GCSE science: An evaluation of the expected difficulty of items

February 2017

Ofqual/17/6163

Contents

Executive Summary

In 2015, exam boards submitted draft science specifications (biology, chemistry and physics single award, and combined science double award) to Ofqual for the purposes of accreditation for first teaching in 2016. We use accreditation to decide whether new GCSEs, AS and A level qualifications produced by exam boards can be awarded.

The accreditation of reformed GCSE science specifications included the evaluation of overall qualification demand which is determined by many features, of which the difficulty of items is just one part. Compared to legacy science qualifications, the reformed qualifications have increased subject content, a linear assessment structure and only examined assessment (i.e. they do not use any form of non-exam assessment). All of these factors were considered by the accreditation panel in 2015/2016, alongside the expected difficulty of items estimated by comparative judgement reported here.

Ofqual carried out 2 phases of comparative judgement studies of the relative expected difficulty of science items from the 2014 legacy specifications together with items from the sample assessment materials (SAMs) for the first and second submissions of the reformed science specifications. The purpose of this was to inform discussions and recommendations made by the accreditation panels regarding the likely difficulty of future live examinations. Unlike previous work looking at GCSE mathematics assessments[1], this was not an inter-board comparability exercise. Comparisons were focussed on the relative expected difficulty of items from the 2014 papers and the SAMs within each specification.

Overall the distribution of expected difficulty of items was very similar between the legacy and reformed specifications. Figures 1 to 4 on the following pages show the distributions of expected item difficulties for each of the science subjects aggregated across all of the specifications. These graphs give an overview of the expected difficulty of the items pre and post-reform for 2 phases of accreditation submission. The small levels of variation between the expected difficulty distributions of the reformed specifications is very similar to the variability observed in the legacy specifications. Such small differences can easily be accounted for in the setting of grade boundaries during awarding, and are therefore of no substantive impact.

The data presented only covers the reformed sample assessments up to and including the second submission for accreditation. All of the specifications went

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

through additional submissions which included some changes to their SAMs, for example where there were concerns about their level of demand and so the phase 2 study does not precisely represent the final accredited specifications. However, the scope of change requested to items during the later phases of accreditation was not large and a subsequent phase of comparative judgement was not deemed necessary.

Finally, a new approach to accounting for some subtle, unconscious but systematic biases in the judgements made by the expert judges was used to adjust the expected difficulty estimates. This process is described in detail in Appendix A.

Biology – Overall

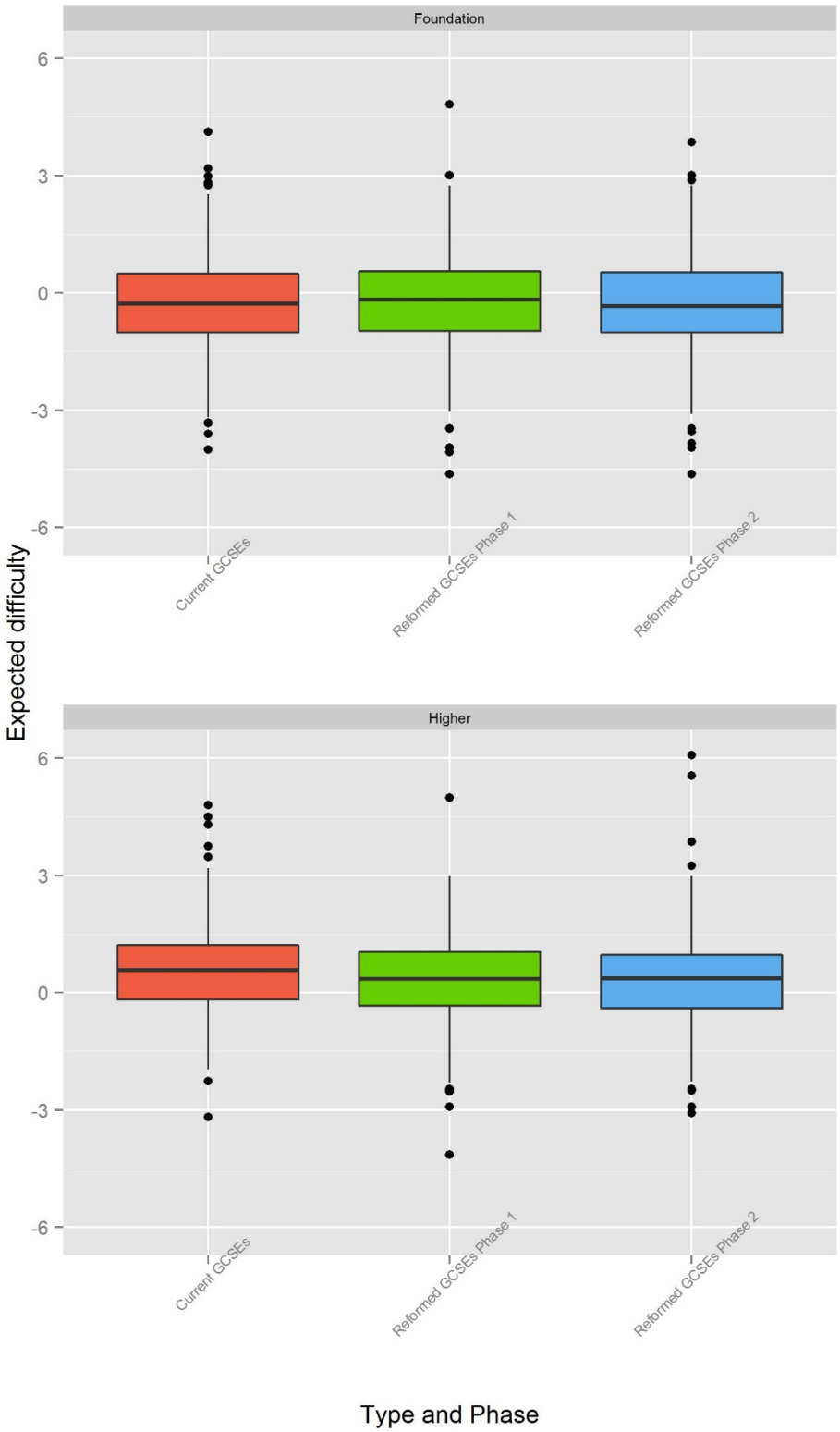


Figure 1: *Box plots showing median and interquartile ranges of expected item difficulties for the aggregated 2014 assessments and sample assessments for biology.*

Chemistry – Overall

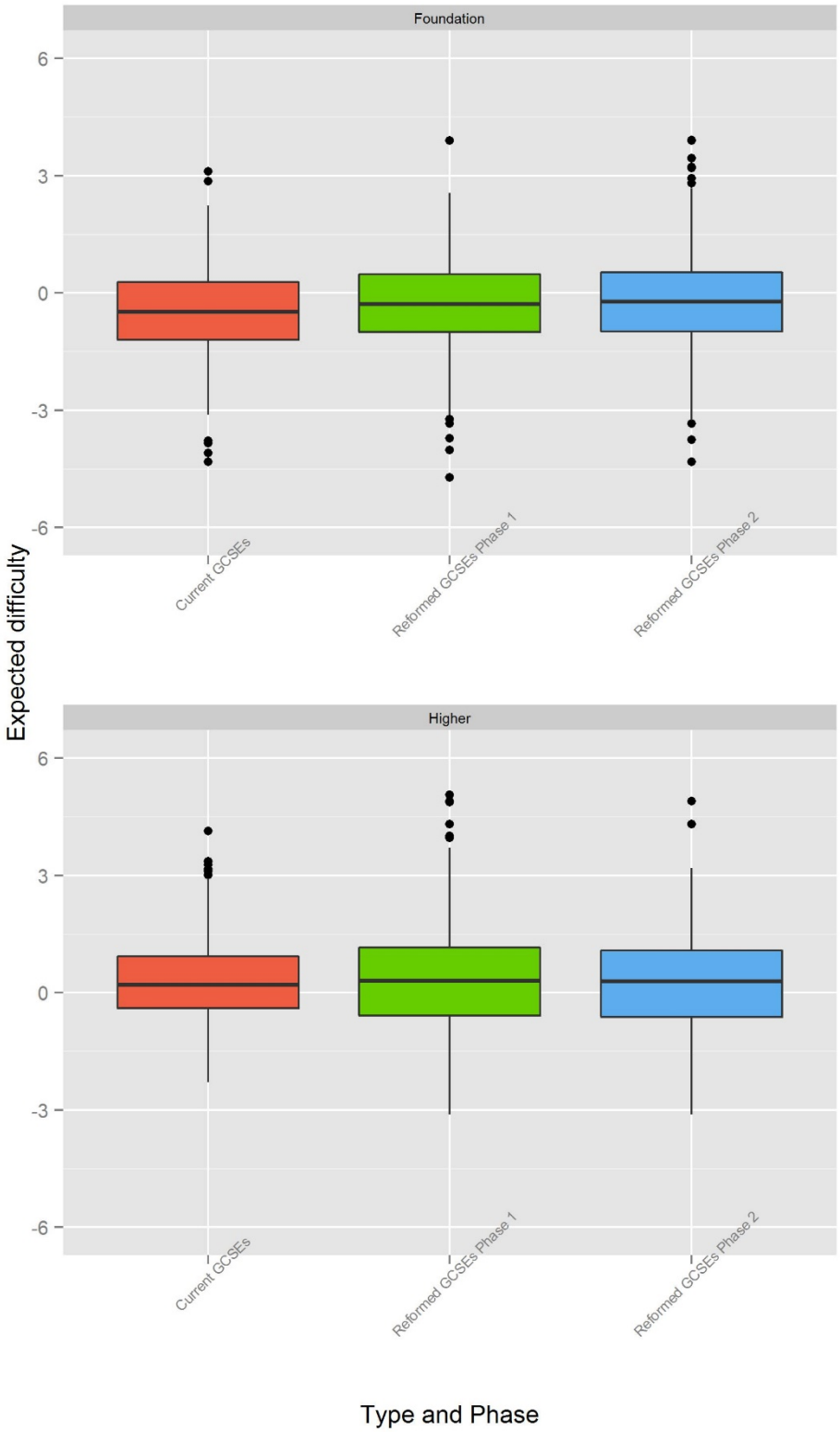


Figure 2: *Box plots showing median and interquartile ranges of expected item difficulties for the aggregated 2014 assessments and sample assessments for chemistry.*

