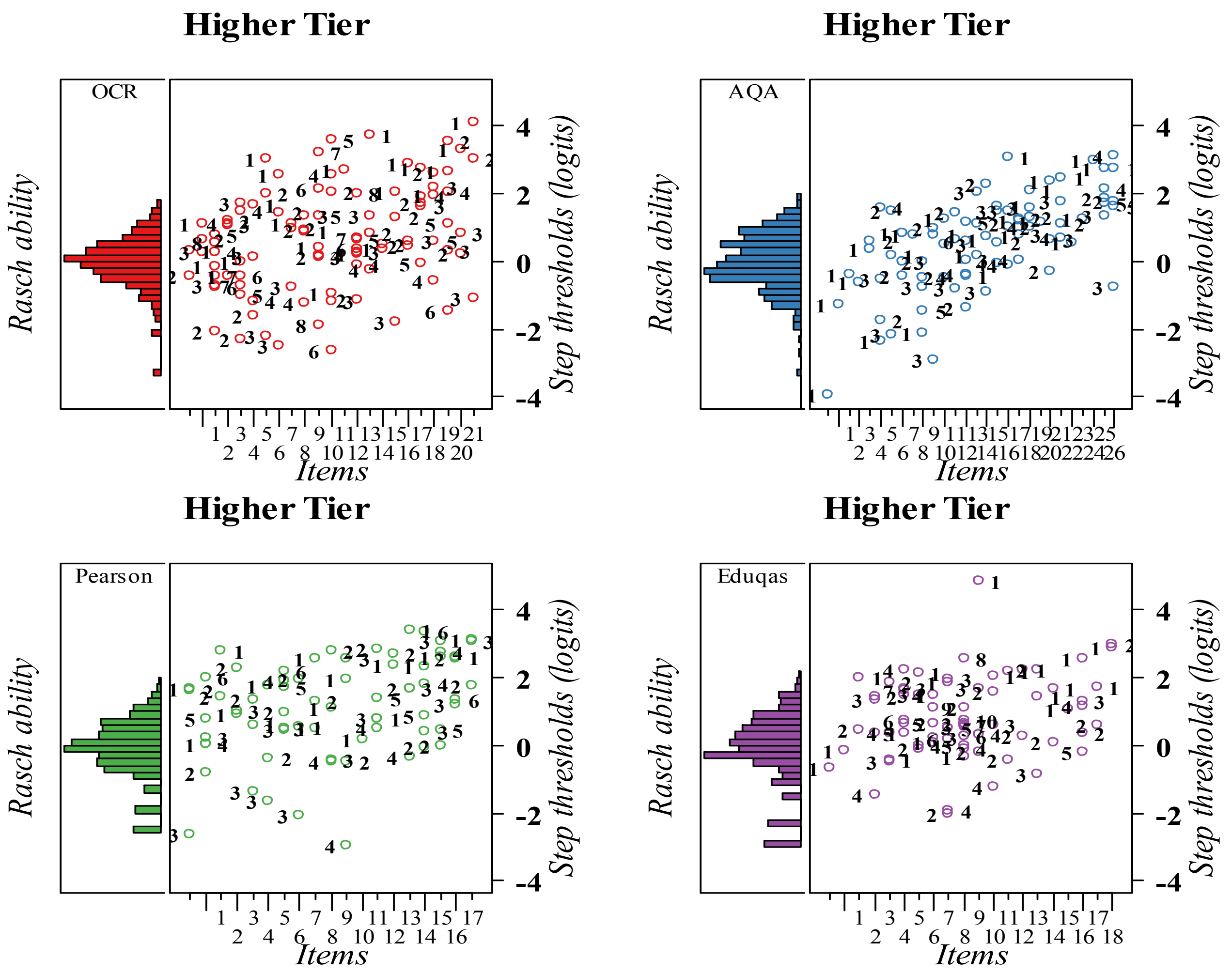


Figure 23: Wright Maps for the higher tier

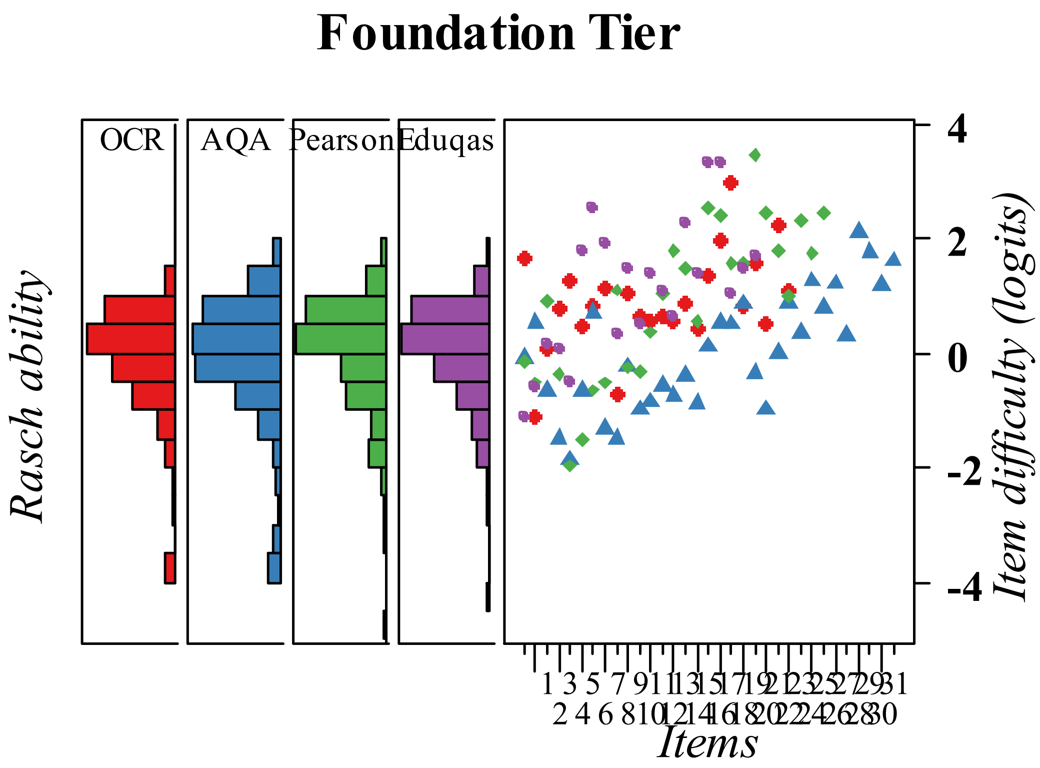


Figure 24: Combined Wright Maps for all papers on the foundation tier

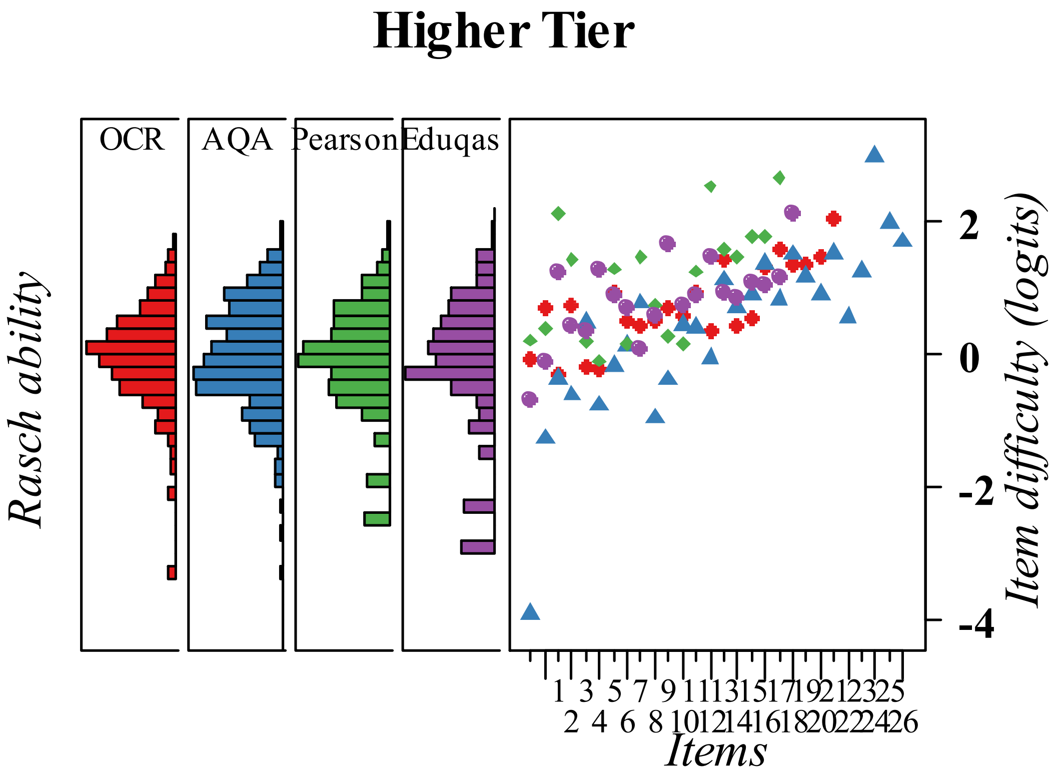


Figure 25: Combined Wright Maps for all papers on the higher tier

2.3.11 Detailed item-by-item analysis

A detailed item-by-item analysis is included in appendix F. While space does not allow discussion of each item, one obvious issue was that the response rate (as opposed to zero scores) by item was very low on certain items, and it fell substantially towards the end of every question paper. On Pearson's higher tier paper, for example, question 4, early on in the paper, had a non-response rate of 21.69 per cent, while Eduqas's foundation tier paper question 14, later on in the paper, had a non-response rate of 79.77 per cent. Generally, it seemed that some items were unfamiliar in format or content, and that either there was not enough time for students to complete the papers or that motivation fell as students progressed through the question papers. The pattern of non-responses by question paper reflects the general pattern of question paper difficulty, but this relationship still does not clarify whether the non-response rate was due to the time required, differing by paper, or the motivation, differing by paper.

Table 11: Non-response rate (weighted by marks)

Exam board	Tier	Non-response rate (weighted by marks)
OCR	F	19.63
AQA	F	14.92
Pearson	F	27.68
Eduqas	F	26.18
OCR	H	22.30
AQA	H	18.82
Pearson	H	27.68
Eduqas	H	22.17

2.3.12 The performance of the highest and lowest performing schools

The poor performance of students on the papers was of concern. It was impossible to know whether this was due to a lack of student motivation, a lack of preparedness for the new content and style of the assessments or because the papers were simply too difficult. To try to gain some insight into this conundrum, a plot of the item facilities of the two highest and two lowest performing schools can be seen in figure 26. The item facilities of the same two schools for all the higher tier booklets and the same two schools for all the foundation tier booklets are shown. The pattern of item facilities for the highest and lowest performing schools mirror each other, and even the students at the highest performing schools found some of the items extremely challenging.

Students at the schools which performed best on each paper scored the following means on the foundation tier papers: OCR 39.5 per cent (n=13), AQA 56.1 per cent (n=20), Pearson 34.0 per cent (n=17) and Eduqas 37.5 per cent (n=32). On the higher tier papers the students scored: OCR 55.9 per cent (n=23), AQA 52.9 per cent (n=23), Pearson 42.9 per cent (n=16) and Eduqas 53.4 per cent (n=8). In sum, even the students at the highest performing schools did not do well. It is interesting to note that for the higher tier, the school that performed best on three out of four of the papers was an independent school that had converted to be an academy with 100 per cent of students achieving A* to C, including English and maths, in 2014.

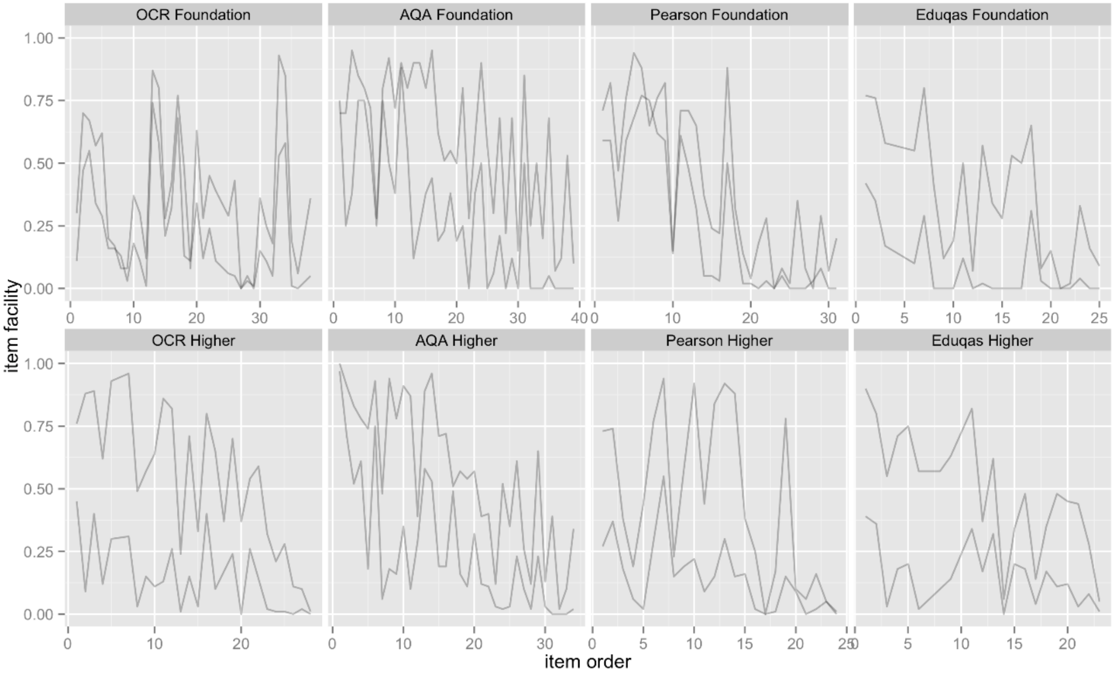


Figure 26: Item facility for the highest and lowest performing schools

2.3.13 Difficulty by assessment objective

Figure 27 shows that, in general, the pattern of difficulty across exam boards followed the trend observed for the whole paper. AQA items tended to be the easiest whatever the assessment objective they were intended to measure. The differences in difficulty between boards were more pronounced on the foundation tier and were greater for AO1 and AO2 than for AO3.

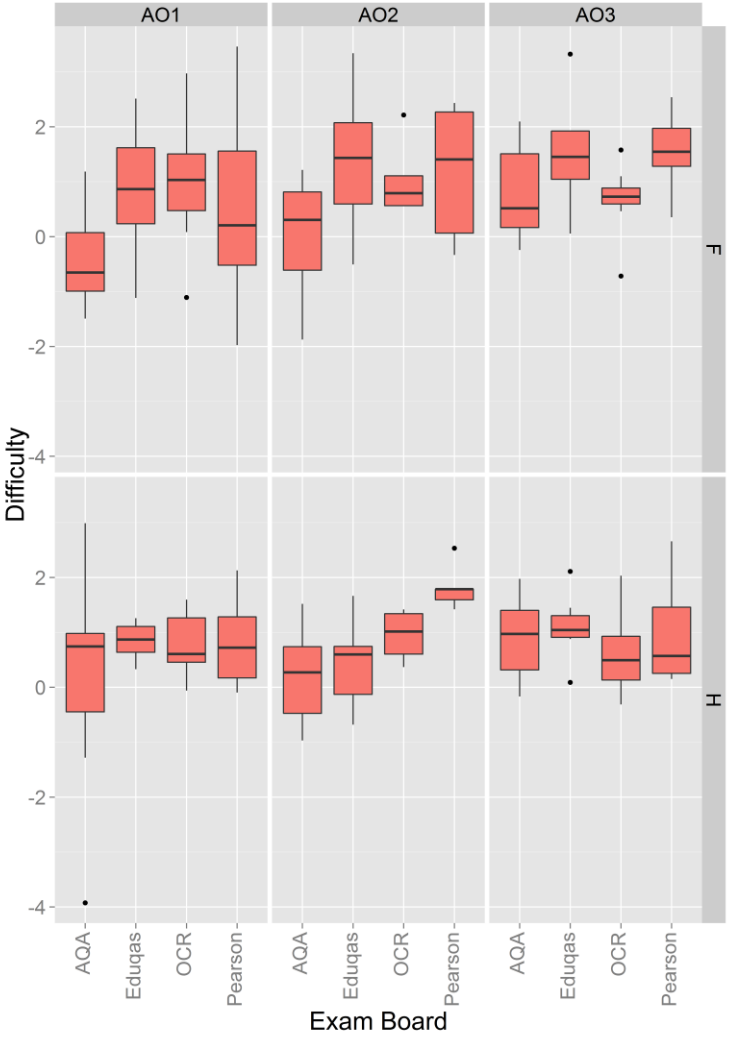


Figure 27: Difficulty by assessment objective

2.3.14 Item expected and actual difficulty relationship

Figure 28 shows that there was a moderately strong correlation[28] between the expected difficulty of the items and the difficulty as experienced by students (r=0.66). The disattenuated correlation, which estimates what the correlation would be if the measurement of expected and actual difficulty had been more precise, was reasonably high (r=0.76).

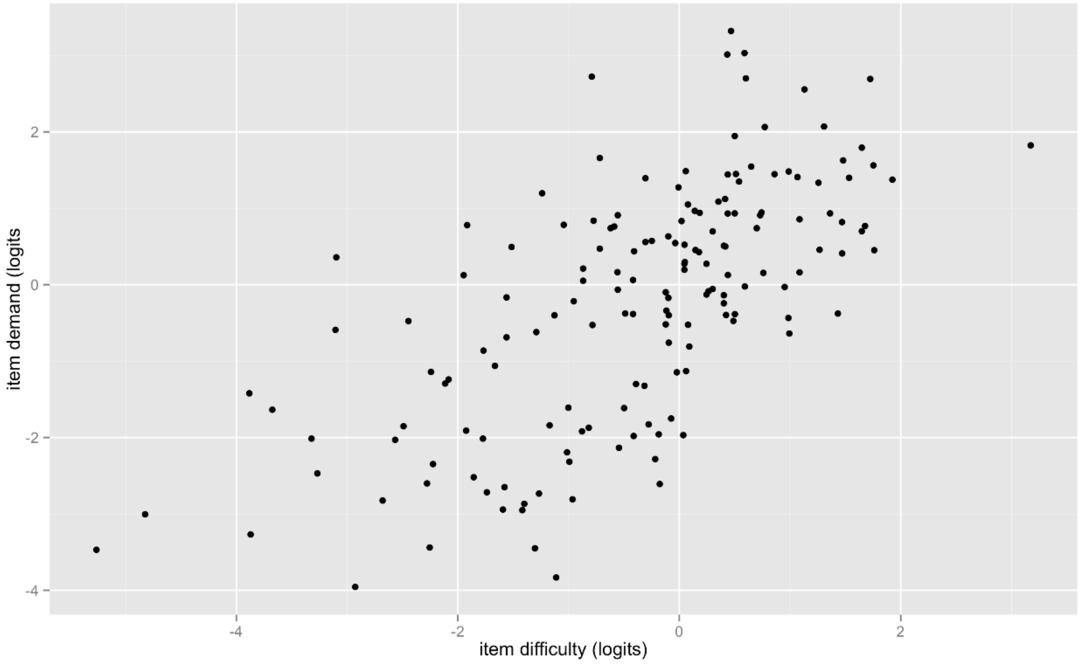


Figure 28: A scatter plot to show the relationship between expected and actual difficulty of items

2.3.15 Residual analysis of the relationship between expected and actual difficulty

Analysis of the residuals of a linear model between expected and actual difficulty revealed no systematic pattern between the independent variable (item difficulty) and the residuals. However, there is a correlation between item order and the residuals

[28] This correlation is between the study 1 difficulty parameters and the Rasch model parameters from study 2. The correlations between the study 1 parameters and study 2 item facility values were 0.56 for foundation tier and 0.68 for higher tier. Unlike the Rasch parameters which can be equated, the facility values for the two tiers cannot be combined to obtain one correlation.

(0.32). On examination of a scatter plot (figure 29), it is clear that foundation tier students found the questions at the start of the question papers more difficult than predicted by the general relationship between expected and actual difficulty.

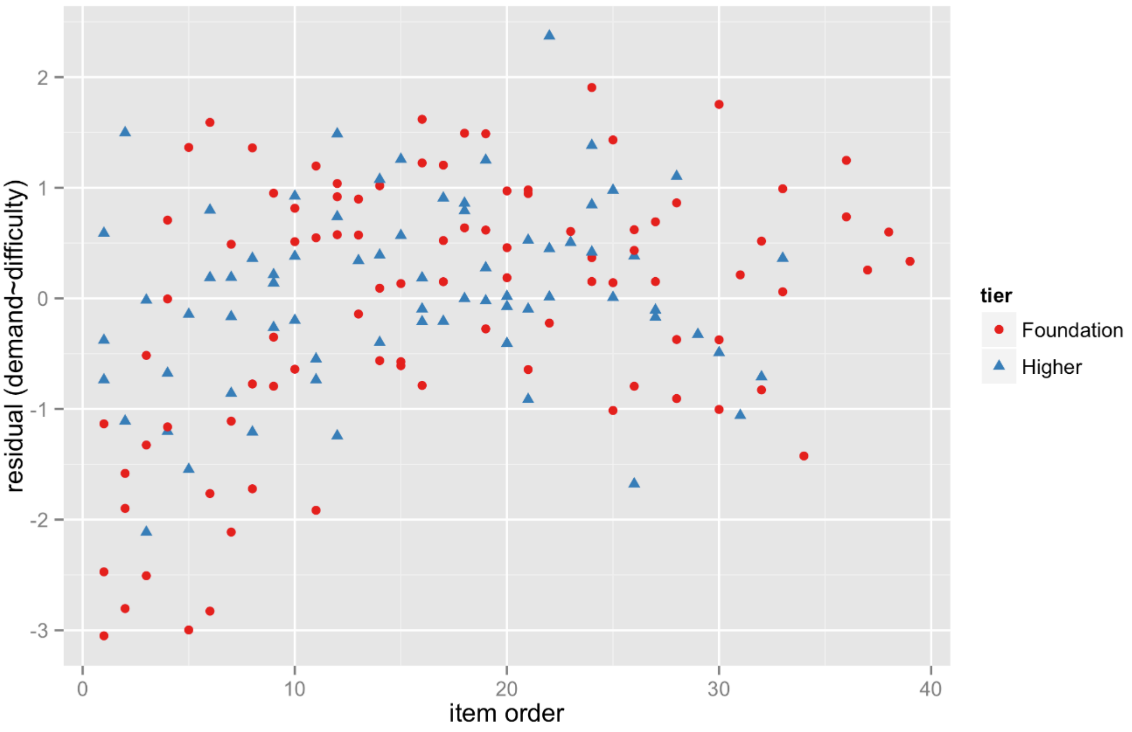


Figure 29: A scatter plot of the expected versus actual difficulty residual by item order

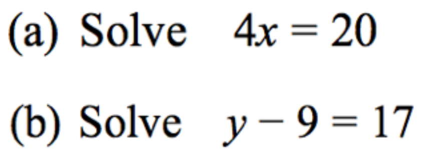
The item with the largest residual, for example, was the first question on OCR's foundation tier paper (figure 30). This proved relatively difficult for the students (0.36 logits) but was judged to be of low expected difficulty (-3.10 logits).

1 **(a)** Work out.

$$3 + 2 \times (3 - 1)$$

Figure 30: An item that proved to be more difficult than expected

The items with the second and third largest residuals formed the second question on Pearson's foundation tier paper (figure 31). These items were relatively easy for most students (-3.47 logits and -3.00 logits) but were judged to be among the lowest expected difficulty questions (-5.26 logits and -4.82 logits) (i.e. slightly more difficult than the expected difficulty would predict).



(a) Solve $4x = 20$

(b) Solve $y - 9 = 17$

Figure 31: Questions with among the lowest expected difficulty

There were no obvious patterns in the items with the highest positive residuals, with the outlying item a higher tier Pearson item that was a little less difficult than predicted by the linear relationship with expected difficulty.

2.3.16 Analysis of item word count

Analyses were performed to investigate the potential impact of the amount of reading students were required to do in order to respond to each item. These analyses were based on the number of words contained within an item rather than a reflection on the complexity of the language used. The word count also took into account the presence of a common stem relevant to different item parts. For example, where both parts a) and b) of an item used the same common section of text to outline the information required, this common section was included in the word count for both item a) and item b). Mathematical expressions were counted as a single word.

Figure 32 shows that there is a statistically significant relationship between the number of words in an item and the actual difficulty of the items. This does not, however, necessarily indicate the presence of construct irrelevance.

To consider whether the number of words in an item had a systematic impact on the relationship between the expected difficulty and the actual difficulty, the relationship between the item residuals and number of words was analysed. This relationship is shown in figure 33 broken down by exam board. It is apparent from figures 32 and 33 that there is a lower word count in the AQA items (mean = 22.8) relative to the OCR (mean = 48.3) and Pearson (mean = 38.2) items. Only OCR had a statistically significant relationship (F(1,55) = 5.76, p = 0.020) with a slight tendency for items with a higher word count to have a higher actual difficulty than expected:
(residual = -0.12×word count + 0.35).

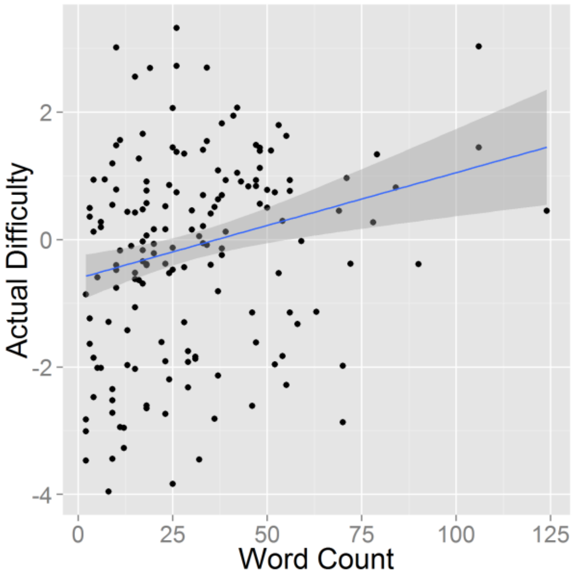


Figure 32: Relationship between actual item difficulty and word count

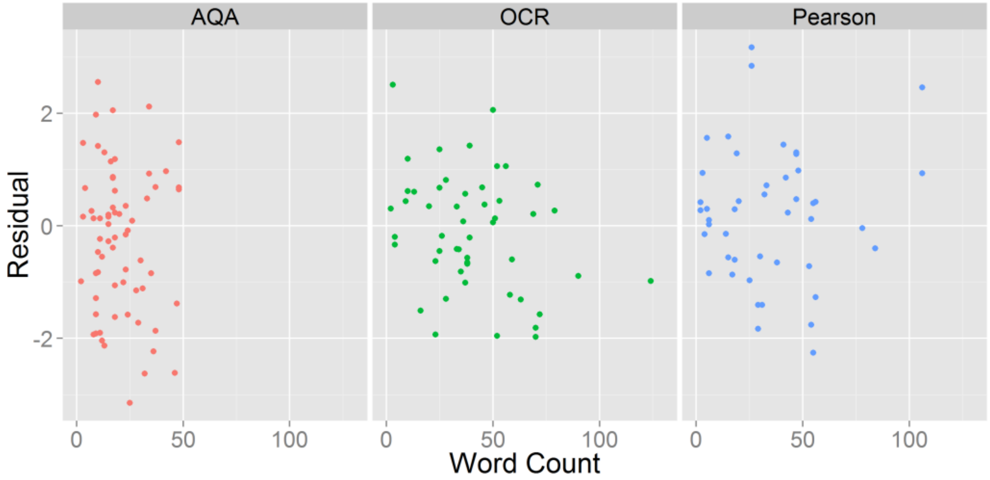


Figure 33: Relationship between expected difficulty vs actual difficulty residual by word count for each exam board.

2.4 Summary of findings

Marking reliability was adequate and there was no evidence of bias in the marking. The assessments had good internal reliability, although they were more difficult than ideal given the ability of the students sitting them and as such they did not function optimally as measurement instruments. It is highly questionable whether the level of difficulty would be appropriate if students were fully prepared (i.e. taught the content of the specifications) and motivated.

Students performed better on AQA's papers, on both items testing old and new content. On the foundation tier, the AQA paper was easier than other papers across the full range of student ability. On the higher tier, the Pearson paper was considerably more difficult than the other papers. The AQA higher tier paper was slightly easier for students whose ability was less than one standard deviation above average. But for students whose ability was greater than one standard deviation above the average, the AQA paper was slightly harder than OCR's.

There was a high non-response rate, which worsened as the students progressed through the papers. Even the students at the highest performing schools found some of the items extremely challenging. Even on the old content the average scores were lower than one would expect on a real examination.

The distributions of marks showed that, with the possible exception of the AQA foundation tier paper, the assessments failed to sufficiently differentiate between students of differing levels of ability. It would not be feasible to reliably grade students on the basis of these assessments. This may, however, be due to a lack of student motivation and preparation. It is impossible to disentangle the effects of motivation and difficulty on students' scores. However, an estimate of the pre-test effect on students' scores on the current content suggests that the papers would require lower than usual grade boundaries, especially on the higher tier.

The basis upon which the PhD students (in study 1) made their judgements as to the relative expected difficulty of items is unknown. However, their judgement of the difficulty of items proved to be a surprisingly good predictor of the actual difficulty experienced by students. This supports both the use of PhD students and the comparative judgement methodology in work of this kind.

3. Study 3 – A comparison across exam boards of the extent to which items are judged as eliciting problem solving as defined by AO3

3.1 Design

The assessment of problem solving has increased in the reformed maths GCSEs relative to the existing GCSEs. Discourse around the sample assessment materials produced by the exam boards suggested that there was variation in the way in which problem solving had been operationalised which would affect the difficulty of the papers.

This study involved GCSE maths teachers judging the degree to which items elicited mathematical problem-solving abilities as described in assessment objective 3 (AO3, see below). This exercise was conducted for 33 items that predominantly assessed AO3. The allocation of items to assessment objective was done by the exam boards and had been scrutinised as part of the accreditation process.

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.

To support the teachers' judgements, authentic exemplar responses with descriptions of the students' thinking were presented alongside the items. These were obtained from very able Year 11 students.

To help validate the basis upon which judgements were made, four 'authentication' AO1/2 items (one from each exam board) were also included. These authentication items were selected to appear superficially similar to AO3 items and so were taken from the higher tier only.

The comparative-judgement framework from study 1 was again used. This exact kind of judgement has not been made in previous studies but comparative judgement of problem solving has previously been shown to produce robust data.[29]

3.2 Method

3.2.1 Item selection

All items from the sample assessments for the four exam boards were considered. Using the assignment of marks to each assessment objective produced by the exam boards, 33 items were selected,[30] which included four or more AO3 marks. The items are summarised in table 12.

Table 12: Summary of items used in study 3

Exam board	Number of items	Foundation /Common/ Higher	Minimum AO3 mark	Maximum AO3 mark	Minimum total mark	Maximum total mark
AQA	9	2 / 1 / 6	4	5	4	8
Pearson	8	1 / 2 / 5	4	5	5	9
OCR	9	3 / 1 / 5	4	7	5	10
Eduqas	7	1 / 2 / 4	4	6	5	9

Four authentication AO1/2 higher tier items were also included – one from each exam board.

3.2.2 Materials

Authentic model responses to the 33 AO3 items and the four AO1/2 items were produced by high-achieving Year 11 GCSE maths students. The top set maths class from two schools participated, with approximately 20 and 30 students in the classes. The aim was to capture the best possible mathematical problem solving that the items could elicit.

The students were asked to give as full a description of their thinking as they could. They were asked to explain each step, stating what they were doing and why, and

[29] Jones, I., Swan, M. and Pollitt, A. (2014) Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, 13(1), 151–177.

[30] An ‘item’ included all parts of a numbered item. In all items selected, the parts followed on from one another, rather than being unrelated items grouped together.

they were discouraged from giving their answer with only mathematical workings. They were asked to imagine that they were describing how the problem should be solved to a younger sibling or a less-able student, who would not necessarily be able to follow their workings alone.

The students were also asked to explain different approaches to solving the problem if they could see more than one (in practice this rarely occurred, students were usually satisfied with just one correct approach and were resistant to working through a second). They were encouraged to take their time producing their response and to produce the clearest, most detailed explanation of their thinking possible.

The students worked in pairs, to produce richer responses through discussion and also to increase the probability of the correct answer being reached. Each pair worked on a different item, with the items distributed randomly. Once a pair had completed their item, they were given another randomly selected item. This continued until the class ended.

The aim was to collect two exemplar responses for each item. In total, 84 responses were collected. However, 18 were judged to be incorrect or incomplete, resulting in 66 usable responses. Therefore, eight items had only one exemplar response (three items from Pearson, two each from OCR and Eduqas, and one from AQA). Where possible, incorrect working where the students had made more than one attempt to tackle the problem was removed. Exemplar responses were then uploaded to the online CJ platform No More Marking.[31]

3.2.3 Participants

Thirty-three GCSE maths teachers were recruited as judges from the schools and markers that participated in study 2. Each judge was given 50 judgements to make. They were informed that the purpose of the study was to judge the relative merit of items in eliciting problem solving as defined by AO3.

3.2.4 Procedure

Judges were provided with the definition of AO3 (as shown earlier) and asked to think about what they would expect of items assessing these skills. They were told that the exemplar responses were there to help them consider the thinking the item required of students. However, they were to judge the problem solving elicited by the item, not the quality or correctness of the response.

[31] Wheadon, C. and Jones, I. (2014, June 1) Online Comparative Judgement. Retrieved April 21, 2015, from www.nomoremarking.com

Pairs of items were presented side by side online, and the judges were prompted to indicate:

'Which item best elicits mathematical problem solving as described by AO3?'

Judges were free to carry out the 50 judgements however they preferred and exactly how they interpreted 'best' is unknown. The only restriction was a date by which to complete all judgements.

3.3 Results

3.3.1 Analysis

The node package, Comparative-Judgement,[32] was used to estimate all model parameters for this study. The foundation and higher tier items were analysed together.

3.3.2. Reliability of the judging

The inter-rater reliability of the judging was 0.66, while the Rasch separation reliability was 0.83. While we do not know the cognitive processes underlying the judgements, a reliable scale was established. That said, these reliability figures are lower than usual and suggest that there was more disagreement among the judges for this study than for study 1. Six out of the 33 judges had an outfit greater than 1.2 (the worst being 1.45), but given the number of judges and the fairly modest outfit all judges were included in the analysis. Only one of the item exemplars had a fit worse than 1.2 – an authentication AO1/2 item exemplar from AQA's higher tier paper. While most judges rated the response as the lowest on the scale, several of the judges rated the response as more indicative of AO3 than items above average on the scale. The extent of the disparity of the judgement is indicative of the difficulty judges can have in reaching consensus in this domain.

3.3.3. AO3 parameter estimates

The AO3 parameter estimates are summarised in figure 34 (and reported in tables 13 and 14). The authentication items testing AO1/2 appeared, with one exception (reproduced in figure 35) to be judged lower in their AO3 qualities than the genuine AO3 items (see table 15). On the whole, therefore, it is likely that the judges were making judgements on the AO3 quality of the items rather than any other quality. While the range between the parameters of the different exemplars of the items varied between 0 and 1.25, the mean difference between exemplars was 0.35, which

[32] Wheadon, C. (2014, Sept) Comparative Judgement Algorithms. Retrieved April 21, 2015, from www.npmjs.com/package/comparative-judgement

is only slightly higher than the mean standard error of the parameters (0.32). The small difference between exemplars suggests that judges were mainly judging the item rather than the response, although the response did to some extent influence their judgements.

Differences between exam boards were more substantial on the foundation tier papers than they were on the higher tier papers. Indeed the differences were significant for the foundation tier (F(3,20)=5.903, p=0.005, η2 =0.470, power=0.389) but not the higher tier (F(3,42)=0.512, p=0.677, η2 =0.035, power=0.053). The Eduqas AO3 foundation tier items were considered to be better at eliciting mathematical problem solving as defined by AO3 than the AO3 foundation tier items from the other exam boards (p≤0.023, although Eduqas items were no longer significantly different from Pearson items once a Bonferroni correction for multiple comparisons was applied). Indeed, the second exemplar of the Eduqas item 17 was judged to be the item that best elicited an AO3 response for both higher and foundation tiers. The response is reproduced in figure 36. It should be noted that Eduqas and Pearson had chosen the majority of their largely AO3 foundation tier items to be common to both tiers.

