A level and AS mathematics: An evaluation of the expected item difficulty

January 2018

Ofqual/18/6344

Authors

This report was written by Stephen D Holmes and Stephen Rhead, from Ofqual's Strategy, Risk and Research directorate.

Contents

Executive summary

In 2016-2017, exam boards submitted draft AS and A level mathematics specifications to Ofqual for the purposes of accreditation for first teaching in 2017. We use accreditation to decide whether new GCSEs, AS and A level qualifications produced by exam boards can be awarded. To inform discussions and recommendations made by the accreditation panels regarding the likely difficulty of future live examinations, Ofqual carried out several phases of comparative judgement studies of the relative expected difficulty of items from the 2015 legacy specifications together with items from the sample assessment materials (SAMs).

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

The overall objective of this exercise was to be able to compare the profile of item difficulty within the SAMs with that of the corresponding 2015 assessments. A variety of other factors such as assessment structure (modular or linear) and changes to content were also considered by the accreditation panel in 2016/2017, alongside the expected difficulty of items estimated by comparative judgement reported here.

Prior to carrying out the main comparative judgement study, we piloted several different study designs, where the effect of judge type, inclusion (or not) of mark schemes, and approaches to judging difficulty were tested. The results suggested that teachers were slightly better than PhD maths students at judging the difficulty of an item for candidates, that mark schemes helped in making the judgements, and that estimating the overall difficulty of items was more closely related to average candidate score on the item than estimating the difficulty of giving a completely accurate answer. All of these findings were consistent with expectations and informed the main study design; teachers were recruited, mark schemes were included and the appropriate judging criteria was selected.

In the main study we took into consideration the change from a modular to a linear design in the A level and applied an empirically-derived adjustment to the expected difficulty of some of the legacy items. Having done so, overall the distribution of expected difficulty of items in the main studies was very similar between the legacy and reformed specifications for both AS and A level, with only a small increase in average difficulty for the reformed assessments. This was in broad agreement with the intention to keep item difficulty equal between the legacy and reformed AS/A level mathematics as the legacy qualifications were considered to be of appropriate demand. The small levels of variation between the expected difficulty distributions of the reformed specifications is 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.

1 Background

Alongside the formal accreditation process for reformed A level and AS mathematics specifications for first teaching in 2017, Ofqual carried out a series of comparative judgement studies on the expected difficulty of items from the reformed A level and AS sample assessment materials (SAMs), together with items from the legacy 2015 A level and AS mathematics assessments. The purpose of this was to inform discussions and decisions 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 2015 papers and the SAMs within each specification. 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 the assessment. There was no intention in the reform of A level/AS mathematics to change the level of difficulty of the items –the legacy assessments were already considered to be of an appropriate level of demand. Therefore we would not expect to see any major changes in item difficulty.

Prior to the main series of studies, a pilot study was run using a subset of the 2015 items to help inform the design of the main studies. Appendix A describes this pilot in detail. Following the pilot, there were several phases of studies. In the first phase of submissions, all the items from the sample assessment materials submitted for 4 specifications (AQA, MEI, OCR and Pearson, here anonymised as Specifications 1-4, not in that order) in June 2016 were judged together for difficulty alongside the 2015 assessment items from the corresponding legacy specifications. Subsequently, each exam board re-submitted their sample assessments at different times, and so each submission was judged independently. In order to retain the same difficulty scale, all of the subsequent studies included a number of anchor items from the first submission, or phase 1, study.

It is worth noting that the accreditation process considers a wide variety of factors, only one of which is difficulty. The reasons for rejection and resubmission may not always have been related to difficulty, but sometimes another phase of comparative judgement was required to confirm that other requested changes did not impact upon difficulty.

2 Choice of legacy (2015) units for comparison

Given the modular nature of the legacy A level and AS, the choice of units to include from these specifications was important. Two considerations were uppermost – content coverage, and representativeness of route (in terms of candidate numbers).