Physics – Overall

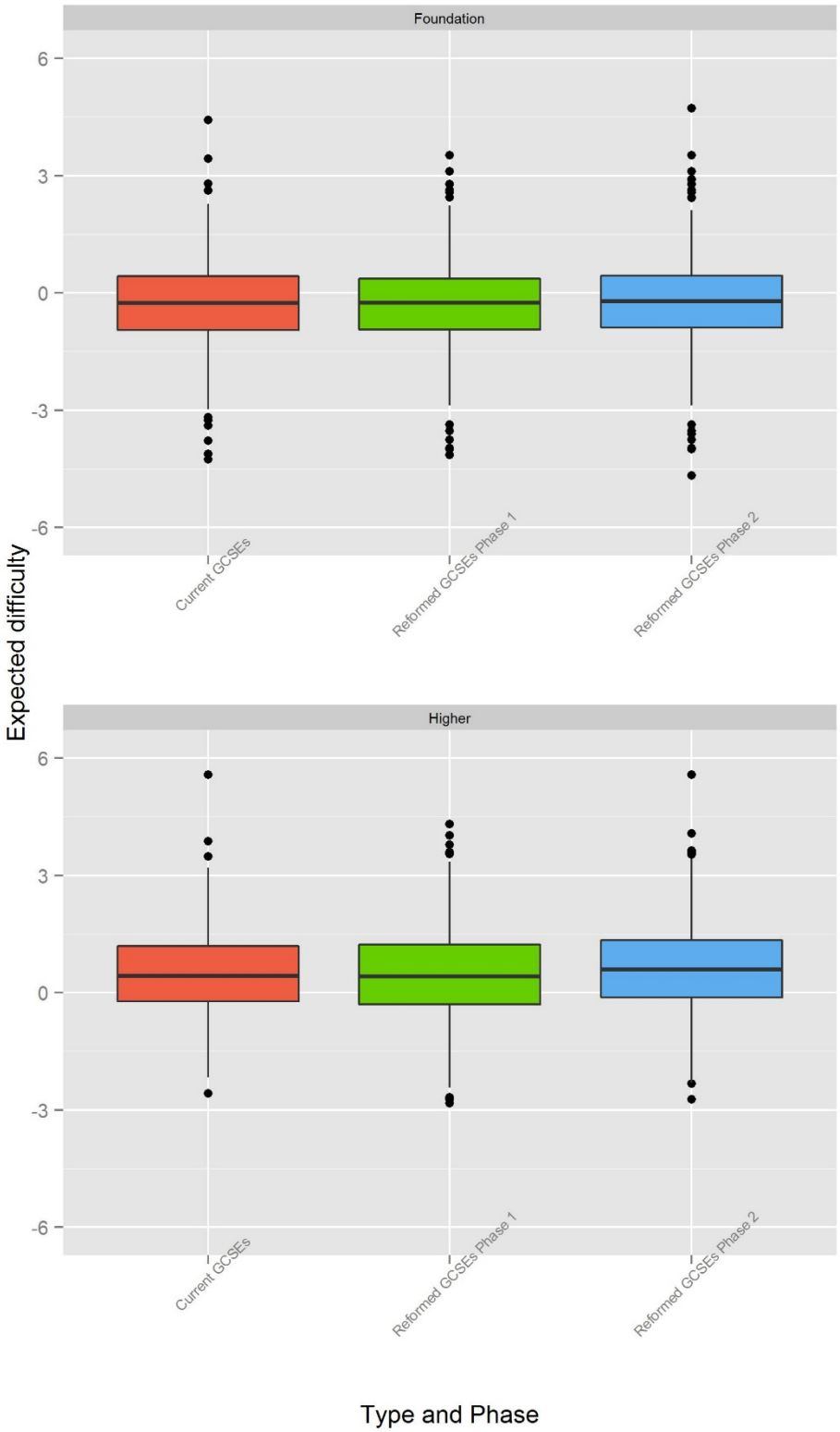


Figure 3: *Box plots showing median and interquartile ranges of expected item difficulties for the aggregated 2014 assessments and sample assessments for physics.*

Combined science – Overall

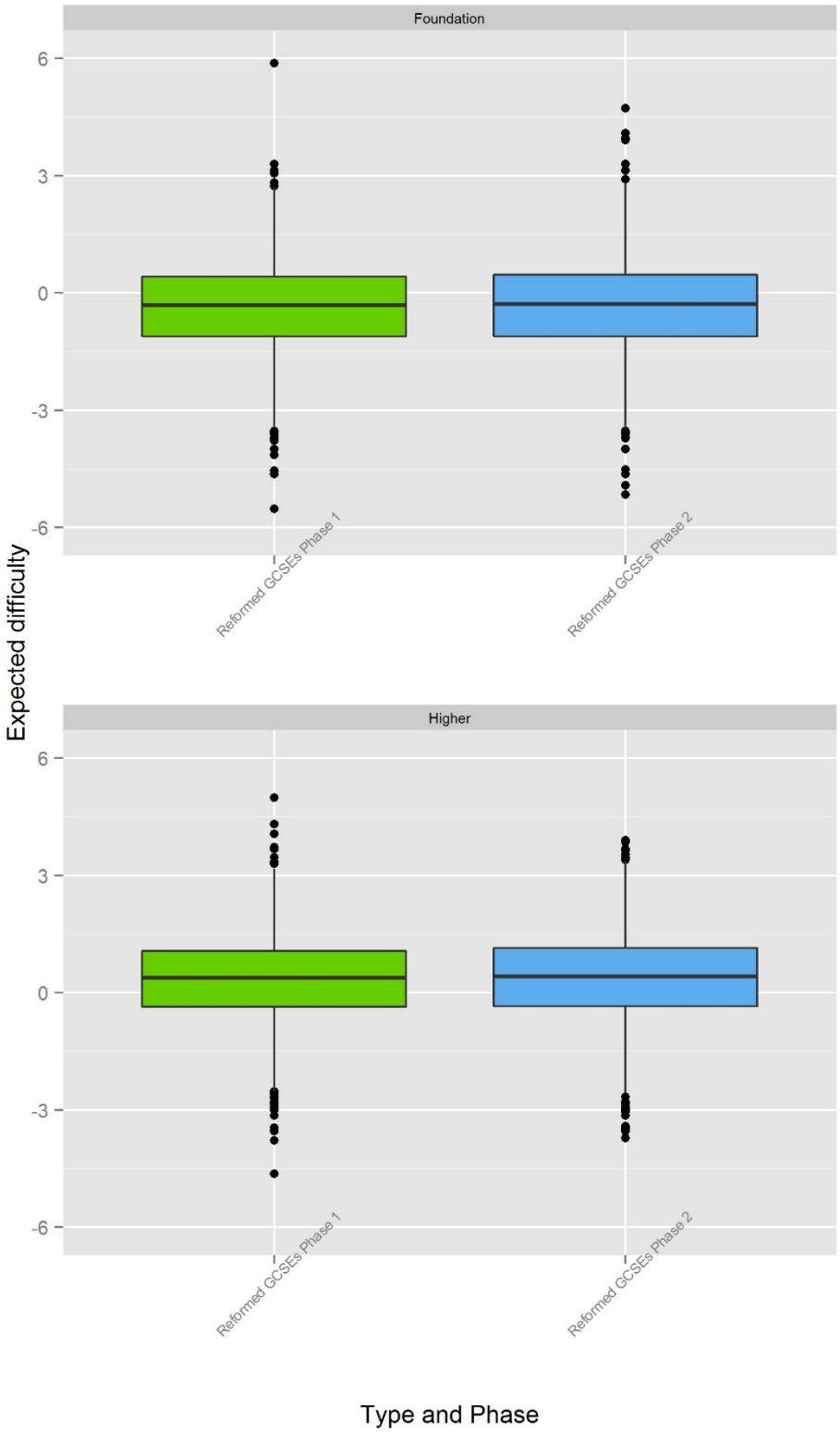


Figure 4: *Box plots showing median and interquartile ranges of expected item difficulties for the aggregated sample assessments for combined science.*

1 Background

Alongside the formal accreditation process for reformed GCSE science specifications for first teaching in 2016, Ofqual carried out 2 phases of comparative judgement studies on the expected difficulty of items from the reformed GCSE sample assessment materials (SAMs), together with items from the legacy 2014 GCSE science assessments. The purpose of this was to inform discussions and recommendations made by the accreditation panels regarding the likely difficulty of future live examinations. Comparisons were focussed on the relative expected difficulty of items from the 2014 papers and the SAMs within each specification. Unlike previous work looking at GCSE mathematics assessments[2], this was not an inter-board comparability exercise. When considering the findings, it is worth noting that the approach used focused only on one aspect of demand – the difficulty of items. The accreditation panel considered the data on expected item difficulties alongside other features of demand such as the subject content and linear structure of assessment. These factors are described further in the Discussion section.

In phase 1, all the items from the sample assessment materials submitted in July 2015 were judged for difficulty alongside the 2014 assessment items. There were 3 separate judging studies, 1 for items from each of biology, chemistry and physics. In phase 2, the new and modified items from sample assessments resubmitted in December 2015 were judged in 3 more subject-specific studies. The data from these phase 2 studies was then combined with the data for the unmodified questions from phase 1 to assemble the complete assessments.

2 Method

The comparative judgement method broadly followed the method used in the previous piece of research into the difficulty of GCSE maths questions. This study into item difficulty in science, involved a number of science teachers using an online system to remotely select the more difficult question for students to answer from pairs of questions presented side by side on screen. Each judge saw a random selection of questions, so each question was judged against many other questions by many judges. Only the items were presented and not the mark schemes. For this set of assessments, it was considered that what the presentation of the mark scheme might add to the validity of the judgements of item difficulties, would be more than offset by the additional cognitive load.