The AQA AO3 foundation tier items, on the other hand, were more varied. The common item between tiers was highly rated on its AO3 quality, while the foundation tier only items were rated in line with the AO1/2 authentication items. The judges believed that these two foundation tier only items were relatively poor in eliciting AO3 responses. The item with the lowest parameter estimate from these items is reproduced in figure 37. On average, the AQA foundation tier items were judged to be worse than other boards’ items in eliciting AO3.

On the higher tier papers the median AO3 parameter values were far more in line across exam boards. Once again, however, there were a number of items that were judged to be quite poor at eliciting AO3 responses as they fell below the parameter estimates of two of the AO1/2 items. The items with the highest AO3 parameter value (Pearson paper 3, question 12) and the lowest parameter value (OCR paper 3, question 17) are reproduced in figures 38 and 39.

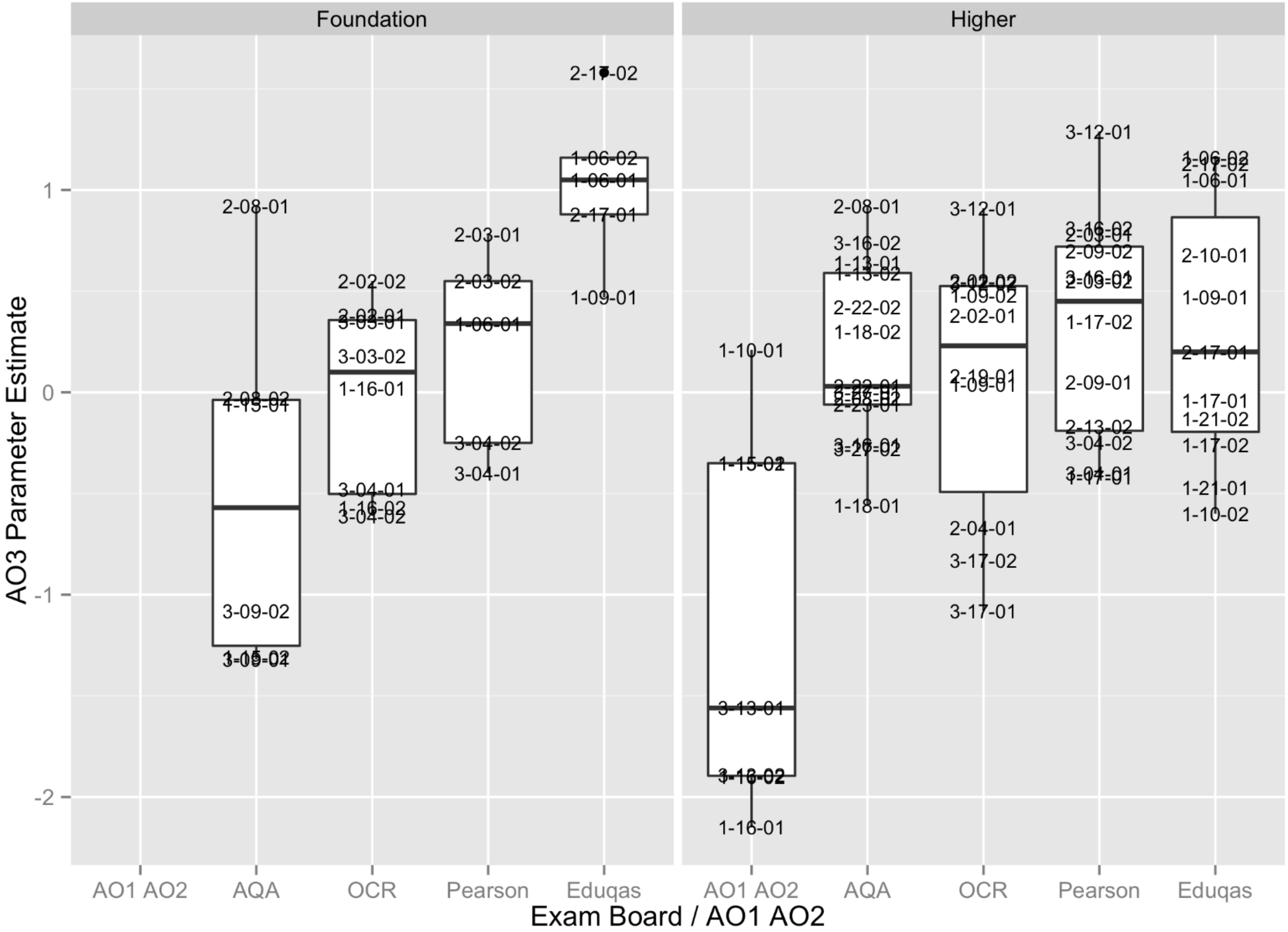


Figure 34: AO3 parameter estimate by question paper; items common to both tiers are included in both plots

Table 13: Foundation tier AO3 parameter estimates

Exam board	Question paper	Item – exemplar	AO3 parameter estimate	AO3 parameter estimate se
AQA	1	15–01	-0.06	0.29
	1	15–02	-1.31	0.37
	2	08–01	0.92	0.33
	2	08–02	-0.03	0.30
	3	09–01	-1.32	0.36
	3	09–02	-1.08	0.34
OCR	1	16–01	0.02	0.30
	1	16–02	-0.57	0.32
	2	02–01	0.38	0.31
	2	02–02	0.55	0.31
	3	03–01	0.35	0.30
	3	03–02	0.18	0.30
	3	04–01	-0.48	0.31
	3	04–02	-0.61	0.31
Pearson	1	06–01	0.34	0.31
	2	03–01	0.78	0.31
	2	03–02	0.55	0.31
	3	04–01	-0.40	0.32
	3	04–02	-0.25	0.30
Eduqas	1	06–01	1.05	0.32
	1	06–02	1.16	0.34
	1	09–01	0.47	0.32
	2	17–01	0.88	0.31
	2	17–02	1.58	0.36

Items that are common across tiers are indicated by shading.

Table 14: Higher tier AO3 parameter estimates

Exam board	Question paper	Item – exemplar	AO3 parameter estimate	AO3 parameter estimate se
AQA	1	13–01	0.64	0.31
	1	13–02	0.59	0.31
	1	18–01	-0.56	0.31
	1	18–02	0.30	0.30
	2	08–01	0.92	0.33
	2	08–02	-0.03	0.30
	2	22–01	0.03	0.29
	2	22–02	0.42	0.30
	2	23–01	-0.06	0.30
	3	16–01	-0.26	0.30
	3	16–02	0.74	0.32
	3	27–01	0.00	0.30
	3	27–02	-0.28	0.31
OCR	1	09–01	0.04	0.31
	1	09–02	0.48	0.30
	2	02–01	0.38	0.31
	2	02–02	0.55	0.31
	2	04–01	-0.67	0.31
	2	19–01	0.08	0.30
	3	12–01	0.91	0.32
	3	12–02	0.54	0.31
	3	17–01	-1.08	0.33
	3	17–02	-0.83	0.32
Pearson	1	17–01	-0.42	0.31
	1	17–02	0.35	0.31
	2	03–01	0.78	0.31
	2	03–02	0.55	0.31
	2	09–01	0.05	0.32
	2	09–02	0.70	0.30
	2	13–02	-0.17	0.29
	3	04–01	-0.40	0.32
	3	04–02	-0.25	0.30
	3	12–01	1.29	0.34
	3	16–01	0.57	0.30
	3	16–02	0.81	0.32
Eduqas	1	06–01	1.05	0.32
	1	06–02	1.16	0.34

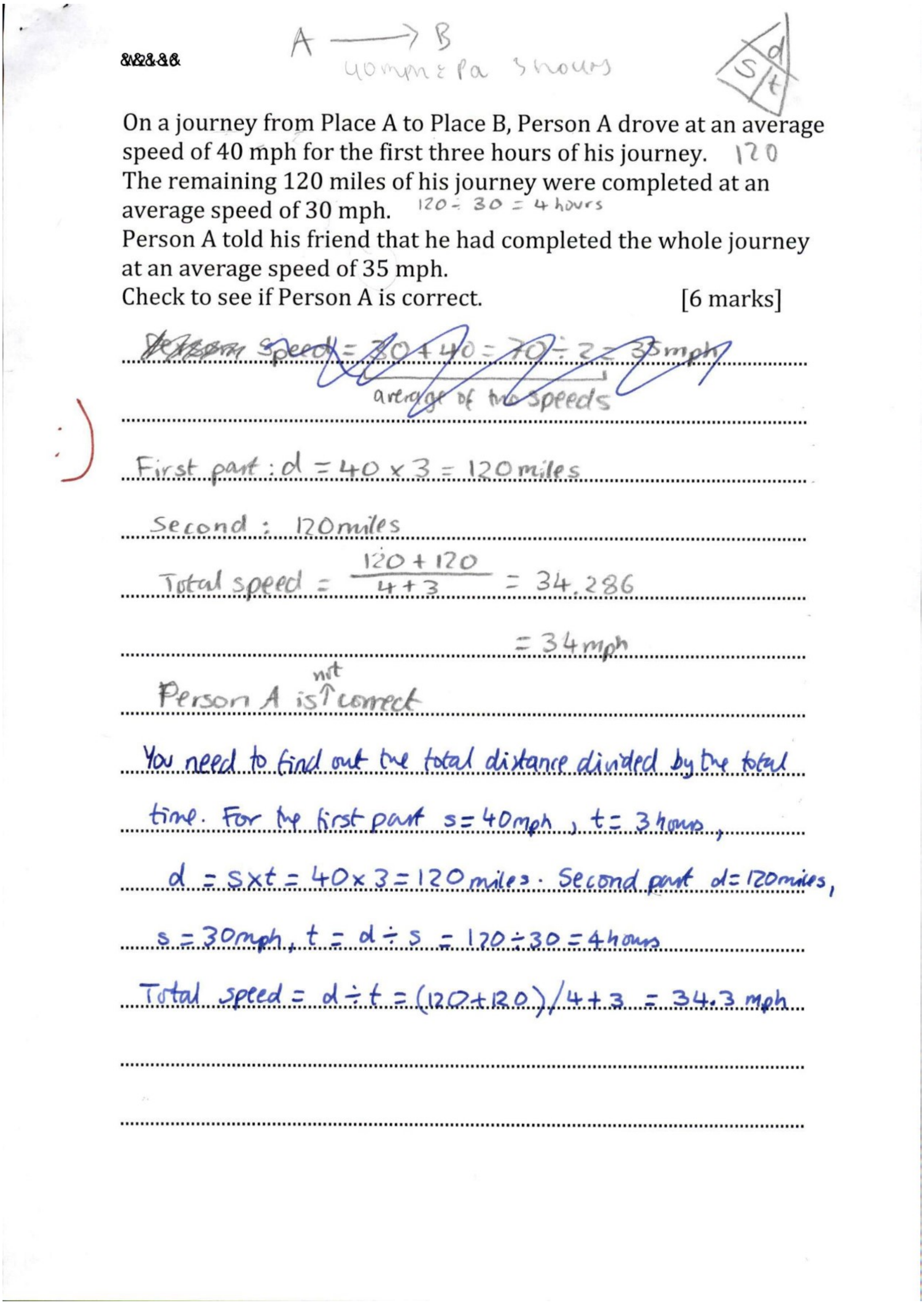
Exam board	Question paper	Item – exemplar	AO3 parameter estimate	AO3 parameter estimate se
	1	09–01	0.47	0.32
	1	10–02	-0.60	0.32
	1	17–01	-0.04	0.31
	1	17–02	-0.26	0.31
	1	21–01	-0.47	0.31
	1	21–02	-0.13	0.29
	2	10–01	0.68	0.31
	2	17–01	0.20	0.30
	2	17–02	1.13	0.34

Items that are common across tiers are indicated by shading.

Table 15: AO3 parameter estimates for AO1/2 items

Exam board	Tier	Question paper	Item–exemplar	AO3 parameter estimate	AO3 parameter estimate se
AQA	Higher	1	16–01	-2.15	0.45
	Higher	1	16–02	-1.90	0.41
OCR	Higher	3	13–01	-1.56	0.37
	Higher	3	13–02	-1.89	0.40
Pearson	Higher	1	15–01	-0.35	0.30
	Higher	1	15–02	-0.35	0.30
Eduqas	Higher	1	10–01	0.21	0.30

Figure 35: AO1/2 item with a high AO3 parameter value (Eduqas, higher tier, paper 1, question 10)



&&&&&

A → B

40mph for 3hours

On a journey from Place A to Place B, Person A drove at an average speed of 40 mph for the first three hours of his journey. 120

The remaining 120 miles of his journey were completed at an average speed of 30 mph. 120 ÷ 30 = 4 hours

Person A told his friend that he had completed the whole journey at an average speed of 35 mph.

Check to see if Person A is correct. [6 marks]

~~Person speed = 30 + 40 = 70 ÷ 2 = 35mph~~

~~average of the speeds~~

First part : d = 40 x 3 = 120 miles

Second : 120miles

Total speed = $\frac{120+120}{4+3}$ = 34.286

= 34mph

Person A is not correct

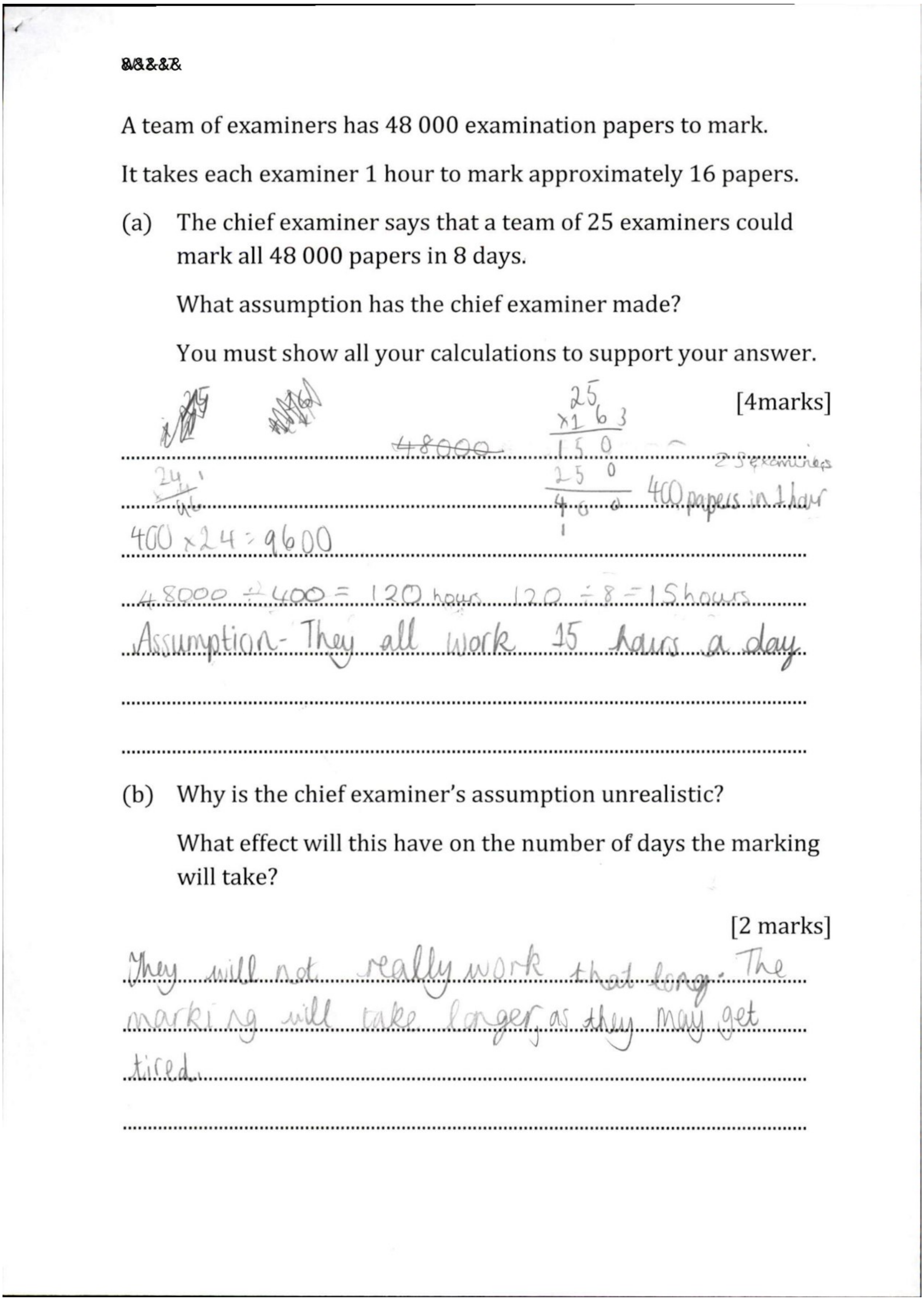
You need to find out the total distance divided by the total time. For the first part s = 40mph , t = 3 hours ,

d = s x t = 40 x 3 = 120 miles. Second part d = 120miles,

s = 30mph, t = d ÷ s = 120 ÷ 30 = 4 hours

Total speed = d ÷ t = (120+120)/4+3 = 34.3 mph

Figure 36: Item with the highest AO3 parameter on the foundation tier (Eduqas paper 2, question 17)



&&&&&

A team of examiners has 48 000 examination papers to mark.

It takes each examiner 1 hour to mark approximately 16 papers.

(a) The chief examiner says that a team of 25 examiners could mark all 48 000 papers in 8 days.

What assumption has the chief examiner made?

You must show all your calculations to support your answer.

[4marks]

~~48000~~

25
×16
150
250
400

25 examiners
400 papers in 1 hour

24
×4
96

400 × 24 = 9600

48000 ÷ 400 = 120 hours 120 ÷ 8 = 15 hours

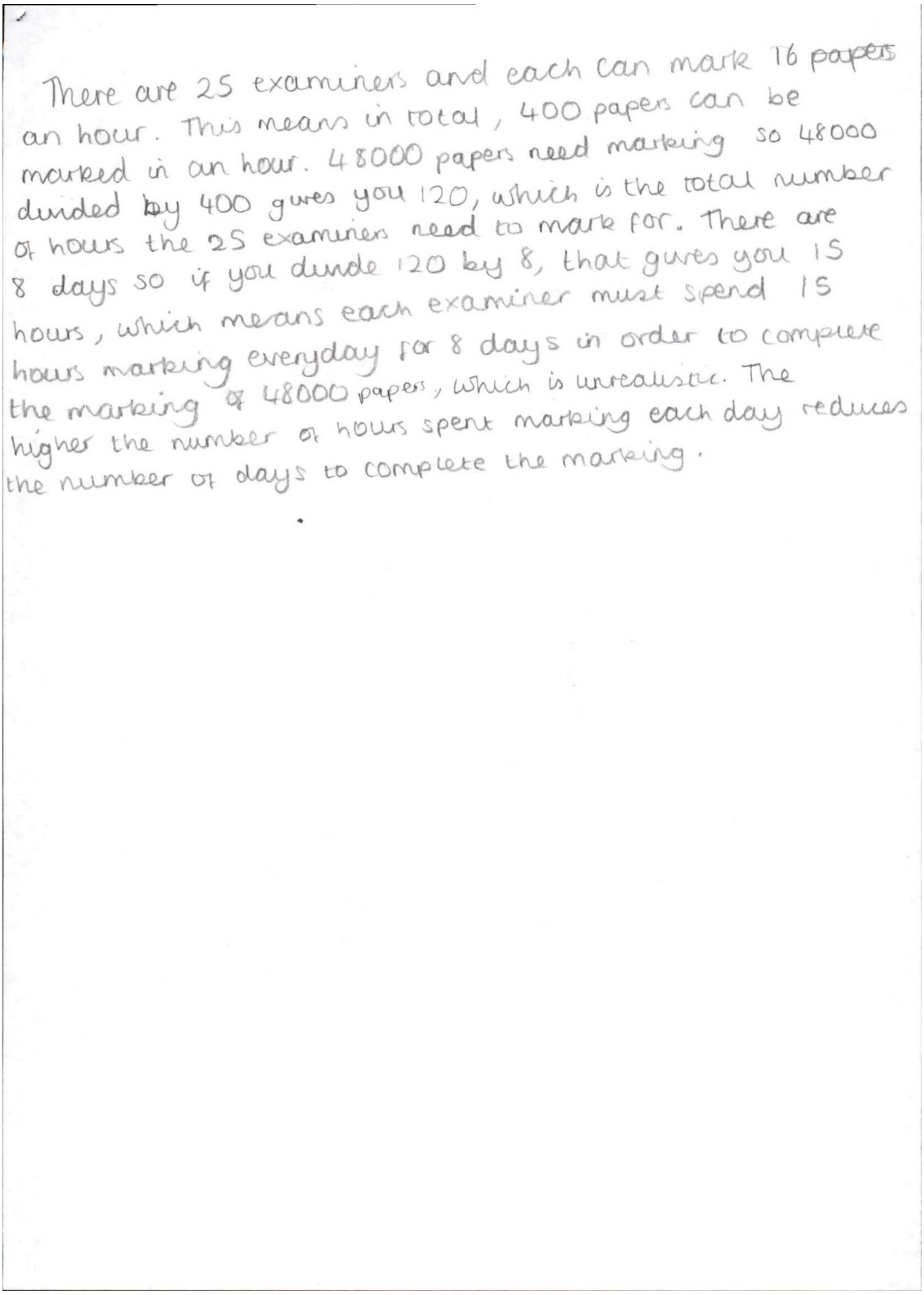
Assumption - They all work 15 hours a day

(b) Why is the chief examiner's assumption unrealistic?

What effect will this have on the number of days the marking will take?

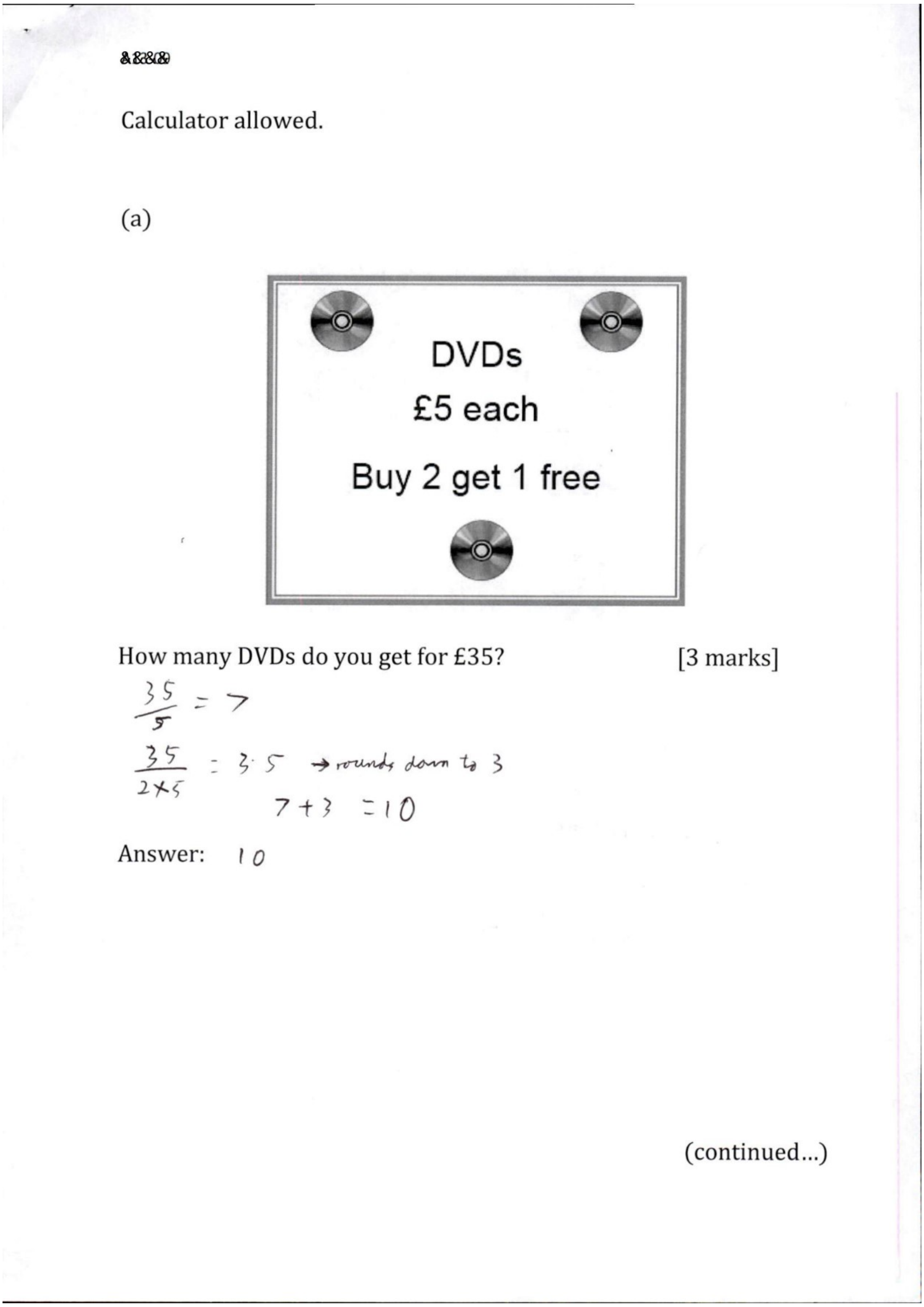
[2 marks]

They will not really work that long. The marking will take longer, as they may get tired.



There are 25 examiners and each can mark 16 papers an hour. This means in total, 400 papers can be marked in an hour. 48000 papers need marking so 48000 divided by 400 gives you 120, which is the total number of hours the 25 examiners need to mark for. There are 8 days so if you divide 120 by 8, that gives you 15 hours, which means each examiner must spend 15 hours marking everyday for 8 days in order to complete the marking of 48000 papers, which is unrealistic. The higher the number of hours spent marking each day reduces the number of days to complete the marking.

Figure 37: Item with the lowest AO3 parameter value on the foundation tier (AQA paper 3, question 9)



&8&&&

Calculator allowed.

(a)

DVDs

£5 each

Buy 2 get 1 free

How many DVDs do you get for £35? [3 marks]

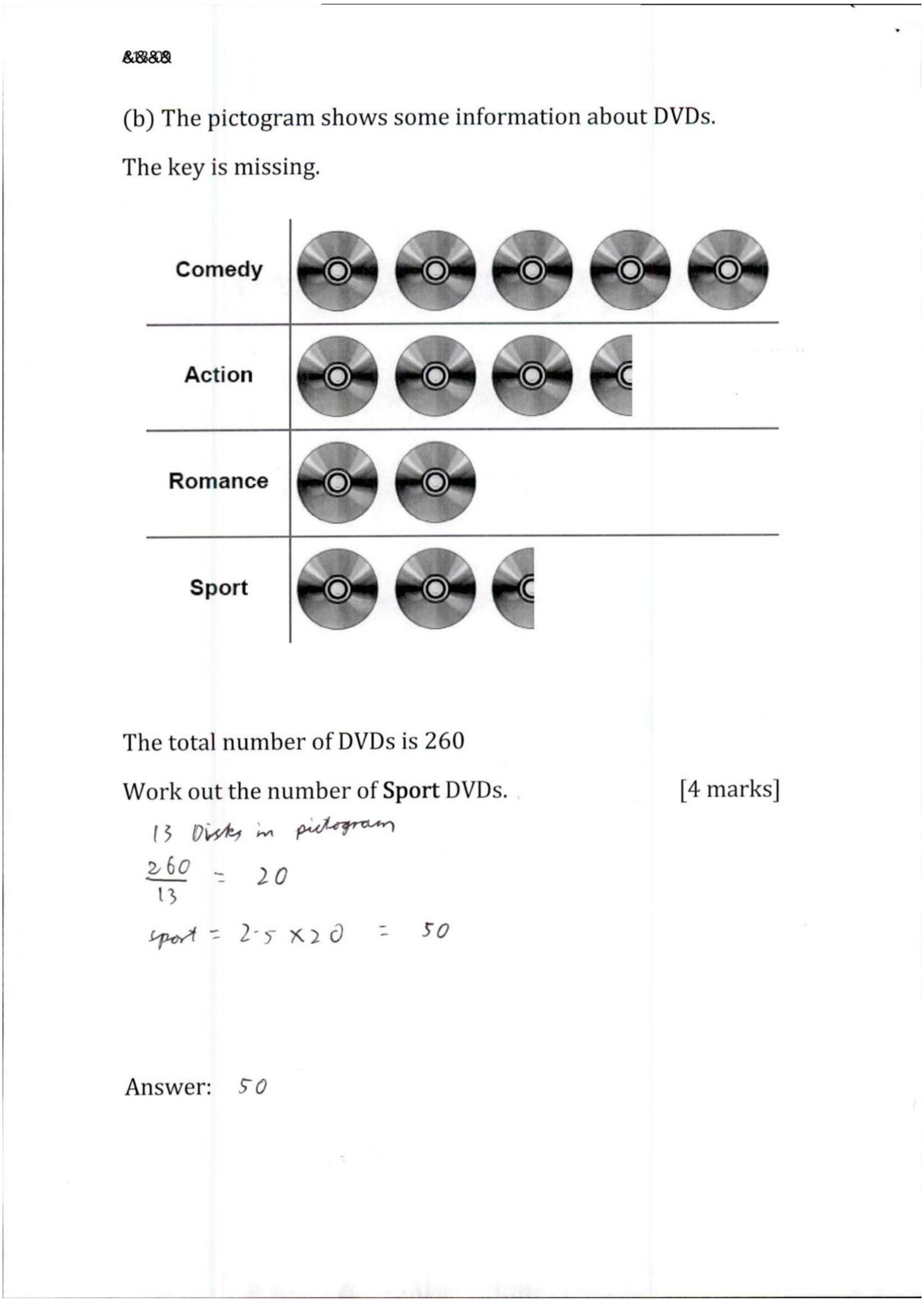
$\frac{35}{5} = 7$

$\frac{35}{2 \times 5} = 3.5$ → rounds down to 3

$7 + 3 = 10$

Answer: 10

(continued...)



&&&&

(b) The pictogram shows some information about DVDs.

The key is missing.

Comedy	●●●●●
Action	●●●◐
Romance	●●
Sport	●●◐

The total number of DVDs is 260

Work out the number of **Sport** DVDs. [4 marks]

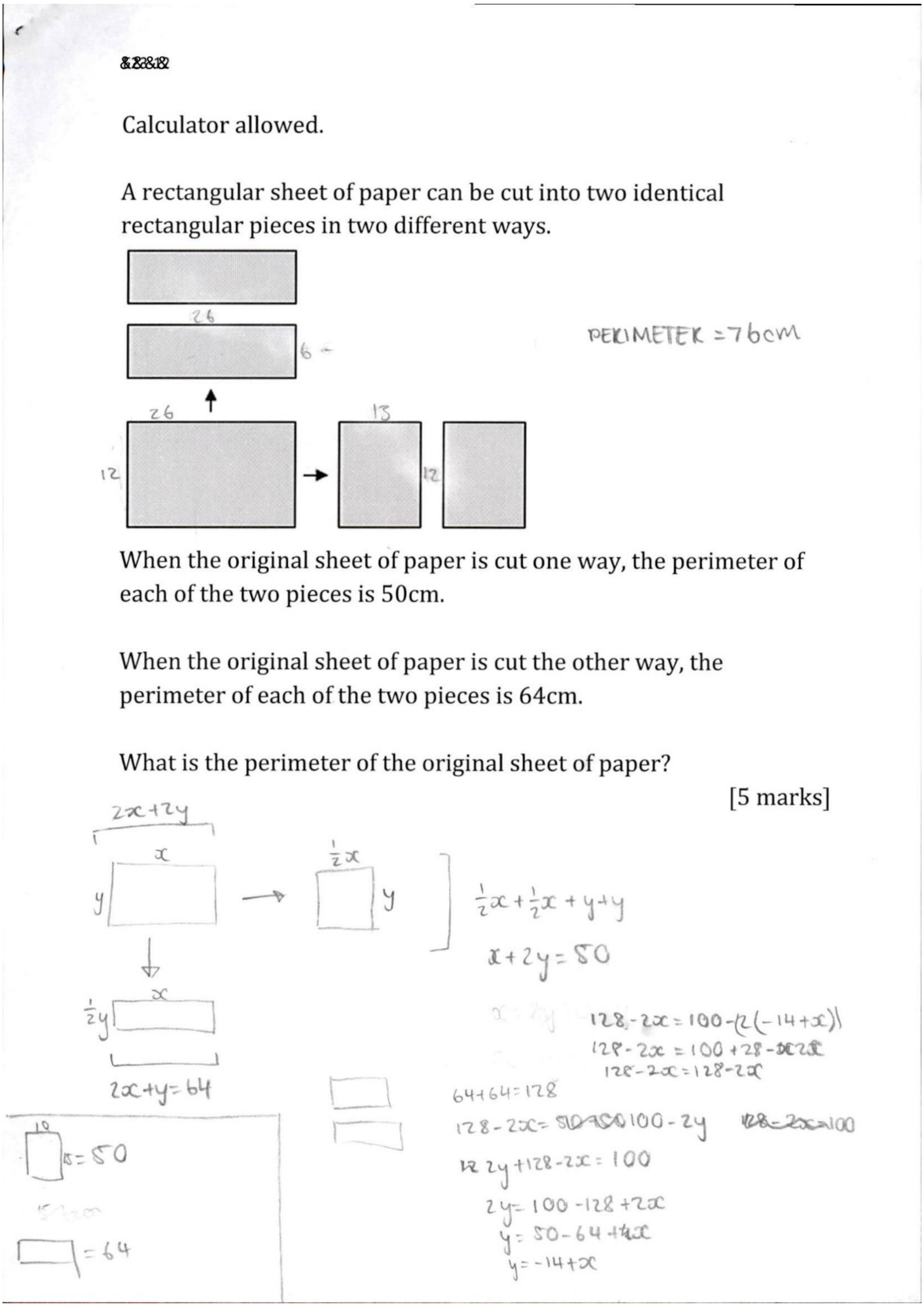
13 Disks in pictogram

$\frac{260}{13} = 20$

sport = $2 \cdot 5 \times 20 = 50$

Answer: 50

Figure 38: Item with the highest AO3 parameter value on the higher tier (Pearson paper 3, question 12)



Calculator allowed.

A rectangular sheet of paper can be cut into two identical rectangular pieces in two different ways.

26
6
PERIMETER = 76cm
26
12
13
12

When the original sheet of paper is cut one way, the perimeter of each of the two pieces is 50cm.

When the original sheet of paper is cut the other way, the perimeter of each of the two pieces is 64cm.

What is the perimeter of the original sheet of paper?