In consultation with the exam boards, the following 6 units were chosen as representative of the legacy A level:

- Core 1 (C1)
- Core 2 (C2)
- Core 3 (C3)
- Core 4 (C4)
- Mechanics 1 (M1)
- Statistics 1 (S1)

These units make up a very frequently-chosen route through the modular legacy A level, and also match the planned content of the reformed A level, which includes both statistics and mechanics content alongside pure mathematics. This representative route comprises 4 AS units and 2 A2 units, and Section 4.2 and Appendix B describe an adjustment we applied to the estimated difficulty of the AS unit items.

For the 3 AS units, 2 equally representative routes were chosen: core 1, core 2 and either mechanics 1 or statistics 1. The overall difficulty of both these routes through the 2015 AS are presented as comparators to the single reformed AS in the results that follow.

3 Method

The comparative judgement method broadly followed the method used in earlier research into the difficulty of GCSE mathematics and GCSE science[1] questions. Briefly, the current study involved a number of A level and AS mathematics 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. The items were presented with their mark schemes, as it was possible that changes in the design of mark schemes in the reformed assessments could have an effect on item difficulty. Pilot work also showed that inclusion of mark schemes improved the correspondence between the judged difficulty and item facility from the 2015 series.

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

https://www.gov.uk/government/publications/gcse-science-an-evaluation-of-the-expected-difficulty-of-items

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.

3.1 Materials

In the first phase, all items from the sample assessments submitted in June 2016 were included in the comparative judgement exercise, together with items from the summer 2015 A level and AS assessment units described in Section 2 (see Table 1, showing the first phase counts to give a sense of the number of items in the reformed assessments).

Table 1. *Items included in the first phase study.*

2015 papers							
Specification	**C1**	**C2**	**C3**	**C4**	**M1**	**S1**	**Total**
Specification 1	22	22	26	21	17	33	141
Specification 2	23	29	17	19	23	25	126
Specification 3	28	25	23	22	23	29	150
Specification 4	22	22	19	19	24	31	137
							554

Phase 1 sample assessments		**AS**			**A level**		
Specification	**S1**	**S2**	**Total**	**A1**	**A2**	**A3**	**Total**
Specification 1	32	20	52	40	32	33	105
Specification 2	25	29	54	29	32	25	86
Specification 3	32	36	68	31	35	35	101
Specification 4	28	26	54	31	33	27	91
			228				383

Subsequent phases were carried out as the submissions were received from each exam board (in one instance they could be combined). The judging was carried out on items that were either new in the submission, or modified sufficiently from the previous submission to justify re-judging. This decision was made by Ofqual researchers, and by default items were re-judged, unless the change was very minor such as a layout change. Table 2 lists the number of items that were included in the phase 2 and later studies.

3.1.1 Anchor items

To ensure that the modelled scale of expected difficulty was the same across all phases of this work, a number of items from the phase 1 study were included in the comparisons for phases 2-4. Their expected difficulty parameters were fixed at the value obtained in phase 1 when the phase 2-4 models were fitted. These items are referred to as anchor items.

We used the same 50 anchor items for most of the phase 2-4 studies (see brackets in the first column of Table 2) in order to cover the full extent of the difficulty scale. These were drawn randomly from the output of phase 1, and included items from all specifications and both reformed and legacy assessments. Where the number of new items to be judged was very low, rather than collect hundreds of (effectively meaningless[2]) comparisons between anchor items, only 20 anchor items were used. These were drawn from the original 50 anchors at roughly equal spacing along the expected difficulty scale.

3.1.2 Item format

A standardised format was used so that any formatting and layout features which might have enabled judges to identify the specification were removed. However, note that the mark schemes were copied as images from the published/submitted mark schemes. Although every attempt was made to select only the parts of the mark schemes that contained the detailed mark scheme information, exam boards used slightly different columns and layouts in their mark schemes.

[2] Because the anchor item expected difficulties were fixed in the model fitting, comparisons between them contributed nothing to the analysis. Only comparisons between new items, and between new and anchor items conveyed any information.