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

A model was then fitted to the judgement data which gave an estimate of difficulty for each item which best explained the pattern of judgements made.

2.1 Materials

2.1.1 Phase 1

All items from the sample assessments submitted in July 2015 were included in the comparative judgement exercise, together with items from the summer 2014 GCSE assessments (see Table 1). A standardised format was used so that any formatting and layout features which might have enabled judges to identify the exam board were removed.

Combined science items were split across the 3 separate subject studies based on the classification of items provided by the exam boards. Common items across tiers were created for the higher tier only. These items were judged in the study, their item parameters were calculated, and then this record was duplicated to create the foundation tier item. Similarly, Pearson's combined science papers consisted of items shared with the single science papers. These were judged as items on the single science papers, then their item parameters were duplicated to create the combined science item records. The numbers in Table 1, therefore, do not include the duplicate foundation tier or combined science common items but represent the number of unique questions.

2.1.2 Phase 2

New and modified items from the resubmitted sample assessments were created in the same way as before and entered into the comparative judgement exercise for phase 2 (for numbers of items see Table 1). Some items with very minor modifications were not included where the change was not likely to alter the perceived difficulty of the item (bearing in mind that in these instances the uncertainty in the estimated difficulty parameter is likely to be larger than any small change in expected difficulty). Examples of such modifications include corrections to typographical errors or grammar, small changes to the preamble to the question, where the question remained the same, or changes to bolding of text. Although some of these may have marginally reduced the chance of students misunderstanding the question in exam conditions, they would have minimal effect on the judged difficulty. The expected difficulty value from phase 1 was therefore used for these items.

Some items from phase 1 were included in phase 2 as 'anchor items' and were judged together with the phase 2 items. Their estimated difficulty parameters from phase 1 were entered into the model fitting process (see Analysis section below) as fixed values, and the parameters for the phase 2 items were fitted around them. This ensured that the phase 1 and 2 studies were both on the same scale of difficulty, allowing direct comparison of the results. Anchors were randomly drawn from all

phase 1 items and made up 20% of the items in each phase 2 subject study. The number of anchors used were: biology – 107; chemistry – 134; physics – 128.

Table 1. *Items included in the studies. OCR G = OCR Gateway Science Suite; OCR 21 = OCR Twenty-First Century Science Suite; H = Higher Tier; F = Foundation Tier.*

Biology	**Phase 1**						**Phase 2**		
	2014 papers			**Sample assessments**			**Sample assessments**		
Board	**H**	**F**	**Total**	**H**	**F**	**Total**	**H**	**F**	**Total**
AQA	87	88	175	218	218	436	77	57	134
Eduqas	94	69	163	137	115	252	54	30	84
OCR G	79	67	146	162	123	285	11	24	35
OCR 21	71	59	130	142	121	263	43	30	73
Pearson	85	84	169	96	83	179	61	41	102
			783			1415			428

Chem	**Phase 1**						**Phase 2**		
	2014 papers			**Sample assessments**			**Sample assessments**		
Board	**H**	**F**	**Total**	**H**	**F**	**Total**	**H**	**F**	**Total**
AQA	94	102	196	182	200	382	100	89	189
Eduqas	84	78	162	131	120	251	48	55	103
OCR G	77	72	149	158	141	299	44	39	83
OCR 21	76	67	143	117	115	232	33	25	58
Pearson	89	96	185	95	90	185	60	43	103
			835			1349			536

Physics	**Phase 1**						**Phase 2**		
	2014 papers			**Sample assessments**			**Sample assessments**		
Board	**H**	**F**	**Total**	**H**	**F**	**Total**	**H**	**F**	**Total**
AQA	88	78	166	213	211	424	100	77	177
Eduqas	72	70	142	123	119	242	32	49	81
OCR G	72	65	137	160	135	295	42	65	107
OCR 21	74	63	137	125	104	229	26	28	54
Pearson	82	95	177	89	78	167	53	40	93
			759			1357			512

Items from phase 1 that were not modified in the sample assessments submitted in December 2015 were merged with the new items from phase 2 to form each complete sample assessment.

2.2 Participants

2.2.1 Phase 1

One hundred and five science teachers were recruited as judges. All were current science teachers or had teaching experience within the last 3 years. Each judge completed 1000 or 1500 judgements (depending on their availability) for the subject studies they were specialists in. Most judges completed all their judgements in just one subject study, but some completed judgements in 2 subject studies in order to align the number of judgements per item across the 3 subject studies. Due to 2 biology judges not starting and others not completing their judgement allocation the total judgements per item were not exactly equal across studies. Table 2 shows a summary of the judges and numbers of judgements made across the three studies. Judges were paid for their time.

2.2.2 Phase 2

Sixty-five of the judges from phase 1 were recruited to take part in phase 2. Again, judges were allocated to the subject studies according to specialism and also to balance the number of judgements per item. Each judge was allocated 500 judgements. For some judges this was split equally between 2 subject studies. As in phase 1, not every judge started or completed their allocated judgements so the final number of judgements per item (see Table 2) were not exactly equal.

Table 2: *Number of judges and number of judgements made for each study*

	Phase 1			Phase 2		
Subject	**Number of judges**[3]	**Mean judgements per item**	**Total judgements**	**Number of judges**	**Mean judgements per item**	**Total judgements**
Biology	34	32.5	35718	21	35.5	9500
Chemistry	35	36.6	39933	25	32.1	10750
Physics	36	37.0	39131	23	32.7	10450

[3] These totals include only judges who made some judgements, and excludes a small number who were recruited but did not take part in the judging.

2.3 Procedure

Comparisons were conducted using the online comparative judgement platform, No More Marking[4]. Judges were given detailed instructions on how to access the platform and how to make their judgements. Pairs of items were presented side by side on the screen and the judges were prompted on screen to indicate:

'Which question is more difficult to achieve full marks on?'

Additional clarification was given in written instructions to the judges, which stated:

'Read the questions and decide which you think is the more difficult of the two questions for a 16-year-old student to achieve full marks on.

The marks available for the part are shown by a number in square brackets (e.g. '[3]') – use this as a guide to the depth of answer required. You should consider the precise wording of the question and think about what would be required to receive all the marks on the question, and judge the difficulty for a student based on these factors.'

There was no matching of items; any item could be paired with any other item (including, for example, 1-mark items with 6-mark 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. The items were randomly distributed among judges so that the items were all seen a similar number of times.

3 Analysis

The R package sirt [5]was used to estimate expected difficulty parameters for each item under the Bradley-Terry model. The node package, Comparative-Judgement[6], which implements the same Bradley-Terry model as sirt, using the same estimation procedure, was used to estimate item and judge infit, scale-separation reliability (SSR) and inter-rater reliability.

For each study the expected difficulty values for items are distributed along a different scale. Each difficulty scale is measured in logits, a probabilistic scale based

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

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

[6] https://www.npmjs.com/package/comparative-judgement

on the log odds of one item being judged more difficult than another item. The values are entirely arbitrary, with the scale centred on a 0 value which simply represents the mean difficulty of the set of items. The spread of items along the logit scale is also determined by the discriminability of the items; if items are more discriminable for a subject, the items will be spread over a wider numerical range. So no meaning can be attributed to a specific expected difficulty value, other than as a relative difficulty compared to other items on the same scale.

For phase 2 the expected difficulty values of the anchor items from phase 1 were fixed and the best fit of the Bradley-Terry model obtained under this constraint. This ensured that all new phase 2 items were fitted onto the same difficulty scale as the equivalent subject study in phase 1.

3.1 Judge consistency and exclusion

3.1.1 Phase 1

For the biology study, 2 judges were excluded due to high infit values[7]. The range of median judgement times for the 32 included judges was 10 to 38 seconds (mean = 23 seconds).

For the chemistry study, 3 judges were excluded due to high infit values. The range of median judgement times for the 32 included judges was 7 to 37 seconds (mean = 18 seconds).

For the physics study, 3 judges were excluded due to high infit. The range of median judgement times for the 33 included judges was 4 to 82 seconds (mean = 19 seconds). The upper end of this range was an outlier who only completed a small number of judgements. The second highest median judgement time was 38 seconds. The infit for the judge with a mean judging time of 4 seconds was close to the mean infit, so although their judgements were fast they were included in the analysis.

The median inter-rater reliability was assessed by repeatedly allocating judges to 2 groups, fitting the Bradley-Terry model independently for each group and correlating the 2 rank orders of the item parameters. Across 100 replications the Pearson correlations for the three studies were: biology = 0.78 (sd=0.01); chemistry = 0.82 (sd=0.01); physics = 0.80 (sd=0.01).

[7] Infit is a measure of the consistency of the judgements made by a judge compared to the overall model. A high infit indicates that the judge was either inconsistent within their own judgements, or was applying different criteria to the overall consensus. The usual threshold for establishing outlying judges is that their infit value is more than two standard deviations above the mean infit value.

Reliability is quantified in comparative judgement studies by an SSR statistic that is derived in same way as the person separation reliability index in Rasch analyses. It is interpreted as the proportion of 'true' variance in the estimated scale values. The SSRs were biology = 0.89, chemistry = 0.90, physics = 0.90, showing good reliability and also consistency across the three studies.

3.1.2 Phase 2

For the biology study one judge was excluded due to a high infit value. The range of median judgement times for the 20 included judges was 8 to 30 seconds (mean = 20 seconds).

For the chemistry study one judge was excluded due to a high infit value. The range of median judgement times for the 24 included judges was 7 to 38 seconds (mean = 20 seconds).

For the physics study one judge infit was marginal, approximately 2 standard deviations away from the mean infit, but they were retained. The range of median judgement times for the 23 included judges was 8 to 50 seconds (mean = 24 seconds).

