



A comparative judgement study of MEI GCSE exam items

Dr Patrick Barmby and Dr Chris Wheadon, No More Marking Ltd.

Report date: 18 September 2019

1. Introduction

In this report, we present the results of a comparative judgement study of the difficulty of GCSE mathematics exam items and papers developed by MEI as part of their funded project 'A new mathematics GCSE curriculum for post-16 resit students'. We will begin by detailing the comparative judgement methodology used in the study. We then present the difficulties of the MEI items and the items they are compared against, and the calculated weighted means of the difficulties of the MEI and comparison papers. The research question to be answered in this study was "Are the new MEI GCSE exam papers comparable in difficulty to existing GCSE papers?"

2. Comparative judgement

Comparative judgement is an analytical process in which judges use their professional judgement to compare two scripts at a time, and to decide which of these scripts is 'better' in each case. Repeated comparisons result in a measurement scale showing the relative quality of the scripts (Pollitt, 2012). In this study, comparative judgement was used to compare individual mathematics items from various examination papers. The judgements were carried out by PhD/Masters mathematics students, deciding in each case which item was the 'more difficult'. The mathematics items examined in this way are placed on a scale of difficulty with higher scores for items denoting more difficult questions.

To carry out the comparative judgement process, No More Marking's¹ online platform was used. Individual mathematics items were uploaded to the platform which then presented items side by side to judges who clicked 'left' or 'right' to decide the more difficult item.

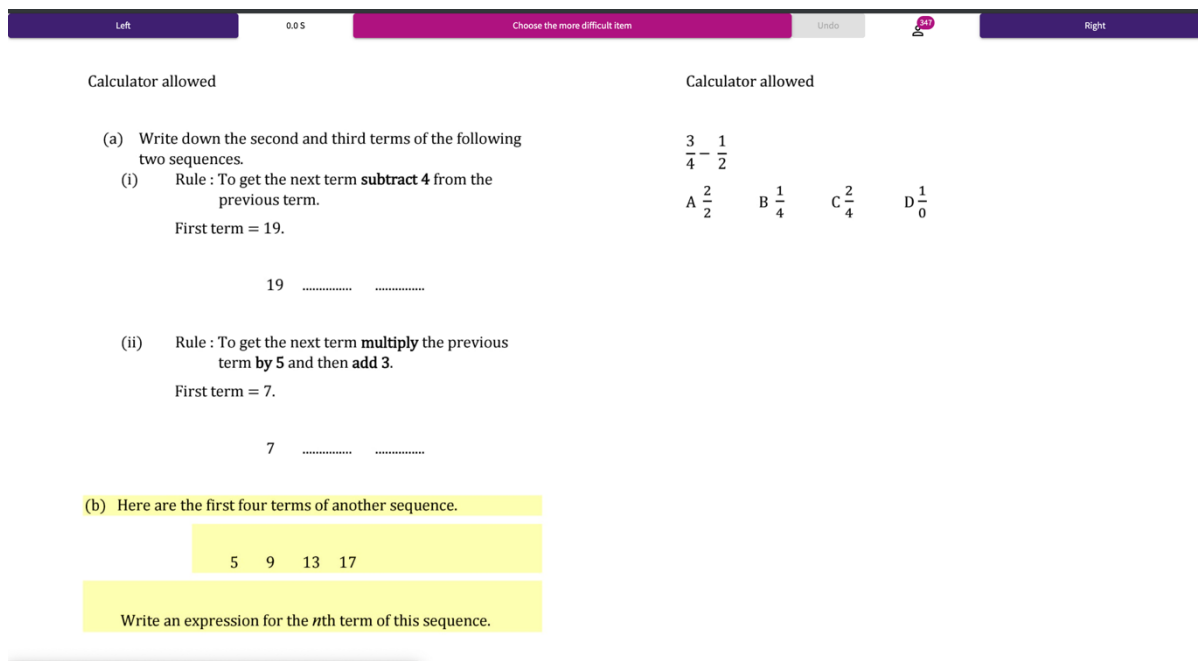


Figure 1: No More Marking's online comparative judgement platform

The built-in analytics of the platform were used to calculate the resulting difficulties of the mathematics items.

3. Method

3.1. Materials

In the study, the complete set of items from the following papers were included:

MEI Sample papers:

- Paper 1 (Calculator allowed)
- Paper 2 (No calculator allowed)
- Paper 3 (Calculator allowed)

¹ See www.nomoremarking.com

AQA Functional Skills (8362) Level 2 Specimen papers:

- Paper 1 (Non-Calculator paper)
- Paper 2 (Calculator paper)

The AQA Functional Skills items were included simply as another set of comparison items (we also compared against other GCSE papers through the anchoring process – detailed below in section 3.2).

Each MEI and AQA Functional Skills paper included the following number of individual items:

Table 1: The papers included in the analysis and the number of items in each

| Paper | No. Items |
|-------------------------------|------------------|
| MEI Paper 1 | 40 |
| MEI Paper 2 | 37 |
| MEI Paper 3 | 30 |
| AQA Functional Skills Paper 1 | 9 |
| AQA Functional Skills Paper 2 | 21 |

By item, this meant individual parts of the paper. Therefore, a question with, for example, parts 1(a), 1(b)(i) and 1(b)(ii) would be considered as three separate items. Therefore, the papers above included a total of 137 items.