[5 marks]

$\frac{1}{2}x+\frac{1}{2}x+y+y$

$x+2y=50$

$2x+y=64$

$128-2x=100-(2(-14+x))$

$128-2x=100+28-\cancel{2x}2x$

$128-2x=128-2x$

$64+64=128$

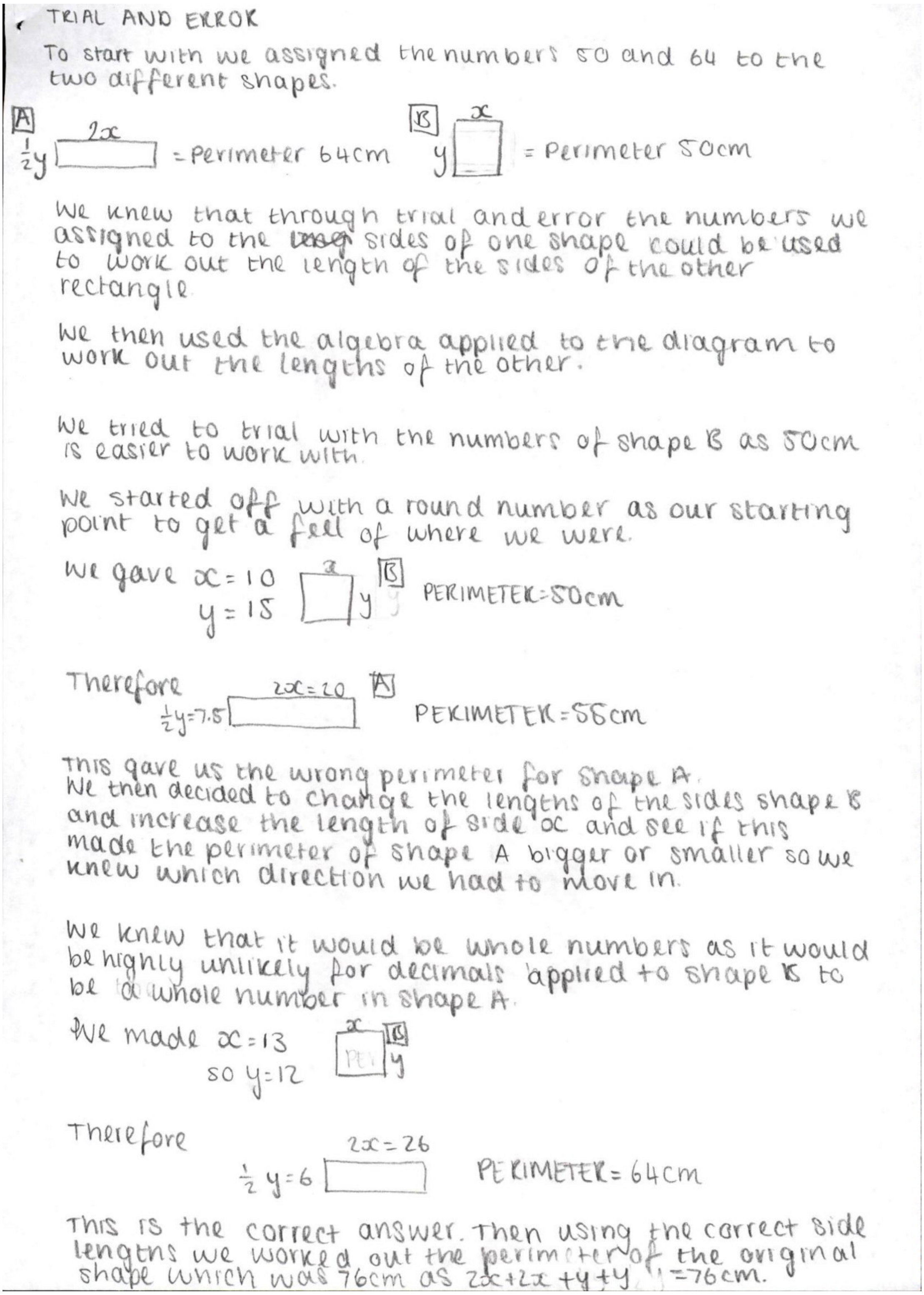
$128-2x=\cancel{50}\ \cancel{50}\ 100-2y$ $\cancel{128-2x=100}$

$\cancel{12}\ 2y+128-2x=100$

$2y=100-128+2x$

$y=50-64+\cancel{2}x$

$y=-14+x$



TRIAL AND ERROR

To start with we assigned the numbers 50 and 64 to the two different shapes.

A $2x$ $\frac{1}{2}y$ = Perimeter 64cm   B x y = Perimeter 50cm

We knew that through trial and error the numbers we assigned to the sides of one shape could be used to work out the length of the sides of the other rectangle

We then used the algebra applied to the diagram to work out the lengths of the other.

We tried to trial with the numbers of shape B as 50cm is easier to work with.

We started off with a round number as our starting point to get a feel of where we were.

We gave $x = 10$
$y = 15$ x B y PERIMETER=50cm

Therefore $2x = 20$ A
$\frac{1}{2}y = 7.5$ PERIMETER=55cm

This gave us the wrong perimeter for shape A.
We then decided to change the lengths of the sides shape B and increase the length of side x and see if this made the perimeter of shape A bigger or smaller so we knew which direction we had to move in.

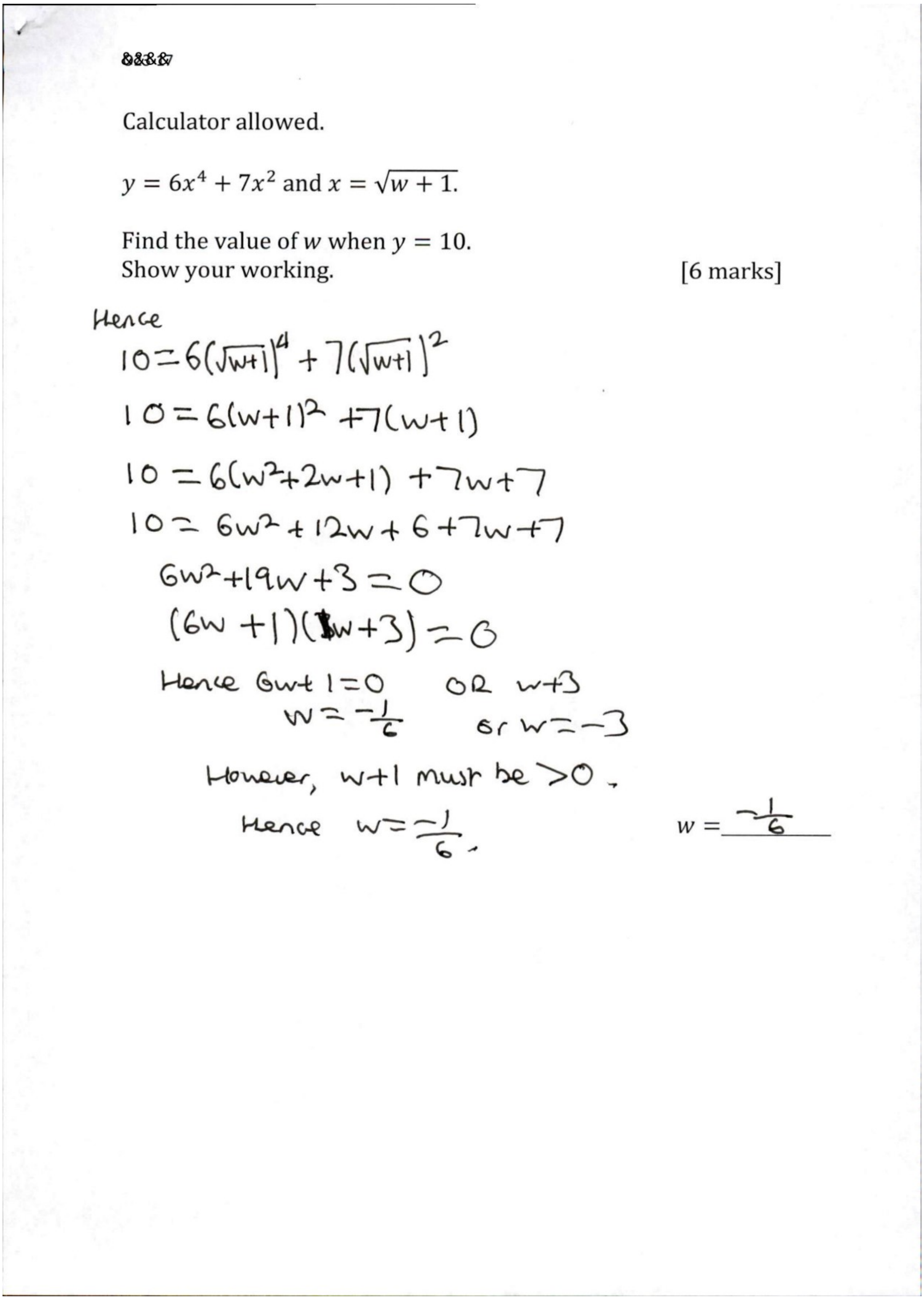
We knew that it would be whole numbers as it would be highly unlikely for decimals applied to shape B to be a whole number in shape A.

We made $x = 13$
so $y = 12$ x B y

Therefore $2x = 26$
$\frac{1}{2}y = 6$ PERIMETER=64cm

This is the correct answer. Then using the correct side lengths we worked out the perimeter of the original shape which was 76cm as $2x + 2x + y + y = 76$cm.

Figure 39: Item with the lowest AO3 parameter on the higher tier (OCR paper 3, question 17)



&&&&7

Calculator allowed.

$y = 6x^4 + 7x^2$ and $x = \sqrt{w+1}$.

Find the value of w when $y = 10$.
Show your working. [6 marks]

Hence

$10 = 6(\sqrt{w+1})^4 + 7(\sqrt{w+1})^2$

$10 = 6(w+1)^2 + 7(w+1)$

$10 = 6(w^2+2w+1) + 7w + 7$

$10 = 6w^2 + 12w + 6 + 7w + 7$

$6w^2 + 19w + 3 = 0$

$(6w+1)(w+3) = 0$

Hence $6w+1=0$ OR $w+3$

$w = -\frac{1}{6}$ or $w = -3$

However, w+1 must be >0.

Hence $w = \frac{-1}{6}$.

$w = \frac{-1}{6}$

There was no correlation between the extent to which the items were judged to elicit AO3 and the expected difficulty of the items derived from study 1 (foundation tier: r=-0.14 and higher tier: r=0.03). Indeed, the item judged to be best at eliciting AO3 (Pearson paper 3, question 12) was only of moderate expected difficulty (logit=0.69).

Further, there was only a weak correlation between the extent to which the items were judged to elicit AO3 and the item tariff (r=0.21), as can be seen in figure 40. This suggests that it is possible to create valid items testing AO3 worth a small number of marks.

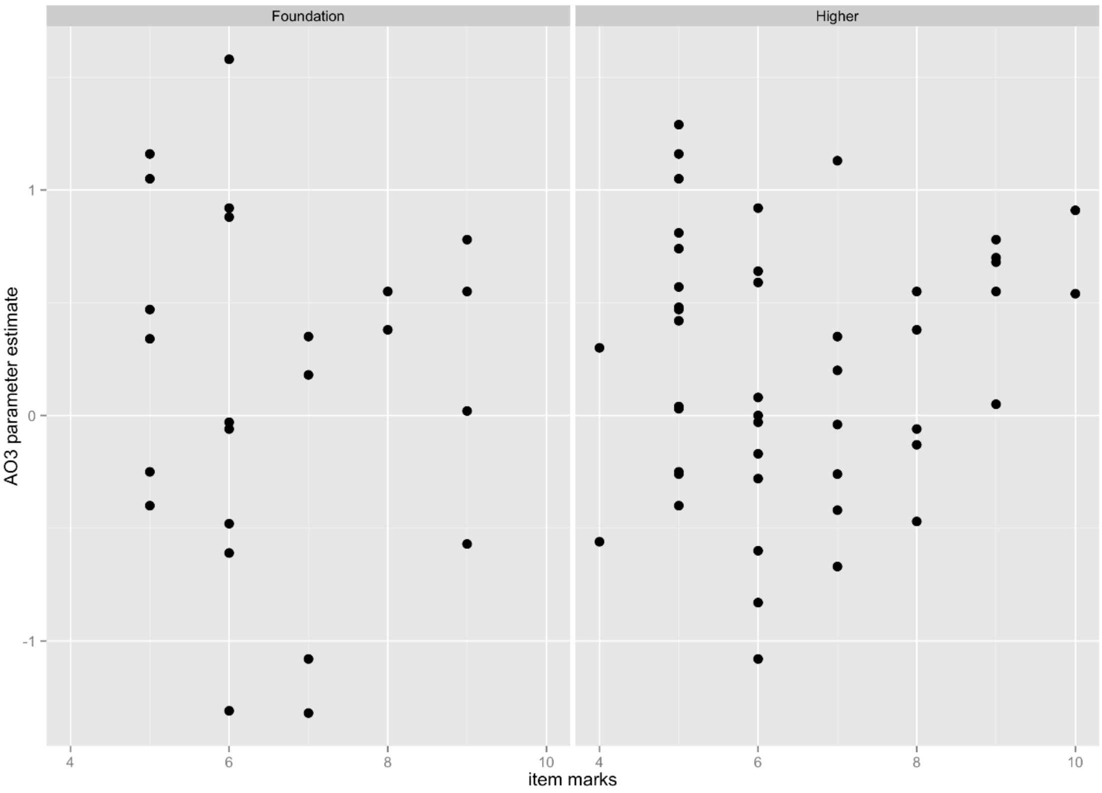


Figure 40: The relationship between AO3 parameter and item tariff

3.4 Summary of findings

The lower inter-rater reliability of the judging compared with that in study 1 shows that there was less consensus between the judges as to which items best elicited problem solving (as defined by AO3). However, the inter-rater reliability was adequate and the judgements of the authentication AO1/2 items suggested that judges were making decisions on the AO3 quality of the items rather than other attributes. The small difference in judgements between exemplars suggests that the response did have some slight effect on judgements but that judges were mainly judging the item.

The differences between exam boards were more substantial on the foundation tier papers than on the higher tier papers. Considering the foundation tier first, the judges considered the Eduqas AO3 items to be better at eliciting mathematical problem solving than the other exam boards' items. Judgements of the AQA items, on the other hand, were more varied. The common item between tiers was highly rated but the foundation tier only items were rated in line with the AO1/2 authentication items. In other words, the judges believed that these two items were poor in eliciting AO3 responses.

Turning to the higher tier papers, the judgements were far more similar across exam boards. Once again, however, there were a number of items that were judged to be quite poor at eliciting AO3 responses, as they fell below the parameter estimates of two of the AO1/2 items.

There was no relationship between the extent to which the items were judged as eliciting AO3 and their expected difficulty. This has important implications for the validity with which AO3 can be tested because it suggests that the papers included some problem-solving items which were both accessible and valid (as defined by AO3). Further, there was only a weak correlation between the extent to which the items were judged to elicit AO3 and the item tariff, which suggests that it is possible to create valid items worth a small numbers of marks.

4. Study 4 – A study of the ways in which problem solving (AO3) items vary across exam boards' sample assessments

4.1 Design

This study collected maths experts' views of the dimensions along which mathematical problem-solving items from the sample assessments vary. This work was conducted in the context of the definition of problem solving as articulated by AO3, and the same 33 problem-solving items used in study 3 were investigated here.

A variant of Kelly's Repertory Grid[33] was used to obtain the dimensions. Although the repertory grid is a method devised for use within personality psychology, it has proven effective in many contexts for eliciting the (unknown) constructs which people use to classify their experience of the world around them. This can be applied to almost any kind of stimuli where a person unconsciously classifies objects in order to distinguish them. The repertory grid technique allows people to share their tacit knowledge because it assumes people use their construct systems to make sense of the world.

In an educational context, this method has been used to define implicit models of how children learn[34] or, more relevantly in this context, to elicit the constructs that formed the concept of item demand in GCSE and A level history, geography and chemistry papers.[35]

The repertory grid was designed to be used in an interview where personality constructs were identified through discussion. The constructs along which items vary can be elicited via a process in which three items are presented and participants are asked to identify a feature which allows them to pair two items as 'similar', in contrast with a third 'different' item.

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

[34] Parsons, J. M., Graham, N. and Honess, T. (1983) A teacher's implicit model of how children learn. British Educational Research Journal 9(1) 91-101.

[35] Hughes, S., Pollitt, A. and Ahmed, A. (1998) The development of a tool for gauging the demands of GCSE and A level exam items. Talk at British Educational Research Association conference, Queens University Belfast, August. Retrieved on 02/04/15 from www.cambridgeassessment.org.uk/images/109649-the-development-of-a-tool-for-gauging-the-demands-of-gcse-and-a-level-exam-items.pdf

Once the dimensions along which the 33 AO3 items varied were identified and refined, the participants rated the items according to these dimensions. These ratings were combined and analysed to explore any systematic differences in items across the exam boards. The extent to which experts' ratings were consistent was also considered.

4.2 Method

4.2.1 Materials

All exam board items that predominantly test AO3 (the 33 items containing four or more AO3 marks, between seven and nine for each exam board) were used. These were identical to the items used in study 3.

4.2.2 Participants

Five experienced GCSE maths teachers were recruited to take part. They were identified through requests for volunteers sent to professional mathematics bodies and/or involvement in Maths Hubs.[36] All were experienced teachers, holding senior positions with an active involvement in mathematics.

4.2.3 Procedure

Working in groups (a pair and a group of three, with the participants rotated across groups) the participants were given randomly selected triplets of items. They were asked to pick the two items that were most similar and to specify how, and in particular how the third item was different. Debate was encouraged to help clarify the thinking and draw out additional dimensions. The output of the discussion was a list of contrasted attributes (for example 'item context likely to be familiar to the candidate' versus 'novel context unlikely to be familiar to the candidate').

No suggestions were made by the researchers as to what was an appropriate dimension upon which to split the items, the participants were free to pick out any feature except difficulty or mathematical domain (the former was explored directly in study 2 and the latter was already captured). The researchers only prompted the participants to produce succinct descriptions of the poles of each dimension. For each triplet, there was no limit to the number of dimensions allowed. When no more dimensions were identified, another random triplet was drawn for consideration. This procedure was repeated until no more new dimensions were identified.

After working in two groups, participants worked together to consider the two separate lists. These were combined, with a group discussion over any overlapping

[36] www.mathshubs.org.uk

dimensions, and final wording of the chosen dimensions and their poles. Similar but slightly different dimensions were retained, as the aim was to gather a rich data set for analysis, not to evaluate the quality or value of the dimensions. Participants then independently rated each item against every dimension. One pole represented a rating of 1, the other pole represented a rating of 5.

4.3 Results

4.3.1 Analysis

The 23 dimensions identified are shown in table 16. The participants' ratings were combined to obtain a mean rating for each item on each dimension. Mean standard deviations were calculated across all items on each dimension to determine which dimensions were more problematic for the participants to consistently rate. As expected, the variability of the ratings corresponded to the subjectivity of the judgement. Dimensions with low variability were easily judged surface features. For example, lined versus unlined response areas (mean sd = 0.11) and diagrams versus just text (mean sd = 0.15).

Dimensions upon which there was less agreement included the relevance (or otherwise) of text and/or diagrams (mean sd = 1.45), and the salience of the parameters needing to be selected to solve the problem (mean sd = 1.33). Even the most difficult-to-rate dimension had a mean standard deviation of less than 1.5 on a scale of 5.

Given that AO3 determines that students ought to be able to 'evaluate solutions to identify how they may have been affected by assumptions made', it is worth noting that the participants' mean rating of the extent to which items required students to evaluate assumptions was 4.08, where a score of 5 meant that the item did not make this requirement. Of those items considered to most require the evaluation of assumptions, two received a mean rating of 1.0 on this scale, figure 41 shows one of these items. Students are clearly required to consider their answer.

Further, AO3 determines that students should be able to 'make and use connections between different parts of mathematics'. Yet participants' mean rating was 3.86, where a score of 5 meant that the item did not make this requirement. In other words, on average, participants did not believe that the items required students to evaluate assumptions or make connections between parts of mathematics. However, some items were considered to capture this requirement, the one most thought to do so is shown in figure 42. It received a mean rating of 1.6.

Each dimension was analysed separately to see whether there were any systematic differences between exam boards. Table 16 shows the mean rating of all items for each exam board. These mean ratings were compared using one-way independent-samples ANOVAs, for each dimension. There was a significant effect of exam board for six of the dimensions (highlighted in bold). For these, pairwise comparisons (t-tests) were conducted to explore the source of the difference. All significant effects are reported given the relatively low sample size and resultant restricted power of these analyses. It is worth noting, however, that comparisons with p values < 0.009 would be significant even with a conservative Bonferroni correction.

There was a significant effect for blank versus lined response areas. Eduqas was significantly more likely to use lines (mean = 5.00) than the other boards which use unlined response spaces (mean = 1.09 or less, all comparisons p < 0.001). The other dimensions upon which there were significant differences between the boards were more interesting in relation to differences in approach to the construction of problem-solving items. Eduqas's items were significantly more likely to require an open-ended written answer rather than a numerical/mathematical answer (Eduqas: mean = 3.00; AQA: mean = 1.78, p = 0.014; Pearson: mean = 1.00, p < 0.001; OCR: mean = 1.58, p = 0.005). An example of such an open-ended item is shown in figure 43. It is an Eduqas item, and received the joint highest mean rating of 4.8. The differences between the other boards on this dimension were not significant. Eduqas's items were also significantly more likely to require candidates to justify their answers and methods compared with those of other boards (Eduqas: mean = 2.89; AQA: mean = 4.24, p = 0.007; Pearson: mean = 4.98, p < 0.001; OCR: rating = 4.47, p = 0.002). Again, it was an Eduqas item (shown in figure 44) that received the lowest rating on this dimension with a mean rating of 1.0, indicating strong justification is required. The differences between the other boards' items on this dimension were not significant.

Moving on to the linguistic difficulty of the items as judged by the participants, AQA's items (mean = 4.36) had significantly less text than OCR's (mean = 2.98, p = 0.005) or Eduqas's (mean = 2.97, p = 0.008). Pearson's items (mean = 3.40) were not significantly different from any other boards' items on this dimension. The item judged to have the greatest amount of text (mean rating = 1.4) is shown in figure 42. Eduqas's items (mean = 3.83) were significantly more likely to use demanding language (including the use of unusual words) than either AQA's items (mean = 4.80, p = 0.004) or Pearson's items (mean = 4.60, p = 0.021). OCR's items (mean = 4.33) were not significantly different from any other boards' items on this dimension. The most demanding item to read, according to our experts, is shown in figure 45, with a mean rating of 2.4. It contains some relatively difficult words and phrases. Unsurprisingly, this item was also rated the second lowest (1.6) on the dimension for the greatest amount of text to read. Of course, the valid testing of problem solving often requires some reading. The careful use of natural language ensures that the reading demand remains construct relevant.

An example of an item that scored very low on several of the above dimensions is shown in figure 46. This item received a mean rating of 5.0 on little or no text to be read and a low level of language demand, and a mean rating of 1.0 on numerical/mathematical answer required. Note that for all three dimensions there were other items which were similarly rated. This item has been picked for illustrative purposes.

Finally, Eduqas's items were significantly more likely to allow multiple approaches to solving a problem (mean = 2.37) than all three other boards' items (AQA: mean = 1.51, p = 0.001; Pearson: mean = 1.70, p < 0.001; OCR: mean = 1.69, p = 0.008). There was no significant difference between the latter three boards' items on this dimension. Overall the questions tended to offer few opportunities for alternative approaches. The two items rated most highly on this dimension have already been presented, being Eduqas items shown in figures 43 and 45. Both items received a mean rating of 3.2.

Although there were no statistically significant differences between boards on any of the other dimensions, consideration of examples at the extremes of the scales can be informative. Figure 47 shows the item that was judged to have the 'hardest numbers', with a mean rating of 4.6 compared with the overall mean of 2.2.

Figure 48 shows the item judged to most exemplify a 'twist in the tail', with a mean rating of 2.2. The answer required the number of losing tickets rather than the total number of tickets, and it is this feature that was considered a twist. Most items were rated closer to, or at, 5 on this dimension, with an overall mean rating of 4.22.

The item shown in figure 45 was also the strongest exemplification of the need to 'select parameters to do the calculation'. This item received a mean rating of 1.2 on this dimension, due to the two question parts requiring different sets of values to be selected from those presented. There was a wide range of item ratings on this dimension, with other items where few parameters were presented and all were needed, hence the overall mean rating of 3.48.

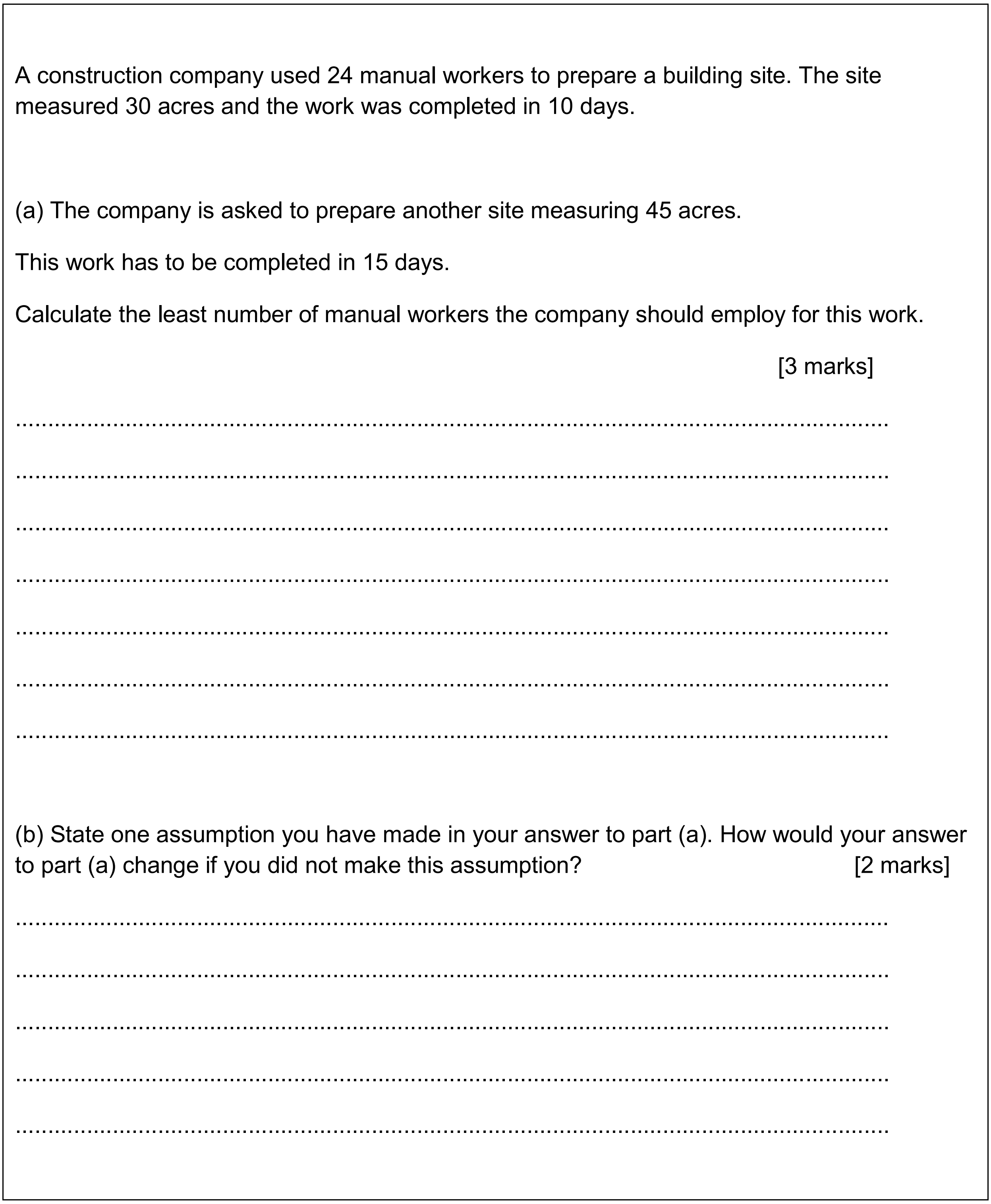
Finally, some items were judged to contain some irrelevant or arbitrary context (including diagrams), although the overall mean rating of 3.64 suggested that most text and diagrams were relevant. Figure 49 shows the item with the most extreme mean rating of 1.8. The specific context used here is no more relevant or informative than any other context used to frame this ratio question, and it does not help in answering the question.

Table 16: Summary of dimensions obtained, showing for each dimension the mean standard deviation of the participants' ratings for each item, and the mean rating by exam board and for all items

Pole (score of 1)	Pole (score of 5)	Mean SD	AQA mean rating	Pearson mean rating	OCR mean rating	Eduqas mean rating	All mean rating	
Includes useful diagrams	Text only	0.15	1.87	2.98	3.69	2.74	2.82	F(3,29)=1.388, p=0.266, $\eta^2 =$ 0.126, power=0.329
Exact answer required	Approximation required	0.41	1.96	1.00	1.82	2.40	1.78	F(3,29)=1.638, p=0.202, $\eta^2 =$0.145, power=0.384
Justification for answer and methods required	No justification required	0.61	4.24	4.98	4.47	2.89	4.19	**F(3,29)=6.840, p=0.001, $\eta^2 =$0.414, power=0.958**
Requires student to evaluate assumptions made	Does not require evaluation of assumptions	0.35	4.13	5.00	3.91	3.20	4.08	F(3,29)=2.105, p=0.121, $\eta^2 =$0.179, power=0.482
Requires working through standard procedures in reverse	Requires working through standard procedures in usual order	1.18	4.04	3.30	4.31	3.37	3.79	F(3,29)=2.300, p=0.098, $\eta^2 =$0.192, power=0.521
Multi-part	Single-part	0.70	3.62	3.88	2.58	3.71	3.42	F(3,29)=1.174, p=0.337, $\eta^2 =$0.108, power=0.282
Blank space for response	Lines given for response	0.11	1.09	1.00	1.09	5.00	1.90	**F(3,29)=727.197, p<0.001, $\eta^2 =$0.987, power=1.000**

Pole (score of 1)	Pole (score of 5)	Mean SD	AQA mean rating	Pearson mean rating	OCR mean rating	Eduqas mean rating	All mean rating	
Easy numbers	Hard numbers	1.01	1.96	2.40	2.02	2.51	2.20	F(3,29)=1.166, p=0.340, η^2 =0.108, power=0.280
Mixed units	Single unit	0.64	4.78	4.35	3.89	4.17	4.30	F(3,29)=1.175, p=0.336, η^2 =0.108, power=0.282
General knowledge needed	General knowledge not needed	0.57	4.64	4.60	4.56	3.77	4.42	F(3,29)=1.957, p=0.143, η^2 =0.168, power=0.452
High quantity of text to be read	Little or no text to be read	0.86	4.36	3.40	2.98	2.97	3.45	**F(3,29)=3.905, p=0.019, η^2 =0.288, power=0.772**
A ‘twist’ in the response required	No ‘twist’ in response required	0.96	4.33	3.73	4.58	4.20	4.22	F(3,29)=2.374, p=0.091, η^2 =0.197, power=0.535
Requires selection of parameters to do the calculation	No selection of parameters to do the calculation	1.33	4.02	3.20	3.62	2.91	3.48	F(3,29)=2.367, p=0.091, η^2 =0.197, power=0.534
Unit conversion required	Unit conversion not required	0.57	4.82	4.30	3.73	4.11	4.25	F(3,29)=1.378, p=0.269, η^2 =0.125, power=0.327
Obvious first step	Non-obvious first step	1.19	2.24	2.95	2.13	2.51	2.44	F(3,29)=1.264, p=0.305, η^2 =0.116, power=0.302

Pole (score of 1)	Pole (score of 5)	Mean SD	AQA mean rating	Pearson mean rating	OCR mean rating	Eduqas mean rating	All mean rating	
High level of language demand (unusual words used)	Low level of language demand	0.76	4.80	4.60	4.33	3.83	4.42	**F(3,29)=3.636, p=0.024, η^2 =0.273, power=0.739**
Intermediate steps given or implied	Intermediate steps not obvious	1.23	3.33	3.90	3.20	4.03	3.58	F(3,29)=1.862, p=0.158, η^2 =0.162, power=0.432
Real-world context	Pure maths	0.67	2.64	2.95	2.27	2.63	2.61	F(3,29)=0.285, p = 0.836, η^2 =0.029, power=0.098
Context (inc. diagrams) irrelevant/arbitrary	All text and diagrams relevant	1.45	3.93	3.55	3.47	3.57	3.64	F(3,29)=0.427, p=0.735, η^2 =0.042, power=0.125
Numerical/mathematical answer	Open-ended written answer	0.57	1.78	1.00	1.58	3.00	1.79	**F(3,29)=6.121, p=0.002, η^2 =0.388, power=0.934**
Single approach	Multiple possible approaches	1.04	1.51	1.70	1.69	2.37	1.79	**F(3,29)=4.712, p=0.008, η^2 =0.328, power=0.851**
Requires using obvious standard method	No obvious standard method	1.18	1.64	2.30	2.04	2.23	2.04	F(3,29)=1.428, p=0.255, η^2 =0.129, power=0.338
Requires connections between different parts of maths	Does not require connections between different parts of maths	0.96	3.76	3.25	4.22	4.21	3.86	F(3,29)=2.093, p=0.123, η^2 =0.178, power=0.480



A construction company used 24 manual workers to prepare a building site. The site measured 30 acres and the work was completed in 10 days.

(a) The company is asked to prepare another site measuring 45 acres.

This work has to be completed in 15 days.

Calculate the least number of manual workers the company should employ for this work.

[3 marks]

....

....

....

....

....

....

....

(b) State one assumption you have made in your answer to part (a). How would your answer to part (a) change if you did not make this assumption? [2 marks]

....

....

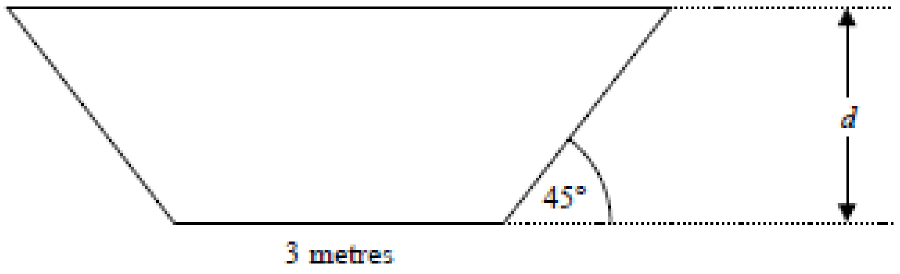
....

....

....

Figure 41: Example of a question rated most highly on the dimension ‘Requires students to evaluate assumptions made’, from Eduqas higher tier paper 1, Q9

The diagram shows the cross-section of the water in a drainage channel.



The cross-section is in the shape of a trapezium with one line of symmetry.

The base of the drainage channel is horizontal.

The two equal sides of the trapezium are each inclined at 45° to the horizontal.

The length of the base of the trapezium is 3 metres.

The depth of the water is d metres.

The area of the cross-section is $A\ \mathrm{m}^2$.

(a) Write a formula for A in terms of d.

Give your answer in its simplest form. [3 marks]

The depth of the water in the drainage channel is 1.5 metres.

(b) Find the area of the cross-section of the water. [2 marks]

The water flows along the drainage channel at a rate of 486,000 litres per minute. The depth of the water is constant.

(c) Work out the speed of the water.

Give your answer in metres per second. [4 marks]

Figure 42: Question rated most highly on the dimension ‘Requires connections between different parts of maths’; also rated most highly on ‘High quantity of text to be read’, from Pearson higher tier paper 2, Q9

A cylinder is made of bendable plastic.

A dog's toy is made by bending the cylinder to form a ring.

The two circular ends of the cylinder are joined to form the ring.

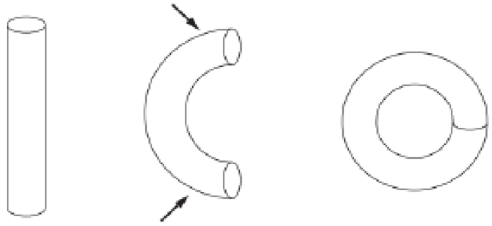


Diagram not drawn to scale

The inner radius of the dog's toy is 8 cm.

The outer radius of the dog's toy is 9 cm.

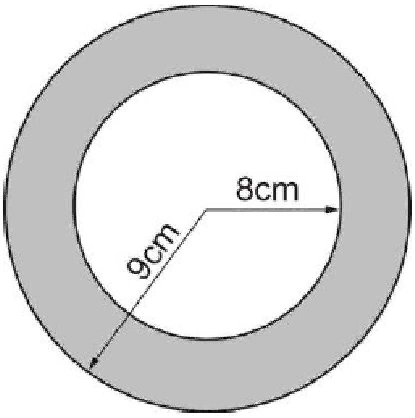


Diagram not drawn to scale

Calculate an approximate value for the volume of the dog's toy.

State and justify what assumptions you have made in your calculations and the impact they have had on your results.

[7 marks]

....

....

Figure 43: Question rated most highly on the dimension 'Open ended written answer', also rated most highly on 'Multiple possible approaches', from Eduqas higher tier paper 2, Q17

The diagram below shows a composite shape formed by joining two rectangles.

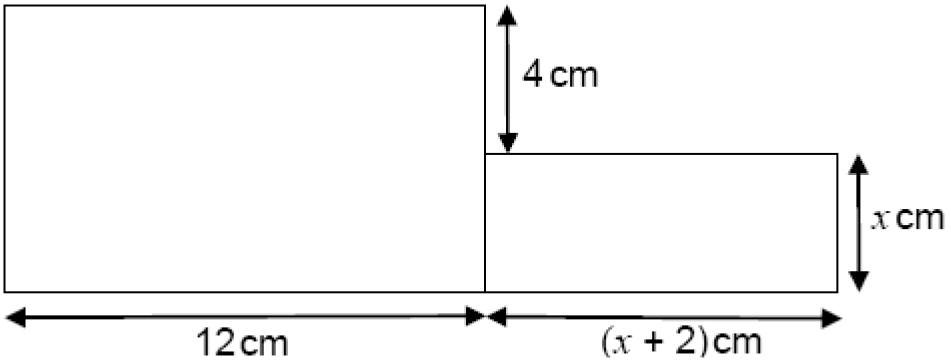


Diagram not drawn to scale

The area of the larger rectangle is $4y\ cm^2$.

The area of the smaller rectangle is $y\ cm^2$

Calculate the dimensions of the smaller rectangle.