The median inter-rater reliability across 100 replications for the 3 studies were: Biology = 0.85 (sd=0.01); Chemistry = 0.86 (sd=0.01); Physics = 0.86 (sd=0.01). These correlation coefficients are higher than in phase 1, probably due to the stabilising effect of the fixed anchor items on the rank order[8]. The SSRs were 0.92 for all 3 subject studies.

3.2 Bias correction

The expected difficulty parameters for each item were adjusted using a multiple regression model. Initial analysis of the relationship between expected difficulty and facility (the average performance of students when taking the item in live testing) for the 2014 items suggested that there were differences between items of particular type in how they fitted the relationship. The multiple regression model was used to correct for biases in the judging of items, and this process is described in detail in Appendix A.

[8] The Spearman correlation between the rank order of items when the model is fitted with fixed anchor items and when the model is fitted with the anchor items free to vary was 0.96. This shows that using anchor items does not distort the final model fit.

4 Results

Figures 5 to 8 show the distributions of expected item difficulties broken down by specification within each science domain. The exam board and/or specification has been anonymised on the figures. The order of the exam board/specifications on each figure is also random. The boxplots show the distribution of estimated item difficulties unweighted by item marks - a 1-mark item and a 6-mark item contribute equally to the boxplot.

Note that the scales for the 3 separate sciences are independent. Although they are centred at 0 with a similar spread, and so look similar, there is no linking between items from different domains and so the same numerical expected difficulty value may mean different things for each subject. For combined science, the data from the 3 science subjects have been combined as the proportion of questions from each domain are almost equal, and so there will be little resulting bias in the aggregated parameter distributions.

Appendix B contains additional data tables for these studies.

Biology – by specification

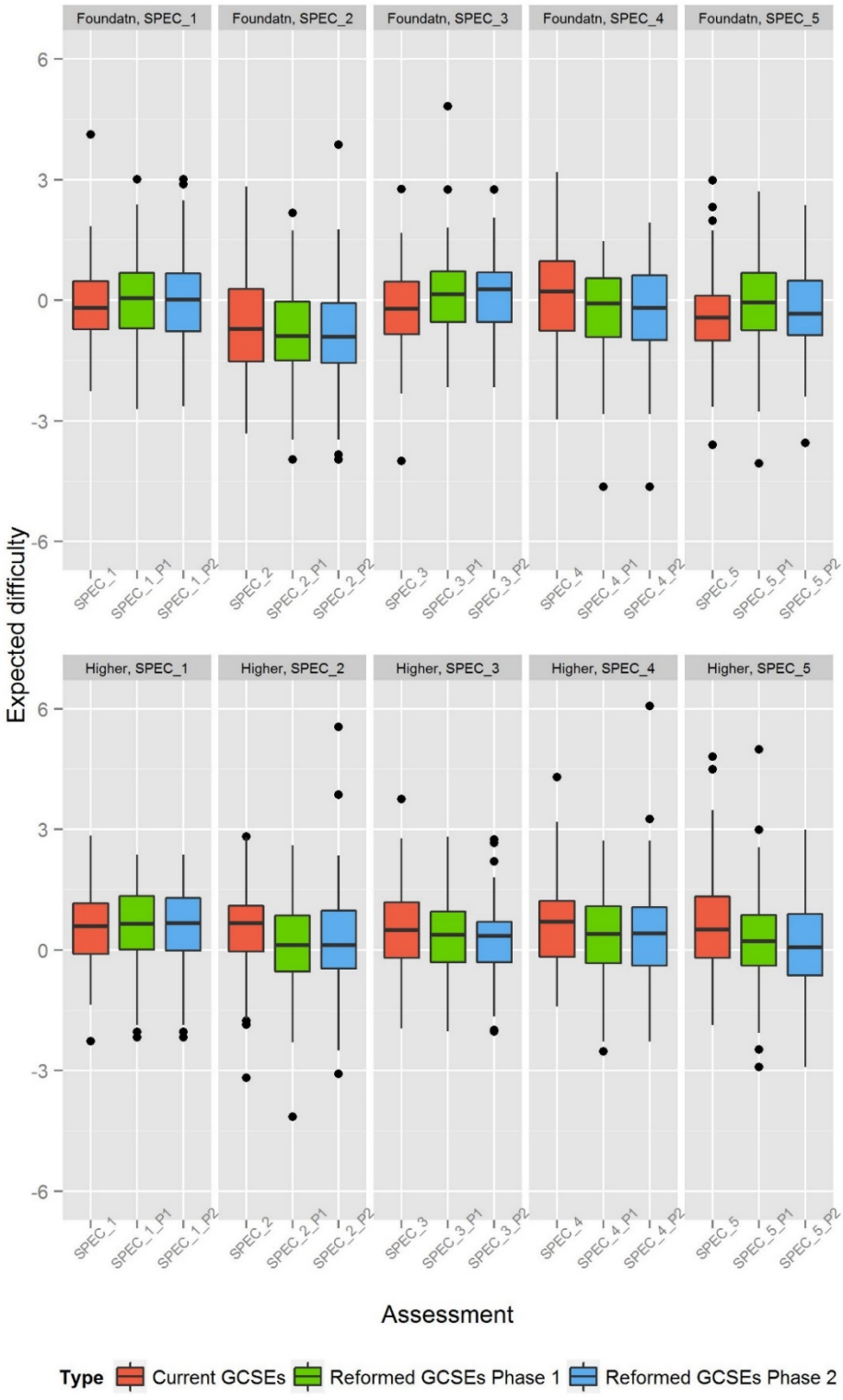


Figure 5: *Box plots showing median and interquartile ranges of expected item difficulties for each specification from the 2014 assessments and sample assessments for biology.*

Chemistry – by specification

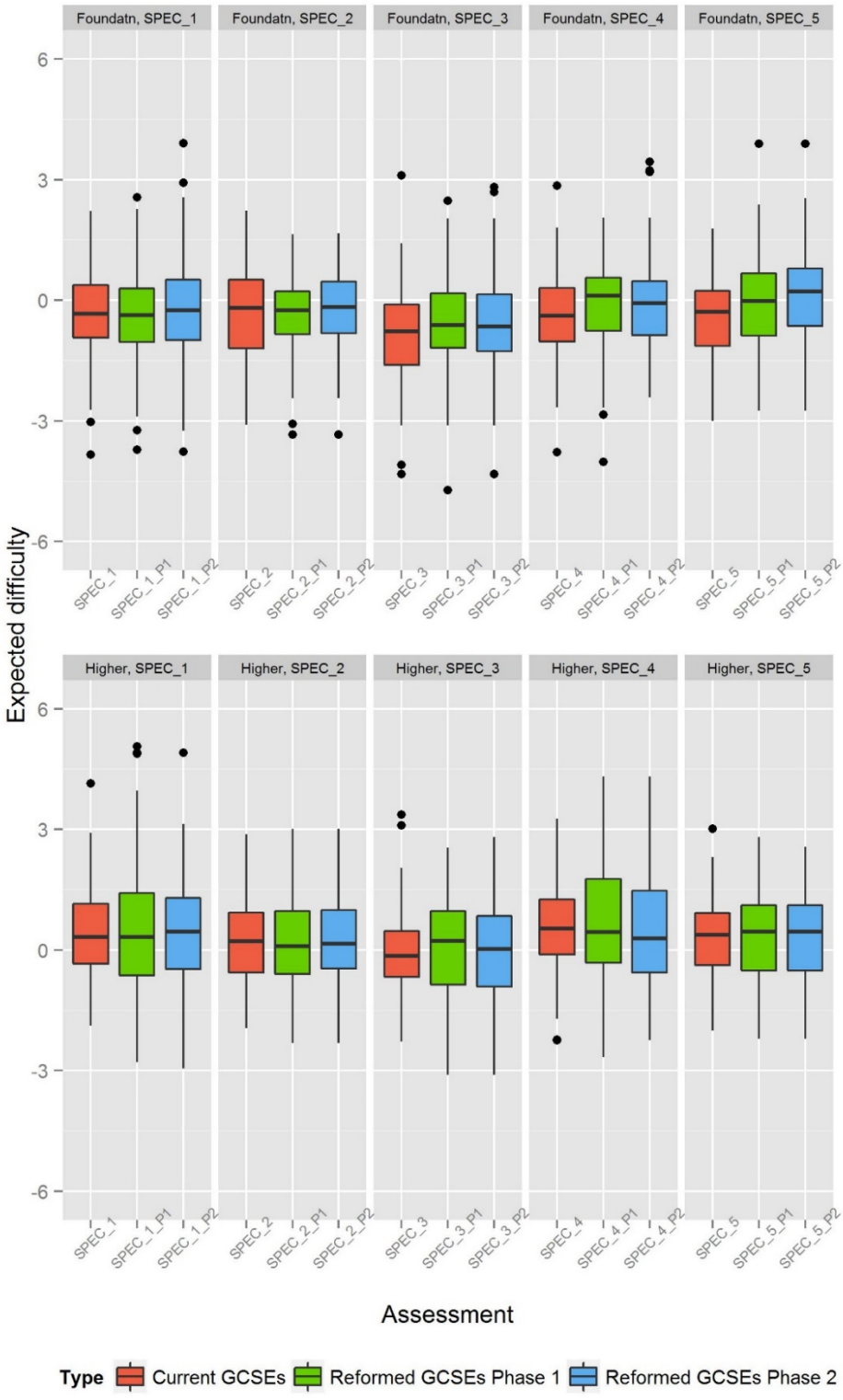


Figure 6: *Box plots showing median and interquartile ranges of expected item difficulties for each specification from the 2014 assessments and sample assessments for chemistry.*