3.2. Anchor items

In addition to the 137 items included from the above papers, a further 60 anchor items were included from the 2017 Ofqual analysis of English reformed GCSE items (Ofqual, 2017). These anchor items spanned the range of item difficulties found in that study. The inclusion of these anchor items allowed us to place the results of the present study on the same scale as this previous Ofqual study. As a result, direct comparisons could be made between the papers included in this analysis, and the papers included in the Ofqual study including previous English GCSE papers. Therefore, in total, 197 items were included in the present analysis.

3.3. Transcription of items

All the items included in the analysis were transcribed according to a common typeset, and any indication of where each item was from was removed from the items. In addition, for multi-item questions, it was made clear which part of the question was to be considered during the judging process. For example, when considering the following question containing numerous parts or items, the two items were presented separately in this way:

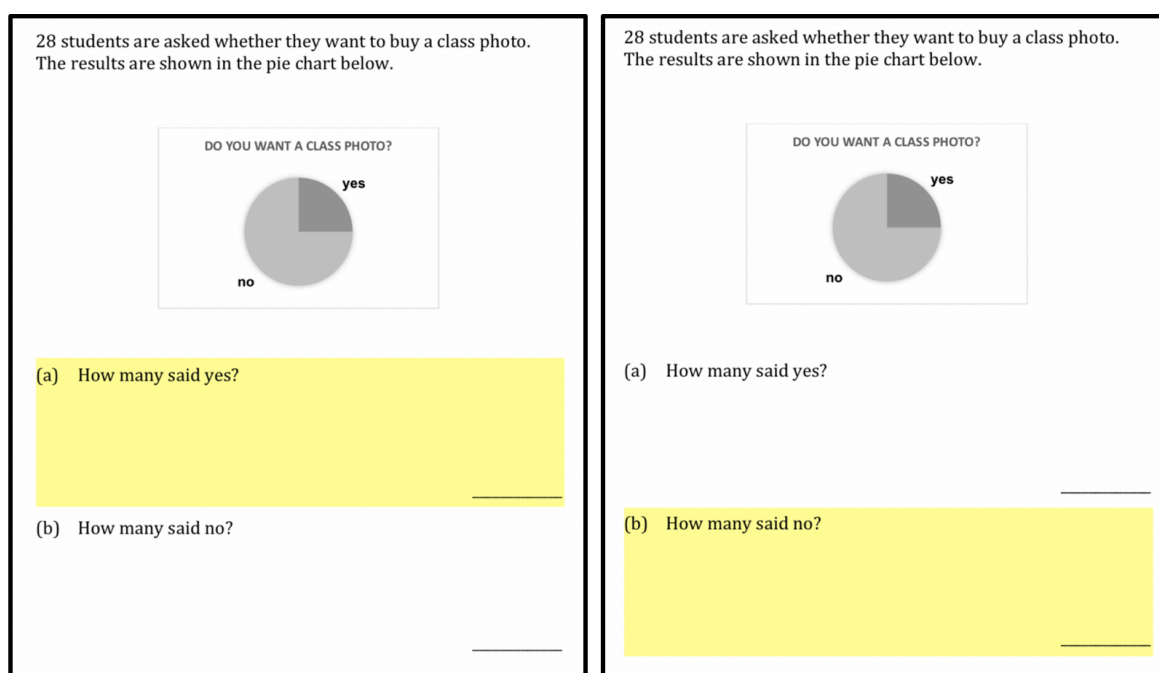


Figure 2: Example of presenting separate items within a question

3.4. Judges

12 judges were recruited to carry out the comparative judgement process. These were PhD students in mathematics who had previously taken part in the 2017 Ofqual analysis, and one new judge with a Masters in mathematics. The judges were paid for their time. Detailed instructions were provided to the judges including the following:

In your judging, you will see two maths questions side by side. Decide on which is the more difficult question

When the judges had signed up for the task, they were presented with two items side by side, and chose the more difficult item by clicking the left or right buttons. This procedure was repeated until the judges had completed their allocated number of judgements.

4. Analysis

4.1. Method of analysis

The No More Marking online platform employs a statistical routine based on the Bradley-Terry model to estimate item difficulties (Pollitt, 2012). In addition to item difficulties, the routine provides estimates of the overall reliability of the analysis process, judge infit and item infit. The judge infit is a measure of the degree of difference between a given judge and the rest of the cohort of judges in terms of their judgements made, and is an indication of the consistency of the judgements made by that given judge. The item infit is an indication of the degree of disagreement between judges, for whatever reason, regarding that item's difficulty.

4.2. Judge consistency

All the judgements were carried out between the 8th August and 6th September 2019 (Figure 3).