Table 2. *Summary of completed studies. For phase 1, the first column separates the number of reformed and legacy items (in brackets). For phases 2-4, the first column separates the number of new items and the number of phase 1 anchor items (in brackets).*

	Number of new A level and AS items (2015 items in brackets for phase 1, anchors in brackets for phases 2-4)	Number of judges (misfitting judges in brackets)	Planned number of judgements per judge	Total judgements analysed	Judgements per item	Range of median judging time in seconds (mean in brackets)	Split-half reliability (std dev of correlations in brackets)	SSR
Phase 1 Study 1	383 + 228 (+ 554)	43 (-4)	500	19277	33.1	11-84 (31)	0.71 (0.05)	0.91
Phase 2								
Study 2	28 + 20 (+ 50)	26 (0)	65	1603	32.7	13-66 (28)	0.92 (0.02)	0.96
Study 3	35 + 19 (+ 50)	27 (-1)	70	1770	34.0	13-61 (33)	0.91 (0.02)	0.96
Study 4	39 + 19 (+ 50)	24 (0)	75	1729	32.0	9-61 (28)	0.93 (0.01)	0.95
Phase 3								
Study 5	19 + 17 (+ 50)	26 (0)	60	1466	34.1	8-48 (26)	0.90 (0.02)	0.97
Study 6	0 + 4 (+ 20)	21 (0)	25	525	43.8	8-37 (17)	0.98 (0.02)	0.99

	Number of new A level and AS items (2015 items in brackets for phase 1, anchors in brackets for phases 2-4)	Number of judges (misfitting judges in brackets)	Planned number of judgements per judge	Total judgements analysed	Judgements per item	Range of median judging time in seconds (mean in brackets)	Split-half reliability (std dev of correlations in brackets)	SSR
Phase 4								
Study 7	0 + 9 (+ 20)	22 (-1)	30	630	43.4	8-26 (17)	0.90 (0.05)	0.99

3.2 Participants

Across the whole set of studies, 45 current A level / AS mathematics teachers were recruited as judges. Fifteen of these teachers had taken part in the pilot studies (see Appendix A). Initially they were only recruited for phase 1, but many of them were willing and able to continue throughout all of the subsequent phases of judging. This continuity of judges was extremely useful in ensuring comparability across the different studies. The number of judges varied across studies, as shown in Table 2.

For each study all judges were allocated the same number of judgements, calculated to give roughly the same number of judgements per item across the different studies (see Table 2). For the two smallest studies, where only 20 anchor items were included, the number of judgements per item was increased slightly to ensure that the expected item difficulties were still reliable given the greater spacing of anchors along the expected difficulty scale.

Individual judges did not necessarily take part in every study, or complete their full allocation in those studies they started, due to their availability (the studies took place with a few weeks' notice and with limited time windows in which to carry out the judging in order to support accreditation timelines). Judges were paid for the number of judgements they completed.

3.3 Procedure

Comparisons were conducted using the online comparative judgement platform, No More Marking[3]. 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 overall?'*

Additional clarification regarding the prompt was given in written instructions to the judges:

> 'This refers to the average difficulty for students. So thinking about students across the whole ability range, for which question do you think that on average students will achieve the lower proportion of the total marks available. You can think about how a whole range of students might perform on the two questions. Alternatively, you might want to consider a single 'average' student, and how that one student would perform on the two questions. Your

[3] www.nomoremarking.com

benchmark measure for both is the proportion of full marks that would be achieved.

Example: For an 8 mark question you might expect, on average, students to earn around 3 of the marks available. The other question is worth 3 marks, and you might expect students, on average, to earn 2 marks. Therefore, the 8 mark question is more difficult – even though students might be getting more marks, they are earning a smaller proportion (0.375) of the maximum mark available compared to the other question (0.667).'

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.

4 Analysis

The R package sirt[4] was used to estimate expected difficulty parameters for each item under the Bradley-Terry model. R code was also used to estimate item and judge infit, scale-separation reliability (SSR) and split-half reliability.

4.1 Judge consistency and exclusions

After the initial model fit to the set of judgements, judge infit was checked. 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 from the consensus. Outlying judges were identified and excluded using the criteria of an infit more than two standard deviations above the mean infit value for all judges.