Physics – by specification

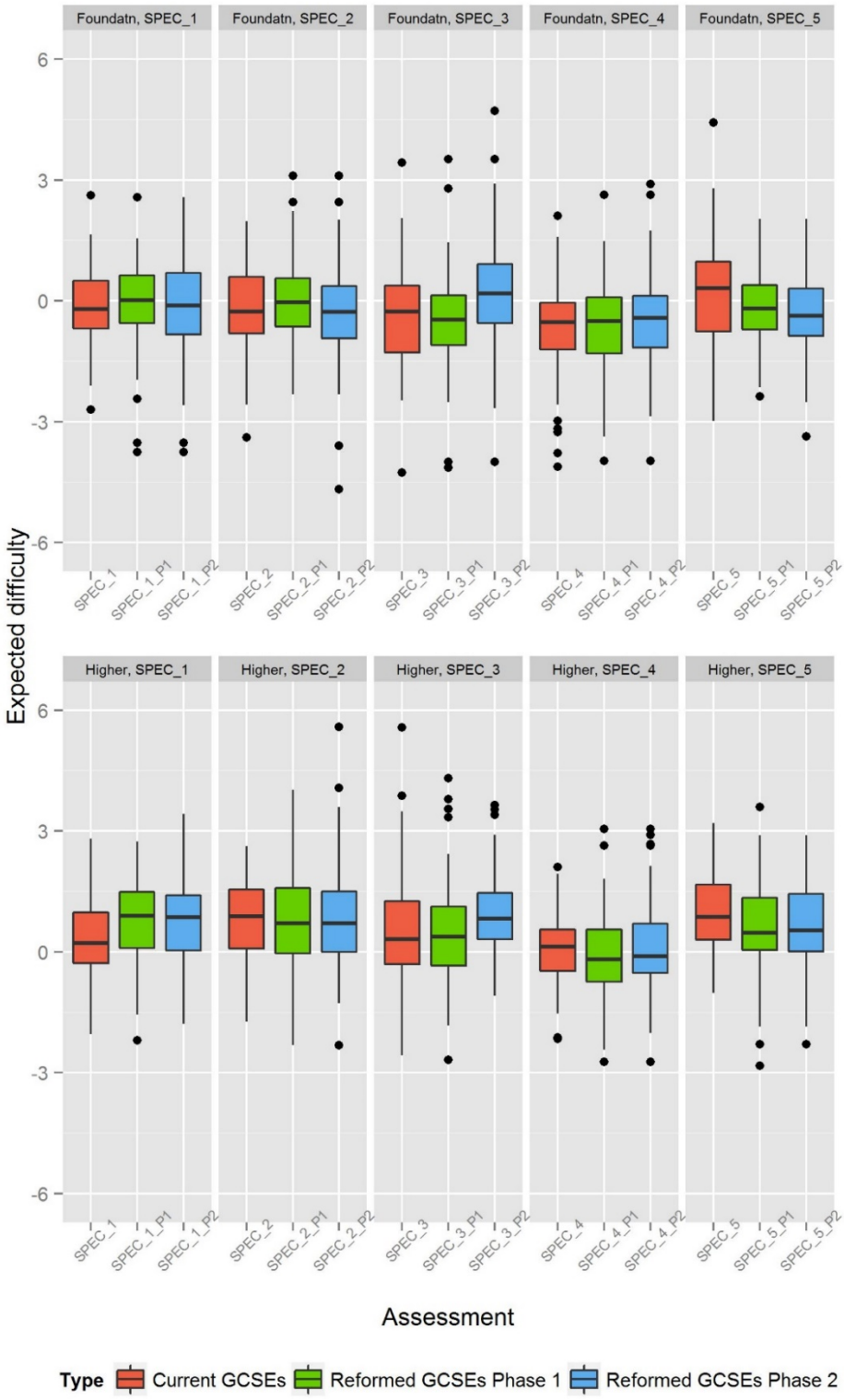


Figure 7: *Box plots showing median and interquartile ranges of expected item difficulties for each specification from the 2014 assessments and sample assessments for physics.*

Combined science – by specification

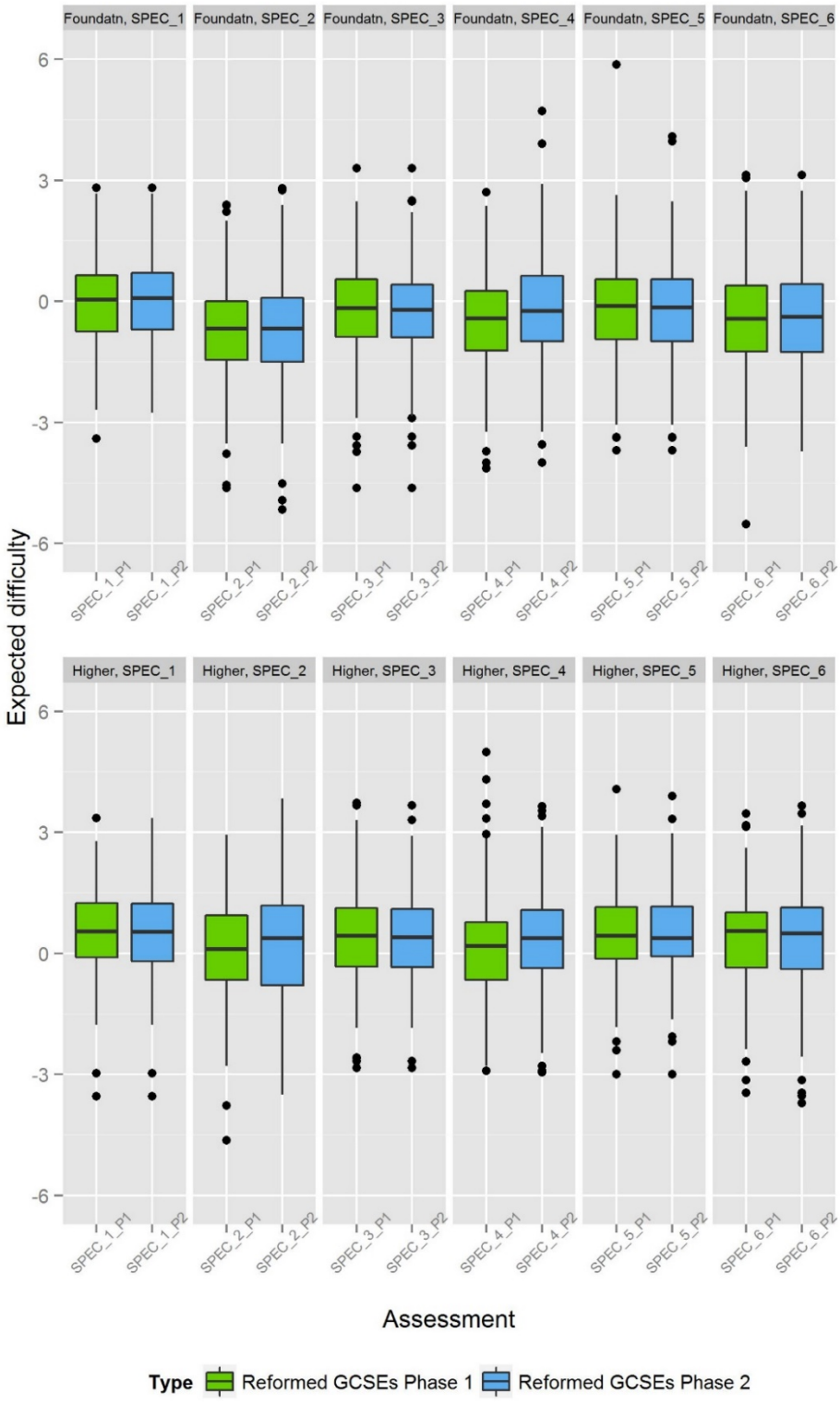


Figure 8: *Box plots showing median and interquartile ranges of expected item difficulties for each specification from the sample assessments for combined science.*

4.1 Biology

For the foundation tier biology assessments, the median difficulty of the 2014 assessments varied from -0.71 to 0.23 (overall median = -0.28), a range of 0.94 logits. After phase 2 of the study, the reformed assessments varied from -0.90 to 0.27 (overall median = -0.34), a slightly wider range of 1.17 logits.

For the higher tier biology assessments, the median difficulty of the 2014 assessments varied from 0.50 to 0.70 (overall median = 0.58), a range of only 0.20 logits. After phase 2 of the study, the reformed assessments varied from 0.07 to 0.67 (overall median = 0.37), a wider range of 0.60 logits.

For the reformed biology sample assessments, relative to the 2014 assessments there appears to be a slight decrease in item difficulty overall, and also slightly greater spread of the assessment medians. However, these differences are not substantive.

4.2 Chemistry

For the foundation tier chemistry assessments, the median difficulty of the 2014 assessments varied from -0.77 to -0.20 (overall median = -0.48), a range of 0.57 logits. After phase 2 of the study, the reformed assessments varied from -0.65 to 0.22 (overall median = -0.23), a slightly wider range of 0.87 logits.

For the higher tier chemistry assessments, the median difficulty of the 2014 assessments varied from -0.15 to 0.53 (overall median = 0.21), a range of 0.68 logits. After phase 2 of the study, the reformed assessments varied from 0.02 to 0.46 (overall median = 0.29), a narrower range of 0.44 logits.

For the reformed chemistry sample assessments, relative to the 2014 assessments there appears to be a slight increase in item difficulty overall, and also slightly greater spread of the foundation tier assessment medians and slightly lesser spread of the higher tier assessment medians. Again, these differences are not substantive.

4.3 Physics

For the foundation tier physics assessments, the median difficulty of the 2014 assessments varied from -0.53 to 0.32 (overall median = -0.25), a range of 0.85 logits. After phase 2 of the study, the reformed assessments varied from -0.43 to 0.18 (overall median = -0.21), a narrower range of 0.61 logits.

For the higher tier physics assessments, the median difficulty of the 2014 assessments varied from 0.13 to 0.89 (overall median = 0.43), a range of 1.02 logits. After phase 2 of the study, the reformed assessments varied from -0.11 to 0.86 (overall median = 0.59), a similar range of 0.97 logits.

For the reformed physics sample assessments, relative to the 2014 assessments there appears to be a slight increase in item difficulty overall, and also slightly lesser spread of the assessment medians. Again, these differences are not substantive.