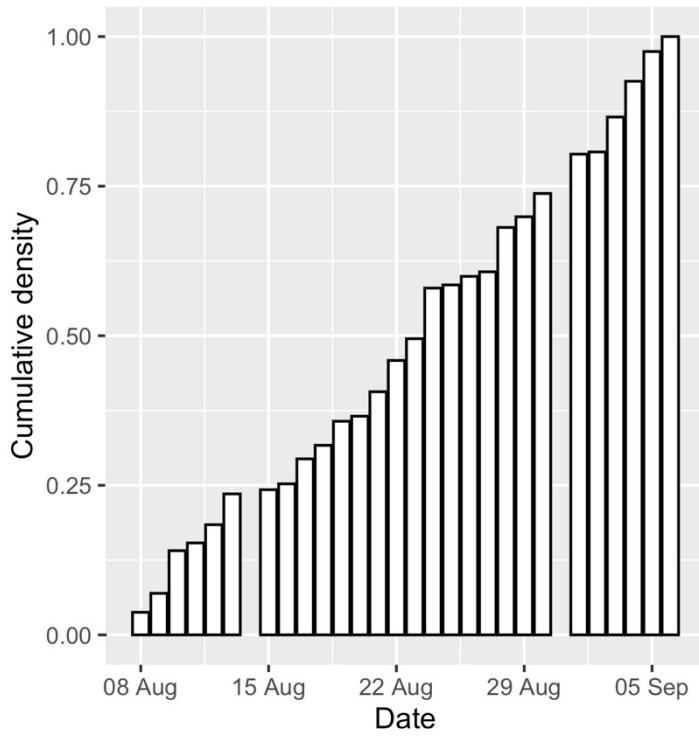


Figure 3: Cumulative frequency graph for when the judgements were carried out

The median time for judgements was 19.5s (Figure 4).

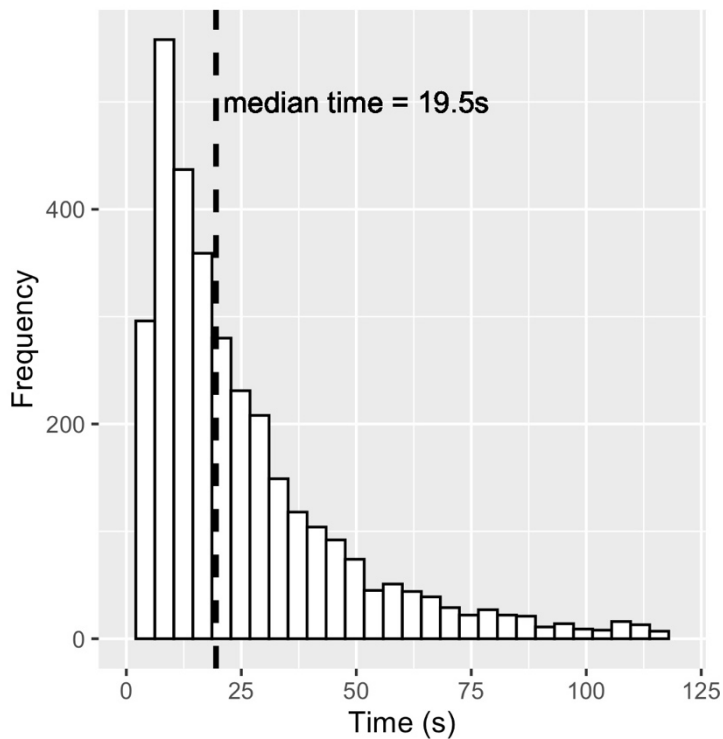


Figure 4: Judgement times (times greater than 120s not shown)

The judge infit values for the 12 judges ranged from 0.7 to 1.6. The greater the judge infit value, the greater the amount of inconsistency indicated. When excluding the judges with the higher infits, the reliability of the task did not change, so all the judges' decisions were retained in the analysis.

4.3. Reliability

In total, there were 3573 judgements for the 197 items considered in the analysis. This meant that on average, as two items were seen in each judgement, each item was seen around 36 times in the analysis. The reliability of the assessment of item difficulties was calculated to be 0.89.

We can also check whether we carried out enough judgements for optimum reliability. Figure 5 shows how the reliability varied for this task as we randomly chose increasing numbers of judgements to include in the comparative judgement statistical calculation.

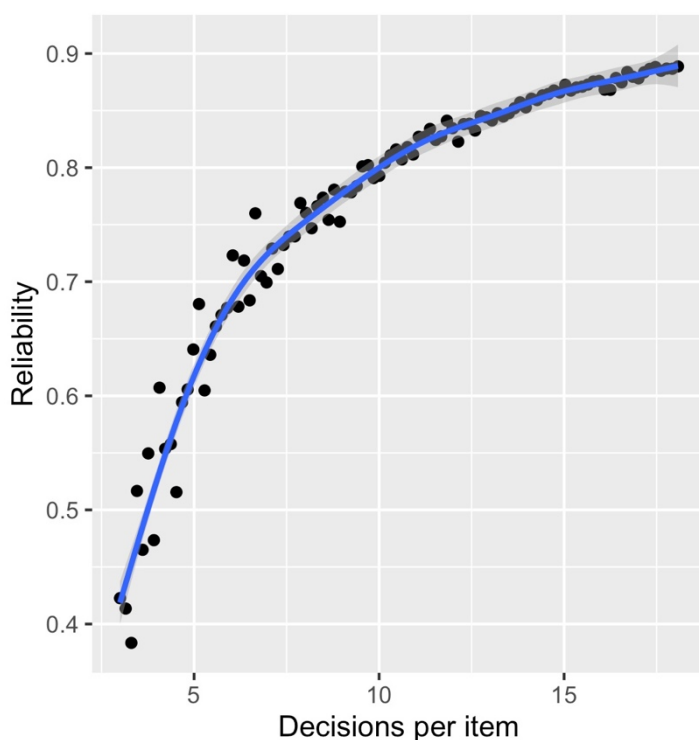


Figure 5: Variation of reliability with decisions per item

The flattening of the above curve shows that we started to reach the optimum reliability for this task, and therefore enough decisions were completed.