You must justify any decisions that you make.

[7 marks]

....

....

....

....

....

....

....

....

....

Figure 44: Question rated most highly on the dimension ‘Justification for answer and methods required’, from Eduqas higher tier paper 1, Q17

Laura has her own car.

During April

- Laura drove a total distance of 560 miles in her car.
- Her car's fuel consumption was 37·8 mpg (miles per gallon).
- Petrol cost £1.48 per litre.
- Laura spent 10 hours 45 minutes driving her car.

(a) Given that 1 gallon is approximately 4·55 litres, calculate the cost of petrol that Laura used during April.

You must show all your working. [5 marks]

(b) Select which of the following best describes the roads on which Laura travelled during April.

You must show working to support your answer.

You must give a reason for your answer.

A. Mainly small narrow country lanes

B. Mainly inner city roads with lots of traffic lights

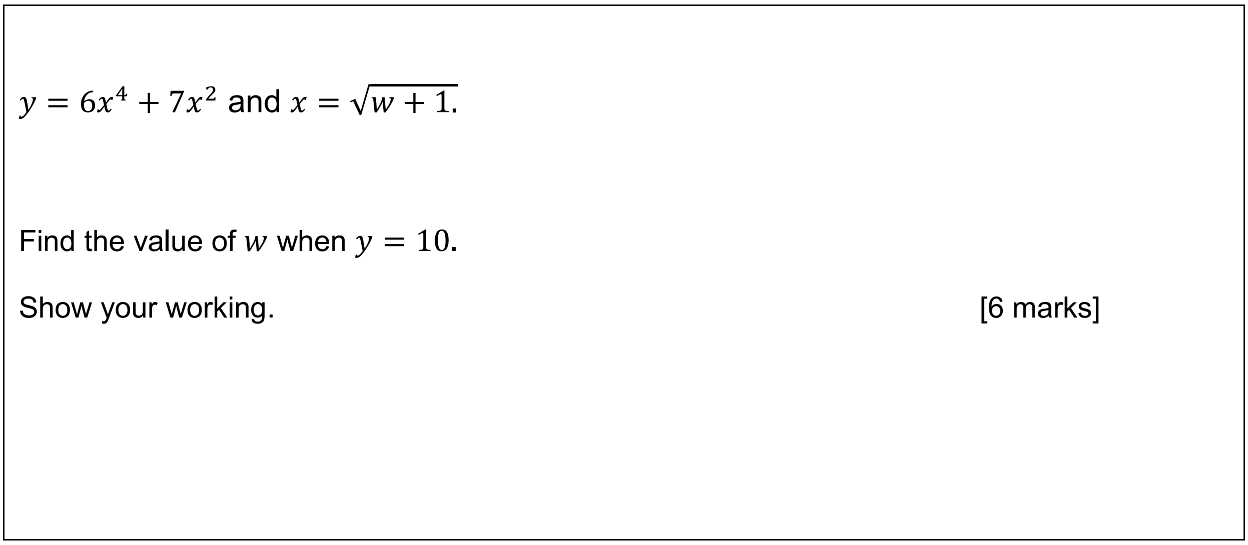
C. Mainly motorways and dual carriageways

D. Mainly steep mountain routes with many sharp bends

E. Mainly roads with speed limits of 30 mph [4 marks]

Reason:

Figure 45: Question rated most highly on the dimension 'High level of language demand (unusual words used)'; also rated most highly on 'Multiple possible approaches' and 'Requires selection of parameters to do the calculation', from Eduqas higher tier paper 2, Q10



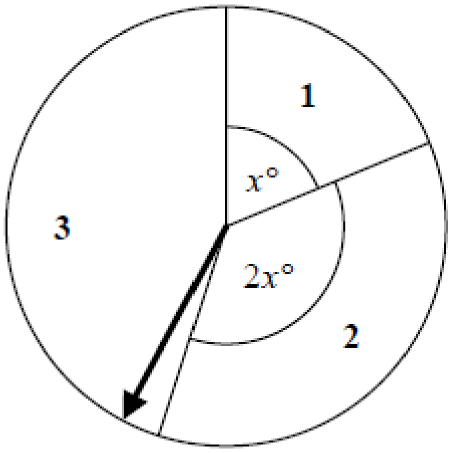
$y = 6x^4 + 7x^2$ and $x = \sqrt{w + 1}$.

Find the value of w when $y = 10$.

Show your working. [6 marks]

Figure 46: Question rated high on the dimensions ‘Low level of language demand’ and ‘Least quantity of text to be read’; also rated high on ‘Numerical answer required’, from OCR higher tier paper 3, Q17

Here is a spinner.



When the arrow is spun once, a 1 or a 2 or a 3 can be scored.

Person A is going to spin the arrow twice.

He will work out his total score by adding the two scores he gets on the two spins.

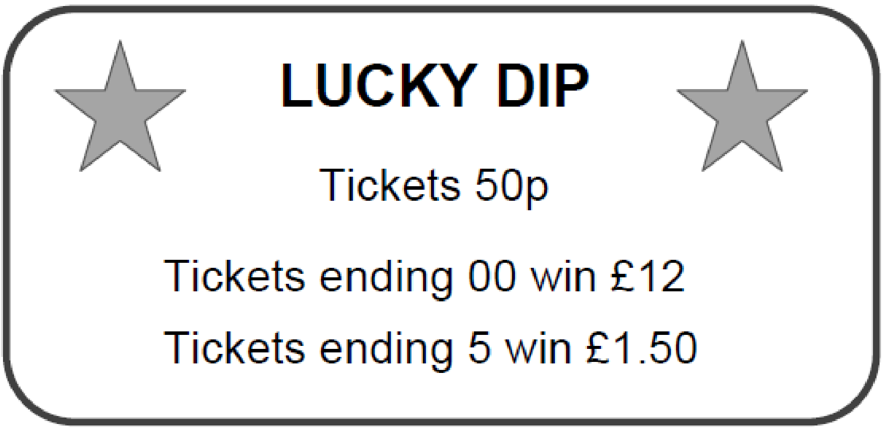
The probability that he will get a total score of 4 is $\frac{16}{81}$

Assuming that the thickness of the three lines between the sectors may be ignored, work out the value of x.

[5 marks]

Figure 47: Question rated most highly on the dimension 'Hard numbers', from Pearson higher tier paper 3, Q16

Person A ran a Lucky Dip stall.



LUCKY DIP

Tickets 50p

Tickets ending 00 win £12

Tickets ending 5 win £1.50

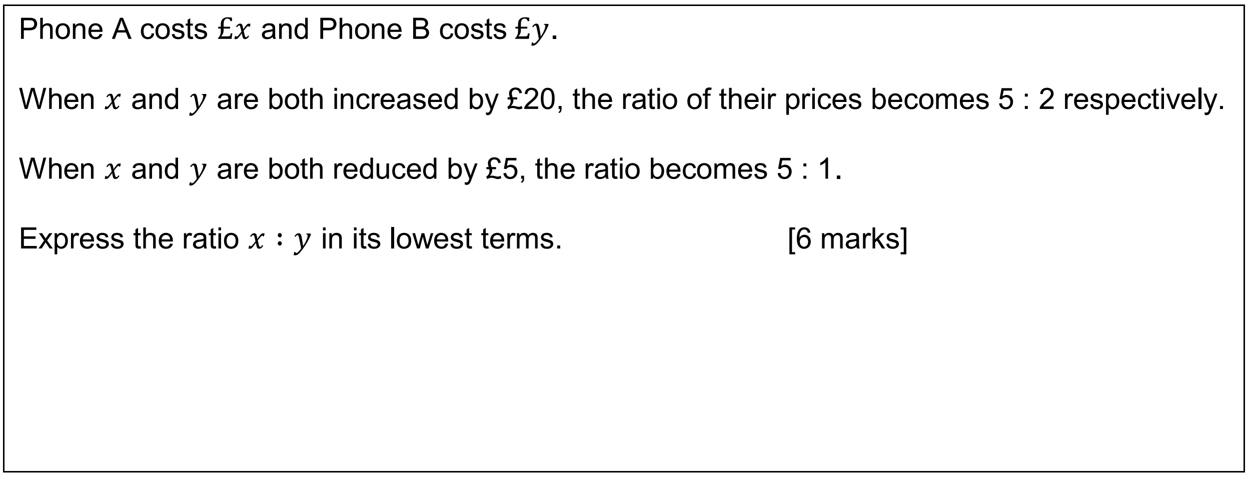
There were 750 tickets, numbered 1 to 750

Person A sold **all** the winning tickets, and **some** of the losing tickets.

They made a profit of £163

How many **losing** tickets did they sell? [6 marks]

Figure 48: Question rated most highly on the dimension ‘A twist in the response required’, from AQA higher tier paper 2, Q8



Phone A costs £x and Phone B costs £y.

When x and y are both increased by £20, the ratio of their prices becomes 5 : 2 respectively.

When x and y are both reduced by £5, the ratio becomes 5 : 1.

Express the ratio $x : y$ in its lowest terms. [6 marks]

Figure 49: Question rated most highly on the dimension ‘Context (inc. diagrams) irrelevant/arbitrary’, from OCR higher tier paper 2, Q19

4.4 Summary of findings

Twenty-three dimensions were elicited from the participants’ scrutiny of the items. There was an acceptable degree of consensus between the participants in their ratings of the items on these dimensions. Given the definition of AO3, it might seem surprising that the items were not generally rated as requiring students to evaluate assumptions or make connections between parts of maths. However, not all items were required to meet all elements of AO3. Indeed, six of the items clearly asked for some consideration of assumptions. Not all items testing AO3 needed to be extended in nature or cover the whole problem-solving cycle.

There were significant differences between the exam boards’ items on only six of the dimensions. Compared with the other boards’ items, Eduqas’s items were more likely to require open-ended answers, to require candidates to justify their answers and methods, to allow multiple approaches to solving a problem and to use demanding language (although not more so than OCR’s items). AQA’s items included less text than the other boards’ items (although not significantly more so than Pearson’s items). Eduqas was more likely to use lines to set out the response area for students.

5. Discussion

In considering whether the evidence suggests that exam boards need to change the difficulty or approach of future assessments, it is important to consider the policy objectives of the reform of GCSE maths. The Department for Education (DfE) has set out the subject aims for GCSE maths.[37] These state that specifications,

> should provide a broad, coherent, satisfying and worthwhile course of study. They should encourage students to develop confidence in, and a positive attitude towards mathematics and to recognise the importance of mathematics in their own lives and to society. They should also provide a strong mathematical foundation for students who go on to study mathematics at a higher level post-16. (p.3)

The subject aims go on to say:

> Students can be said to have confidence and competence with mathematical content when they can apply it flexibly to solve problems.
>
> The expectation is that:
>
> - all students will develop confidence and competence with the content identified by standard type
> - all students will be assessed on the content identified by the standard and the underlined type; more highly attaining students will develop confidence and competence with all of this content
> - only the more highly attaining students will be assessed on the content identified by bold type. The highest attaining students will develop confidence and competence with the bold content. (p.4)

The document goes on to set out the scope of study by mathematical domain, using the type face to indicate expectations as set out above.

[37] Mathematics GCSE subject content and assessment objectives (2013) www.gov.uk/government/uploads/system/uploads/attachment_data/file/254441/GCSE_mathematics_subject_content_and_assessment_objectives.pdf

It is also worth considering the then Secretary of State, Michael Gove's letter to Ofqual (6th February 2013) setting out the policy steer for reforming Key Stage 4 qualifications.[38]

> The reformed GCSEs should remain universal qualifications of about the same size as they are currently, and accessible, with good teaching, to the same proportion of pupils as currently sits GCSE exams at the end of Key Stage 4. At the level of what is widely considered to be a pass (currently indicated by a grade C), there must be an increase in demand, to reflect that of high-performing jurisdictions. This is something we believe the vast majority of children with a good education should be able to achieve. At the top end the new qualification should prepare pupils properly to progress to A levels or other study. This should be achieved through a balance of more challenging subject content and more rigorous assessment structures. We know that employers and others are keen for greater reassurance that pupils who achieve that level of performance in English and mathematics are literate and numerate.

The subject matter of the new GCSE maths covers more content than the current GCSE. As the DfE said, the reformed GCSE,

> will provide greater coverage of areas such as ratio, proportion and rates of change; it will require all students to master the basics, and will be more challenging for those aiming to achieve top grades.[39]

The judgements of expected difficulty collected in the first study were done on an item-by-item basis. This means that comparisons between the current papers and the new sample assessments do not take into account the change in the breadth of the qualification. In one sense, the new maths GCSE is immediately more challenging in that more material must be covered. An appropriate increase in teaching time will be needed to account for this change.[40] It is also worth noting that the recent change to linear assessment (taken at the end of the course of study rather than in bite-sized chunks through the course), which took effect in summer

[38] www.gov.uk/government/uploads/system/uploads/attachment_data/file/278308/sos_ofqual_letter_060213.pdf

[39] www.gov.uk/government/policies/reforming-qualifications-and-the-curriculum-to-better-prepare-pupils-for-life-after-school/supporting-pages/gcse-reform

[40] www.gov.uk/government/speeches/reformed-gcses-in-english-and-mathematics

2014, and the consequent reduction in re-sitting opportunities will also be challenging for some students.

The first study found that, for all exam boards, the expected difficulty of the reformed GCSE sample assessments was higher than that of the current GCSE papers aggregated across boards. Care needs to be taken in drawing strong conclusions from comparisons with other international jurisdictions, as the context in which the assessments operate will have an effect on the actual difficulty that students experience when they attempt the papers. Indeed, it is helpful to think of the content and expected difficulty of international papers as an indication of the curriculum aspirations of different jurisdictions. That said, the expected difficulty of OCR and Pearson's higher tier assessments was more in line with those of international jurisdictions such as Shanghai and Japan than with the expected difficulty of the current papers.

The size of the difference in average expected difficulty between the exam boards' sample assessments was greater than the difference in expected difficulty between current papers. AQA's sample assessments were perceived as the easiest, then OCR's assessments, and Pearson's assessments were perceived to be the hardest. In general, this pattern across exam boards was replicated whatever the mathematical domain or the assessment objective being tested. The differences between boards were more pronounced for items testing AO1 and AO2, than for items testing AO3. Differences in expected difficulty of AO3 items did not account for the differences in overall expected assessment difficulty. This was contrary to the views expressed by a number of stakeholders prior to the research being conducted. The expected difficulty of the OCR and Pearson's sample assessments was higher than that of their current papers, the expected difficulty of AQA's sample assessments was very similar to that of their current papers.

The spread of how hard the items were expected to be was greater on the foundation tier than the higher tier. This was the case for all exam boards' sample assessments. This is of concern as the foundation tier assessment in the new GCSE covers grades 1 to 5, whereas the higher tier assessment supports more grades – from an allowed 3 grade to grade 9. Given the range of ability covered by each tier, we might expect a wider spread of expected item difficulty on the higher rather than the foundation tier. It may be that the higher tier assessments will fail to sufficiently discriminate between students to allow reliable grading.

Reassuringly, there was a moderately strong relationship between the expected difficulty of the items and the actual difficulty values gathered from the second study. We would not predict a perfect relationship between the expected mathematical difficulty and the actual difficulty of the items. The research literature shows that, beyond the mathematics involved, the wording and context of items have an effect on actual difficulty. Indeed, there was a slight tendency for OCR items with a higher word count to have a higher actual difficulty than expected. It would be useful for exam boards to conduct a more sophisticated analysis of the impact of the quality of the language (as opposed to mere number of words) on difficulty.

There are differences in the difficulty of exam boards' current papers. It is impossible for exam boards to precisely control the difficulty of papers and so the adjustment of grade boundaries is necessary to ensure fairness. The comparable outcomes approach[41] to setting boundaries controls for any differences in difficulty. The correlation between expected and actual difficulty meant it was possible to estimate whether differences in difficulty of this magnitude could be accounted for in the setting of grade boundaries. Simulations suggested that the resultant differences in grade boundaries across boards were not extraordinary. Indeed, the differences in grade boundaries might be less than those currently observed.

The adjustment of grade boundaries to compensate for differences in difficulty is a necessary feature of an exams system in which the (resource intensive) pre-testing of items is not conducted. This research is in effect a pre-test of the items and so creates the opportunity to ameliorate differences in difficulty without recourse to the adjustment of grade boundaries. In other words, just because the system can use the awarding process to deal with the observed differences in difficulty, does not mean that it should when evidence of differences in difficulty is available. There is no need to continue with these differences in difficulty now that they have been quantified prior to the live exams. Seeking to better align the difficulty of exam boards' papers will be a complex task, but the information gathered through this research will provide boards with a good starting point.

Moreover, to continue with the differences in expected difficulty risks wash-back effects on teaching and learning. The wash back may be such that students being prepared for the easier papers would have a poorer learning experience than students being prepared for harder papers. Alternatively, the wash back may be such that students being prepared for the hardest papers might have a negative experience of maths which is damaging to their confidence. Either scenario could potentially undermine the policy intentions behind the reform of GCSE maths and raises issues of fairness.

Exam papers need to be sufficiently accessible to students who have been prepared, so as to allow the reliable setting of grade boundaries. Indeed, study 2 found that students performed better on AQA's papers, and for this (ill-prepared and relatively unmotivated) cohort, AQA's papers performed better as assessment instruments.

[41] http://webarchive.nationalarchives.gov.uk/20141031163546/http:/ofqual.gov.uk/standards/summer-2014-exams/#our-approach-to-summer-2014-awarding

There were differences of detail, of course, between exam boards and between tiers. On the foundation tier, the AQA paper was considerably easier than other papers across the full range of student ability. On the higher tier, the Pearson paper was considerably more difficult than the other papers (although the Eduqas paper was almost as difficult for the majority of the ability range). The AQA higher tier paper was slightly easier for lower ability students but at highest levels of ability the AQA paper was slightly harder than OCR's.

The mean marks of students on the sample assessments were very low compared with those which we would expect in a real GCSE maths exam. Indeed, even the students from the best performing schools scored poorly. Unfortunately, it is difficult to disentangle the extent to which this is due to a lack of motivation on the part of the students, unfamiliarity with the style of these papers, a lack of preparation for the content of the papers or the assessments being too difficult for students to access. The high non-response rate, which worsened as the students progressed through the papers, could reflect a lack of motivation or the increasingly inaccessible nature of the items. What we can be sure of is that mark distributions such as those observed in this study would not allow the setting of reliable grade boundaries, would undermine confidence in the exam system and would not support the policy intentions behind the reform of GCSE maths.

Based on previous estimates of the pre-test effect (the impact of lack of motivation and incomplete preparation) it was possible to roughly adjust the facility values for the items testing old content. This gave an estimate of how difficult the papers might be if they had been sat by motivated, prepared students and whether the more difficult papers would allow the reliable setting of grade boundaries. The estimate of the pre-test effect was very approximate and based upon that reported for Key Stage 2 testing. It is possible, of course, that the effect of low motivation on scores might be more or less for the students aged 15 to 16 years old. It is certainly the case that the pre-test effect would not in reality be linear across the ability range.

That said, the analysis suggested that the boundaries would be much lower than ideal. For example, the grade A boundaries for the Eduqas and Pearson higher tier papers were less than half marks. Low boundaries lead to unreliable grading because they are based on little evidence of what the students know and can do, and they are likely to be close together such that small numbers of marks can make large differences to the grade achieved.

Exams have increasingly come to represent the curriculum and all the tested assessments covered the curriculum sufficiently well. However, it is crucial that exams also function well as measurement instruments. Of course, it is impossible to know the extent to which good teaching, exam preparation and more teaching time could mitigate the risk of low boundaries. Nonetheless, the evidence suggests that if these assessments were live exams, they would not function well as measurement instruments and there is a significant risk that the setting of boundaries in summer 2017 could be problematic. In particular, it is worth bearing in mind that even students from the best performing schools, one of which usually had 100 per cent of its students achieving at least a grade C in GCSE maths, performed badly in this study.

In developing the maths GCSE, gaining a common understanding of problem solving and its operationalisation was a new challenge for exam boards. Discourse around the variation in exam boards' sample assessments suggested that different approaches to measuring students' problem solving might be a source of differences in difficulty. This was not found to be the case; the differences in difficulty between boards were greater for AO1 and AO2 than for AO3.

Nonetheless, study 3 showed that there was a difference between the exam boards in their approach to testing AO3. The differences were more substantial for the foundation tier items than for the higher tier items. On the foundation tier, the judges considered the Eduqas AO3 items to be best at eliciting mathematical problem solving. While judgements of the AQA items were more varied, in general, the judges believed that these items were not as good at eliciting AO3 responses compared with other boards' items. On the higher tier, the judgements were far more similar across exam boards. Once again, however, there were a number of items that were judged to be relatively poor at eliciting AO3 responses.

Compared with the other boards' items, Eduqas's AO3 items were more likely to require open-ended answers, to require candidates to justify their answers and methods, to allow multiple approaches to solving a problem, and to require understanding of demanding language (although in this case, not more so than OCR's items). AQA's items included less text than the other boards' items (although not significantly more so than Pearson's items).

While it wasn't possible to investigate the relationship between the extent to which the items were judged as eliciting AO3 and their actual difficulty, there was no relationship between the quality of the items and their expected difficulty. This suggests that the assessments included some problem-solving items which were considered both accessible and valid. There was also only a very weak correlation between the extent to which an item was judged to be eliciting AO3 and the maximum mark for the item. It is possible, therefore, to create valid, short items. From this evidence, it would be wrong to presume that low mark tariffs are problematic for the testing of AO3.

5.1 Summary

Overall, the sample assessments are more difficult than the current papers and the difficulty of the higher tier assessments is more in line with that of international jurisdictions. However, while AQA's current exam papers were judged to be the most difficult of exam boards' GCSE papers, AQA have not increased the difficulty of their sample assessments. There is, therefore, a difference in difficulty across the boards, which could have negative consequences for teaching and learning.

There is also a significant risk that all but AQA's assessments will be too difficult for the full range of ability of the cohort for which the qualification is intended. This is likely to prevent the reliable grading of students. The additional challenge will be beneficial for the most able students but the assessments also need to support a positive experience for the rest of the cohort so as to ensure that all students become more confident and competent as mathematicians.

Adjustments to the expected difficulty of the sample assessments and the associated live exam papers can be made before teaching begins in September 2015. This wealth of information regarding the functioning of items and papers will enable exam boards to design assessments so as to better deliver the policy intentions behind the reform of GCSE maths.

Appendix A: Assessments in study 1

Assessments included in study 1, detailing the purpose of the assessment, the age of the cohort taking it and the paper year from which the items were drawn.

Jurisdiction / awarding organisation	Assessment	Use	Cohort age and proportion taking part	Year of papers used
Cambridge International Examinations	IGCSE	Qualification awarded Can control entry to upper secondary education	14–16 Full cohort	2011
	O Level	Qualification awarded Can control entry to upper secondary education	14–16 Full cohort	2011
England	GCSE (+ SAMs for reformed GCSEs)	Qualification awarded Controls entry to upper secondary education	14–16 Full cohort	2011–2012
Hong Kong (China)	Hong Kong Certificate of Education Examination (HKCEE)	Qualification awarded Controls entry to upper secondary education Superseded by HKDSE from 2012	16 Full cohort	2010
Hungary	National Assessment of Basic Competence (NABC) – Grade 10	No qualification awarded Provides schools and teachers with student performance data	16 Full cohort (but only a sample are centrally marked)	2011
Japan	National Assessment of Academic Ability (NAAA) – Lower Secondary Year 3 Maths	No qualification awarded Taken only by a sample of students to provide national and regional performance data	14–15 30% of cohort take test	2012

Jurisdiction / awarding organisation	Assessment	Use	Cohort age and proportion taking part	Year of papers used
Massachusetts (USA)	Massachusetts Comprehensive Assessment System (MCAS) – Grade 10 Mathematics	No qualification awarded Required for high-school graduation (at 18 yrs old)	15–16 Full cohort	2011
Netherlands	VMBO TL/GL	Qualification awarded Controls entry to upper secondary education	15–16 Full cohort (one of three options available)	2011
New Zealand	National Certificate of Educational Achievement (NCEA) Level 1	Qualification awarded Controls entry to upper secondary education	16 Full cohort	2011
Ontario (Canada)	Grade 9 Assessment of Mathematics – Academic and Applied papers	No qualification awarded Provides student-level data to monitor progress	14-15 Full cohort	2012
Scotland – SQA	Standard Grade	Qualification awarded Controls entry to upper secondary education	14–16 Full cohort	2011
Shanghai (China)	Zhong Kao - Junior High School Joint Graduation and Academic Examination – Mathematics Exam	No qualification awarded Controls entry to upper secondary education	14–15 Full cohort	2011

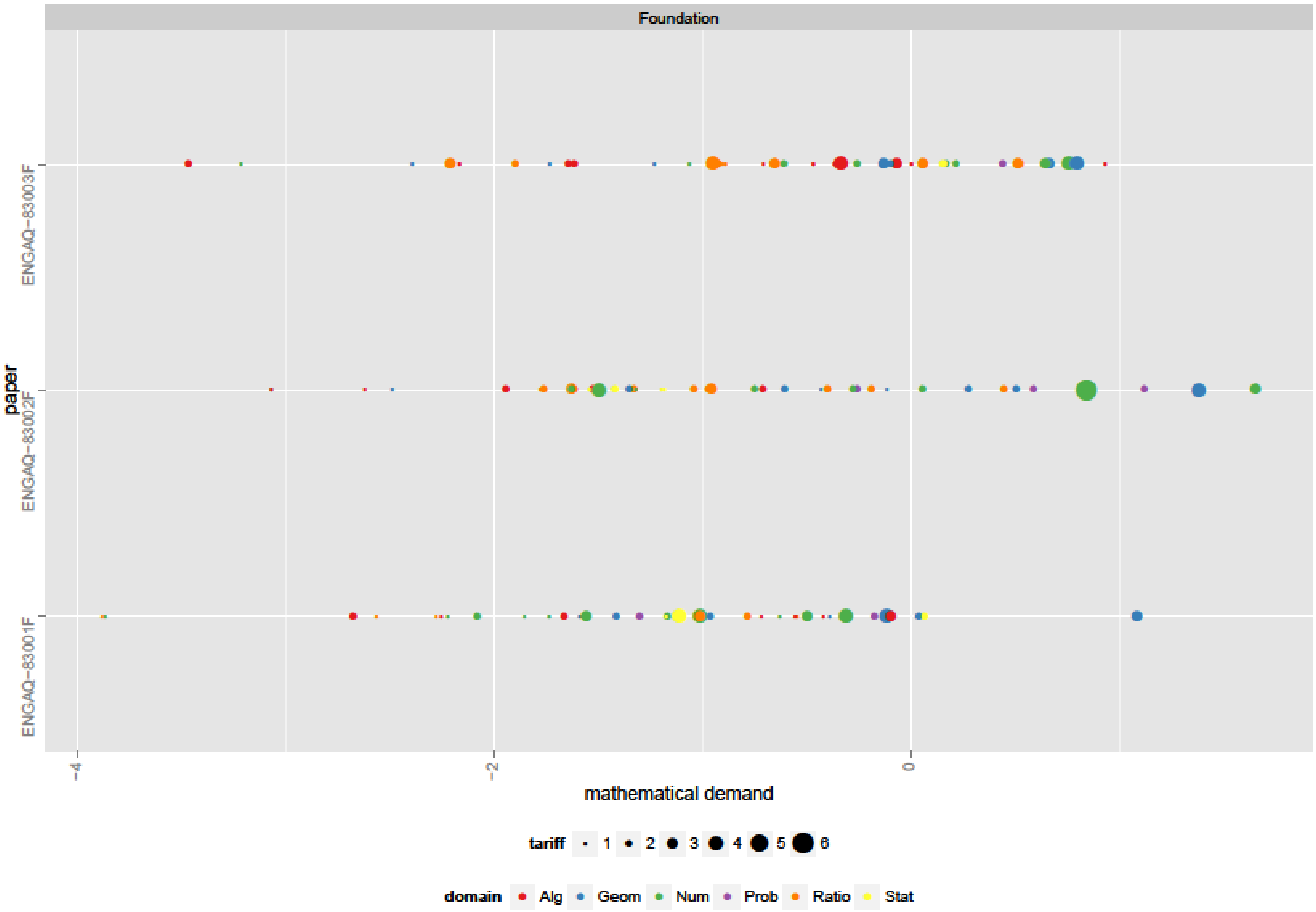
Jurisdiction / awarding organisation	Assessment	Use	Cohort age and proportion taking part	Year of papers used
South Korea	National Assessment of Educational Achievement (NAEA) – 9th Grade Mathematics	No qualification awarded Provides national performance data	15 0.5-1.0% of cohort tested	2011

Appendix B: Additional study 1 analysis

Analysis of individual questions on English GCSE papers from AQA, OCR and Pearson, by domain and tariff

The following pages show plots and tables for Rasch expected difficulty parameter, mathematical domain and tariff.

AQA SAMs – foundation tier

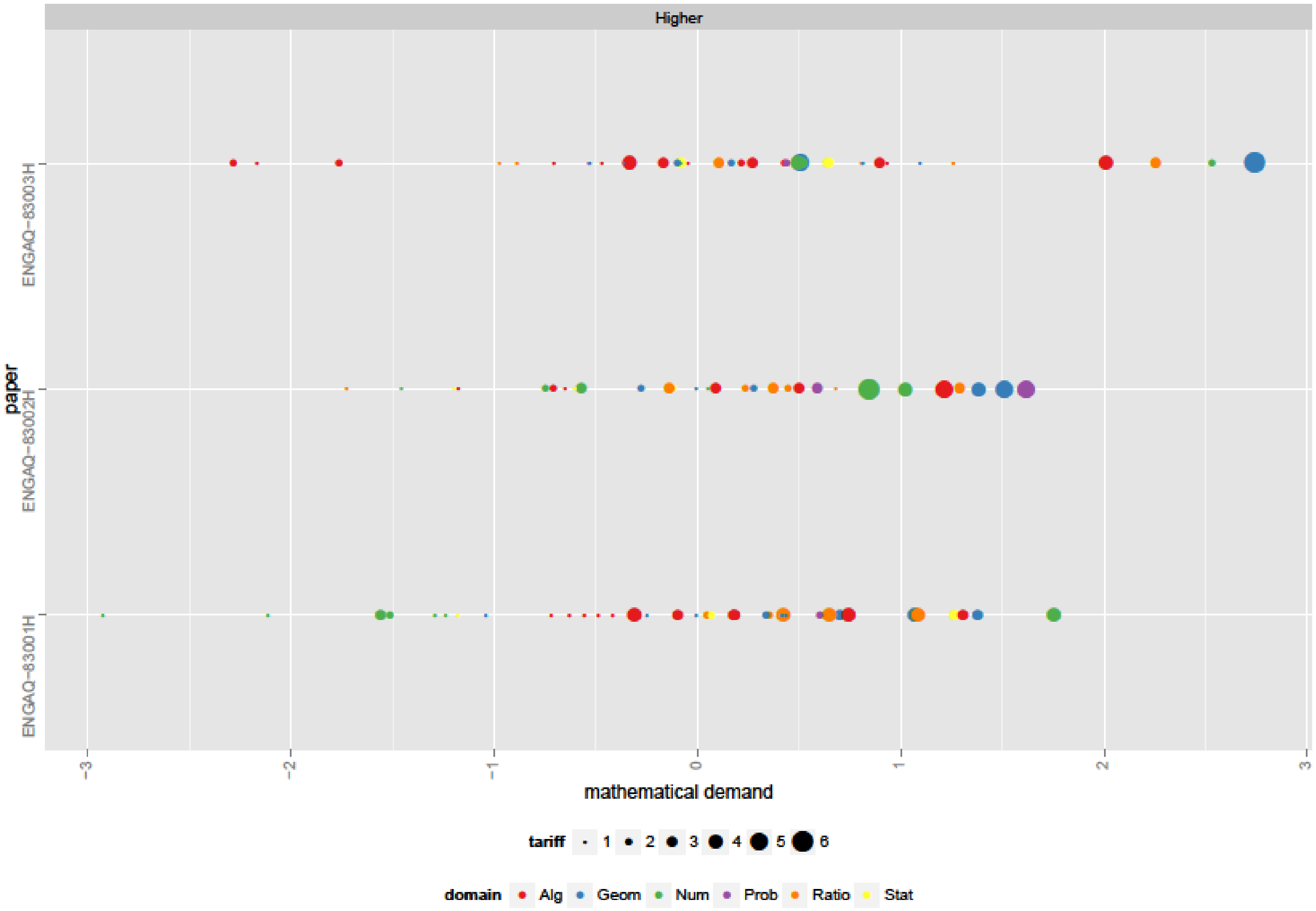


Paper 83001F				Paper 83002F				Paper 83003F			
Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff
Q_26	1.09	Geom	3	Q_14	1.66	Num	3	Q_28c	0.93	Alg	1
Q_23c	0.07	Stat	2	Q_30	1.38	Geom	4	Q_21	0.80	Geom	4
Q_19b	0.04	Geom	2	Q_8b	1.12	Prob	2	Q_19	0.76	Num	4
Q_28	-0.10	Alg	3	Q_25	0.84	Num	6	Q_31	0.66	Geom	3
Q_27	-0.10	Geom	3	Q_22	0.59	Prob	2	Q_29	0.65	Num	3
Q_29	-0.12	Geom	4	Q_29a	0.50	Geom	2	Q_26	0.52	Ratio	3
Q_15c	-0.12	Num	1	Q_24b	0.45	Ratio	2	Q_24b	0.44	Prob	2
Q_15a	-0.14	Num	1	Q_29b	0.27	Geom	2	Q_20	0.22	Num	2
Q_5	-0.18	Prob	2	Q_18	0.06	Num	2	Q_10	0.17	Num	2
Q_23d	-0.31	Stat	1	Q_16	-0.11	Geom	1	Q_7	0.16	Stat	2
Q_25b	-0.31	Num	4	Q_12b	-0.19	Ratio	2	Q_18a	0.06	Ratio	3
Q_23a	-0.33	Ratio	1	Q_8a	-0.26	Prob	2	Q_11c	0.01	Alg	1
Q_16a	-0.39	Geom	1	Q_23	-0.28	Num	2	Q_27b	-0.07	Alg	3

Q_24c	-0.42	Alg	1	Q_12a	-0.40	Ratio	2	Q_25	-0.10	Geom	2
Q_21	-0.50	Num	3	Q_17a	-0.43	Geom	1	Q_13	-0.13	Geom	3
Q_17a	-0.54	Ratio	1	Q_26	-0.61	Geom	2	Q_6	-0.26	Num	2
Q_24b	-0.56	Alg	1	Q_28	-0.71	Alg	2	Q_30	-0.33	Alg	4
Q_25a	-0.63	Num	1	Q_27	-0.75	Num	2	Q_24a	-0.36	Prob	1
Q_24d	-0.72	Alg	1	Q_31	-0.96	Ratio	3	Q_11b	-0.37	Alg	1
Q_17b	-0.78	Ratio	2	Q_21	-0.97	Ratio	2	Q_11a	-0.47	Alg	1
Q_11	-0.96	Geom	2	Q_15	-1.04	Ratio	2	Q_28b	-0.47	Alg	1
Q_18	-0.99	Num	1	Q_7a	-1.18	Stat	1	Q_16	-0.61	Num	2
Q_16b	-1.00	Geom	2	Q_24a	-1.19	Stat	1	Q_12	-0.65	Ratio	3
Q_9	-1.01	Ratio	3	Q_4	-1.32	Num	1	Q_28a	-0.71	Alg	1
Q_15b	-1.01	Num	4	Q_10	-1.33	Ratio	2	Q_23	-0.89	Ratio	1
Q_4	-1.11	Stat	4	Q_13	-1.35	Geom	2	Q_18b	-0.90	Ratio	1
Q_14	-1.17	Num	2	Q_7b	-1.42	Stat	2	Q_17	-0.92	Ratio	2
Q_23b	-1.18	Stat	1	Q_6	-1.50	Num	4	Q_9b	-0.95	Ratio	4
Q_8	-1.30	Prob	2	Q_11a	-1.53	Alg	2	Q_3	-1.06	Num	1
Q_10	-1.42	Geom	2	Q_7c	-1.54	Stat	1	Q_8a	-1.23	Geom	1