4.4 Combined science

No comparison with 2014 specifications was included for combined science specifications. However, the median difficulty of the foundation tier sample assessments was -0.68 to 0.07 (overall median = -0.29), a range of 0.75 logits. For the higher tier sample assessments, the median difficulty was 0.37 to 0.53 (overall median = 0.42), a narrower range of 0.16 logits.

5 Discussion

Overall this comparative judgement analysis shows similar levels of expected difficulty for items from the 2014 assessments and the sample assessments (see Table B1, Appendix B). The biology sample assessments have very slightly lower difficulty than the 2014 assessments, while the chemistry and physics sample assessments have slightly higher difficulty than the 2014 assessments. The mean and maximum of the range of assessment median difficulties are very similar for the 2014 and reformed specifications, indicating that there will not be substantial differences in overall difficulty of items between specifications. Such small differences can be easily accommodated by the setting of grade boundaries at awarding.

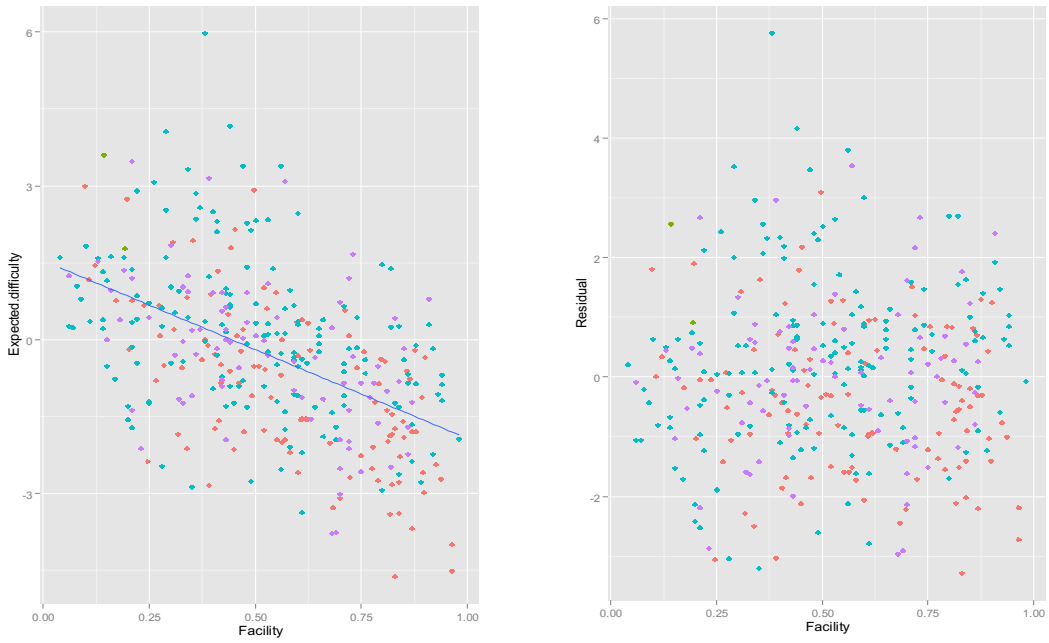
This research exercise helped to inform the accreditation panel's review of the overall demand of the submissions. It is worth noting that the correlation between expected difficulty and item facility for the 2014 items was lower than that obtained in the previous GCSE maths study, likely as a result of the judging criteria that was used and the more varied types of items in science relative to maths (see Appendix C). The panel were informed of this lower correlation and they considered the difficulty of items in the context of a variety of other factors which can influence the overall demand. Such factors include the removal of controlled assessment from the reformed specifications and the linear structure of the assessment. In addition, the subject content in the reformed specifications has been increased in both quantity and demand relative to the legacy specifications, and new requirements for mathematics and synoptic assessment have been introduced. All of these factors were also taken into consideration during the accreditation decision-making process alongside the information on item demand.

Finally, note that this data only covers the reformed sample assessments up to the second submission for accreditation. Several of the specifications went through additional submissions which included some changes to their sample assessments, for example where there were concerns about their level of demand, and so the data

from phase 2 does not necessarily represent the final expected difficulty distributions of the accredited specifications.

Appendix A – Adjustment of expected difficulty by multiple regression model

Expected item difficulty was regressed onto item facility. This relationship is discussed and investigated in Appendix C. We assumed comparability between the cohorts taking the assessments with each exam board, such that we treated the facility values from each board as equivalent. An example of a scatterplot with the regression line overlaid is shown in Figure A1a). The linear regression residual for each item is shown in Figure A1b). These residuals do not seem to be equally distributed around zero for all of the exam boards (boards are plotted in different colours on Figure A1).



a) b)

Figure A1: a) *Expected difficulty plotted against facility, with the calculated linear regression line and* b) *model residual derived from the deviation of each point from the regression line for all biology foundation tier items. Exam boards are differentiated by colour on both plots.*

The mean residual of all items for each exam board was calculated. Table A1 shows the mean residual for each board for each subject. A positive value indicates that the expected difficulties are higher than predicted by the regression equation – the items are judged harder than the facility values predict. A negative value indicates that items are judged less difficult than the facility values predict. The mean residuals are quite consistent for each board across the 3 science subjects and 2 tiers. One exam

board is not included in this analysis as facility values were only available for a small number of extended response questions that were not broken into sub-questions (a, b, c, etc).

Table A1: *Mean residual in the regression of expected item difficulty onto facility, split by board, subject and tier (F or H) for 2014 items*

Board	Biology (F/H)	Chemistry (F/H)	Physics (F/H)
Board A	-0.436/-0.370	-0.470/-0.197	-0.363/-0.338
Board B	0.306/0.227	0.426/0.119	0.343/0.182
Board C	0.001/-0.022	-0.132/0.003	-0.170/0.040

Board A's 2014 items appear to be underestimated in difficulty by the judges while Board B's 2014 items appear to be slightly overestimated in difficulty. Board C's items lie in-between with a generally small mean residual.

Given the offset pattern of residuals for each board, the estimated difficulty parameters obtained appear to contain an element of 'bias'. This is not to suggest conscious bias from the judges in any way. The bias, almost certainly unconscious, could not have arisen from knowledge of the board associated with an item. However, given that different boards have different profiles of item types and item characteristics, we investigated whether biases in judging particular item types or characteristics could explain the pattern of residuals as shown in Table A1.

A.1 Multiple linear regression analysis

All the items were coded on a set of features and a multiple linear regression was used to determine which features were significant predictors of bias. Although item tariff could also be a predictor of bias, it was not used as a factor since it would be correlated with question type, and classifying items by specific type of question should account for more variance in the bias than the tariff alone.

Items were coded as one of the following five question types:

- Multiple choice
- Constrained response (this covers any other kind of question that does not require writing – ticking more than one option, drawing lines between diagram parts, circling options etc)
- Short answer (one line response area)
- Short answer (2 to 3 lines response area)
- Extended response (any written response requiring more than 3 lines of writing)

The other coded features were:

- Word count (for each item this includes any contextual/scene setting text that applies to more than one item)
- Numerical equation/calculation question (any question that requires a specific numerical answer, so excluding estimates)
- Table included
- Graph included
- Line drawing/diagram included (relevant)
- Line drawing/diagram included (irrelevant) (if drawing/diagram is included but is used purely to illustrate the topic area and is not absolutely necessary to answer the question)
- Photograph included (relevant)
- Photograph included (irrelevant) (if photograph is included but is used purely to illustrate the topic area and is not absolutely necessary to answer the question)

All, except word count, were coded as '0' (feature not present) or '1' (feature present). Dummy variables were created for the categorical question-type variable, with extended response held constant as the reference question type (the constant term in the regression analysis). Interaction variables were created for all of the 2-way interactions. As these were correlated with the top-level variables from which they were formed, they were made orthogonal by regressing the interaction term onto the 2 original variables, and using the resulting residuals as the value to represent the interaction between the 2 terms[9]. This then characterises the interaction with the effect of the two primary variables factored out.

Having coded all items, a multiple linear regression model was fitted to the features with bias as the predicted variable, for each subject study (independent models were obtained for biology, chemistry and physics items).

A.2 Model development

The regression model was developed from those 2014 items for which item expected difficulty parameters and facility from the 2014 summer exams were available (approximately 600 to 700 items per subject).

[9] Aiken, L. S., and West, S. G. (1991). Multiple regression: Testing and interpreting interactions. (Newbury Park: Sage Publications)

To model the sources of bias we wanted to develop a single regression model per subject rather than one model for each tier. Facility values for the foundation and higher tier items are not directly comparable due to the very different abilities of the cohorts. The foundation and higher tier items, therefore, had to be combined onto a single dimension representing student performance. The common items between tiers were available to equate the facility values, and the mean offset in facility between the two tiers was calculated for each subject. This offset (around 0.2 for all three subjects) was added to the facility values for all foundation tier items to equate the two tiers. The foundation tier common items were then removed to avoid these duplicate items unduly influencing the model fit.

After making these adjustments, the bias values were recalculated from a new regression equation fitted to the expected difficulty parameters regressed onto the adjusted facility values for items across both tiers. The bias values with the adjusted facility scale are shown in Table A2 and are consistent with the individual tier values shown in Table A1.

Table A2: *Mean bias in the expected difficulty calculated using a regression equation fitted to the combined facility scale across both tiers split by board and subject for 2014 items*

Board	Biology	Chemistry	Physics
Board A	-0.437	-0.414	-0.348
Board B	0.246	0.287	0.231
Board C	0.055	-0.014	-0.032

Before running the final regression analysis, the data was checked for multicollinearity and normality of the residuals. The residuals were normally distributed and showed no relationship with predicted values (see Figure A2 for an example). To check for multicollinearity, variance inflation factors (vif) were calculated for each independent variable. For the biology model, 2 of the interaction terms had values around 6, while all other values were below 2.5. As these two interaction variables were only marginally collinear they were retained in the initial model. They did not feature in the final developed model. All factors had vifs below 2.7 for the chemistry model and 2.2 for the physics indicating no significant multicollinearity.