4.4. Reliability of anchors

To check on how well the process was anchoring the scores for the MEI and AQA Functional Skills items on to the scale previously used by Ofqual, the correlation between the item difficulties for the anchor items and the difficulty values when the anchoring was switched off was calculated.

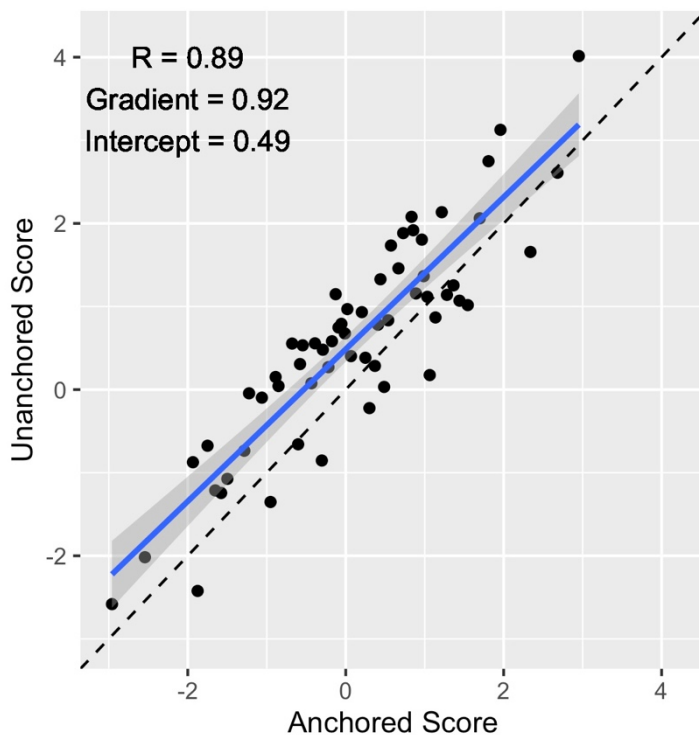


Figure 6: Scatter plot of Annotated Scores vs. Unanchored scores for the anchor items used

Ideally, there should be a high correlation between the anchored and unanchored values, and the gradient for the relationship should be 1. For this analysis, the correlation was 0.89 and the gradient of the graph was 0.92. There was therefore very good agreement between the anchored and unanchored scores, providing confidence in the anchoring process.

5. Results

Having established the reliability of the analysis process, we now present the results of the comparative judgement process.

5.1. Performance by paper

Firstly, the mean difficulties for each individual paper were calculated by averaging over the item difficulties for the given paper. In this averaging process, the weighting of items was included based on the number of marks (tariff) for each item. The resulting mean item difficulties for all the papers included in the analysis are plotted below with 95 % confidence intervals. Figure 7 also includes all the original papers from the 2017 Ofqual analysis of English reformed GCSE items.