Q_22	-1.56	Num	3	Q_11b	-1.62	Alg	2	Q_15	-1.62	Alg	2
Q_24a	-1.58	Alg	1	Q_19	-1.63	Ratio	3	Q_27a	-1.65	Alg	2
Q_19a	-1.59	Geom	1	Q_5	-1.63	Num	2	Q_8b	-1.74	Geom	1
Q_20	-1.66	Alg	2	Q_20	-1.76	Ratio	2	Q_14	-1.90	Ratio	2
Q_12b	-1.74	Num	1	Q_9	-1.78	Ratio	1	Q_22	-2.17	Alg	1
Q_12a	-1.86	Num	1	Q_11c	-1.95	Alg	2	Q_4	-2.20	Num	1
Q_6	-2.08	Num	2	Q_17b	-2.49	Geom	1	Q_9a	-2.21	Ratio	3
Q_12c	-2.22	Num	1	Q_3	-2.62	Alg	1	Q_2	-2.40	Geom	1
Q_3	-2.25	Alg	1	Q_1	-3.07	Num	1	Q_1	-3.22	Num	1
Q_2	-2.28	Ratio	1	Q_2	-3.07	Alg	1	Q_5	-3.47	Alg	2
Q_1a	-2.57	Ratio	1								
Q_13	-2.68	Alg	2								
Q_7	-3.87	Num	1								
Q_1b	-3.88	Ratio	1								

AQA SAMs – higher tier

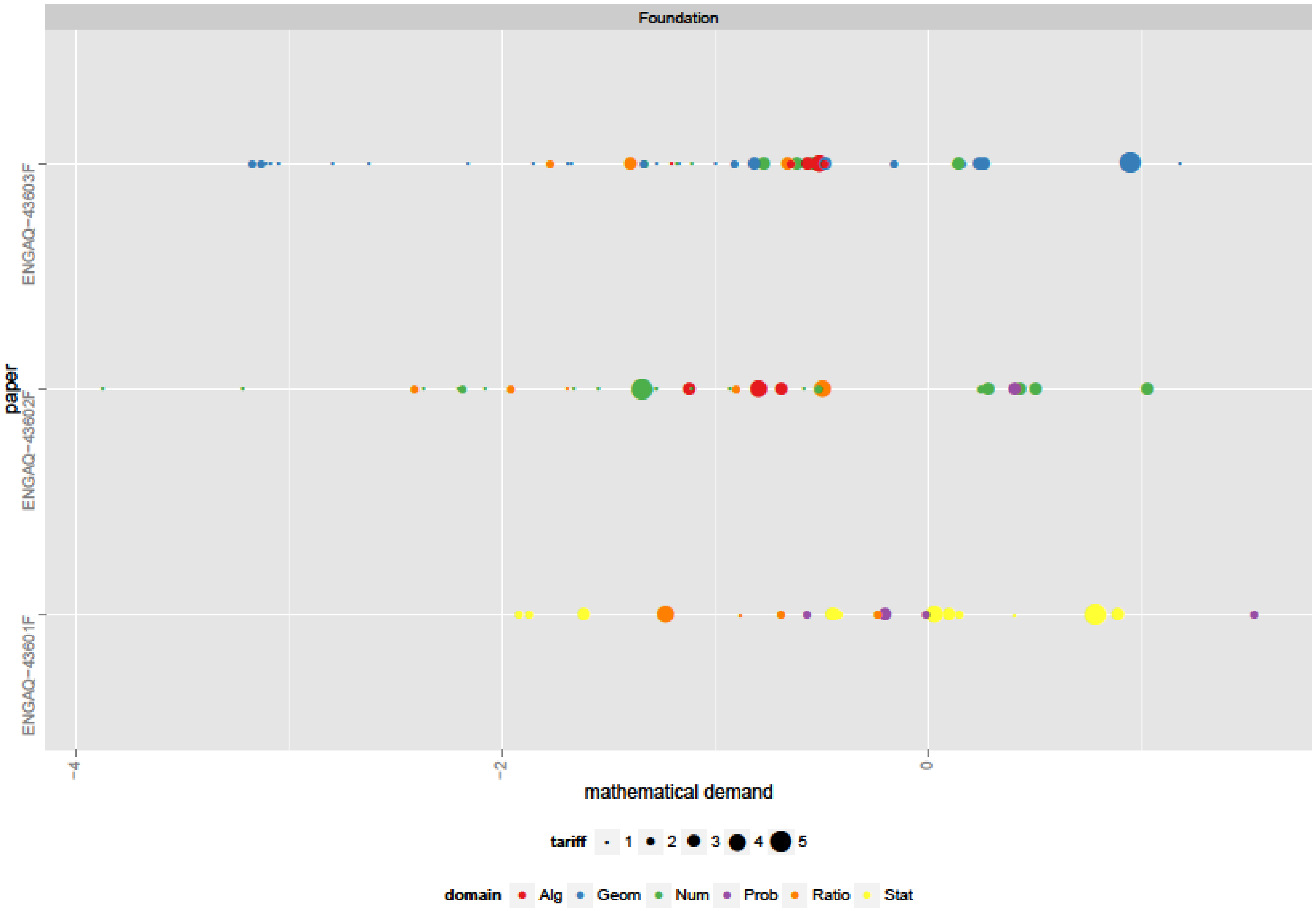


Paper 83001H				Paper 83002H				Paper 83003H			
Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff
Q_22	1.75	Num	4	Q_22	1.62	Prob	5	Q_27	2.74	Geom	6
Q_25a	1.38	Geom	3	Q_21	1.51	Geom	5	Q_21a	2.53	Num	2
Q_24b	1.31	Alg	3	Q_12	1.38	Geom	4	Q_20b	2.26	Ratio	3
Q_11	1.27	Stat	3	Q_23a	1.29	Ratio	3	Q_26b	2.01	Alg	4
Q_18	1.09	Ratio	4	Q_24	1.21	Alg	5	Q_21b	1.26	Ratio	1
Q_17	1.07	Geom	4	Q_23b	1.02	Num	4	Q_24b	1.09	Geom	1
Q_16	0.74	Alg	4	Q_18a	0.87	Stat	3	Q_10c	0.93	Alg	1
Q_15	0.70	Geom	3	Q_8	0.84	Num	6	Q_14	0.90	Alg	3
Q_19	0.65	Ratio	4	Q_19c	0.68	Ratio	1	Q_24d	0.81	Geom	1
Q_24a	0.60	Prob	2	Q_7	0.59	Prob	3	Q_20a	0.81	Ratio	1
Q_23c	0.43	Geom	1	Q_20b	0.58	Alg	2	Q_11	0.65	Stat	3
Q_13a	0.42	Ratio	4	Q_11a	0.50	Geom	2	Q_9	0.52	Num	3
Q_14b	0.41	Geom	1	Q_20a	0.50	Alg	3	Q_16	0.51	Geom	5
Q_14a	0.35	Ratio	2	Q_6b	0.45	Ratio	2	Q_25	0.49	Num	4
Q_25b	0.33	Geom	2	Q_14	0.37	Ratio	3	Q_7b	0.44	Prob	2

Q_21	0.18	Alg	3	Q_11b	0.27	Geom	2	Q_26a	0.43	Alg	2
Q_12	0.18	Alg	3	Q_19a	0.24	Ratio	2	Q_19b	0.28	Stat	2
Q_5b	0.07	Stat	2	Q_16	0.09	Alg	3	Q_23	0.27	Alg	3
Q_13b	0.05	Ratio	2	Q_23c	0.05	Num	1	Q_6	0.22	Alg	2
Q_23b	-0.01	Geom	1	Q_4	-0.01	Geom	1	Q_22	0.17	Geom	2
Q_10	-0.10	Alg	3	Q_17	-0.12	Stat	2	Q_18	0.11	Ratio	3
Q_23a	-0.25	Geom	1	Q_19b	-0.14	Ratio	3	Q_13	-0.05	Alg	1
Q_5c	-0.31	Stat	1	Q_5	-0.28	Geom	2	Q_19a	-0.08	Stat	3
Q_9b	-0.31	Alg	4	Q_15	-0.57	Num	3	Q_24a	-0.08	Geom	1
Q_7b	-0.42	Alg	1	Q_18b	-0.60	Stat	1	Q_8	-0.10	Geom	2
Q_2	-0.49	Alg	1	Q_13	-0.65	Alg	1	Q_17	-0.17	Alg	3
Q_7a	-0.56	Alg	1	Q_10	-0.71	Alg	2	Q_12	-0.33	Alg	4
Q_9a	-0.63	Alg	1	Q_9	-0.75	Num	2	Q_7a	-0.36	Prob	1
Q_7c	-0.72	Alg	1	Q_3	-1.18	Alg	1	Q_10b	-0.47	Alg	1
Q_8	-1.04	Geom	1	Q_6a	-1.19	Stat	1	Q_24c	-0.53	Geom	1
Q_5a	-1.18	Stat	1	Q_2	-1.46	Num	1	Q_2	-0.54	Prob	1
Q_20	-1.24	Num	1	Q_1	-1.73	Ratio	1	Q_10a	-0.71	Alg	1

Q_3	-1.29	Num	1					Q_3	-0.89	Ratio	1
Q_4	-1.51	Num	2					Q_4	-0.97	Ratio	1
Q_6	-1.56	Num	3					Q_15	-1.76	Alg	2
Q_1b	-2.11	Num	1					Q_1	-2.17	Alg	1
Q_1a	-2.93	Num	1					Q_5	-2.28	Alg	2

AQA – old GCSE – foundation tier

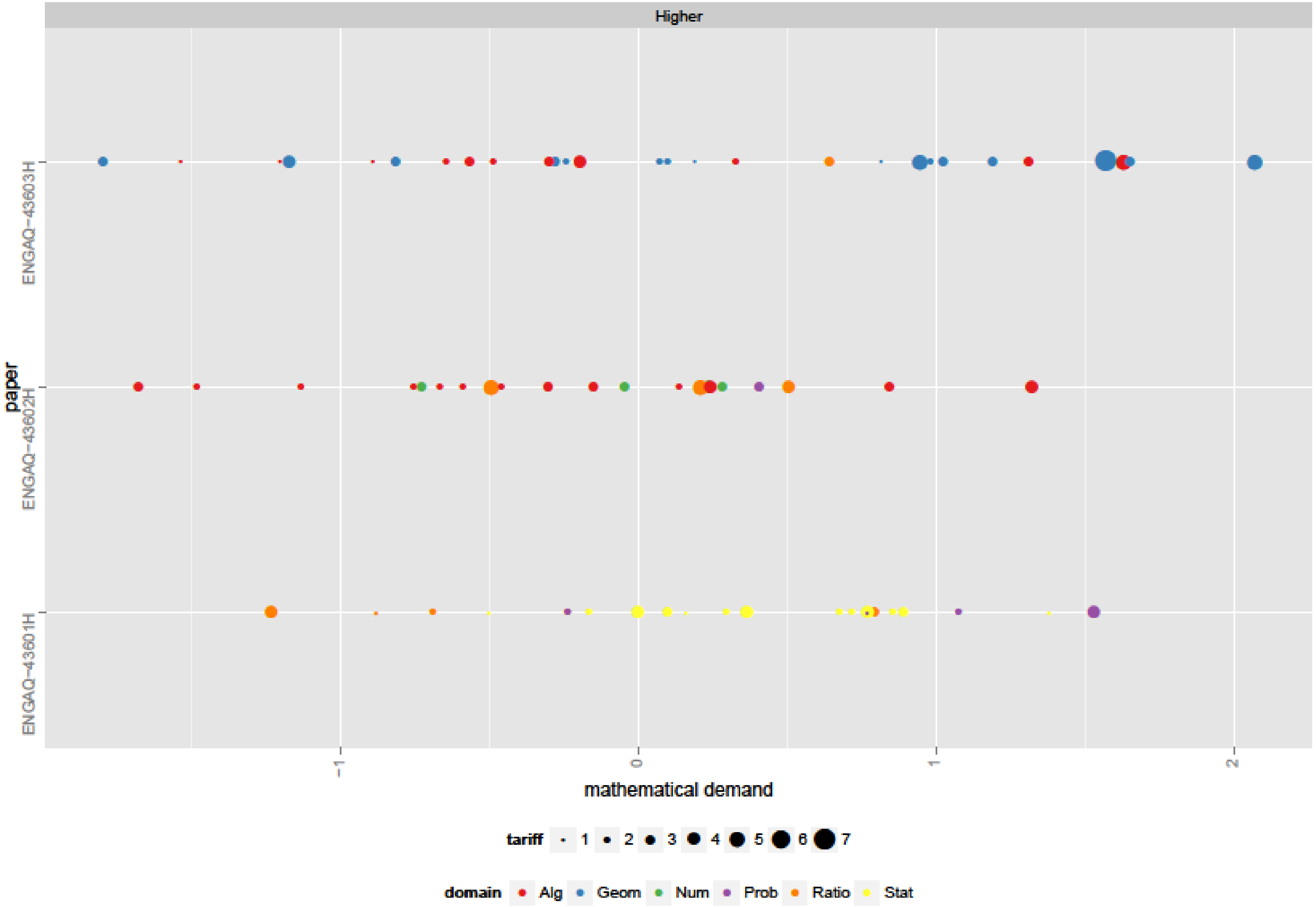


Paper 43601F				Paper 43602F				Paper 43603F			
Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff
Q_2	1.53	Prob	2	Q_5	1.03	Num	3	Q_12c	1.18	Geom	1
Q_9a	0.89	Stat	3	Q_13	0.50	Num	3	Q_19	0.94	Geom	5
Q_5a	0.78	Stat	5	Q_14	0.43	Num	3	Q_15	0.26	Geom	3
Q_5b	0.40	Stat	1	Q_15	0.41	Prob	3	Q_20	0.24	Geom	3
Q_1d	0.15	Stat	2	Q_16b	0.28	Num	3	Q_10	0.16	Geom	2
Q_9b	0.10	Stat	3	Q_16a	0.25	Num	2	Q_6b	0.14	Num	3
Q_1b	0.03	Stat	4	Q_10	-0.50	Ratio	4	Q_6a	0.13	Num	2
Q_4a	-0.01	Prob	2	Q_9c	-0.51	Num	1	Q_12b	-0.16	Geom	2
Q_4b	-0.20	Prob	3	Q_4	-0.52	Num	2	Q_9	-0.48	Geom	3
Q_8b	-0.24	Ratio	2	Q_9a	-0.58	Num	1	Q_18d	-0.49	Alg	2
Q_1aii	-0.42	Stat	2	Q_17b	-0.69	Alg	3	Q_14	-0.51	Alg	4
Q_1c	-0.44	Stat	3	Q_11b	-0.80	Alg	4	Q_11d	-0.55	Num	3
Q_3c	-0.45	Stat	3	Q_11a	-0.81	Alg	2	Q_18b	-0.57	Alg	3
Q_4c	-0.57	Prob	2	Q_3	-0.90	Ratio	2	Q_11c	-0.62	Num	3
Q_8c	-0.69	Ratio	2	Q_1b	-0.93	Num	1	Q_18a	-0.65	Alg	2

Q_8a	-0.88	Ratio	1	Q_9b	-1.11	Num	1	Q_8c	-0.66	Ratio	3
Q_7	-1.23	Ratio	4	Q_17a	-1.12	Alg	3	Q_11b	-0.77	Num	3
Q_1ai	-1.61	Stat	3	Q_2d	-1.28	Num	1	Q_17	-0.82	Geom	3
Q_3b	-1.87	Stat	2	Q_1c	-1.28	Num	1	Q_13b	-0.91	Geom	2
Q_3a	-1.92	Stat	2	Q_6	-1.31	Num	2	Q_7b	-1.00	Geom	1
				Q_2c	-1.33	Num	1	Q_8a	-1.11	Num	1
				Q_12	-1.34	Num	5	Q_12a	-1.17	Geom	1
				Q_1d	-1.55	Num	1	Q_8b	-1.18	Num	1
				Q_1a	-1.67	Num	1	Q_18c	-1.21	Alg	1
				Q_7d	-1.70	Ratio	1	Q_2d	-1.28	Geom	1
				Q_8c	-1.96	Ratio	2	Q_11a	-1.33	Num	2
				Q_1e	-2.08	Num	1	Q_13a	-1.34	Geom	2
				Q_7c	-2.19	Num	2	Q_3b	-1.40	Ratio	3
				Q_2a	-2.21	Num	1	Q_2a	-1.68	Geom	1
				Q_8a	-2.37	Num	1	Q_1b	-1.70	Geom	1
				Q_8b	-2.41	Ratio	2	Q_3a	-1.77	Ratio	2
				Q_7b	-2.41	Num	1	Q_2b	-1.85	Geom	1

				Q_7a	-3.22	Num	1	Q_2c	-2.16	Geom	1
				Q_2b	-3.87	Num	1	Q_4b	-2.63	Geom	1
								Q_1a	-2.80	Geom	1
								Q_5a	-3.05	Geom	1
								Q_5b	-3.09	Geom	1
								Q_4a	-3.11	Geom	1
								Q_7a	-3.13	Geom	2
								Q_16	-3.17	Geom	2

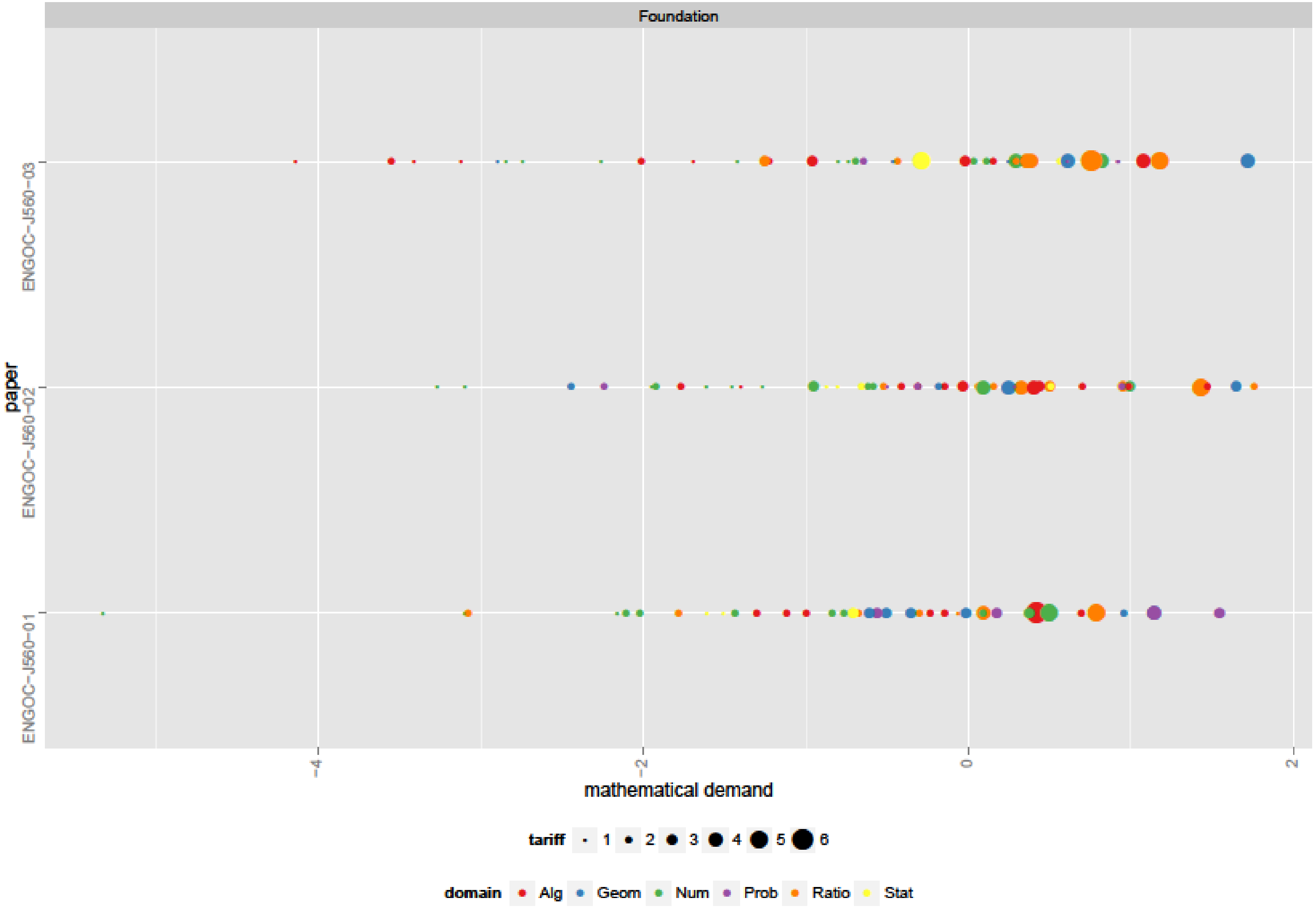
AQA – old GCSE – higher tier



Paper 43601H				Paper 43602H				Paper 43603H			
Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff
Q_8	1.53	Prob	4	Q_12	1.32	Alg	4	Q_10	2.07	Geom	5
Q_1bii	1.38	Stat	1	Q_15b	0.84	Alg	3	Q_5	1.65	Geom	3
Q_7b	1.07	Prob	2	Q_11	0.51	Ratio	4	Q_13	1.63	Alg	5
Q_4a	0.89	Stat	3	Q_4	0.41	Prob	3	Q_18	1.57	Geom	7
Q_5b	0.86	Stat	2	Q_7b	0.28	Num	3	Q_15	1.31	Alg	3
Q_3d	0.79	Ratio	3	Q_7a	0.25	Num	2	Q_14a	1.19	Geom	3
Q_6c	0.77	Stat	4	Q_9	0.24	Alg	4	Q_17	1.02	Geom	3
Q_7a	0.77	Prob	1	Q_14	0.24	Num	4	Q_19b	0.98	Geom	2
Q_5c	0.72	Stat	2	Q_6	0.21	Ratio	5	Q_8	0.94	Geom	5
Q_6aii	0.68	Stat	2	Q_10d	0.14	Alg	2	Q_14b	0.81	Geom	1
Q_4c	0.37	Stat	2	Q_2	-0.05	Num	3	Q_3	0.64	Ratio	3
Q_1bi	0.36	Stat	4	Q_5b	-0.15	Alg	3	Q_11c	0.33	Alg	2
Q_6b	0.30	Stat	2	Q_13	-0.30	Alg	3	Q_6b	0.19	Geom	1
Q_5ai	0.16	Stat	1	Q_8b	-0.46	Alg	2	Q_6a	0.10	Geom	2
Q_4b	0.10	Stat	3	Q_3	-0.49	Ratio	5	Q_6c	0.07	Geom	2

Q_1a	0.00	Stat	4	Q_10c	-0.59	Alg	2	Q_19a	0.07	Geom	1
Q_6ai	-0.17	Stat	2	Q_10b	-0.67	Alg	2	Q_16a	-0.20	Alg	4
Q_3b	-0.24	Ratio	2	Q_15a	-0.73	Num	3	Q_9	-0.24	Geom	2
Q_7c	-0.24	Prob	2	Q_8a	-0.76	Alg	2	Q_12	-0.28	Geom	3
Q_5aii	-0.50	Stat	1	Q_10a	-1.13	Alg	2	Q_16b	-0.30	Alg	3
Q_3c	-0.69	Ratio	2	Q_5a	-1.48	Alg	2	Q_2d	-0.49	Alg	2
Q_3a	-0.88	Ratio	1	Q_1	-1.68	Alg	3	Q_2b	-0.57	Alg	3
Q_2	-1.23	Ratio	4					Q_2a	-0.65	Alg	2
								Q_7	-0.82	Geom	3
								Q_11a	-0.89	Alg	1
								Q_1	-1.17	Geom	4
								Q_2c	-1.21	Alg	1
								Q_11b	-1.54	Alg	1
								Q_4	-1.80	Geom	3

OCR SAMs – foundation tier

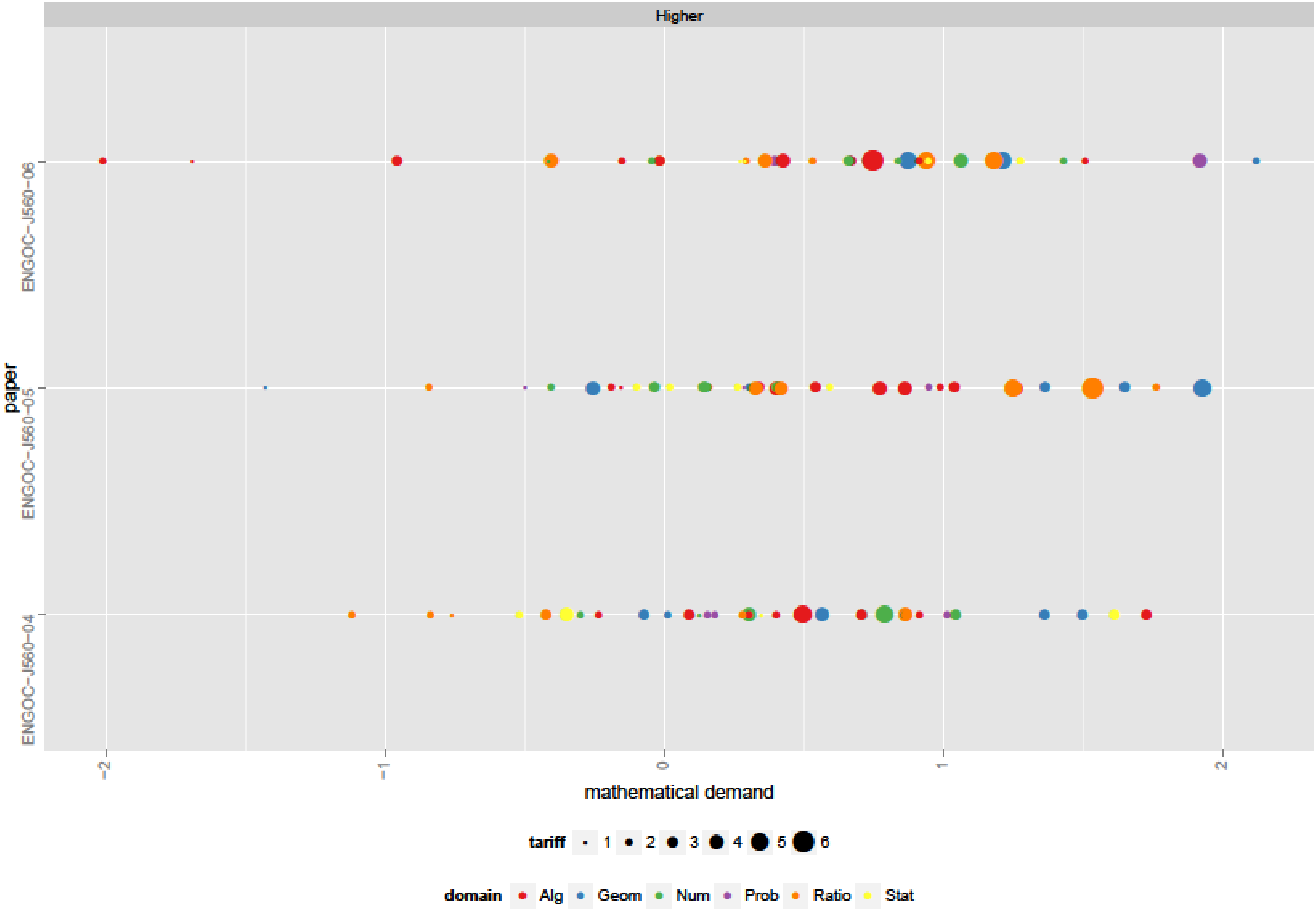


Paper J560-01				Paper J560-02				Paper J560-03			
Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff
Q_4b	1.55	Prob	3	Q_15c	1.76	Ratio	2	Q_6b	1.72	Geom	4
Q_16c	1.14	Geom	4	Q_14b	1.65	Geom	3	Q_14a	1.20	Prob	2
Q_21	1.14	Prob	4	Q_9c	1.47	Alg	2	Q_18a	1.18	Ratio	5
Q_16b	0.96	Geom	2	Q_12c	1.43	Ratio	5	Q_8b	1.08	Alg	4
Q_19	0.79	Ratio	5	Q_8b	0.99	Num	3	Q_14b	0.92	Prob	1
Q_12c	0.70	Alg	2	Q_9b	0.99	Alg	2	Q_9aiii	0.92	Prob	1
Q_15	0.50	Num	5	Q_11bii	0.97	Stat	1	Q_3b	0.82	Num	4
Q_13b	0.49	Alg	4	Q_12b	0.95	Ratio	3	Q_4	0.76	Ratio	6
Q_12d	0.42	Alg	6	Q_16c	0.94	Prob	2	Q_7b	0.61	Geom	4
Q_3	0.38	Num	3	Q_18a	0.70	Alg	2	Q_9aii	0.61	Prob	1
Q_4a	0.17	Prob	3	Q_11biii	0.51	Stat	2	Q_12	0.57	Stat	2
Q_18	0.09	Ratio	4	Q_10b	0.50	Ratio	3	Q_9b	0.39	Ratio	4
Q_14c	0.09	Num	2	Q_13b	0.44	Alg	3	Q_19a	0.37	Geom	3
Q_20	-0.02	Geom	3	Q_19	0.40	Alg	4	Q_16b	0.36	Ratio	4
Q_17a	-0.06	Ratio	1	Q_15b	0.33	Ratio	4	Q_6c	0.35	Geom	3

Q_12b	-0.15	Alg	2	Q_14a	0.30	Geom	2	Q_18b	0.29	Ratio	1
Q_13aii	-0.24	Alg	2	Q_16b	0.28	Prob	1	Q_16a	0.29	Ratio	2
Q_1b	-0.30	Ratio	2	Q_6b	0.25	Geom	4	Q_15	0.29	Num	4
Q_16a	-0.36	Geom	3	Q_15a	0.16	Ratio	2	Q_3c	0.24	Geom	1
Q_5	-0.51	Geom	3	Q_11aii	0.10	Stat	1	Q_8a	0.15	Alg	2
Q_6	-0.56	Prob	3	Q_5	0.09	Num	4	Q_13bii	0.11	Num	2
Q_8	-0.61	Geom	3	Q_10a	0.06	Ratio	2	Q_13bi	0.03	Num	2
Q_17b	-0.68	Ratio	2	Q_17d	-0.03	Alg	3	Q_17c	-0.02	Alg	3
Q_9c	-0.71	Stat	3	Q_18b	-0.15	Alg	2	Q_2	-0.29	Stat	5
Q_14a	-0.76	Num	1	Q_9a	-0.19	Geom	2	Q_19b	-0.44	Ratio	2
Q_7b	-0.77	Num	2	Q_17c	-0.31	Alg	2	Q_7a	-0.47	Geom	1
Q_14b	-0.84	Num	2	Q_2b	-0.31	Prob	2	Q_9ai	-0.64	Prob	2
Q_12a	-1.00	Alg	2	Q_17b	-0.41	Alg	2	Q_3a	-0.70	Num	2
Q_13ai	-1.12	Alg	2	Q_16a	-0.50	Prob	1	Q_11a	-0.74	Num	1
Q_11	-1.30	Alg	2	Q_12a	-0.52	Ratio	2	Q_11b	-0.80	Num	1
Q_10b	-1.44	Num	2	Q_4c	-0.59	Num	2	Q_17ai	-0.96	Alg	3
Q_9b	-1.50	Stat	1	Q_4d	-0.62	Num	2	Q_1bi	-1.23	Alg	2

Q_9a	-1.61	Stat	1	Q_11c	-0.65	Stat	2	Q_5	-1.25	Ratio	3
Q_1c	-1.78	Ratio	2	Q_11ai	-0.80	Stat	1	Q_13a	-1.42	Num	1
Q_2	-2.02	Num	2	Q_11bi	-0.87	Stat	1	Q_17aii	-1.69	Alg	1
Q_7a	-2.11	Num	2	Q_8a	-0.95	Num	3	Q_17b	-2.01	Alg	2
Q_10aiii	-2.16	Num	1	Q_7a	-1.27	Num	1	Q_10b	-2.26	Num	1
Q_1a	-3.08	Ratio	2	Q_17a	-1.40	Alg	1	Q_10c	-2.74	Num	1
Q_10aii	-3.10	Num	1	Q_3a	-1.46	Num	1	Q_10a	-2.84	Num	1
Q_10ai	-5.33	Num	1	Q_3b	-1.61	Num	1	Q_6a	-2.90	Geom	1
				Q_13a	-1.77	Alg	2	Q_1aiii	-3.12	Alg	1
				Q_4b	-1.92	Num	1	Q_1ai	-3.41	Alg	1
				Q_7b	-1.92	Num	2	Q_1bii	-3.55	Alg	2
				Q_4a	-1.95	Num	1	Q_1aii	-4.14	Alg	1
				Q_2a	-2.24	Prob	2				
				Q_6a	-2.45	Geom	2				
				Q_1a	-3.10	Num	1				
				Q_1b	-3.27	Num	1				

OCR SAMs – higher tier

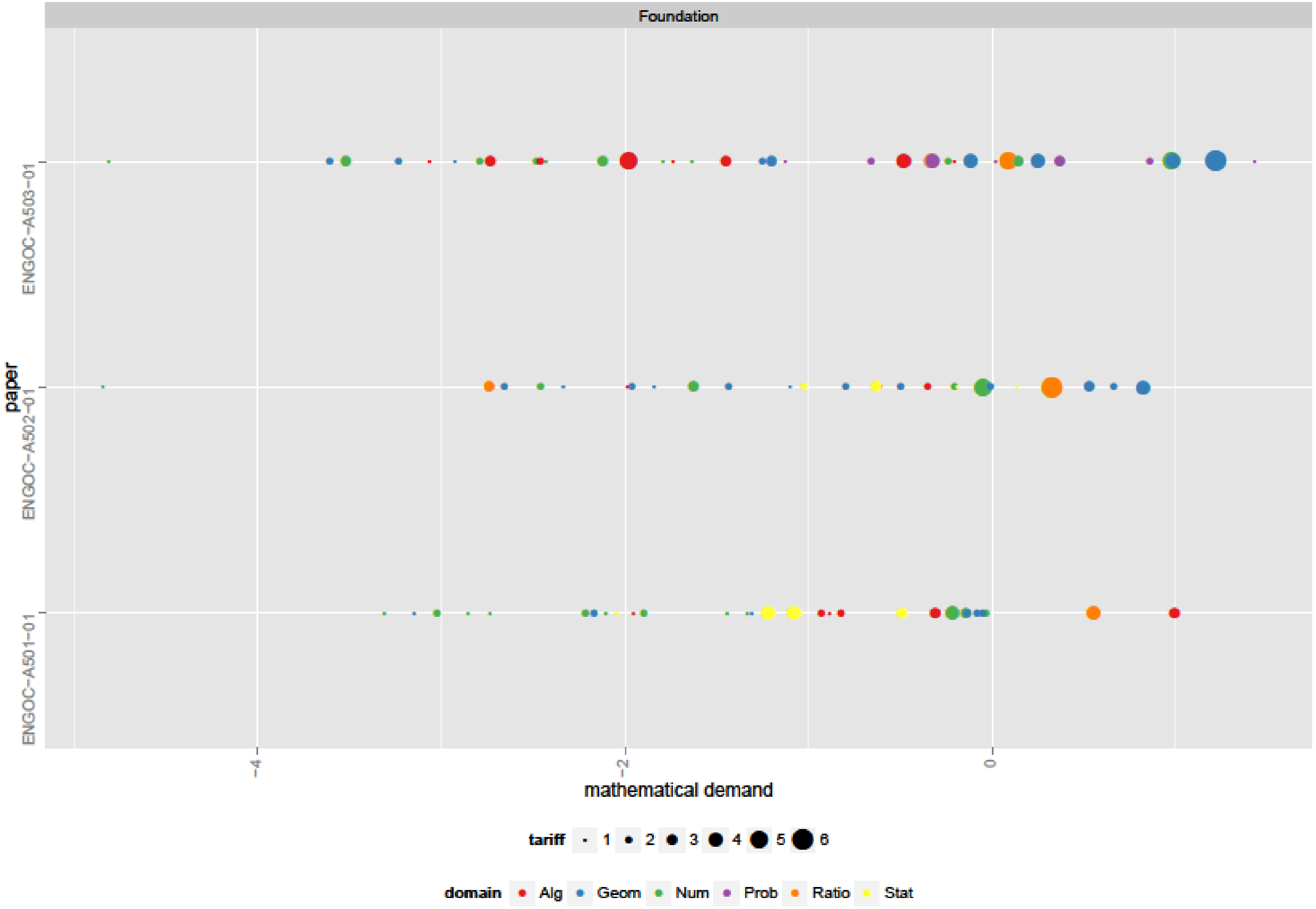


Paper J560-04				Paper J560-05				Paper J560-06			
Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff
Q_12b	1.73	Alg	3	Q_18	1.92	Geom	5	Q_12b	2.12	Geom	2
Q_14b	1.61	Stat	3	Q_2c	1.76	Ratio	2	Q_11b	1.92	Prob	4
Q_9a	1.50	Geom	3	Q_1b	1.65	Geom	3	Q_8bii	1.51	Alg	2
Q_16b	1.36	Geom	3	Q_19	1.53	Ratio	6	Q_5b	1.43	Num	2
Q_11b	1.04	Num	3	Q_12	1.36	Geom	3	Q_6aii	1.28	Stat	2
Q_7c	1.01	Prob	2	Q_10d	1.26	Alg	4	Q_6b	1.27	Stat	2
Q_20b	0.91	Alg	2	Q_4b	1.25	Ratio	5	Q_12aii	1.21	Geom	5
Q_19b	0.86	Ratio	3	Q_11b	1.04	Alg	3	Q_4a	1.18	Ratio	5
Q_15	0.86	Ratio	4	Q_20a	0.99	Alg	2	Q_5c	1.06	Num	4
Q_16a	0.85	Geom	2	Q_3c	0.94	Prob	2	Q_1c	0.94	Stat	2
Q_6	0.79	Num	5	Q_14	0.86	Alg	4	Q_13	0.94	Ratio	5
Q_17	0.71	Num	3	Q_20b	0.77	Alg	4	Q_8bi	0.91	Alg	2
Q_8b	0.71	Alg	3	Q_13d	0.59	Stat	2	Q_14	0.87	Geom	5
Q_18	0.56	Geom	4	Q_10c	0.55	Alg	1	Q_5aii	0.84	Num	2
Q_19a	0.50	Alg	2	Q_16	0.54	Alg	3	Q_17	0.75	Alg	6