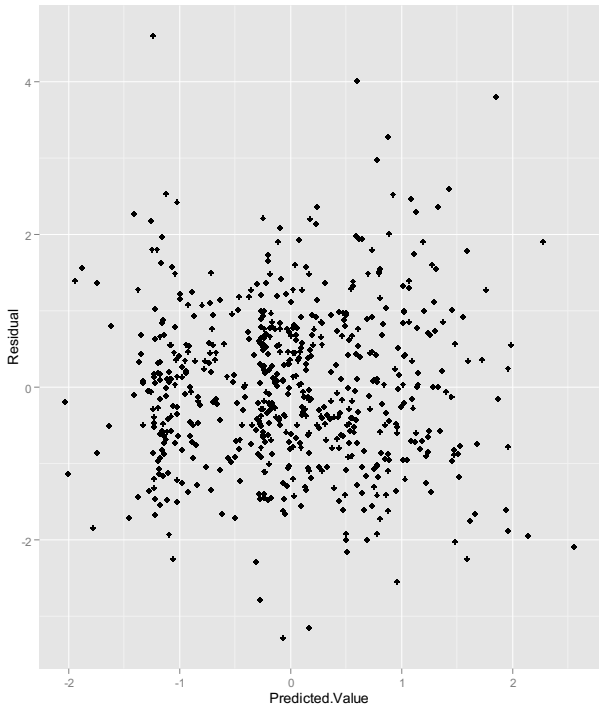


Figure A2: *Initial regression model residuals plotted against predicted values*

The full model of all the top level variables and 2-way interactions was analysed by stepwise multiple regression in R statistical software, then the significant variables from the output were re-run as the final multiple regression model. The significant predictor variables are shown in Tables A3 to A5 for each subject in turn. The Spearman's rho rank order correlation coefficients between each variable and the linear regression residual are also shown. These correlations do not account for the effect of any of the other factors in combination with the variable, unlike the regression coefficients. For the interaction factors the correlations are based on the original factor values before they were made orthogonal.

Table A3: *Final biology multiple regression model: correlations with residuals and regression coefficients of significant predictor variables*

Factor	Spearman correlation with residuals	Unstandardised coefficient	Std error	t-value
Constant		0.496	0.110	4.514***
MCQ	-0.229	-1.896	0.154	12.294***
Constrained	-0.008	-1.275	0.168	7.566***
Short 1	-0.264	-1.762	0.129	13.652***
Short 2 to 3	0.023	-1.082	0.118	9.168***
Word count	0.268	0.011	0.001	7.778***
Calculation	0.126	0.740	0.186	3.986***
Diagram (Irrelevant)	-0.088	-0.509	0.177	2.872**
Photograph (Irrelevant)	0.089	0.473	0.220	2.154*
Short 2 to 3 by word count	-0.124	-0.010	0.003	3.371***
Short 1 by diagram (Irrelevant)	0.137	-1.032	0.383	2.693**
Constrained by diagram (Irrelevant)	0.032	-1.279	0.601	2.127*
Summary statistics		$R^2 = 0.374$ Adjusted $R^2 = 0.363$		
Model significance		$F_{11, 602} = 32.69$, $p < 0.001$		

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A4: *Final chemistry multiple regression model: correlations with residuals and regression coefficients of significant predictor variables*

Factor	Spearman correlation with residuals	Unstandardised coefficient	Std error	t-value
Constant		0.971	0.126	7.689***
MCQ	-0.343	-2.701	0.154	17.505***
Constrained	-0.001	-1.570	0.174	9.049***
Short 1	-0.263	-1.898	0.123	15.375***
Short 2 to 3	0.102	-1.309	0.124	10.523***
Word count	0.297	0.010	0.001	7.103***
Calculation	0.182	0.691	0.174	3.974***
Diagram	-0.040	-0.234	0.111	2.114*
Short 1 by calculation	0.242	0.717	0.341	2.102*
Summary statistics		$R^2 = 0.442$ Adjusted $R^2 = 0.435$		
Model significance		$F_{8, 647} = 64.12$, $p < 0.001$		

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A5: *Final physics multiple regression model: correlations with residuals and regression coefficients of significant predictor variables*

Factor	Spearman correlation with residuals	Unstandardised coefficient	Std error	t-value
Constant		0.798	0.117	6.823***
MCQ	-0.300	-2.368	0.151	15.732***
Constrained	-0.041	-1.551	0.177	8.759***
Short 1	-0.339	-2.113	0.126	16.730***
Short 2-3	0.096	-1.193	0.113	10.582***
Word Count	0.190	0.008	0.001	5.687***
Calculation	0.236	0.639	0.126	5.080***
Short 1 by Calculation	0.127	1.006	0.292	3.446***
Summary statistics		$R^2 = 0.435$ Adjusted $R^2 = 0.429$		
Model significance		$F_{7, 606} = 66.65$, $p < 0.001$		

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

These models were all significant and explained between 36 and 44 per cent of the variance in the residuals. Given that the analysis did not include any deeper features of the questions such as topic, context or complexity which may also be prone to judging bias, this is a substantial amount of variance explained.

Relative to the reference 'Extended response' category, all of the other question types reduce the bias value (or introduce negative bias). This is particularly true for MCQ and one line short answer questions. These were judged to be easier for students than the item facility would suggest. Constrained answer questions and intermediate length short answer questions have less strong effects on bias. Extended response questions introduce a positive bias as indicated by the constants and are judged to be harder than the facility predicts. It should be remembered that our prediction of difficulty was based on the difficulty of achieving full marks, and so difficulty is somewhat related to question tariff, whilst the measure of facility factors out tariff.

Of the other factors, word count is a strong predictor of increased bias, with longer written questions over-judged on difficulty relative to their difficulty for students. Similarly, calculation questions are perceived by judges as more difficult than they actually turn out to be for students.

Weaker, less consistent effects across the three models are seen for diagrams, irrelevant diagrams and photos, with varying bias predictions. Some interactions were also significant, although it is worth noting that in many cases the frequency of questions sharing these interaction features is quite low.

A.3 Using the regression model to reduce the bias

Although the 3 models are based upon the relationship between item characteristics and residual only for the 2014 items, we can assume that the same relationship holds for the items drawn from the sample assessment materials, given that they were judged together. So we can use the regression coefficients to correct the estimates of difficulty for the discovered judging bias for every single item, not just the 2014 items.

To apply the correction, the modelled bias for each item was subtracted from its expected difficulty. This gives an adjusted expected difficulty value for each item with any bias caused by the surface features of the items removed. The effect of this correction can be seen in Figure A3 showing the regression residuals for the same 2014 biology items shown in Figure A1. Whereas before the model correction there were vertical offsets from 0 for the distribution of item residuals for two of the boards shown, after correction, they are more evenly centred around 0. The spread of residuals has also been reduced following the correction.

Figure A3: *Residuals for individual 2014 Biology items.* a) *Original model parameters.* b) *After applying multiple regression model correction. Exam boards are differentiated by colour on both plots.*

A.4 Conclusion

Using the relationship of expected difficulty and item facility for all of the 2014 items, between-board biases were discovered in the expected difficulty parameters. Using a multiple regression model to explore which item features predicted bias, a little under half of the bias could be explained by superficial features of the questions, such as the type, word count, or use of visual aids in the questions. The model parameters could then be used to calculate a correction factor for every item, to correct for unconscious sources of bias in the judging.

The regression modelling shows that certain question types and questions with some features were consistently judged as either more or less difficult than would be expected, based upon the way students actually performed on them. The need for the judgements to be made relatively quickly in the comparative judgement exercise may have encouraged the use of heuristics rather than deep consideration of the questions. However, it is known that expert judges sometimes misjudge the difficulty of items for students, and these biases may be inherent in any judgement of difficulty, not just unique to this comparative judgement study. Research is needed to investigate the occurrence of bias in different types of difficulty judging processes.

Appendix B – Additional data tables

Table B1: *Median expected item difficulty for all items from the 2014 and reformed GCSE assessments split by entry tier.*

	Phase 1		Phase 2
	2014 assessment	**Sample assessments**	**Sample assessments**
Foundation Tier	-0.33	-0.27	-0.27
Higher Tier	0.42	0.37	0.41

Table B2: *Median expected item difficulty for all items from the 2014 and reformed GCSE assessments split by subject domain and entry tier.*

	Phase 1		Phase 2
	2014 assessment	**Sample assessments**	**Sample assessments**
Biology			
Foundation tier	-0.28	-0.17	-0.34
Higher tier	0.58	0.35	0.37
Chemistry			
Foundation tier	-0.48	-0.28	-0.23
Higher tier	0.21	0.30	0.29
Physics			
Foundation tier	-0.25	-0.25	-0.21
Higher tier	0.43	0.42	0.59
Combined science			
Foundation tier		-0.31	-0.29
Higher tier		0.38	0.42

Table B3: *Median, mean and standard deviation of expected item difficulty for all single science biology specifications from the 2014 and reformed GCSE assessments split by tier.*

Biology	Phase 1						Phase 2		
	2014 assessments			Sample assessments			Sample assessments		
Board	**median**	**mean**	**sd**	**median**	**mean**	**sd**	**median**	**mean**	**sd**
Spec 1									
Foundation	-0.19	-0.10	1.07	0.05	-0.01	1.07	0.01	-0.01	1.08
Higher	0.59	0.54	1.04	0.66	0.54	1.03	0.67	0.54	1.02
Spec 2									
Foundation	-0.71	-0.60	1.23	-0.90	-0.82	1.08	-0.90	-0.88	1.21
Higher	0.67	0.51	1.11	0.11	0.06	1.17	0.12	0.21	1.27
Spec 3									
Foundation	-0.22	-0.23	1.01	0.15	0.16	1.04	0.27	0.12	0.96
Higher	0.50	0.48	1.03	0.37	0.34	1.12	0.34	0.19	0.96
Spec 4									
Foundation	0.23	0.19	1.18	-0.08	-0.21	1.06	-0.19	-0.26	1.09
Higher	0.70	0.61	1.08	0.40	0.38	1.04	0.41	0.43	1.25
Spec 5									
Foundation	-0.43	-0.41	1.08	-0.06	-0.10	1.14	-0.33	-0.20	1.02
Higher	0.50	0.58	1.29	0.22	0.25	1.22	0.07	0.15	1.15