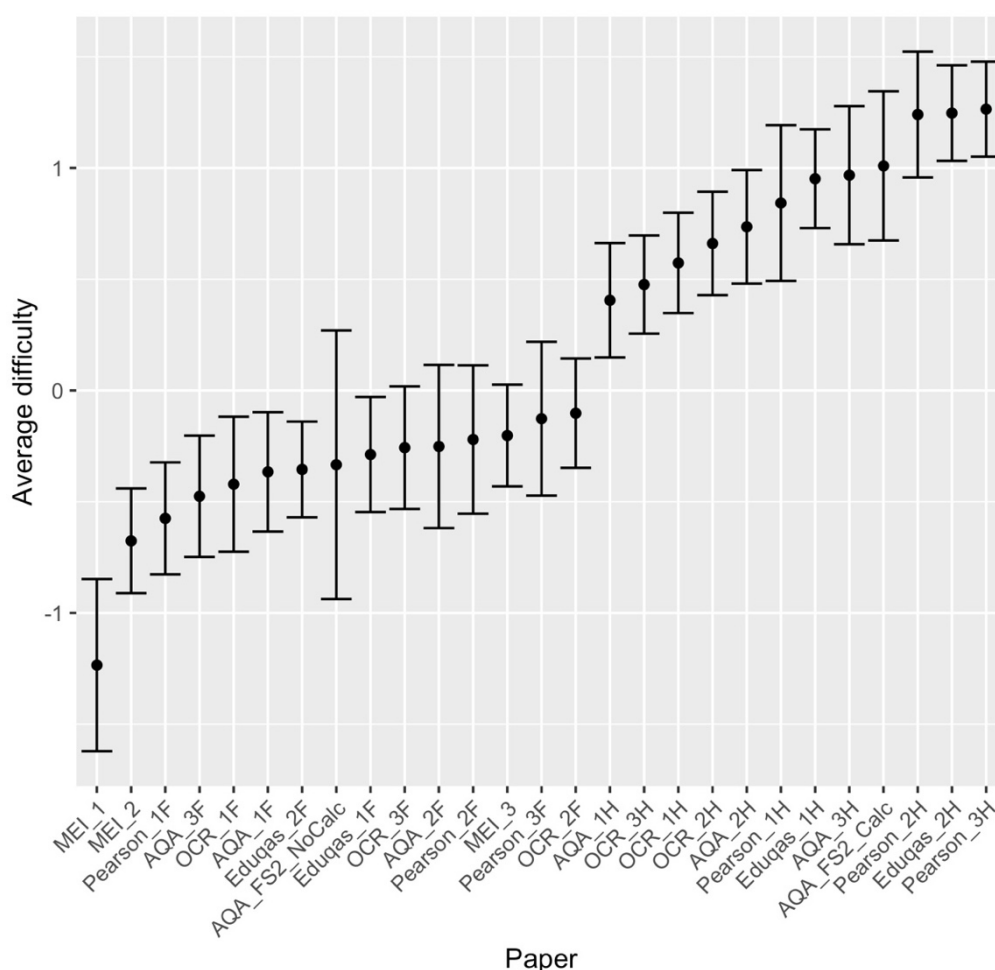


Figure 7: Average difficulties of different exam papers

We can see that the MEI Paper 1 was judged to be the easiest paper in the selection, whilst MEI Paper 3 lay towards the middle of the range of papers.

At this point, we can focus down our analysis to comparison papers that are particularly relevant for this study of the MEI papers. The papers being developed by MEI are aimed at the GCSE Foundation level. Therefore, we can just examine the difficulties of the past Foundation papers examined by Ofqual, as well as the MEI papers (Figure 8):

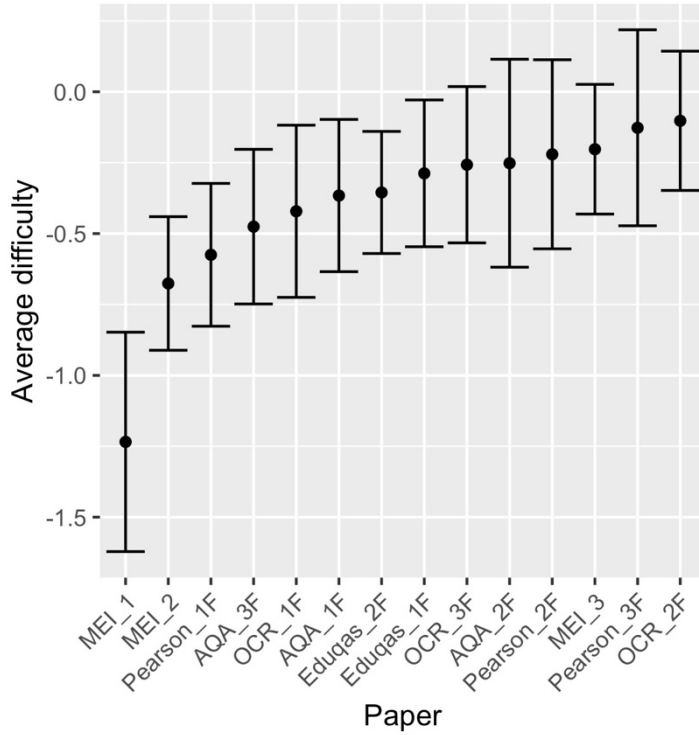


Figure 8: Average difficulties of the MEI and comparison GCSE Foundation papers

All the papers are comparable in difficulty except for MEI paper 1 which appears to be easier than the others. We therefore suggest that MEI paper 1 could be brought in line more with the other papers in terms of difficulty.

If we combine each group of papers (e.g. papers 1, 2 and 3) together:

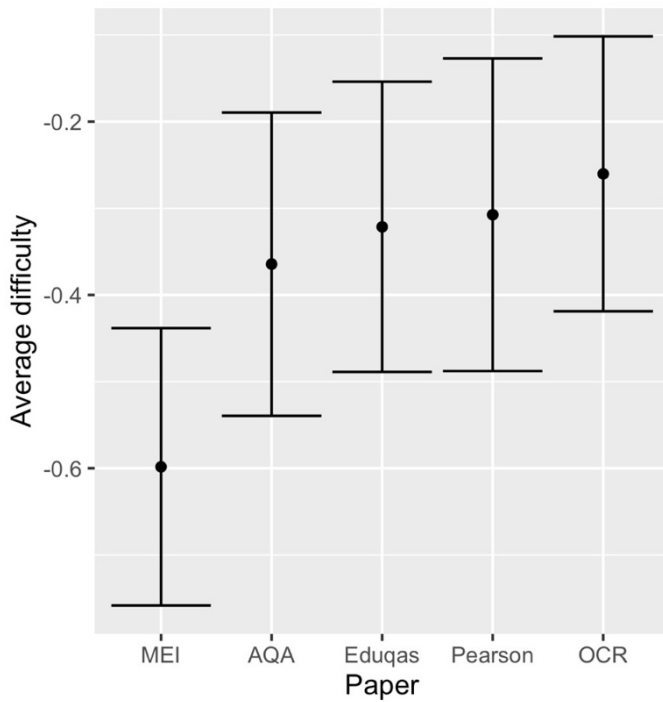


Figure 9: Average difficulties of the MEI and comparison GCSE Foundation papers (combined)

From the overlapping of the error bars above, we concluded that there was no significant difference in difficulty between the MEI papers and the comparison GCSE Foundation papers. However, ideally, the MEI papers could be more difficult so that they are more in line with the other comparison Foundation papers.