Q_5	0.50	Alg	5	Q_9	0.42	Ratio	4	Q_12ai	0.73	Geom	3
Q_4b	0.49	Alg	4	Q_7	0.40	Num	3	Q_15	0.67	Alg	3
Q_8ai	0.40	Alg	2	Q_5	0.40	Alg	4	Q_16	0.66	Num	3
Q_14c	0.35	Stat	1	Q_11a	0.34	Alg	3	Q_10a	0.53	Ratio	2
Q_13	0.30	Num	4	Q_2b	0.33	Ratio	4	Q_9	0.42	Alg	4
Q_20a	0.30	Alg	2	Q_1a	0.30	Geom	2	Q_11a	0.40	Prob	3
Q_9b	0.28	Ratio	2	Q_3b	0.28	Prob	1	Q_10c	0.38	Ratio	2
Q_7a	0.18	Prob	2	Q_13a	0.26	Stat	2	Q_2b	0.36	Ratio	4
Q_7b	0.15	Prob	2	Q_2a	0.16	Ratio	2	Q_4b	0.29	Ratio	1
Q_11a	0.12	Num	1	Q_8	0.15	Num	3	Q_2a	0.29	Ratio	2
Q_1a	0.09	Geom	2	Q_6	0.14	Num	3	Q_6ai	0.29	Stat	1
Q_3c	0.09	Ratio	2	Q_13c	0.02	Stat	2	Q_1b	0.27	Stat	1
Q_8aii	0.09	Alg	3	Q_15b	-0.04	Num	3	Q_1a	-0.02	Stat	3
Q_1b	0.01	Geom	2	Q_13b	-0.10	Stat	2	Q_3c	-0.02	Alg	3
Q_10	-0.07	Geom	3	Q_10b	-0.16	Alg	1	Q_7	-0.05	Num	2
Q_4aii	-0.24	Alg	2	Q_10a	-0.19	Alg	2	Q_5ai	-0.15	Alg	2
Q_12a	-0.30	Num	2	Q_17b	-0.26	Geom	4	Q_10b	-0.41	Ratio	4

Q_14aii	-0.35	Stat	4	Q_15a	-0.41	Num	2	Q_8a	-0.42	Num	1
Q_2	-0.42	Ratio	3	Q_3a	-0.50	Prob	1	Q_3ai	-0.96	Alg	3
Q_14ai	-0.52	Stat	2	Q_4a	-0.84	Ratio	2	Q_3aii	-1.69	Alg	1
Q_3a	-0.76	Ratio	1	Q_17a	-1.43	Geom	1	Q_3b	-2.01	Alg	2
Q_3b	-0.84	Ratio	2								
Q_4ai	-1.12	Ratio	2								

OCR – old GCSE – foundation tier

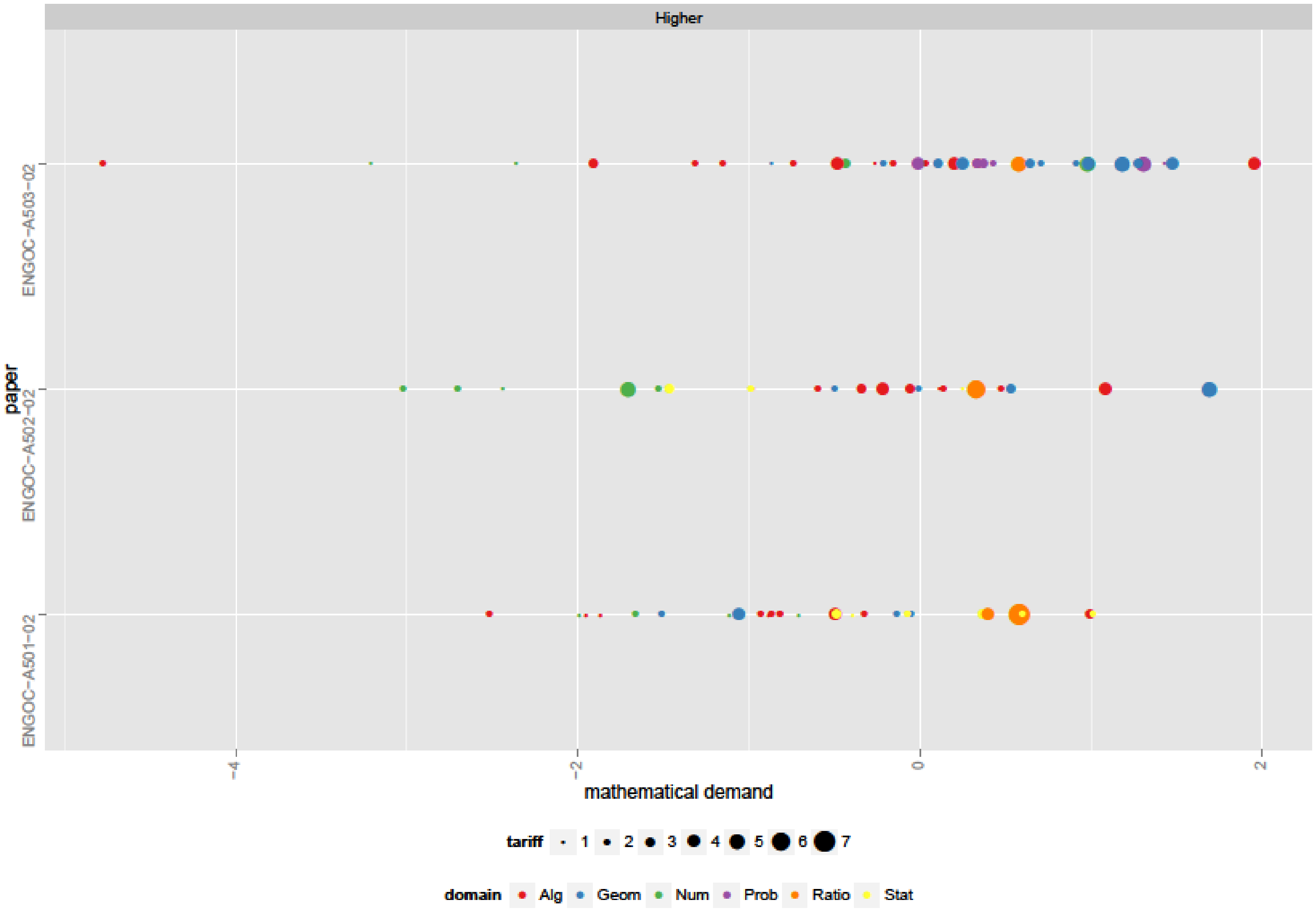


Paper A501-01				Paper A502-01				Paper A503-01			
Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff
Q_10b	0.99	Alg	3	Q_6cii	0.82	Geom	4	Q_20a	1.43	Prob	1
Q_8	0.55	Ratio	4	Q_6ci	0.66	Geom	2	Q_15	1.22	Geom	6
Q_6b	-0.03	Num	2	Q_9a	0.53	Geom	3	Q_18b	0.98	Geom	4
Q_9b	-0.06	Geom	2	Q_10a	0.33	Ratio	6	Q_17	0.98	Num	5
Q_4d	-0.08	Geom	2	Q_8biii	0.14	Stat	1	Q_5b	0.86	Prob	2
Q_9a	-0.14	Geom	2	Q_9b	-0.01	Geom	2	Q_20b	0.37	Prob	3
Q_4b	-0.14	Num	3	Q_4b	-0.05	Num	5	Q_18a	0.25	Geom	4
Q_6a	-0.22	Num	4	Q_8bii	-0.19	Stat	1	Q_7b	0.14	Num	3
Q_4a	-0.31	Alg	3	Q_8biv	-0.19	Stat	1	Q_16	0.09	Ratio	5
Q_11b	-0.49	Stat	3	Q_7	-0.21	Num	2	Q_8c	0.02	Prob	1
Q_12b	-0.82	Alg	2	Q_5c	-0.35	Alg	2	Q_3a	-0.12	Geom	4
Q_12aii	-0.88	Alg	1	Q_10b	-0.50	Geom	2	Q_6c	-0.20	Alg	1
Q_10a	-0.93	Alg	2	Q_5b	-0.62	Alg	2	Q_7a	-0.24	Num	2
Q_11a	-1.04	Stat	1	Q_8a	-0.63	Stat	3	Q_5a	-0.32	Prob	4
Q_5b	-1.08	Stat	4	Q_3c	-0.80	Geom	2	Q_8b	-0.33	Prob	2

Q_4c	-1.22	Stat	4	Q_8bi	-1.02	Stat	2	Q_13	-0.33	Ratio	4
Q_2b	-1.31	Geom	1	Q_3d	-1.10	Geom	1	Q_14b	-0.47	Prob	3
Q_7a	-1.33	Num	1	Q_6a	-1.44	Geom	2	Q_19	-0.48	Alg	4
Q_7b	-1.44	Num	1	Q_1c	-1.63	Num	3	Q_8a	-0.66	Prob	2
Q_3c	-1.90	Num	2	Q_3b	-1.84	Geom	1	Q_14a	-1.13	Prob	1
Q_12ai	-1.95	Alg	1	Q_4a	-1.96	Geom	2	Q_9b	-1.20	Geom	3
Q_5a	-2.04	Stat	1	Q_5a	-1.99	Alg	1	Q_2a	-1.25	Geom	2
Q_1b	-2.10	Num	1	Q_3a	-2.34	Geom	1	Q_10a	-1.45	Alg	3
Q_2c	-2.17	Geom	2	Q_1b	-2.46	Num	2	Q_4d	-1.64	Num	1
Q_3b	-2.22	Num	2	Q_6b	-2.66	Geom	2	Q_12b	-1.74	Alg	1
Q_1c	-2.73	Num	1	Q_2a	-2.72	Num	1	Q_4b	-1.79	Num	1
Q_1a	-2.85	Num	1	Q_2b	-2.74	Ratio	3	Q_12a	-1.98	Alg	5
Q_1d	-3.02	Num	2	Q_1a	-4.84	Num	1	Q_11a	-2.12	Num	3
Q_2a	-3.15	Geom	1					Q_4c	-2.43	Num	1
Q_3a	-3.31	Num	1					Q_6b	-2.46	Alg	2
								Q_1b	-2.48	Num	2
								Q_10b	-2.73	Alg	3

								Q_11b	-2.79	Num	2
								Q_2b	-2.93	Geom	1
								Q_6a	-3.06	Alg	1
								Q_3b	-3.23	Geom	2
								Q_1a	-3.52	Num	3
								Q_9a	-3.61	Geom	2
								Q_4a	-4.81	Num	1

OCR – old GCSE – higher tier

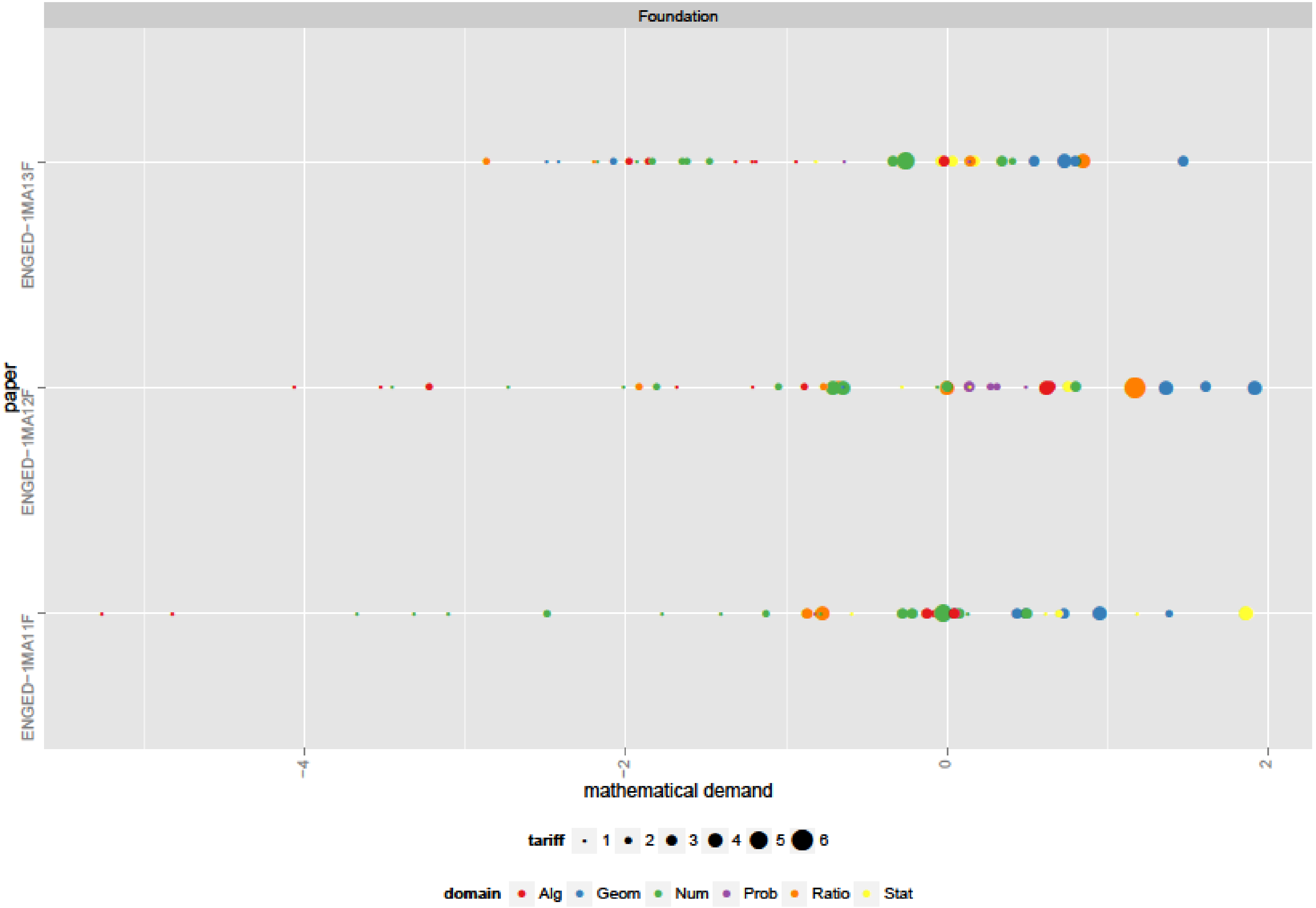


Paper J501-02				Paper J502-02				Paper J503-02			
Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff
Q_9bi	1.01	Stat	2	Q_4	1.69	Geom	5	Q_16	1.96	Alg	4
Q_3b	0.99	Alg	3	Q_7b	1.08	Alg	4	Q_15	1.48	Geom	4
Q_9bii	0.60	Stat	2	Q_6a	0.53	Geom	3	Q_7a	1.43	Prob	1
Q_12	0.58	Ratio	7	Q_10bi	0.47	Alg	2	Q_14	1.30	Prob	5
Q_4	0.40	Ratio	4	Q_1a	0.33	Ratio	6	Q_8c	1.27	Geom	3
Q_11	0.37	Stat	3	Q_3b	0.25	Stat	1	Q_17	1.18	Geom	5
Q_2b	-0.06	Geom	2	Q_2cii	0.14	Alg	2	Q_2b	0.98	Geom	4
Q_9aii	-0.07	Stat	2	Q_7a	0.12	Alg	1	Q_1	0.98	Num	5
Q_2a	-0.14	Geom	2	Q_6b	-0.01	Geom	2	Q_12c	0.91	Geom	2
Q_10bii	-0.33	Alg	2	Q_10bii	-0.06	Alg	3	Q_8b	0.71	Geom	2
Q_9ai	-0.39	Stat	1	Q_9	-0.22	Alg	4	Q_8a	0.64	Geom	3
Q_5b	-0.49	Stat	3	Q_2a	-0.34	Alg	3	Q_3	0.58	Ratio	5
Q_10a	-0.50	Alg	4	Q_1b	-0.50	Geom	2	Q_4b	0.43	Prob	2
Q_1b	-0.71	Num	1	Q_2b	-0.60	Alg	2	Q_7b	0.37	Prob	3
Q_6b	-0.82	Alg	2	Q_2ci	-0.98	Alg	1	Q_18	0.34	Alg	3

Q_7b	-0.87	Alg	2	Q_3a	-0.99	Stat	2	Q_4c	0.33	Prob	3
Q_6aii	-0.88	Alg	1	Q_8	-1.46	Stat	3	Q_2a	0.25	Geom	4
Q_3a	-0.93	Alg	2	Q_10aii	-1.53	Num	2	Q_11a	0.25	Alg	3
Q_5a	-1.04	Stat	1	Q_11a	-1.71	Num	5	Q_11b	0.20	Alg	4
Q_8b	-1.06	Geom	4	Q_10ai	-2.44	Num	1	Q_12a	0.10	Geom	3
Q_1ai	-1.12	Num	1	Q_5b	-2.71	Num	2	Q_10c	0.03	Alg	2
Q_8a	-1.51	Geom	2	Q_5a	-3.02	Num	2	Q_4a	-0.01	Prob	4
Q_1c	-1.67	Num	2					Q_10a	-0.16	Alg	2
Q_10bi	-1.87	Alg	1					Q_2c	-0.22	Geom	2
Q_6ai	-1.95	Alg	1					Q_10d	-0.26	Alg	1
Q_1aii	-1.99	Num	1					Q_13b	-0.44	Num	3
Q_7a	-2.52	Alg	2					Q_6	-0.48	Alg	4
								Q_10b	-0.74	Alg	2
								Q_12b	-0.87	Geom	1
								Q_9b	-1.15	Alg	2
								Q_9c	-1.32	Alg	2
								Q_5b	-1.91	Alg	3

								Q_9a	-2.36	Num	1
								Q_13a	-3.21	Num	1
								Q_5a	-4.78	Alg	2

Pearson SAMs – foundation tier

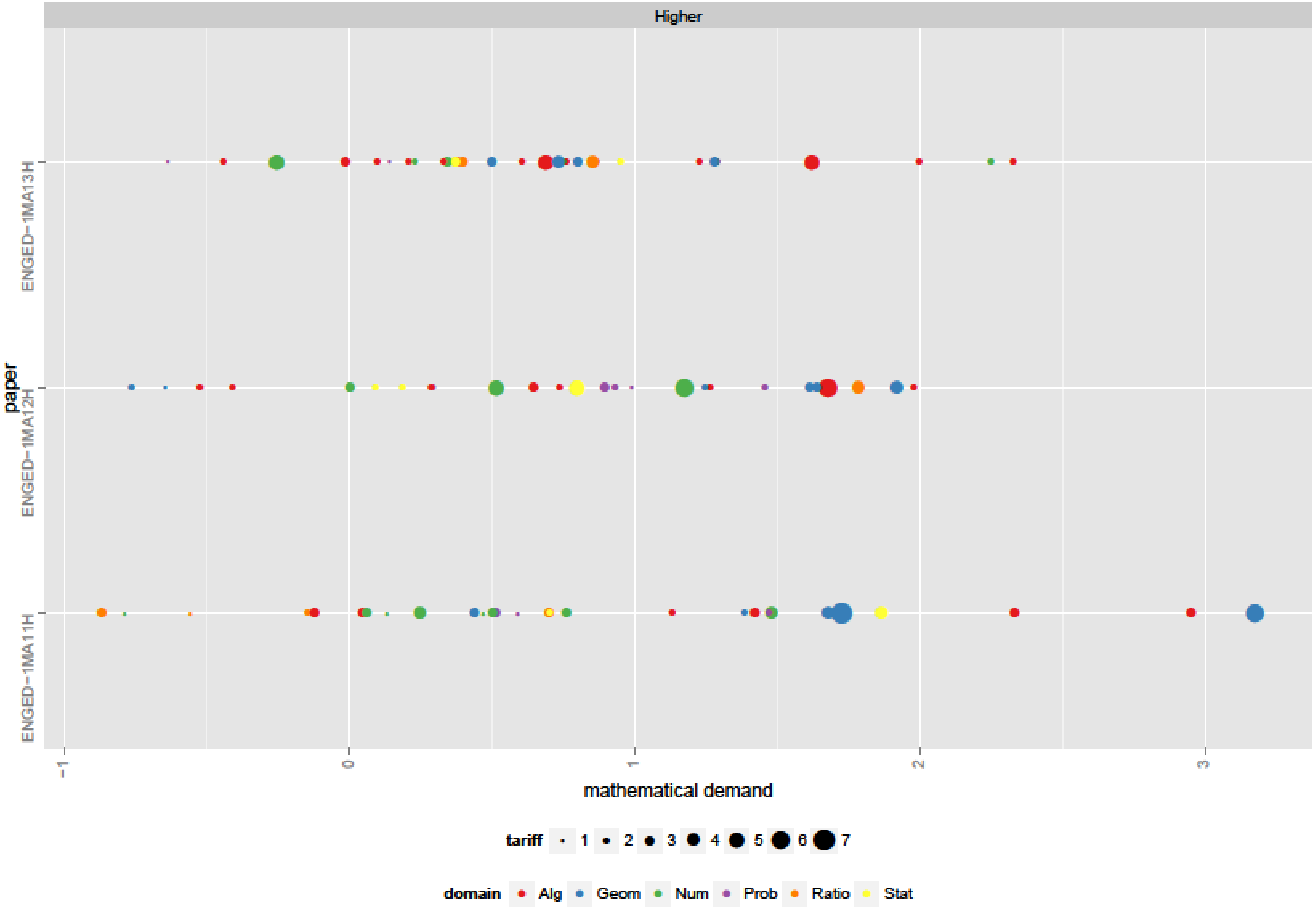


Paper 1MA11F				Paper 1MA12F				Paper 1MA13F			
Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff
Q_14a	1.87	Stat	4	Q_16b	1.92	Geom	4	Q_18	1.47	Geom	3
Q_18ai	1.38	Geom	2	Q_15	1.61	Geom	3	Q_13c	0.87	Prob	2
Q_20bii	1.19	Stat	1	Q_6	1.36	Geom	4	Q_16	0.85	Ratio	4
Q_9	0.95	Geom	4	Q_14b	1.17	Ratio	6	Q_17	0.80	Geom	3
Q_12	0.73	Geom	3	Q_12	0.80	Num	3	Q_14b	0.73	Geom	4
Q_14b	0.70	Stat	2	Q_5ai	0.76	Stat	3	Q_11a	0.54	Geom	3
Q_20bi	0.62	Stat	1	Q_13	0.65	Alg	3	Q_6b	0.41	Num	2
Q_18b	0.50	Num	3	Q_10	0.62	Alg	4	Q_12	0.34	Num	3
Q_7	0.49	Geom	3	Q_5bii	0.49	Prob	1	Q_11b	0.32	Geom	1
Q_15b	0.47	Num	1	Q_5bi	0.31	Prob	2	Q_19a	0.18	Stat	3
Q_16	0.44	Geom	3	Q_4b	0.27	Prob	2	Q_7c	0.15	Ratio	3
Q_18aii	0.13	Num	1	Q_5aii	0.14	Stat	1	Q_13b	0.14	Prob	1
Q_10b	0.08	Prob	2	Q_18	0.14	Prob	3	Q_19b	0.04	Stat	3
Q_11	0.08	Num	3	Q_17	0.00	Ratio	4	Q_14a	-0.01	Alg	3
Q_8	0.05	Geom	3	Q_14a	0.00	Num	3	Q_10b	-0.03	Stat	3

Q_17a	0.05	Alg	3	Q_8b	-0.06	Num	1	Q_9	-0.25	Num	3
Q_19b	0.00	Num	2	Q_11a	-0.27	Stat	1	Q_15	-0.26	Num	5
Q_6	-0.02	Num	5	Q_16aii	-0.65	Geom	1	Q_4b	-0.34	Num	3
Q_4c	-0.07	Alg	2	Q_8a	-0.65	Num	4	Q_13a	-0.64	Prob	1
Q_17b	-0.12	Alg	3	Q_11b	-0.66	Stat	1	Q_10a	-0.81	Stat	1
Q_3b	-0.22	Num	3	Q_4c	-0.67	Ratio	3	Q_3b	-0.94	Alg	1
Q_5	-0.28	Num	3	Q_7	-0.71	Num	4	Q_5d	-1.19	Alg	1
Q_20a	-0.59	Stat	1	Q_16ai	-0.77	Ratio	2	Q_5a	-1.21	Alg	1
Q_19c	-0.77	Ratio	4	Q_3d	-0.89	Alg	2	Q_3a	-1.32	Alg	1
Q_15a	-0.79	Num	1	Q_1b	-1.05	Num	2	Q_4a	-1.48	Num	2
Q_10a	-0.82	Prob	1	Q_9a	-1.21	Alg	1	Q_1c	-1.62	Num	2
Q_13	-0.87	Ratio	3	Q_9b	-1.68	Alg	1	Q_6a	-1.65	Num	2
Q_3a	-0.88	Num	2	Q_1a	-1.81	Num	2	Q_8a	-1.84	Num	2
Q_4b	-1.13	Num	2	Q_4a	-1.91	Ratio	2	Q_1b	-1.85	Num	1
Q_19a	-1.41	Num	1	Q_2a	-2.01	Num	1	Q_5b	-1.86	Alg	2
Q_4a	-1.77	Num	1	Q_2b	-2.73	Num	1	Q_8b	-1.93	Num	1
Q_1d	-2.49	Num	2	Q_3c	-3.22	Alg	2	Q_5c	-1.98	Alg	2

Q_1c	-3.10	Num	1	Q_2c	-3.46	Num	1	Q_2b	-2.08	Geom	2
Q_1b	-3.32	Num	1	Q_3b	-3.53	Alg	1	Q_1a	-2.18	Num	1
Q_1a	-3.68	Num	1	Q_3a	-4.07	Alg	1	Q_7b	-2.20	Ratio	1
Q_2b	-4.82	Alg	1					Q_2ai	-2.42	Geom	1
Q_2a	-5.26	Alg	1					Q_2aii	-2.49	Geom	1
								Q_7a	-2.87	Ratio	2

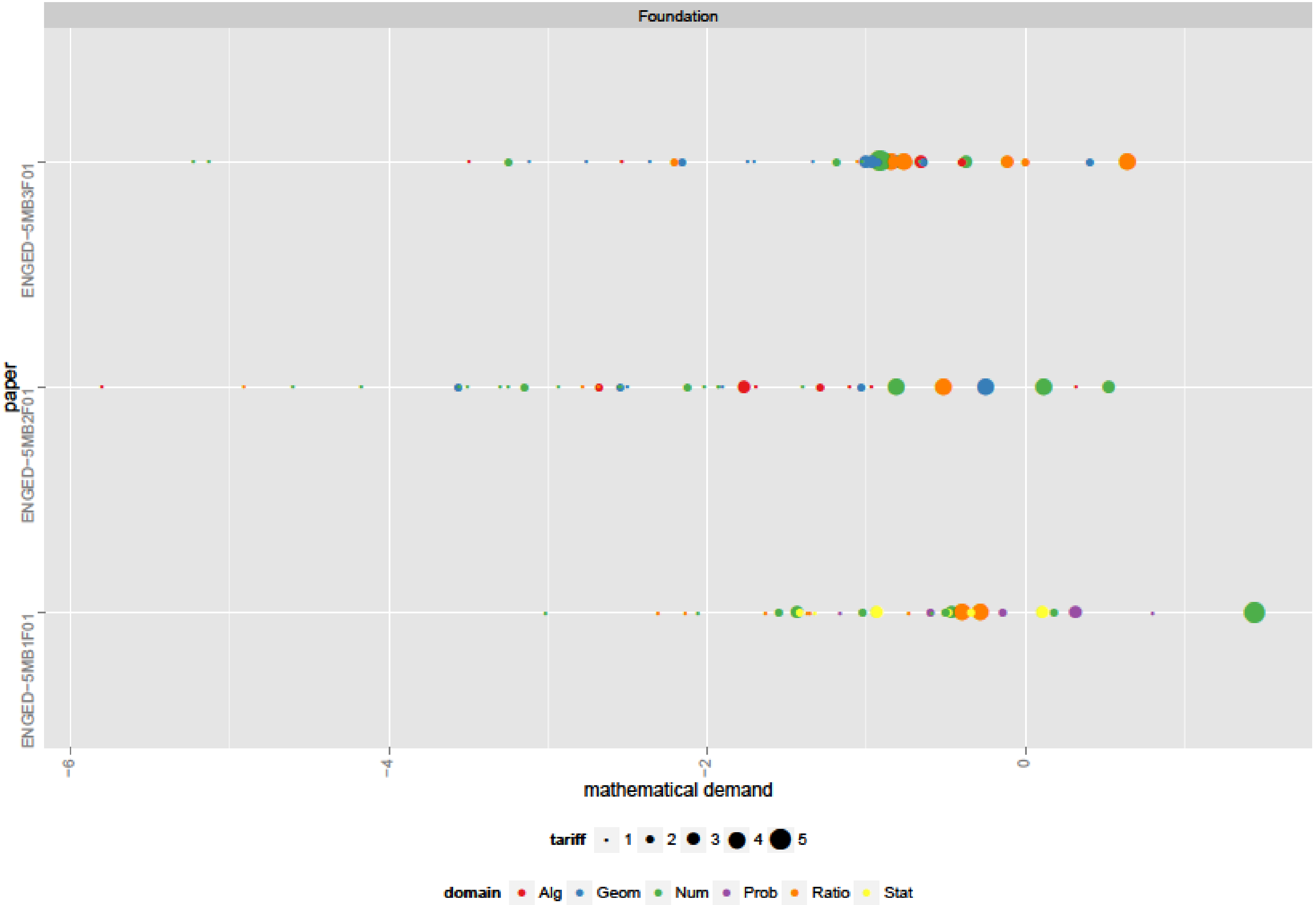
Pearson SAMs – higher tier



Paper 1MA11H				Paper 1MA12H				Paper 1MA13H			
Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff
Q_15	3.17	Geom	6	Q_8c	1.98	Alg	2	Q_14bii	2.33	Alg	2
Q_12b	2.95	Alg	3	Q_5b	1.92	Geom	4	Q_15c	2.25	Num	2
Q_16a	2.33	Alg	3	Q_9c	1.78	Ratio	4	Q_8bii	2.00	Alg	2
Q_2a	1.87	Stat	4	Q_13	1.68	Alg	6	Q_16	1.62	Alg	5
Q_17	1.72	Geom	7	Q_9a	1.64	Geom	3	Q_10c	1.29	Alg	2
Q_9	1.68	Geom	4	Q_4	1.61	Geom	3	Q_7	1.28	Geom	3
Q_13	1.48	Num	4	Q_12b	1.45	Prob	2	Q_8bi	1.23	Alg	2
Q_14a	1.47	Prob	2	Q_8b	1.26	Alg	2	Q_13b	0.95	Stat	2
Q_16b	1.42	Alg	3	Q_9b	1.24	Geom	2	Q_2c	0.87	Prob	2
Q_6ai	1.38	Geom	2	Q_3b	1.17	Num	6	Q_5	0.85	Ratio	4
Q_12a	1.13	Alg	2	Q_10	1.16	Geom	3	Q_6	0.80	Geom	3
Q_7	0.76	Num	3	Q_11b	0.99	Prob	1	Q_14bi	0.76	Alg	2
Q_2b	0.70	Stat	2	Q_11c	0.93	Prob	2	Q_15b	0.75	Num	2
Q_11bii	0.70	Ratio	3	Q_11a	0.89	Prob	3	Q_3b	0.73	Geom	4
Q_14c	0.59	Prob	1	Q_1	0.80	Num	3	Q_12	0.69	Alg	5

Q_14b	0.51	Prob	3	Q_14a	0.80	Stat	5	Q_8ai	0.60	Alg	2
Q_6b	0.50	Num	3	Q_6b	0.74	Alg	2	Q_9	0.50	Geom	3
Q_3b	0.47	Num	1	Q_14b	0.66	Stat	1	Q_11b	0.49	Ratio	1
Q_4	0.44	Geom	3	Q_2	0.65	Alg	3	Q_11a	0.40	Ratio	3
Q_10	0.25	Num	4	Q_15	0.51	Num	5	Q_10b	0.38	Alg	3
Q_6aii	0.13	Num	1	Q_12a	0.29	Prob	2	Q_13a	0.37	Stat	3
Q_8	0.06	Num	3	Q_6c	0.29	Alg	2	Q_1	0.34	Num	3
Q_5a	0.05	Alg	3	Q_7a	0.19	Stat	2	Q_10d	0.33	Alg	2
Q_5b	-0.12	Alg	3	Q_7b	0.09	Stat	2	Q_15a	0.23	Num	2
Q_11bi	-0.15	Ratio	2	Q_3a	0.00	Num	3	Q_10a	0.21	Alg	2
Q_11a	-0.56	Ratio	1	Q_8a	-0.41	Alg	2	Q_2b	0.14	Prob	1
Q_3a	-0.79	Num	1	Q_6a	-0.53	Alg	2	Q_8aii	0.10	Alg	2
Q_1	-0.87	Ratio	3	Q_5aii	-0.65	Geom	1	Q_3a	-0.01	Alg	3
				Q_5ai	-0.77	Geom	2	Q_4	-0.26	Num	5
								Q_14a	-0.44	Alg	2
								Q_2a	-0.64	Prob	1