Table B4: *Median, mean and standard deviation of expected item difficulty for all single science chemistry specifications from the 2014 and reformed GCSE assessments split by tier.*

Chemistry	Phase 1						Phase 2		
	2014 assessments			Sample assessments			Sample assessments		
Board	**median**	**mean**	**sd**	**median**	**mean**	**sd**	**median**	**mean**	**sd**
Spec 1									
Foundation	-0.34	-0.36	1.11	-0.37	-0.39	1.19	-0.27	-0.29	1.32
Higher	0.33	0.46	1.14	0.33	0.50	1.55	0.46	0.38	1.32
Spec 2									
Foundation	-0.20	-0.31	1.17	-0.26	-0.34	0.93	-0.16	-0.24	0.91
Higher	0.22	0.24	1.02	0.10	0.15	1.14	0.16	0.23	1.13
Spec 3									
Foundation	-0.77	-0.86	1.15	-0.62	-0.59	1.18	-0.65	-0.65	1.21
Higher	-0.15	-0.05	0.94	0.23	0.07	1.28	0.02	-0.02	1.28
Spec 4									
Foundation	-0.38	-0.37	1.05	0.11	-0.10	1.15	-0.07	-0.12	1.18
Higher	0.53	0.53	1.16	0.45	0.60	1.49	0.29	0.46	1.52
Spec 5									
Foundation	-0.28	-0.42	1.06	-0.02	-0.01	1.21	0.22	0.12	1.18
Higher	0.37	0.30	0.94	0.45	0.40	1.10	0.45	0.37	1.07

Table B5: *Median, mean and standard deviation of expected item difficulty for all single science physics specifications from the 2014 and reformed GCSE assessments split by tier.*

Physics	Phase 1						Phase 2		
	2014 assessments			Sample assessments			Sample assessments		
Board	**median**	**mean**	**sd**	**median**	**mean**	**sd**	**median**	**mean**	**sd**
Spec 1									
Foundation	-0.21	-0.15	0.95	0.01	-0.03	0.99	-0.11	-0.14	1.13
Higher	0.21	0.29	0.92	0.89	0.77	1.01	0.86	0.75	0.99
Spec 2									
Foundation	-0.26	-0.22	1.12	-0.03	0.01	1.03	-0.26	-0.27	1.17
Higher	0.89	0.84	0.96	0.71	0.71	1.17	0.71	0.85	1.34
Spec 3									
Foundation	-0.26	-0.39	1.20	-0.47	-0.51	1.12	0.18	0.18	1.35
Higher	0.31	0.48	1.36	0.37	0.45	1.23	0.82	0.93	1.00
Spec 4									
Foundation	-0.53	-0.64	1.05	-0.51	-0.63	1.10	-0.43	-0.48	1.07
Higher	0.13	0.09	0.83	-0.18	-0.11	1.03	-0.11	0.13	1.10
Spec 5									
Foundation	0.32	0.12	1.36	-0.19	-0.14	0.82	-0.38	-0.35	0.94
Higher	0.87	0.91	0.92	0.47	0.57	1.13	0.52	0.63	1.05

Table B6: *Median, mean and standard deviation of expected item difficulty for all combined science specifications from the 2014 and reformed GCSE assessments split by tier.*

Combined science	Phase 1			Phase 2		
	Sample assessments			Sample assessments		
Board	**median**	**mean**	**sd**	**median**	**mean**	**sd**
Spec 1						
Foundation	0.04	-0.07	1.07	0.07	-0.03	1.08
Higher	0.54	0.54	1.05	0.53	0.49	1.05
Spec 2						
Foundation	-0.67	-0.71	1.14	-0.68	-0.69	1.26
Higher	0.10	0.08	1.25	0.37	0.26	1.39
Spec 3						
Foundation	-0.17	-0.22	1.24	-0.21	-0.27	1.21
Higher	0.43	0.39	1.15	0.40	0.36	1.12
Spec 4						
Foundation	-0.42	-0.49	1.16	-0.24	-0.21	1.29
Higher	0.18	0.15	1.29	0.38	0.36	1.25
Spec 5						
Foundation	-0.12	-0.18	1.19	-0.15	-0.23	1.24
Higher	0.44	0.53	1.19	0.38	0.49	1.12
Spec 6						
Foundation	-0.43	-0.44	1.20	-0.38	-0.44	1.19
Higher	0.55	0.37	1.15	0.49	0.32	1.30

Appendix C - Relationship of expected item difficulty and live performance data for 2014 items

Facility values[10] were obtained from the summer exams for all of the 2014 items[11]. The Pearson correlation of facility and expected item difficulty is shown in Table C1. An example scatterplot of the relationship between these values for Biology items is shown in Figure A1a. Note that these two variables are measuring slightly different things, facility accounts for the awarding of intermediate marks for multi-mark questions while expected difficulty is based only on achieving full marks.

Table C1: C*orrelation between expected item difficulty and item facility for 2014 items*

Subject	Foundation tier	Higher tier
Biology	-0.500 (n=351)	-0.442 (n=326)
Chemistry	-0.462 (n=373)	-0.388 (n=342)
Physics	-0.448 (n=360)	-0.360 (n=314)

The proportion of students receiving each mark on each item were available for the 2014 OCR items. This proportion receiving full marks was correlated with the expected item difficulty, shown in Table C2, to give an indication of how much the correlations with facility above underestimate the relationship between student performance and expected difficulty. The correlations with facility are re-calculated since the OCR items are a subset of those items correlated in Table C1.

[10] Facility is the average proportion of the total marks available for the item achieved by the cohort of students taking that assessment.

[11] Eduqas collect facility values at the question-level, not item-level, so we were only able to use the facility values for those questions that were not divided into parts.

Table C2: *Correlation between expected item difficulty and item facility or maximum mark proportion for 2014 OCR items*

Subject	Foundation tier		Higher tier	
	Facility	Maximum mark proportion	Facility	Maximum mark proportion
Biology	-0.383	-0.630	-0.345	-0.599
Chemistry	-0.438	-0.667	-0.418	-0.653
Physics	-0.424	-0.637	-0.356	-0.609

When expected difficulty is correlated against maximum mark proportion rather than facility the correlation coefficient increases substantially. This indicates that against an equivalent measure of item difficulty, the predictive power of the expected difficulty parameters is likely to be similar to that in the GCSE maths study, where a correlation of about 0.66[12] was found. This finding is consistent with the high inter-judge reliability reported above in confirming that judges were making consistent judgements, and knew what criteria they were judging against.

It is worth noting that question tariffs are similarly distributed across the different specifications, with similar proportions of low- and high-tariff questions. This means that there is little bias inherent in the use of 'full mark difficulty' as judged here. Although this may lead to high-tariff questions being judged relatively more difficult than low-tariff questions, this effect is similar across specifications.

C.1 Conclusion

The judgement made by judges was based on which question would be harder to achieve full marks on. This judgement contributed to a lower correlation with live marking data than we found with GCSE mathematics questions, probably because the two measures are not assessing quite the same thing. There was a need for the judges to have a clearly defined benchmark with which to compare items. The inter-judge reliability analysis, and the higher correlation found between expected item difficulty and the proportion of students receiving full marks in the live papers suggests that the judges were able to effectively judge against this benchmark. Although some notion of 'overall' difficulty may in theory be more similar to live item facility, a loosely defined criteria would tend to lead to inconsistency between judges

[12] All correlations are unadjusted Pearson correlations. Correction for the uncertainty associated with difficulty estimates (the disattenuated correlation) have not been applied.

who would apply this criteria differently, and would give lower reliability and make fitting a model to the data problematic.

Appendix D - Practical skills questions

The reformed GCSEs in science include questions designed to assess practical skills. These replace the controlled assessment in the legacy GCSEs. The exam boards identified questions designed to assess practical skills, and these are here analysed separately. The same specification numbers used in the main body of the report are used here. Due to the random allocation of specification number across subjects, this means that each row represents sample assessments from different exam boards. The key comparison is down each column, where the range of difficulties within a subject can be seen.

Table D1: *Median expected item difficulty for practical questions on the sample assessment materials by specification and subject.*

Board	Biology		Chemistry		Physics		Combined science	
	Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2
Spec 1								
Foundation	0.39	0.28	-0.45	-0.55	0.30	0.02	0.11	0.07
Higher	0.35	0.66	-0.06	-0.08	0.69	0.50	0.06	0.17
Spec 2								
Foundation	-1.14	-1.15	-0.61	-0.46	-0.14	-0.25	-0.39	-0.60
Higher	0.05	0.01	-0.46	0.02	0.73	0.72	-0.05	0.10
Spec 3								
Foundation	0.05	0.24	-1.25	-1.23	-0.55	0.05	-0.18	-0.28
Higher	-0.12	-0.01	-0.68	0.27	-0.19	0.65	0.33	0.21
Spec 4								
Foundation	-0.33	-0.02	0.14	-0.07	0.08	0.00	-0.62	-0.13
Higher	-0.33	-0.35	0.28	0.40	-0.39	-0.20	-0.19	0.20
Spec 5								
Foundation	0.17	0.01	-0.19	-0.09	-0.12	-0.60	-0.30	-0.38
Higher	-0.08	-0.24	0.23	0.23	0.15	0.38	0.33	0.21
Spec 6								
Foundation							0.22	-0.05
Higher							0.65	0.67

Across all specifications and subjects the median difficulty for practical items after phase 2 was -0.21 for foundation tier and 0.22 for higher tier. This compares to all items where median difficulty was -0.27 for foundation tier and 0.41 for higher tier. The practical skills questions differ slightly less between tiers than the full set of items. There are several instances where the data suggests that there is little

differentiation in difficulty between the tiers, and sometimes less difficulty on the higher tier practical items than the foundation tier, particularly for the biology sample assessments.

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

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

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

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

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