5.2. Performance of individual items - difficulties

Looking at the individual MEI items, we can see the distribution of difficulties:

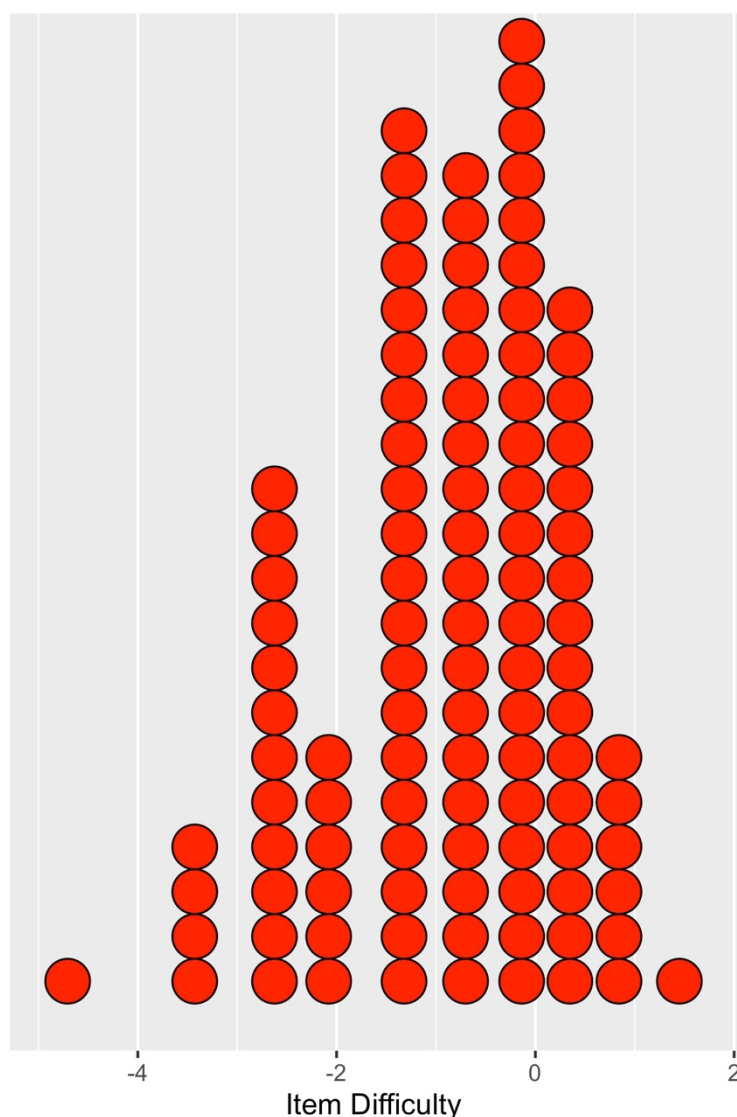


Figure 10: Distribution of individual item difficulties

It may be instructive to look at the characteristics of the easier or harder MEI items.

Looking at the four easiest MEI items:

Calculator allowed

How many grams in a kilogram?

A 10 B 100 C 1000 D 10 000


Calculator allowed

Round the number 0.012851 to 3 decimal places.

A 0.012 B 0.013 C 0.0128 D 0.0129

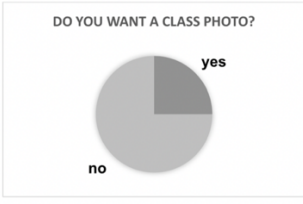
Calculator allowed

Which of the digital clocks could be showing the same time as the clock below?



| | |
|------------|------------|
| A 02:06 | B 03:30 |
| C 06:12 | D 14:30 |

28 students are asked whether they want to buy a class photo. The results are shown in the pie chart below.



(a) How many said yes? _____

(b) How many said no? _____

Figure 11: Four easiest MEI items

In turn, looking at the four hardest MEI items:

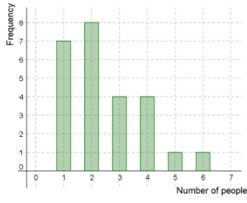
(c) Jill's car depreciates in value by 15% a year. It is worth £5000 when she buys it.

How much will it be worth after 3 years?

Give your answer to the nearest pound

Calculator allowed

The bar chart below shows the number of people living in each house on a particular street.



(a) Find the total number of people living on the street.

(b) 100 years ago the mean number of people in a house on the same street was 4.2.

Calculate the total number of people living on the street 100 years ago.

A carton of fruit juice contains 1.2 litres.