Pearson – old GCSE – foundation tier

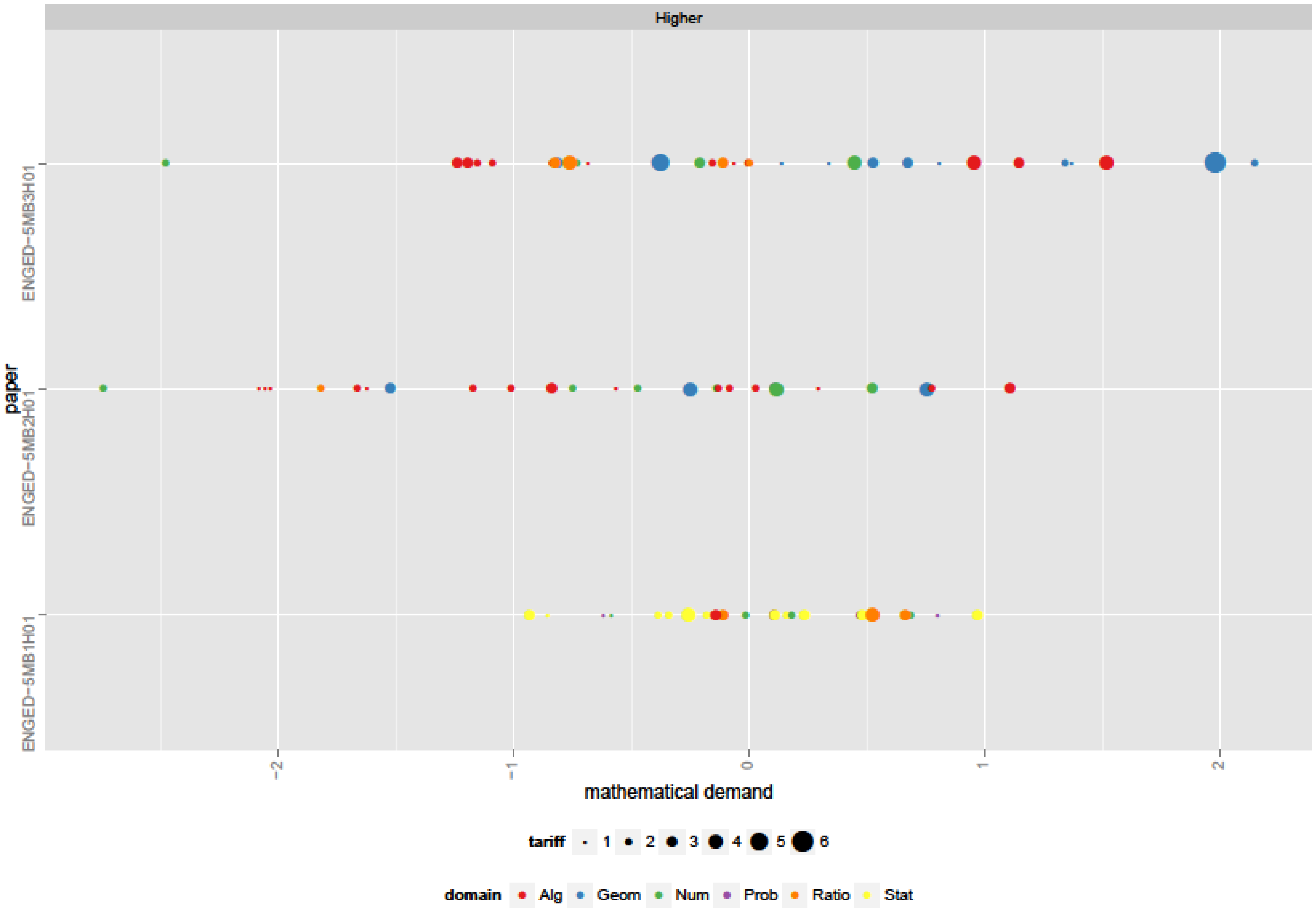


Paper 5MB1F01				Paper 5MB2F01				Paper 5MB3F01			
Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff
Q_10	1.44	Num	5	Q_16	0.52	Num	3	Q_15	0.64	Ratio	4
Q_13b	0.80	Prob	1	Q_14c	0.32	Alg	1	Q_14b	0.40	Geom	2
Q_8	0.32	Prob	3	Q_19	0.12	Num	4	Q_20b	0.00	Ratio	2
Q_12c	0.18	Num	2	Q_15	-0.25	Geom	4	Q_17	-0.11	Ratio	3
Q_11b	0.11	Stat	3	Q_12	-0.51	Ratio	4	Q_9c	-0.12	Geom	2
Q_13a	-0.14	Prob	2	Q_11	-0.81	Num	4	Q_12b	-0.37	Num	3
Q_9c	-0.28	Ratio	4	Q_14b	-0.97	Alg	1	Q_18b	-0.40	Alg	2
Q_14	-0.34	Stat	2	Q_10b	-1.03	Geom	2	Q_1c	-0.64	Geom	2
Q_12b	-0.34	Num	1	Q_14aii	-1.11	Alg	1	Q_18a	-0.65	Alg	3
Q_6b	-0.40	Ratio	4	Q_17	-1.29	Alg	2	Q_2a	-0.67	Num	1
Q_4ai	-0.42	Prob	1	Q_3b	-1.40	Num	1	Q_20c	-0.73	Num	2
Q_3b	-0.46	Num	3	Q_14ai	-1.69	Alg	1	Q_21	-0.76	Ratio	4
Q_5c	-0.48	Stat	2	Q_18	-1.77	Alg	3	Q_16	-0.82	Geom	3
Q_7	-0.50	Num	2	Q_6b	-1.90	Geom	1	Q_20a	-0.83	Ratio	3
Q_12a	-0.59	Num	1	Q_8c	-1.93	Num	1	Q_13	-0.84	Ratio	4

Q_4b	-0.60	Prob	2	Q_2b	-1.93	Num	1	Q_2c	-0.90	Num	1
Q_9a	-0.73	Ratio	1	Q_3aii	-2.02	Num	1	Q_8	-0.91	Num	5
Q_11a	-0.93	Stat	3	Q_2c	-2.13	Num	2	Q_2d	-0.92	Num	2
Q_1b	-1.02	Num	2	Q_6a	-2.50	Geom	1	Q_9a	-0.92	Geom	2
Q_4aii	-1.17	Prob	1	Q_7b	-2.54	Num	1	Q_11bii	-0.95	Ratio	1
Q_5b	-1.32	Stat	1	Q_10a	-2.55	Geom	2	Q_9b	-0.96	Geom	3
Q_2e	-1.35	Ratio	1	Q_13c	-2.55	Num	1	Q_11a	-0.96	Ratio	2
Q_6a	-1.37	Ratio	1	Q_4b	-2.68	Alg	2	Q_14a	-1.00	Geom	3
Q_5a	-1.42	Stat	2	Q_9b	-2.68	Ratio	1	Q_2b	-1.02	Num	1
Q_1a	-1.43	Num	3	Q_9a	-2.78	Ratio	1	Q_11bi	-1.05	Ratio	1
Q_3a	-1.55	Num	2	Q_1a	-2.93	Num	1	Q_12a	-1.19	Num	2
Q_2d	-1.63	Ratio	1	Q_1b	-3.14	Num	1	Q_10b	-1.34	Geom	1
Q_2b	-2.06	Num	1	Q_7a	-3.15	Num	2	Q_1b	-1.71	Geom	1
Q_9b	-2.13	Ratio	1	Q_3a	-3.25	Num	1	Q_10a	-1.75	Geom	1
Q_2c	-2.31	Ratio	1	Q_8b	-3.30	Num	1	Q_10c	-2.16	Geom	2
Q_2a	-3.01	Num	1	Q_2a	-3.50	Num	1	Q_3	-2.17	Geom	1
				Q_8a	-3.56	Num	1	Q_6	-2.20	Ratio	2

				Q_5	-3.56	Geom	2	Q_7b	-2.36	Geom	1
				Q_13b	-4.17	Num	1	Q_5b	-2.53	Alg	1
				Q_13a	-4.60	Num	1	Q_7a	-2.76	Geom	1
				Q_9c	-4.91	Ratio	1	Q_1a	-3.12	Geom	1
				Q_4a	-5.80	Alg	1	Q_19	-3.25	Num	2
								Q_5a	-3.50	Alg	1
								Q_4a	-5.13	Num	1
								Q_4b	-5.23	Num	1

Pearson – old GCSE – higher tier



Paper 5MB1H01				Paper 5MB2H01				Paper 5MB3H01			
Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff	Question	Demand	Domain	Tariff
Q_8c	0.97	Stat	3	Q_16a	1.11	Alg	3	Q_17a	2.15	Geom	2
Q_5c	0.80	Prob	1	Q_12	0.78	Alg	2	Q_18	1.98	Geom	6
Q_1i	0.68	Num	2	Q_11	0.75	Geom	4	Q_7	1.52	Alg	4
Q_3b	0.66	Ratio	3	Q_3b	0.52	Num	3	Q_15ai	1.37	Geom	1
Q_3a	0.52	Ratio	4	Q_5bii	0.29	Alg	1	Q_15bii	1.34	Geom	2
Q_11a	0.48	Stat	3	Q_7	0.12	Geom	2	Q_16b	1.15	Alg	3
Q_10a	0.47	Prob	2	Q_9	0.12	Num	4	Q_14b	0.96	Alg	4
Q_8d	0.24	Stat	3	Q_13	0.11	Geom	3	Q_15bi	0.81	Geom	1
Q_2c	0.18	Num	2	Q_16b	0.03	Alg	2	Q_17b	0.67	Geom	3
Q_11b	0.16	Stat	2	Q_10a	-0.08	Alg	2	Q_6	0.53	Geom	3
Q_4b	0.11	Stat	3	Q_2d	-0.13	Alg	2	Q_10	0.45	Num	4
Q_10c	0.11	Prob	3	Q_14b	-0.14	Num	2	Q_15aiii	0.34	Geom	1
Q_1ii	-0.01	Num	2	Q_6	-0.25	Geom	4	Q_15aii	0.14	Geom	1
Q_10b	-0.10	Prob	2	Q_14a	-0.47	Num	2	Q_4b	0.00	Ratio	2
Q_7	-0.11	Ratio	3	Q_5bi	-0.57	Alg	1	Q_11b	0.00	Alg	2

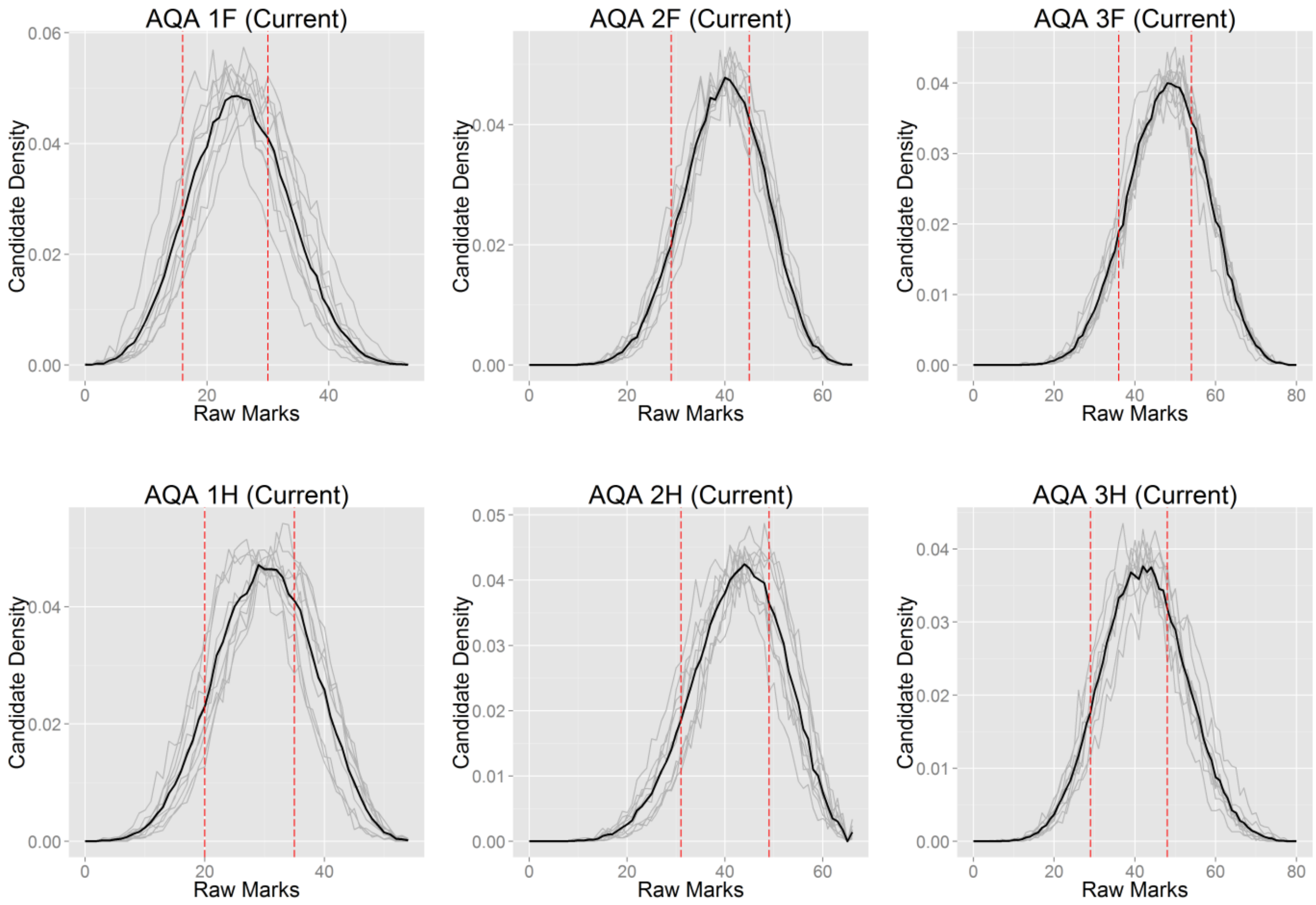
Q_9	-0.14	Alg	3	Q_3a	-0.75	Num	2	Q_8b	-0.06	Alg	1
Q_5a	-0.14	Prob	2	Q_15a	-0.82	Num	1	Q_2	-0.11	Ratio	3
Q_6c	-0.18	Stat	2	Q_5a	-0.84	Alg	3	Q_14a	-0.16	Alg	2
Q_8b	-0.26	Stat	4	Q_10b	-1.01	Alg	2	Q_12	-0.21	Num	3
Q_6a	-0.34	Stat	2	Q_2b	-1.17	Alg	2	Q_9	-0.38	Geom	5
Q_2b	-0.34	Num	1	Q_1	-1.52	Geom	3	Q_13b	-0.68	Alg	1
Q_6b	-0.39	Stat	2	Q_8c	-1.62	Alg	1	Q_4c	-0.73	Num	2
Q_2a	-0.59	Num	1	Q_2c	-1.66	Alg	2	Q_5	-0.76	Ratio	4
Q_5b	-0.62	Prob	1	Q_4	-1.82	Ratio	2	Q_1	-0.82	Geom	3
Q_8a	-0.85	Stat	1	Q_2a	-2.03	Alg	1	Q_4a	-0.83	Ratio	3
Q_4a	-0.93	Stat	3	Q_8a	-2.06	Alg	1	Q_11c	-0.84	Alg	2
				Q_8b	-2.08	Alg	1	Q_8a	-1.09	Alg	2
				Q_15b	-2.75	Num	2	Q_11a	-1.16	Alg	2
								Q_16a	-1.19	Alg	3
								Q_13a	-1.24	Alg	3
								Q_3	-2.48	Num	2

Appendix C: Assessment objectives

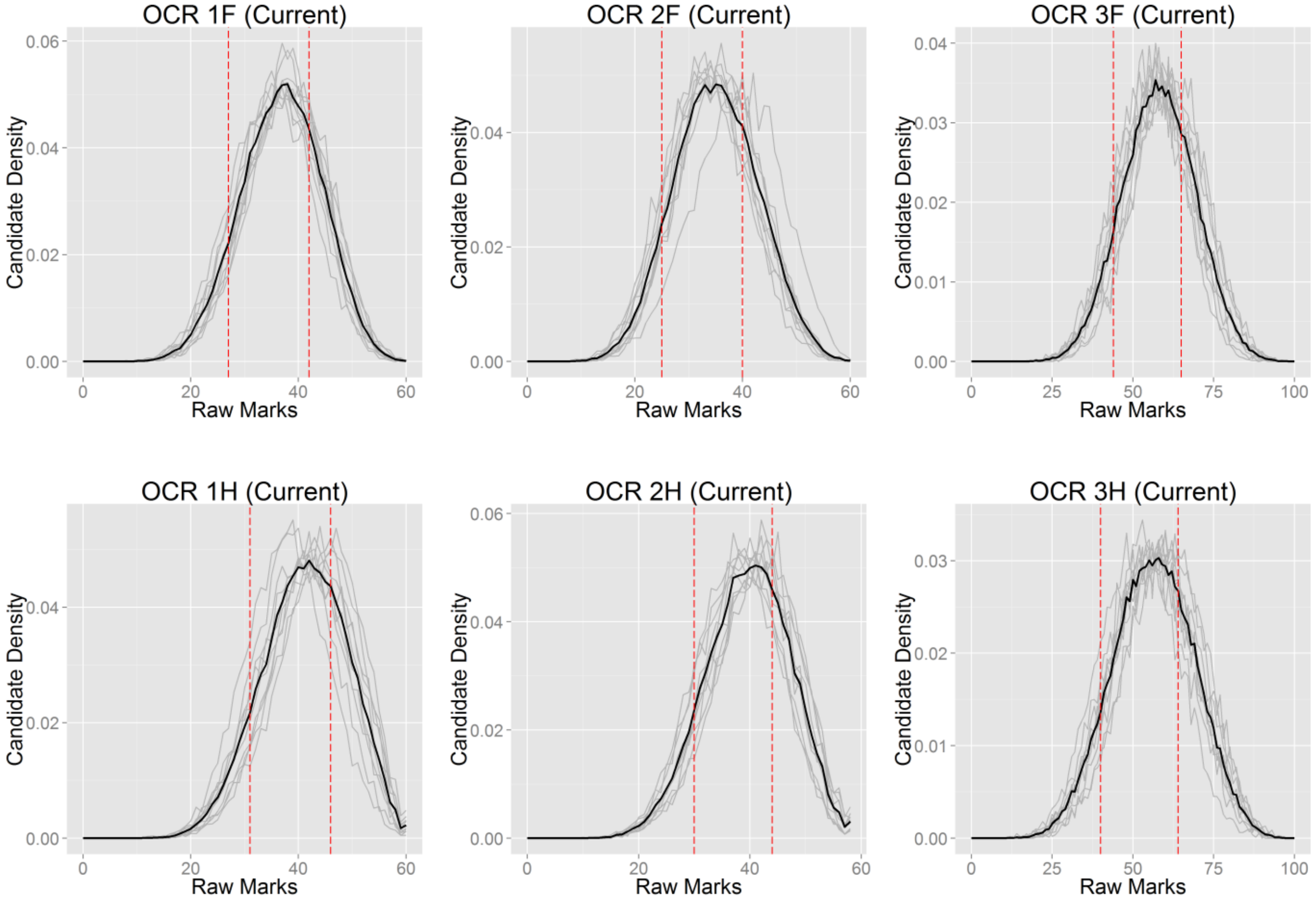
		Weighting	
		Higher	Foundation
AO1	Use and apply standard techniques students should be able to: ■ accurately recall facts, terminology and definitions ■ use and interpret notation correctly ■ accurately carry out routine	40%	50%
AO2	Reason, interpret and communicate mathematically students should be able to: ■ make deductions, inferences and draw conclusions from mathematical information ■ construct chains of reasoning to achieve a given result ■ interpret and communicate information accurately ■ present arguments and proofs ■ assess the validity of an argument and critically evaluate a given way of presenting information Where problems require candidates to 'use and apply standard techniques' or to independently 'solve problems' a proportion of those marks should be attributed to the corresponding assessment objective	30%	25%
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	30%	25%

	Where problems require candidates to 'use and apply standard techniques' or to 'reason, interpret and communicate mathematically' a proportion of those marks should be attributed to the corresponding assessment objective.		

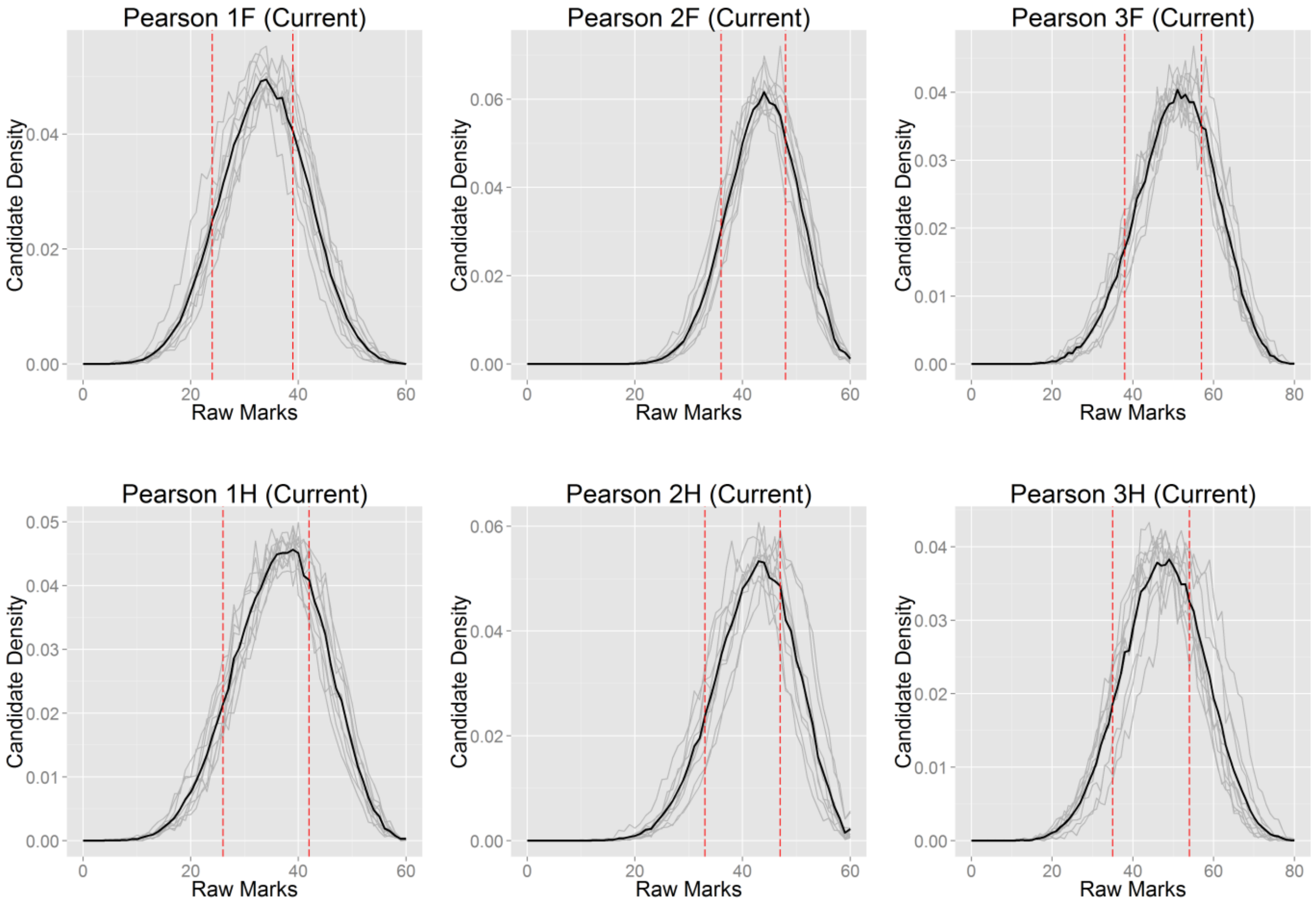
Appendix D: Mark distributions for all current GCSE papers and sample assessment materials



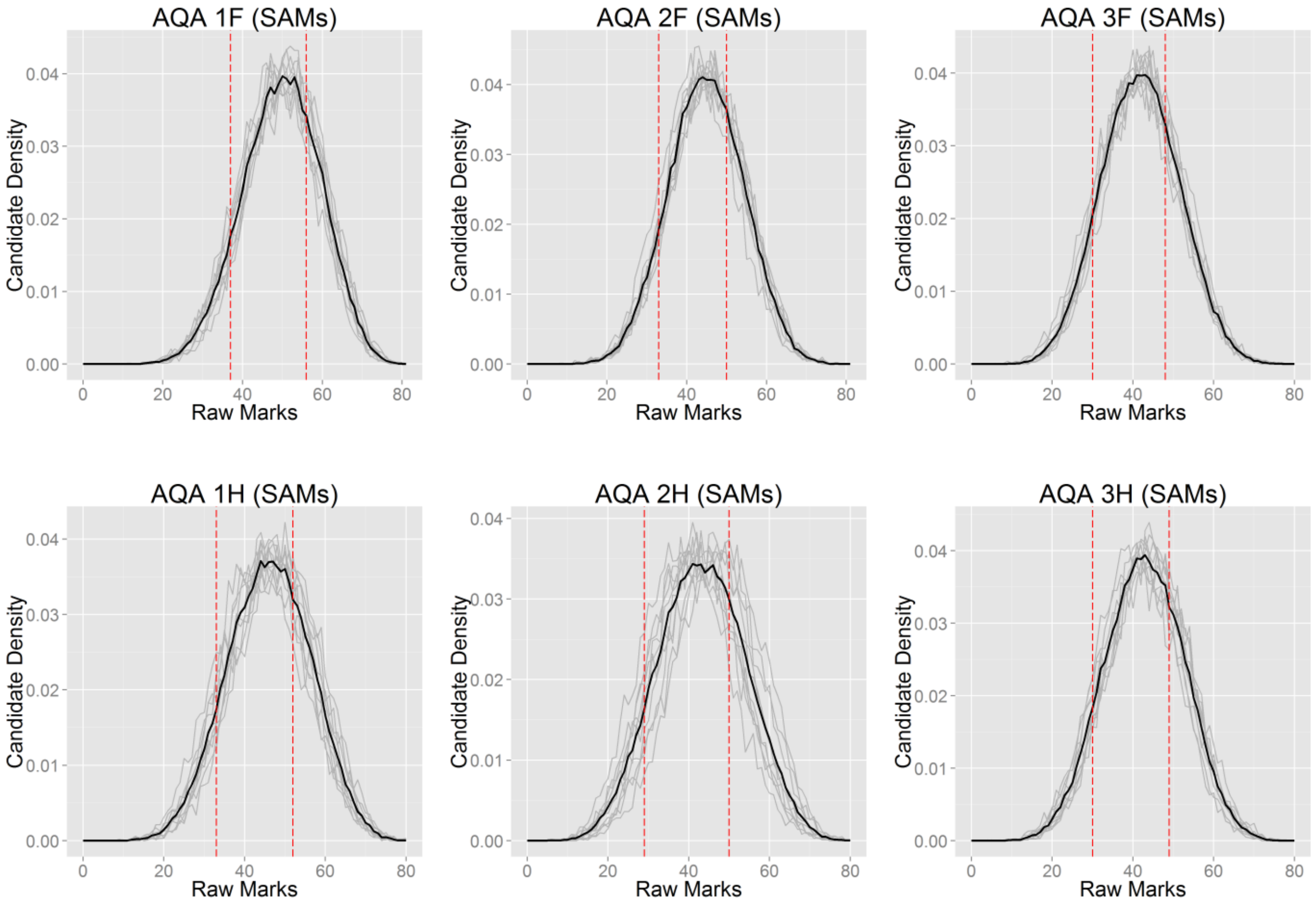
Simulated mark distributions for the current AQA question papers; red lines indicate the notional grade boundary positions



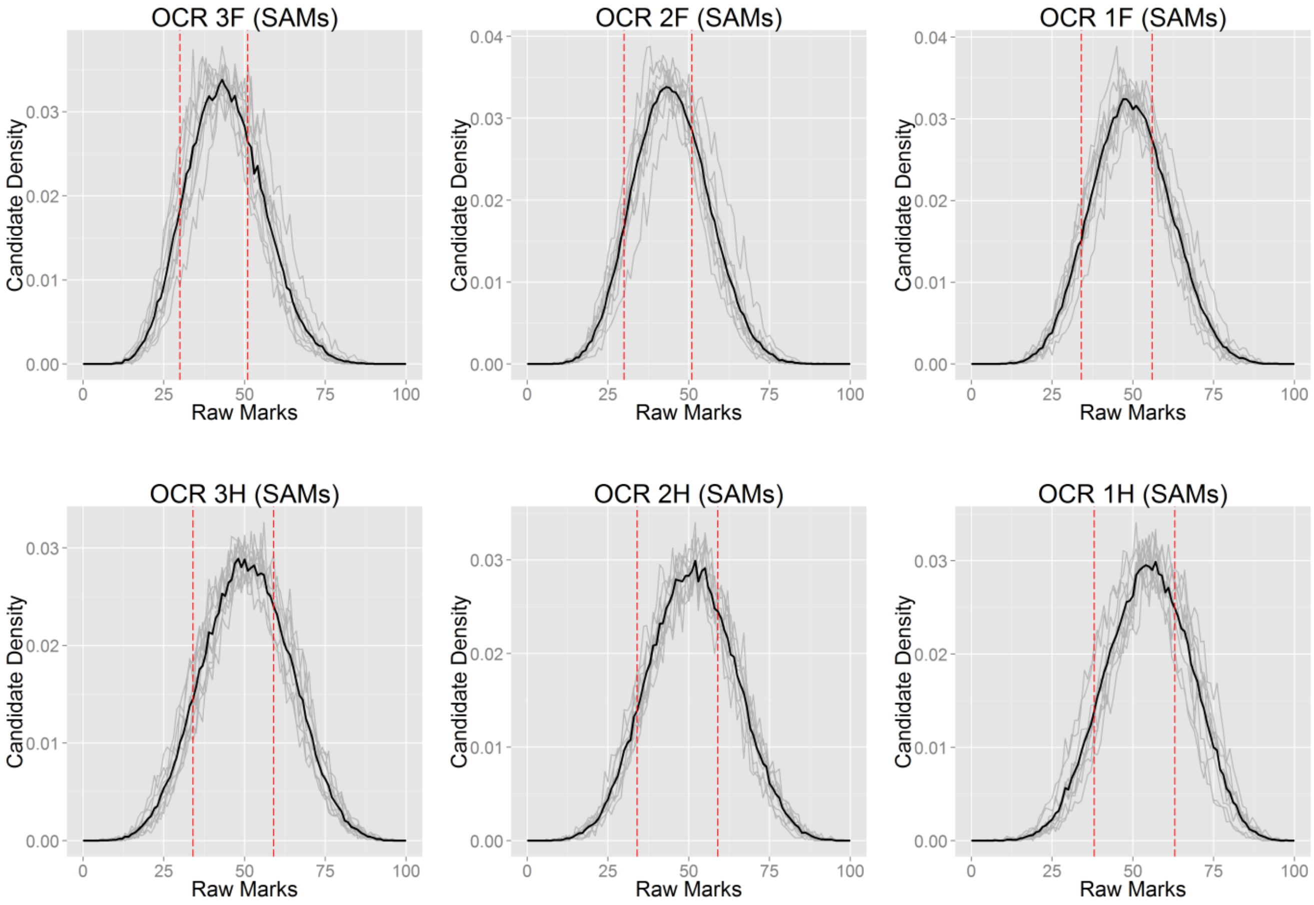
Simulated mark distributions for the current OCR question papers; red lines indicate the notional grade boundary positions



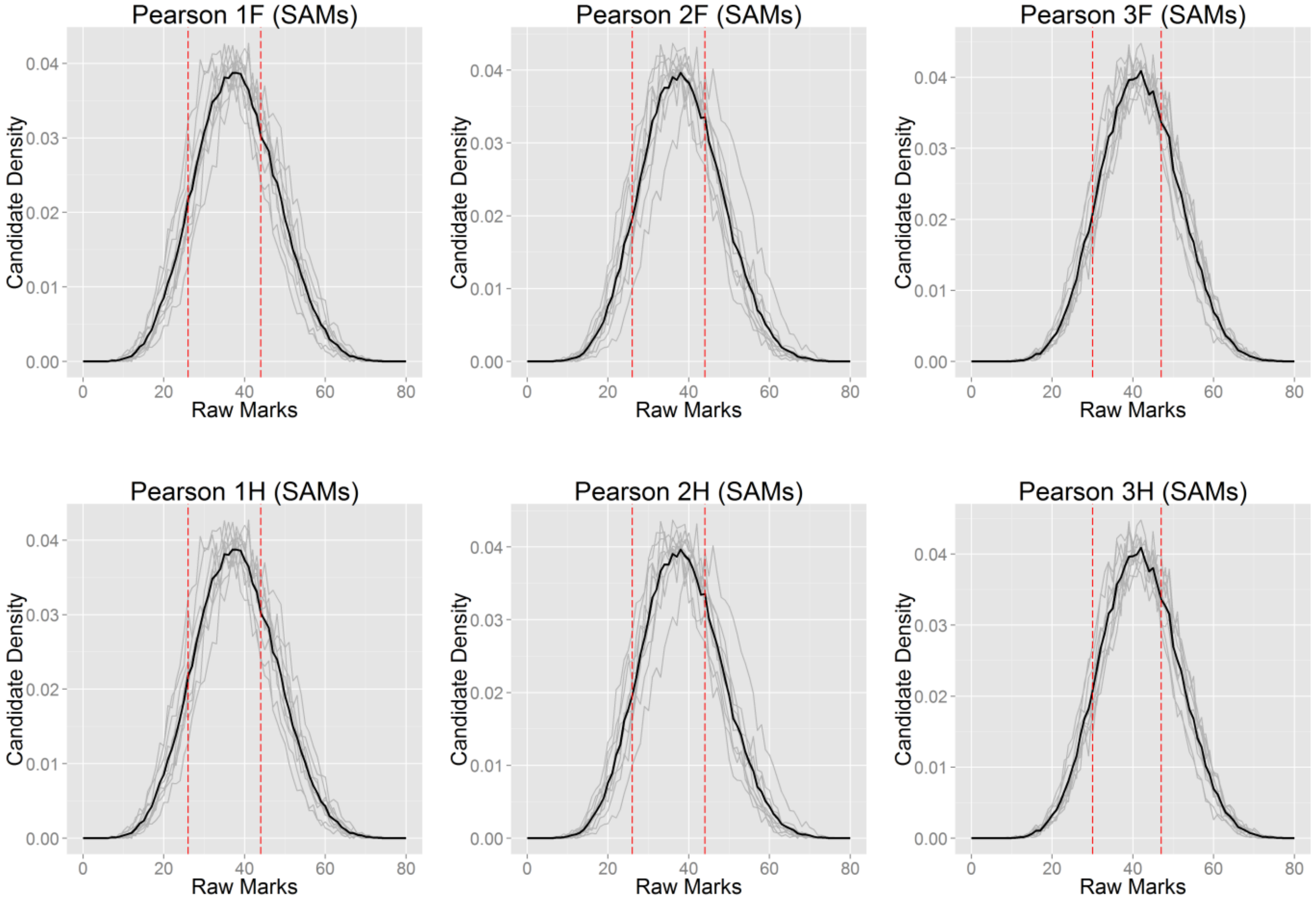
Simulated mark distributions for the current Pearson question papers; red lines indicate the notional grade boundary positions



Simulated mark distributions for the AQA SAMs; red lines indicate the notional grade boundary positions.



Simulated mark distributions for the OCR SAMs; red lines indicate the notional grade boundary positions



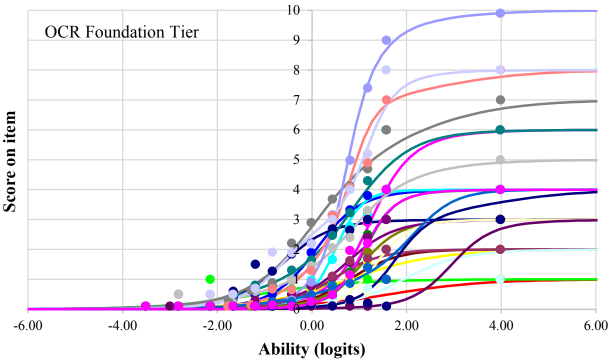
Simulated mark distributions for the Pearson SAMs; red lines indicate the notional grade boundary positions

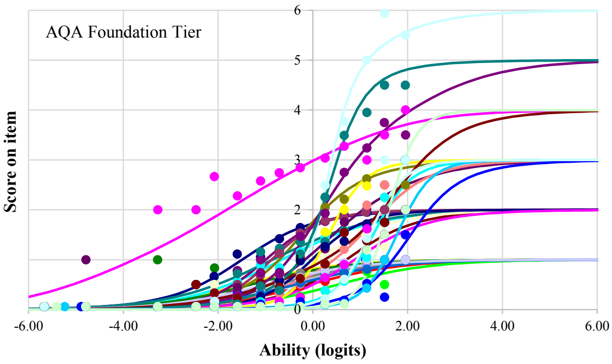
Appendix E: Rasch model fit for study 2

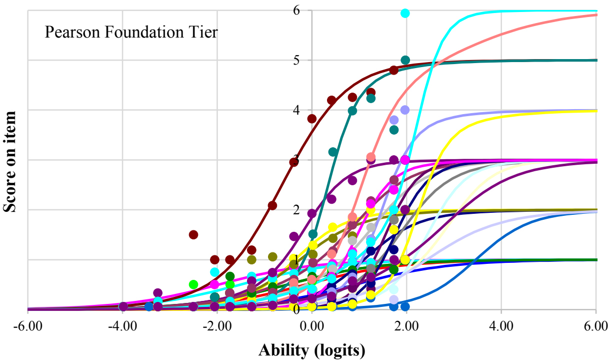
Dimensionality, model assumptions and model fit.