(a) How many 200 ml glasses will it fill? _____

The carton is a cuboid made of thin card. David suggests a new design for the carton. David's design is a cuboid with square base 5 cm long, 5 cm wide.

(b) Work out the height of David's design. _____ cm

(c) Give one reason why a cuboid with smaller height would be a better design.

Calculator allowed

Hayley is going on holiday. The maximum weight her suitcase can be is 20 kg. Hayley's weighing scales give 19.4 kg as the weight of the case. Her weighing scales are not accurate; they can be wrong by up to 5%.

Which of the following statements about the real weight of the suitcase is true?

A It could be up to 19.9 kg.
 B It must be more than 20 kg.
 C It must be exactly 20.37 kg.
 D It could be more than 20 kg but is less than 21 kg.

Figure 12: Four hardest MEI items

Two main properties of the easy/hard items can be seen. The easier MEI items tend to be multiple choice items and are single step questions largely involving factual recall (e.g. how to convert to 24-hour time). The more difficult MEI items are mostly multiple step questions with room for working out the calculations, however one of the difficult

questions is a multiple choice question. In order to increase the difficulty of MEI paper 1 (multiple choice) as suggested previously, it is suggested that more questions with multiple steps such as the one identified could be included in paper 1.

The difficulties of all the individual MEI items, along with the marks tariff for each item, are given in the appendix of this report.

5.3. Performance of individual items – item infit

Looking at the infit for the MEI items, the two items with the highest infit (1.6), and therefore items more likely to have possible disagreement between judges, were:

Calculator allowed

Nadia's target is to walk at least 10 000 steps each day.

The numbers of steps for each day in one week are shown in the table below.

| Mon | Tues | Weds | Thurs | Fri | Sat | Sun |
|------|------|------|--------|--------|--------|------|
| 8119 | 6156 | 9006 | 12 051 | 10 000 | 13 949 | 5910 |

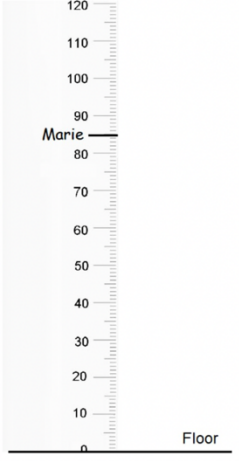
On how many days did Nadia meet her target?

A 1 B 2 C 3 D 4

Part of a vertical measuring scale is shown on the right.

The scale is in centimetres and starts at floor level.

It is used to measure children's heights.



(a) The mark on the scale shows Marie's height.

How tall is Marie?

Figure 13: MEI items with the highest item infit values

It is not entirely clear why these items may have been more likely to cause disagreements amongst judges; with the left-hand item, it may be due to assumptions needed for the question (that Nadia started her walking on Monday), but this remains speculative. The highest infit value of 1.6, which is not very high, suggests that there was not great disagreement amongst the judges regarding the MEI items.

6. Conclusion

In this report we have reported the details of the comparative judgement process used to estimate the difficulty of MEI and comparison GCSE maths items. The process was found

to be highly reliable, with confidence in the accuracy of the judging and the anchoring methods used.

In answer to the research question “Are the new MEI GCSE exam papers comparable in difficulty to existing GCSE papers?”, from the results of the analysis, we have concluded that overall the MEI papers are not significantly different in difficulty to previous GCSE Foundation papers. However, we did find that ideally, the difficulty of the MEI papers could be increased a little so that they are more in line with the comparison Foundation papers. In particular, we suggest that the difficulty of MEI paper 1 could be increased, perhaps by including multiple choice questions containing multiple steps rather than those with more straightforward factual recall.

7. References

Ofqual (2017). *An evaluation of the difficulty of the assessments and the characteristics of the problem-solving (AO3) items*. Coventry: Ofqual.

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: principles, policy & practice*, 19(3), 281-300.

Appendix: Difficulties of MEI items

| Item | Difficulty | Tariff |
|---------------|-------------------|---------------|
| MEI_1_Q_1 | -4.71 | 1 |
| MEI_2_Q_5b | -3.64 | 2 |
| MEI_1_Q_5 | -3.56 | 1 |
| MEI_1_Q_14 | -3.38 | 1 |
| MEI_1_Q_3 | -3.22 | 1 |
| MEI_2_Q_5a | -2.85 | 2 |
| MEI_1_Q_6 | -2.84 | 1 |
| MEI_3_Q_4b | -2.77 | 1 |
| MEI_1_Q_17 | -2.76 | 1 |
| MEI_1_Q_2 | -2.74 | 1 |
| MEI_2_Q_9a | -2.72 | 2 |
| MEI_1_Q_8 | -2.64 | 1 |
| MEI_3_Q_4a | -2.58 | 2 |
| MEI_2_Q_6 | -2.53 | 2 |
| MEI_2_Q_1a | -2.51 | 2 |
| MEI_2_Q_1b | -2.41 | 1 |
| MEI_1_Q_24 | -2.40 | 1 |
| MEI_1_Q_15 | -2.22 | 1 |
| MEI_1_Q_16 | -2.18 | 1 |
| MEI_3_Q_1 | -2.05 | 2 |
| MEI_2_Q_1c | -1.96 | 2 |
| MEI_3_Q_6a | -1.95 | 3 |
| MEI_3_Q_4c | -1.94 | 1 |
| MEI_3_Q_6c | -1.56 | 1 |
| MEI_3_Q_6b | -1.56 | 2 |
| MEI_2_Q_10a | -1.52 | 2 |
| MEI_1_Q_22 | -1.45 | 1 |
| MEI_1_Q_18 | -1.45 | 1 |
| MEI_2_Q_3b | -1.43 | 2 |
| MEI_3_Q_8bi | -1.42 | 2 |
| MEI_2_Q_7b | -1.37 | 1 |
| MEI_1_Q_30 | -1.37 | 1 |
| MEI_2_Q_10c | -1.32 | 1 |
| MEI_1_Q_20 | -1.31 | 1 |
| MEI_1_Q_4 | -1.28 | 1 |
| MEI_2_Q_7a | -1.27 | 1 |
| MEI_3_Q_8a | -1.21 | 2 |
| MEI_1_Q_31 | -1.19 | 1 |
| MEI_1_Q_37 | -1.18 | 1 |
| MEI_1_Q_13 | -1.18 | 1 |
| MEI_3_Q_8bii | -1.12 | 3 |
| MEI_2_Q_12a | -1.11 | 1 |
| MEI_2_Q_13b | -1.08 | 2 |
| MEI_1_Q_11 | -0.94 | 1 |
| MEI_1_Q_36 | -0.94 | 1 |
| MEI_1_Q_7 | -0.93 | 1 |
| MEI_2_Q_4 | -0.93 | 2 |
| MEI_2_Q_3a | -0.92 | 2 |
| MEI_2_Q_13e | -0.92 | 1 |
| MEI_1_Q_12 | -0.91 | 1 |
| MEI_3_Q_11a | -0.89 | 3 |
| MEI_2_Q_2a | -0.73 | 2 |
| MEI_3_Q_13a | -0.70 | 2 |
| MEI_1_Q_33 | -0.64 | 1 |
| MEI_2_Q_8 | -0.63 | 3 |
| MEI_1_Q_25 | -0.62 | 1 |
| MEI_2_Q_10b | -0.55 | 1 |
| MEI_2_Q_2b | -0.54 | 1 |
| MEI_3_Q_3 | -0.54 | 4 |
| MEI_1_Q_10 | -0.50 | 1 |
| MEI_2_Q_13a | -0.48 | 4 |
| MEI_1_Q_19 | -0.46 | 1 |
| MEI_2_Q_12b | -0.39 | 2 |
| MEI_1_Q_35 | -0.37 | 1 |
| MEI_2_Q_13c | -0.34 | 2 |
| MEI_1_Q_40 | -0.31 | 1 |
| MEI_1_Q_23 | -0.26 | 1 |
| MEI_1_Q_9 | -0.25 | 1 |
| MEI_2_Q_14c | -0.20 | 3 |
| MEI_2_Q_12c | -0.20 | 2 |
| MEI_3_Q_5 | -0.19 | 2 |
| MEI_1_Q_26 | -0.19 | 1 |
| MEI_3_Q_2 | -0.17 | 4 |
| MEI_2_Q_15b | -0.09 | 3 |
| MEI_1_Q_28 | -0.09 | 1 |
| MEI_1_Q_32 | -0.09 | 1 |
| MEI_1_Q_29 | -0.06 | 1 |
| MEI_3_Q_15a | -0.05 | 2 |
| MEI_1_Q_34 | -0.04 | 1 |
| MEI_3_Q_10a | -0.03 | 2 |
| MEI_1_Q_27 | 0.04 | 1 |
| MEI_2_Q_11 | 0.04 | 6 |
| MEI_1_Q_38 | 0.05 | 1 |
| MEI_3_Q_12cii | 0.11 | 2 |

| | | |
|--------------|------|---|
| MEI_3_Q_12b | 0.12 | 3 |
| MEI_2_Q_9b | 0.18 | 3 |
| MEI_2_Q_14a | 0.20 | 3 |
| MEI_2_Q_9d | 0.23 | 3 |
| MEI_2_Q_7c | 0.24 | 2 |
| MEI_2_Q_15a | 0.27 | 4 |
| MEI_1_Q_21 | 0.29 | 1 |
| MEI_2_Q_15c | 0.29 | 2 |
| MEI_2_Q_13d | 0.32 | 2 |
| MEI_3_Q_10b | 0.37 | 5 |
| MEI_3_Q_7 | 0.38 | 4 |
| MEI_3_Q_12ci | 0.38 | 1 |
| MEI_3_Q_15c | 0.38 | 5 |
| MEI_3_Q_11b | 0.46 | 3 |
| MEI_3_Q_15b | 0.56 | 2 |
| MEI_3_Q_12a | 0.57 | 3 |
| MEI_3_Q_9 | 0.63 | 4 |
| MEI_2_Q_14b | 0.65 | 3 |
| MEI_3_Q_14 | 0.81 | 3 |
| MEI_2_Q_9c | 0.83 | 1 |
| MEI_1_Q_39 | 0.88 | 1 |
| MEI_3_Q_13b | 1.05 | 3 |
| MEI_3_Q_11c | 1.45 | 4 |