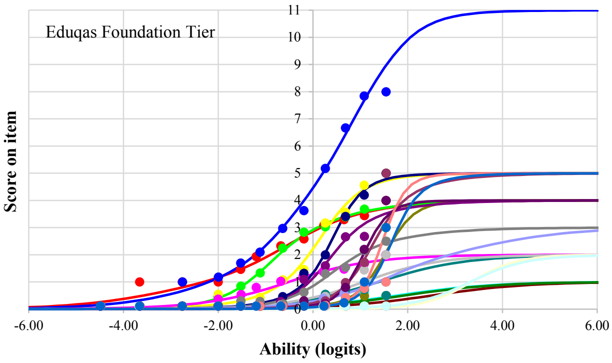
	OCR		AQA		Pearson		EDUQAS	
	F	H	F	H	F	H	F	H
Ratio of first eigenvalue to second eigenvalue frm (EFA)	4.2	4.3	3.7	4.5	3.6	2.9	4.3	6.1
Variances explained								
Variance explained by Rasch model (%)	49.8	58.5	54.3	60.2	60.8	48.8	63.1	55.9
Unexplained variance (%)	50.2	41.5	45.7	39.8	39.2	51.2	36.9	44.1
Explained by the first contrast (%)	4.1	3.4	2.8	2.5	2.5	4.6	3.2	4.1
Explained by the second contrast (%)	3.5	2.7	2.7	2.2	2.4	3.6	2.7	3.0
Correlation between item residuals								
Mean	-0.03	-0.03	-0.02	-0.02	-0.02	-0.04	-0.03	-0.03
Standard deviation	0.08	0.06	0.07	0.06	0.07	0.06	0.08	0.07
Person separation index	2.42	2.49	2.92	2.60	2.30	1.83	2.26	2.17
Person reliability	0.85	0.86	0.90	0.87	0.84	0.77	0.84	0.82

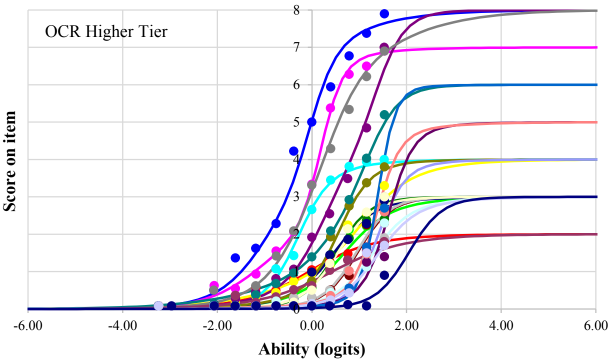
Expected scores on individual items/questions

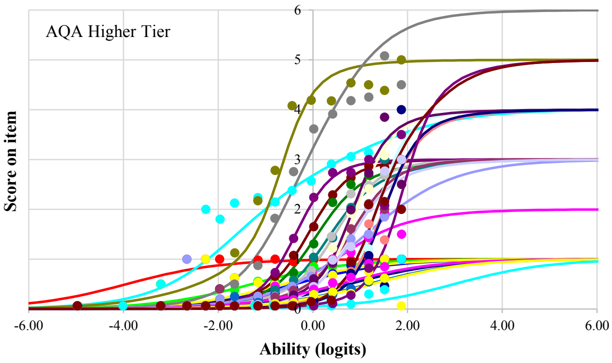


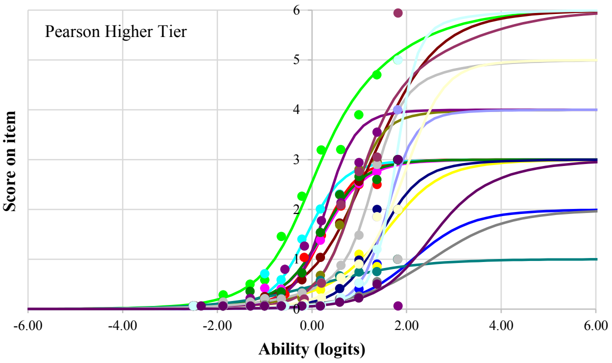


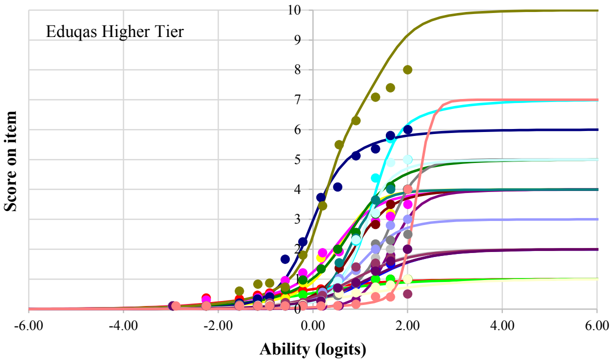




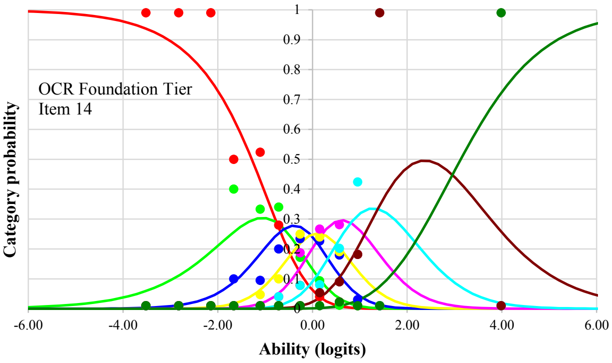


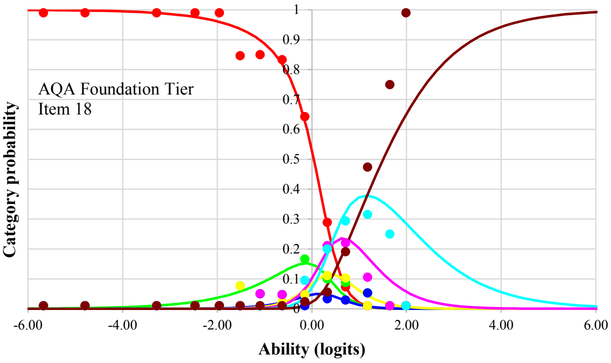


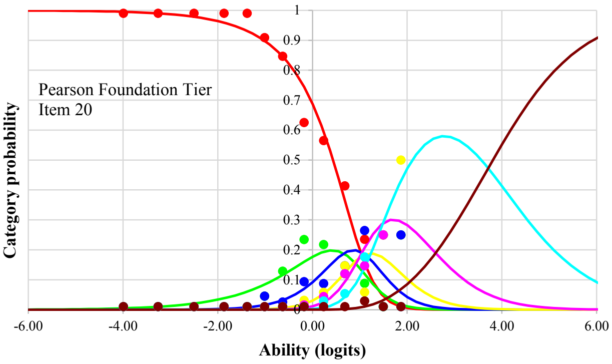


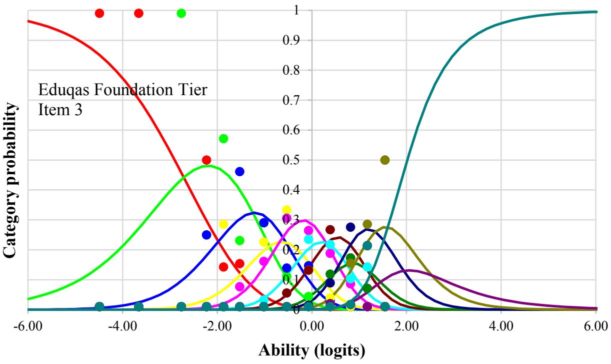


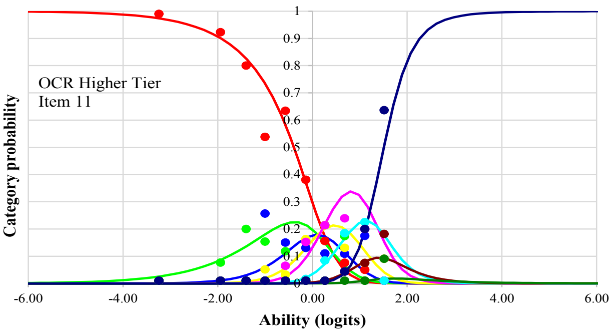
Category probability curves for a selection of questions

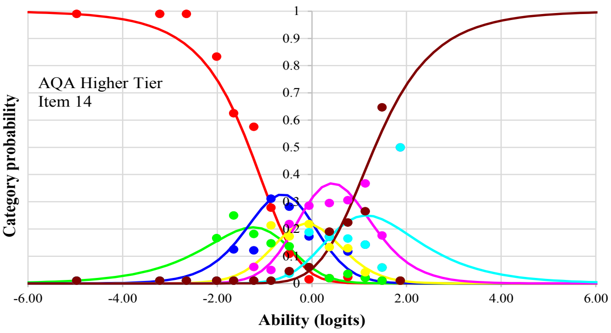


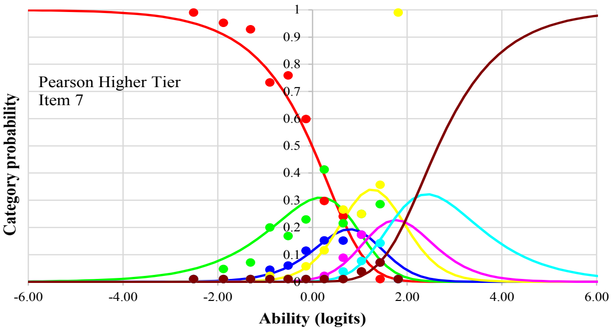


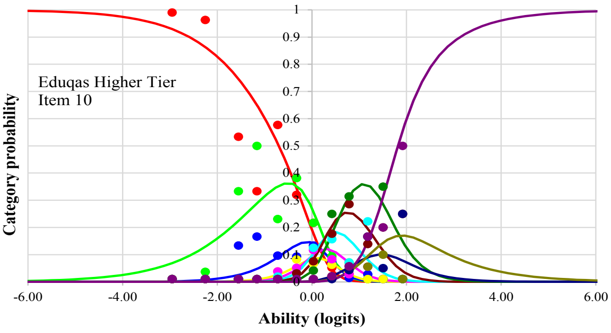












Appendix F: Item-level analysis for study 2

Summary of item-level performance in study 2.

Number of questions and marks in the foundation tier papers

OCR		AQA		Pearson		Eduqas	
Question	**Mark**	**Question**	**Mark**	**Question**	**Mark**	**Question**	**Mark**
1(a)	1	1(a)	1	1(a)	1	1	4
1(b)	1	1(b)	1	1(b)	1	2	4
2(ab)	4	2	1	1(c)	1	3	11
3	2	3	1	1(d)	2	4	5
4(abcd)	6	4	4	2(a)	1	5(a)	2
5	4	5	2	2(b)	1	5(b)	1
6(a)	2	6	2	3(ab)	5	5(c)	1
6(b)	4	7	1	4(a)	1	5(d)	1
7(ab)	3	8	2	4(b)	2	6	5
8(a)	3	9	3	4(c)	2	7	4
8(b)	3	10	2	5	3	8	4
9(abc)	6	11	2	6	5	9(a)(I,ii)	2
10(ab)	5	12(a)	1	7	3	9(b)(I,ii)	2
11	8	12(b)	1	8	3	10	3
12(abc)	10	12(c)	1	9	4	11	6
13(a)	2	13	2	10(ab)	3	12	5
13(b)	3	14	2	11	3	13	6
14(a)	2	15	6	12	3	14	4
14(b)	3	16(ab)	3	13	3	15	4
15(abc)	8	17(ab)	3	14	6	16	5
16	4	18	1	15	2	17	5
17(cabcd)	8	19(a)	1	16	3	Overall test	84
18	4	19(b)	2	17(a)	3		
19	4	20	2	17(b)	3		
Overall test	100	21	3	18(ab)	6		
		22	3	19(abc)	7		
		23(abcd)	5	20	3		
		24(abcd)	4	Overall test	80		
		25	5				
		26	3				

		27	3				
		28	3				
		29	4				
		Overall test	80				

Question performance, OCR foundation tier paper

Question No	Mark	Weight[42]	Achieved weighting[43]	Facility[44]	Standard deviation	Item-total correlation	Item-total corr. (minus item)	Non-response rate[45]
1(a)	1	1.00	0.44	17.68	0.38	0.18	0.16	1.93
1(b)	1	1.00	1.18	70.72	0.46	0.41	0.38	7.18
2(ab)	4	4.00	6.70	46.48	1.63	0.65	0.59	9.39
3	2	2.00	2.17	31.22	0.71	0.49	0.45	14.36
4(abcd)	6	6.00	4.53	12.02	1.24	0.58	0.52	2.76
5	4	4.00	6.26	29.77	1.72	0.58	0.50	10.77
6(a)	2	2.00	2.79	23.20	0.84	0.53	0.49	12.43
6(b)	4	4.00	3.46	10.91	1.06	0.52	0.47	30.66
7(ab)	3	3.00	3.87	69.61	1.19	0.52	0.46	3.31
8(a)	3	3.00	2.92	15.29	0.94	0.49	0.45	19.06
8(b)	3	3.00	3.57	28.08	1.08	0.52	0.47	31.49
9(abc)	6	6.00	6.68	31.45	1.56	0.68	0.62	6.08
10(ab)	5	5.00	4.80	30.50	1.34	0.57	0.50	10.50
11	8	8.00	7.86	34.25	1.83	0.68	0.61	8.56
12(abc)	10	10.00	13.00	20.19	2.76	0.75	0.65	9.94
13(a)	2	2.00	3.76	34.67	0.94	0.63	0.59	24.59
13(b)	3	3.00	1.57	6.91	0.73	0.34	0.30	52.21
14(a)	2	2.00	0.91	6.91	0.44	0.33	0.31	43.09
14(b)	3	3.00	0.48	1.10	0.23	0.33	0.31	73.76
15(abc)	8	8.00	9.60	25.03	2.23	0.68	0.60	18.78
16	4	4.00	2.29	13.47	0.73	0.49	0.46	32.04
17(cabcd)	8	8.00	7.19	31.39	1.84	0.62	0.54	11.05

[42] Defined as the percentage of the maximum available mark on the test.

[43] The achieved weight $W_{i,Achieved}$ (%) for item *i* is defined as follows:

$$W_{i,Achieved} = r_{i,Test} \times \frac{\sigma_i}{\sigma_{Test}} \times 100 = \frac{\sigma(i,Test)}{\sigma_{Test}^2} \times 100$$

where $r_{i,Test}$ is the correlation between the item and the test, σ_i is the standard deviation of item scores, $\sigma(i,Test)$ is the covariance between the item and the test, and σ_{Test} is the standard deviation of test scores.

[44] Facility of an item is defined as: average score / maximum available mark × 100.

[45] Defined as percentage of students who had not attempted the question.

18	4	4.00	0.73	3.38	0.53	0.22	0.19	39.23
19	4	4.00	3.21	11.19	1.09	0.47	0.41	56.35
Overall test	100			24.45	15.82			

Question performance, AQA foundation tier paper

Question No	Mark	Weight	Achieved weighting	Facility	Standard deviation	Item-total correlation	Item-total corr. (minus item)	Non-response rate
1(a)	1	1.25	1.70	50.77	0.50	0.50	0.48	5.54
1(b)	1	1.25	1.36	38.15	0.49	0.41	0.39	4.31
2	1	1.25	1.61	62.15	0.49	0.49	0.46	5.54
3	1	1.25	1.45	76.00	0.43	0.50	0.48	4.92
4	4	5.00	3.20	71.31	0.96	0.49	0.44	3.69
5	2	2.50	2.67	62.00	0.74	0.53	0.50	5.85
6	2	2.50	1.91	32.00	0.79	0.36	0.31	7.69
7	1	1.25	0.98	73.54	0.44	0.33	0.30	3.38
8	2	2.50	2.75	79.23	0.69	0.59	0.56	5.85
9	3	3.75	4.52	56.31	1.17	0.57	0.52	7.38
10	2	2.50	3.09	74.62	0.81	0.57	0.53	5.85
11	2	2.50	2.25	65.08	0.72	0.46	0.42	5.54
12(a)	1	1.25	1.85	60.62	0.49	0.56	0.54	5.23
12(b)	1	1.25	1.91	64.31	0.48	0.59	0.57	5.54
12(c)	1	1.25	1.70	57.23	0.50	0.51	0.48	8.62
13	2	2.50	3.54	73.38	0.85	0.62	0.58	11.08
14	2	2.50	2.97	46.92	0.83	0.53	0.49	9.54
15	6	7.50	10.82	36.15	2.32	0.69	0.60	10.77
16(ab)	3	3.75	3.33	39.18	1.02	0.48	0.43	8.92
17(ab)	3	3.75	3.72	27.18	0.89	0.62	0.58	11.08
18	1	1.25	1.60	56.62	0.50	0.48	0.45	13.23
19(a)	1	1.25	1.79	68.31	0.47	0.57	0.55	13.85
19(b)	2	2.50	3.30	51.23	0.96	0.51	0.46	21.23
20	2	2.50	2.90	28.62	0.74	0.58	0.55	25.85
21	3	3.75	5.68	39.79	1.39	0.60	0.54	16.62
22	3	3.75	2.36	12.10	0.86	0.41	0.36	7.38
23(abcd)	5	6.25	6.63	35.75	1.46	0.67	0.61	2.15
24(abcd)	4	5.00	3.22	23.23	0.86	0.55	0.51	3.69
25	5	6.25	9.29	40.18	1.94	0.71	0.64	18.15
26	3	3.75	0.69	5.54	0.50	0.20	0.17	46.77
27	3	3.75	0.99	3.90	0.53	0.28	0.25	52.62
28	3	3.75	2.69	12.00	0.92	0.43	0.38	30.77
29	4	5.00	1.53	4.08	0.69	0.33	0.28	46.46
Overall test	80			40.16	14.82			

Question performance, Pearson foundation tier paper

Question No	Mark	Weight	Achieved weighting	Facility	Standard deviation	Item-total correlation	Item-total corr. (minus item)	Non-response rate
1(a)	1	1.25	1.85	52.12	0.50	0.39	0.35	3.40
1(b)	1	1.25	1.98	59.77	0.49	0.42	0.38	6.23
1(c)	1	1.25	2.42	30.88	0.46	0.55	0.52	19.26
1(d)	2	2.50	5.50	60.91	0.94	0.62	0.56	11.90
2(a)	1	1.25	1.67	82.15	0.38	0.46	0.43	7.93
2(b)	1	1.25	1.70	76.49	0.42	0.42	0.39	7.08
3(ab)	5	6.25	10.24	65.89	1.67	0.64	0.53	3.97
4(a)	1	1.25	1.60	59.77	0.49	0.34	0.30	9.35
4(b)	2	2.50	2.98	20.40	0.74	0.42	0.36	34.28
4(c)	2	2.50	3.13	56.80	0.89	0.37	0.29	7.65
5	3	3.75	7.55	58.92	1.30	0.61	0.52	7.37
6	5	6.25	14.22	38.98	2.06	0.73	0.61	17.28
7	3	3.75	5.05	22.85	0.89	0.60	0.54	31.16
8	3	3.75	2.77	9.54	0.62	0.47	0.42	29.75
9	4	5.00	4.25	9.07	0.89	0.50	0.44	18.98
10(ab)	3	3.75	5.68	33.52	1.00	0.60	0.53	3.40
11	3	3.75	1.47	5.67	0.40	0.39	0.36	12.75
12	3	3.75	0.90	2.55	0.33	0.28	0.25	47.88
13	3	3.75	2.42	6.04	0.69	0.37	0.31	29.46
14	6	7.50	7.07	15.53	1.43	0.52	0.41	35.98
15	2	2.50	0.19	0.99	0.16	0.13	0.11	44.76
16	3	3.75	1.00	4.82	0.40	0.26	0.23	50.99
17(a)	3	3.75	1.49	4.44	0.57	0.27	0.22	46.74
17(b)	3	3.75	5.45	18.70	1.11	0.51	0.43	52.12
18(ab)	6	7.50	1.15	1.18	0.37	0.32	0.29	41.64
19(abc)	7	8.75	4.20	8.46	0.80	0.55	0.50	35.41
20	3	3.75	2.08	9.54	0.51	0.43	0.39	44.48
Overall test	80			23.62	10.51			

Question performance, Eduqas foundation tier paper

Question No	Mark	Weight	Achieved weighting	Facility	Standard deviation	Item-total correlation	Item-total corr. (minus item)	Non-response rate
1	4	4.76	5.69	69.06	1.12	0.55	0.47	2.05
2	4	4.76	7.33	67.74	1.30	0.61	0.52	5.87
3	11	13.10	18.15	42.63	2.47	0.79	0.68	1.47
4	5	5.95	12.15	48.91	1.90	0.69	0.58	7.62
5(a)	2	2.38	2.97	62.02	0.78	0.41	0.35	5.28
5(b)	1	1.19	1.46	16.72	0.37	0.42	0.39	15.54
5(c)	1	1.19	0.47	9.09	0.29	0.18	0.15	11.73
5(d)	1	1.19	0.89	14.96	0.36	0.27	0.24	26.39
6	5	5.95	12.38	38.71	2.01	0.66	0.54	11.14
7	4	4.76	1.47	5.43	0.72	0.22	0.16	28.45
8	4	4.76	8.92	36.80	1.56	0.62	0.52	20.23
9(a)(i,ii)	2	2.38	2.26	21.11	0.62	0.40	0.35	13.49
9(b)(i,ii)	2	2.38	3.55	24.93	0.70	0.55	0.50	14.08
10	3	3.57	5.83	34.31	1.15	0.55	0.47	31.38
11	6	7.14	2.32	5.96	0.63	0.40	0.35	22.87
12	5	5.95	4.87	9.97	1.11	0.48	0.39	32.55
13	6	7.14	0.05	0.10	0.11	0.05	0.04	73.02
14	4	4.76	0.13	0.59	0.17	0.08	0.07	79.77
15	4	4.76	4.68	11.80	1.19	0.42	0.33	41.06
16	5	5.95	2.09	4.05	0.79	0.29	0.22	37.54
17	5	5.95	2.33	3.40	0.66	0.38	0.33	48.09
Overall test	84			25.67	10.81			

Question performance, OCR higher tier paper

Question No	Mark	Weight	Achieved weighting	Facility	Standard deviation	Item-total correlation	Item-total corr. (minus item)	Non-response rate
1(a)	2	2.00	2.09	51.54	0.88	0.40	0.36	6.64
1(b)	3	3.00	4.25	27.06	1.10	0.65	0.61	35.34
2	8	8.00	9.41	60.51	2.57	0.62	0.51	2.01
3	4	4.00	4.01	27.82	1.09	0.62	0.58	3.86
4	7	7.00	10.73	52.18	2.57	0.71	0.62	2.16
5	4	4.00	6.40	61.23	1.70	0.64	0.57	16.36
6	3	3.00	3.00	14.81	1.02	0.50	0.45	18.21
7	3	3.00	4.34	28.91	1.31	0.56	0.50	23.92
8	3	3.00	4.23	32.77	1.24	0.58	0.53	8.49
9	4	4.00	6.36	29.01	1.53	0.70	0.65	21.60
10(abcd)	8	8.00	7.60	25.69	2.03	0.63	0.55	2.01
11	6	6.00	6.12	27.34	1.55	0.67	0.61	8.18
12	3	3.00	2.14	15.53	1.04	0.35	0.30	33.33
13(abcd)	8	8.00	9.19	41.07	2.37	0.66	0.57	6.64
14	4	4.00	2.28	5.25	0.79	0.49	0.45	53.86
15(a)	2	2.00	2.35	38.04	0.81	0.49	0.46	16.36
15(b)	3	3.00	4.73	28.09	1.27	0.63	0.58	32.25
16	3	3.00	1.85	10.13	0.79	0.40	0.36	28.24
17	5	5.00	1.68	3.83	0.70	0.41	0.37	44.91
18	5	5.00	2.72	6.70	0.95	0.49	0.44	53.40
19	6	6.00	2.33	4.04	0.98	0.40	0.35	42.90
20a	3	3.00	1.83	8.08	0.72	0.43	0.40	49.69
20b	3	3.00	0.38	1.80	0.35	0.18	0.17	71.60
Overall test	100			27.98	16.96			

Question performance, AQA higher tier paper

Question No	Mark	Weight	Achieved weighting	Facility	Standard deviation	Item-total correlation	Item-total corr. (minus item)	Non-response rate
1(a)	1	1.25	0.29	96.44	0.19	0.21	0.20	0.97
1(b)	1	1.25	0.95	75.08	0.43	0.30	0.27	0.97
2	1	1.25	1.74	57.77	0.49	0.47	0.45	1.13
3	1	1.25	0.68	62.78	0.48	0.19	0.15	3.07
4	2	2.50	3.02	37.54	0.80	0.51	0.46	10.68
5(abc)	4	5.00	2.91	65.09	0.93	0.42	0.36	0.97
6	3	3.75	6.20	54.21	1.34	0.62	0.56	4.05
7(abc)	3	3.75	6.48	46.39	1.27	0.69	0.64	3.88
8	1	1.25	0.63	33.33	0.47	0.18	0.15	5.34
9	5	6.25	5.17	76.47	1.66	0.42	0.31	2.10
10	3	3.75	6.33	63.65	1.37	0.62	0.56	7.77
11	3	3.75	5.57	37.38	1.22	0.62	0.56	22.01
12	3	3.75	6.20	35.33	1.23	0.68	0.63	22.01
13(ab)	6	7.50	7.93	52.27	1.83	0.59	0.48	4.85
14(ab)	3	3.75	3.40	25.78	0.85	0.54	0.49	2.91
15	3	3.75	4.64	26.27	1.09	0.57	0.51	11.97
16	4	5.00	6.42	20.06	1.32	0.66	0.60	22.33
17	4	5.00	3.08	8.25	0.96	0.43	0.37	21.68
18	4	5.00	6.48	22.21	1.37	0.64	0.57	26.54
19	4	5.00	3.29	9.22	0.87	0.51	0.46	38.83
20	1	1.25	1.00	25.73	0.44	0.31	0.28	12.14
21	3	3.75	4.78	22.38	1.10	0.59	0.53	15.37
22	4	5.00	2.98	8.05	0.85	0.47	0.42	58.09
23(a)	1	1.25	1.96	37.54	0.48	0.55	0.52	30.10
23(b)	1	1.25	1.04	24.43	0.43	0.33	0.30	27.67
23(c)	1	1.25	0.62	5.99	0.24	0.35	0.34	30.91
24(ab)	5	6.25	2.12	3.59	0.62	0.46	0.42	33.66
25	5	6.25	4.08	7.90	0.99	0.55	0.50	44.01
Overall test	80			33.92	13.51			

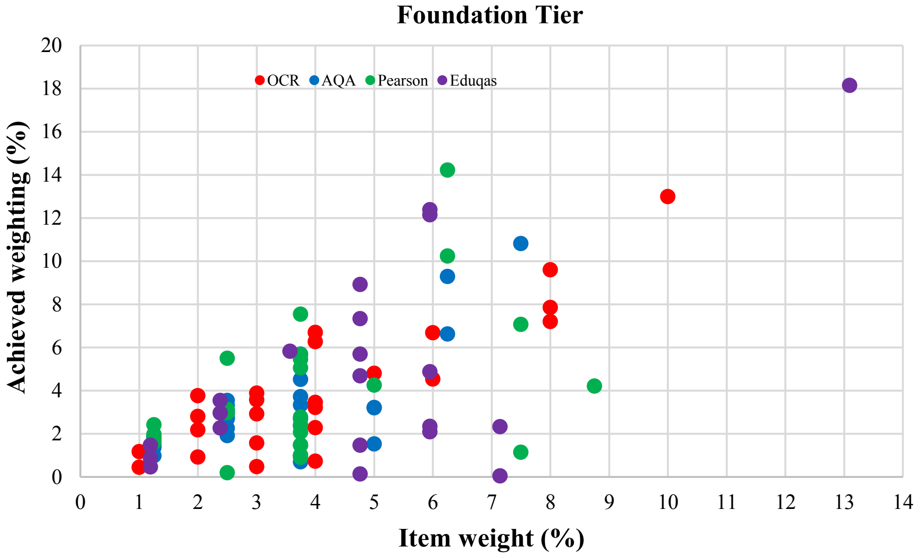
Question performance, Pearson higher tier paper

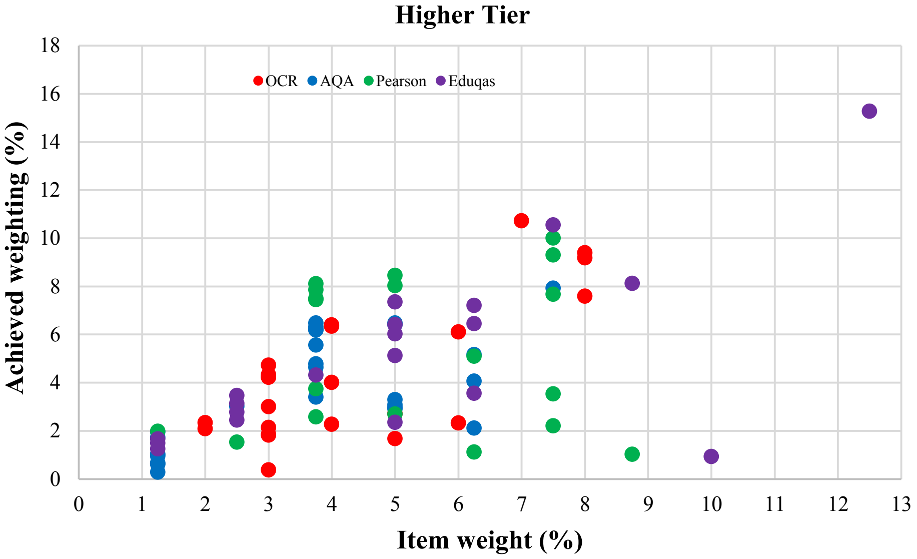
Question No	Mark	Weight	Achieved weighting	Facility	Standard deviation	Item-total correlation	Item-total corr. (minus item)	Non-response rate
1	3	3.75	7.51	39.87	1.40	0.56	0.45	4.47
2	6	7.50	10.01	41.36	1.83	0.57	0.43	7.02
3(i,ii)	2	2.50	1.53	5.66	0.39	0.41	0.38	10.05
4	3	3.75	2.59	14.19	0.71	0.38	0.32	21.69
5(a)	3	3.75	7.87	40.46	1.30	0.63	0.54	12.44
5(b)	3	3.75	7.45	53.06	1.42	0.55	0.44	15.31
6(ab)	6	7.50	7.67	16.80	1.29	0.62	0.53	5.90
7	3	3.75	8.11	41.52	1.32	0.64	0.55	7.18
8	3	3.75	3.74	8.35	0.70	0.56	0.51	24.72
9	4	5.00	8.47	20.22	1.36	0.65	0.56	18.66
10	4	5.00	8.04	36.84	1.78	0.47	0.32	23.76
11(a)	1	1.25	1.98	44.82	0.50	0.41	0.37	6.70
11(b)	5	6.25	5.10	11.67	1.05	0.50	0.42	17.54
12(ab)	5	6.25	1.13	2.36	0.36	0.33	0.30	57.89
13	4	5.00	2.70	3.71	0.64	0.44	0.39	45.45
14(abc)	6	7.50	9.31	13.24	1.40	0.69	0.61	18.50
15	6	7.50	3.54	5.13	0.74	0.50	0.44	34.93
16	6	7.50	2.21	2.45	0.59	0.39	0.34	68.58
17	7	8.75	1.03	1.16	0.34	0.32	0.29	55.98
Overall test	80			18.05	10.40			

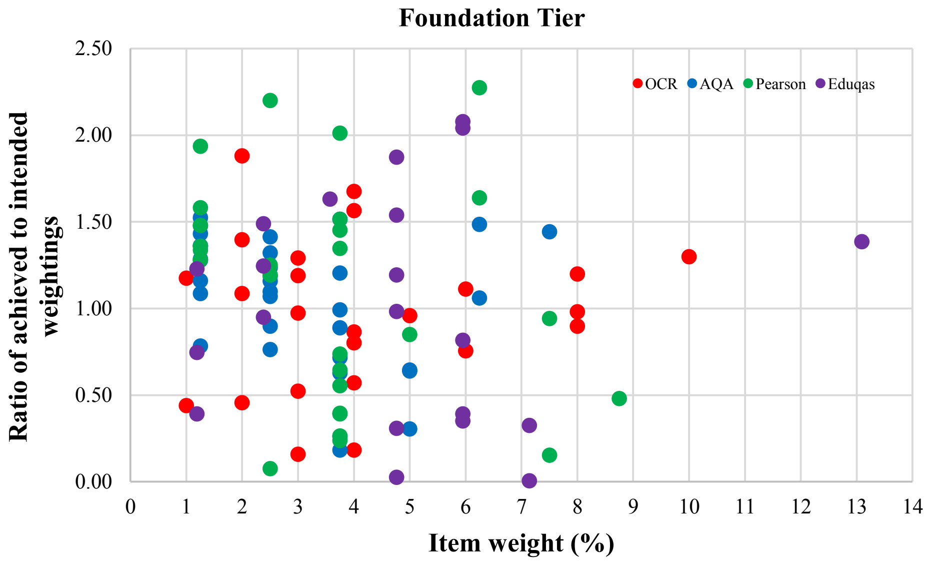
Question performance, Eduqas higher tier paper

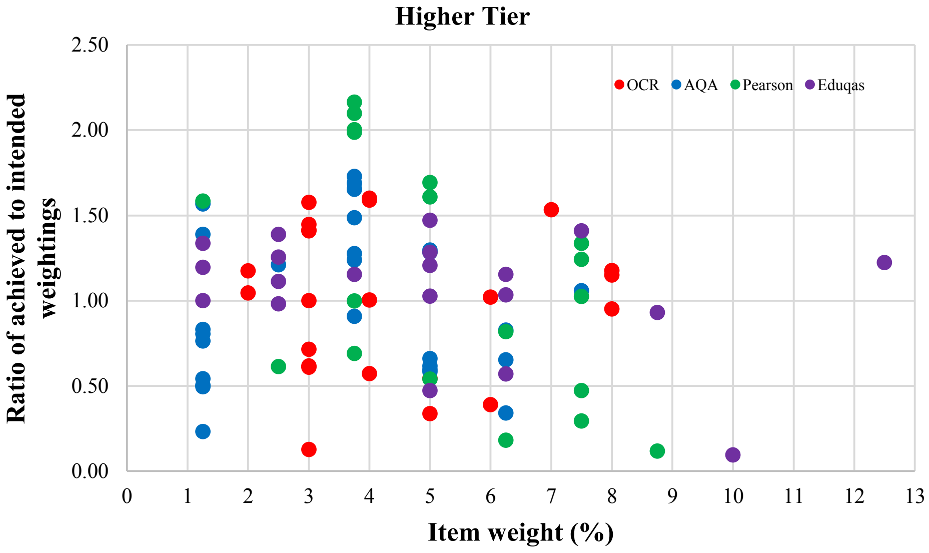
Question No	Mark	Weight	Achieved weighting	Facility	Standard deviation	Item-total correlation	Item-total corr. (minus item)	Non-response rate
1(a)	1	1.25	1.25	61.42	0.49	0.38	0.36	8.63
1(b)	1	1.25	1.67	50.59	0.50	0.50	0.47	5.58
1(c)	2	2.50	2.45	17.77	0.69	0.53	0.50	12.69
2	4	5.00	6.03	32.53	1.58	0.57	0.49	7.11
3	4	5.00	5.13	36.76	1.48	0.52	0.44	13.03
4	7	8.75	8.14	12.16	1.62	0.75	0.70	29.10
5	4	5.00	6.42	21.02	1.38	0.70	0.64	41.46
6	5	6.25	6.45	27.11	1.58	0.61	0.54	7.95
7	6	7.50	10.56	47.86	2.40	0.66	0.55	13.20
8(abcd)	10	12.50	15.28	31.05	3.01	0.76	0.65	3.55
9	4	5.00	2.36	4.53	0.75	0.47	0.43	23.18
10(a)	4	5.00	7.35	21.95	1.54	0.72	0.66	16.75
10(b)	2	2.50	3.47	23.94	0.81	0.64	0.61	38.75
11	5	6.25	3.56	8.66	0.99	0.54	0.49	28.60
12(a)	3	3.75	4.32	19.74	1.03	0.63	0.58	27.75
12(b)	2	2.50	3.14	25.21	0.79	0.59	0.56	32.66
12(c)	1	1.25	1.49	27.07	0.44	0.50	0.48	50.25
13(a)	5	6.25	7.21	14.52	1.52	0.71	0.65	19.29
13(b)	2	2.50	2.78	19.46	0.70	0.59	0.56	34.86
14	8	10.00	0.94	1.10	0.41	0.35	0.32	48.05
Overall test	80			22.27	14.97			

Question intended and achieved weightings






Foundation Tier
Ratio of achieved to intended weightings
Item weight (%)
● OCR ● AQA ● Pearson ● Eduqas
0.00
0.50
1.00
1.50
2.00
2.50
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14


Higher Tier
Ratio of achieved to intended weightings
Item weight (%)
● OCR ● AQA ● Pearson ● Eduqas
0.00
0.50
1.00
1.50
2.00
2.50
0
1
2
3
4
5
6
7
8
9
10
11
12
13

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 2nd Floor
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Glendinning House
6 Murray Street
Belfast BT1 6DN

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