In addition, for phase 1, if the median judging time for a judge was under 10s they were also excluded. Given increasing familiarity with some of the (anchor) items, median judging times slightly below 10s were considered acceptable in phase 2 onwards, providing the judge infit criteria was satisfied.

Table 2 shows that 4 of the 43 judges in phase 1 were excluded, and 1 judge was excluded from 2 of the other studies (shown by the negative number in the second column). The table also shows the range and mean of the median judging times for each judge. Generally, judging became quicker across subsequent studies, due to increasing familiarity with items. This was particularly true for the phase 3 and phase

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

4 studies, where a high proportion of items were anchors. Following the exclusion of judges, the model was refitted and all other statistics are based on this final model fit.

For each study, two separate reliability measures were calculated. The median split-half reliability was assessed by repeatedly allocating judges randomly to 2 groups, fitting the Bradley-Terry model independently for each group and correlating the 2 rank orders of item expected difficulty parameters. This process was repeated 100 times and the median correlation and the standard deviation of the correlations were obtained. Table 2 shows all of the split-half reliability estimates. For phase 1 the median rank order correlation was 0.71, showing reasonable agreement between judges. The correlation was much higher for all of the other phases, due to the effect of anchor items, which force a high degree of consistency between the model fits for the sub-groups of judges.

Reliability is quantified in comparative judgement studies by the scale separation reliability (SSR) statistic that is derived in same way as the person separation reliability index in Rasch analyses. It is interpreted as the proportion of 'true' variance in the estimated scale values. The SSR was 0.91 for the phase 1 study which shows good reliability. It was even higher for the other studies, since there is no variance in the estimated item difficulties for the fixed anchor items.

4.2 Adjustment of 2015 AS item difficulties

Four of the units from the 2015 A level are nominally AS units (core 1, core 2, mechanics 1 and statistics 1). These units are designed to be taken by candidates when they are a year younger than when they take the A2 units (Core 3 and 4) and, importantly, the reformed A level papers. In a straight comparison these items are likely to appear less challenging (although there will be overlap between the AS and A2 items). This could lead to the legacy specification appearing to be easier than it is for students of the appropriate age, due to the design change from modular to linear. Appendix B describes the approach that was taken to mitigate this effect by adjusting the AS unit item difficulties to take into account age effects, when they are considered as part of the A level. This adjustment had the effect of slightly increasing the expected difficulty for these items by around 0.2 on the difficulty scale (items typically ranges from about +5 to -5). No adjustment was applied when considering the items on these papers as part of AS, as they were all targeted at year 12 candidates.

5 Results

Each assessment is shown in the figures in this section as a box plot displaying the median and inter-quartile range of the expected item difficulties on a logit scale on the y-axis. This probabilistic scale describes the log odds of one item being judged more difficult than another item. The absolute value is arbitrary, in this case 0 is set

equal to the mean of all the items included in phase 1. The expected item difficulties have been weighted by the item tariff (maximum mark) by duplicating each item parameter by the number of marks for that item. Each mark on the paper is therefore treated as a 1-mark item, with the same difficulty for all marks within each judged item.

The purpose of this work was to compare the difficulty of each reformed assessment to the difficulty of the legacy specification, so each specification is plotted on separate figures. Specifications have been anonymised as 1-4. For each specification, the distribution of expected item difficulties for the final judged submission of the reformed sample assessment is shown, representing the difficulty of the accredited sample assessments. For some of the specifications minor changes were made following this final judged submission, but these did not substantively affect difficulty.

Appendix C contains additional data tables for these studies.

5.1 AS

Figures 1 to 8 show the distributions of expected item difficulty aggregated by assessment and paper respectively, for the four AS specifications in turn. Figures 1, 3, 5 and 7 combine the data across papers into a whole assessment distribution. For the 2015 assessments, the corresponding specification is shown, together with the combined distribution of all 4 specifications to give a picture of overall qualification difficulty. Two alternative versions of the 2015 assessments are given, one represents the statistics route (C1+C2+S1) and one represents the mechanics route (C1+C2+M1). Figures 2, 4, 6 and 8 plot the individual paper distributions, showing all 4 papers which could be used to form the 2 2015 AS routes.

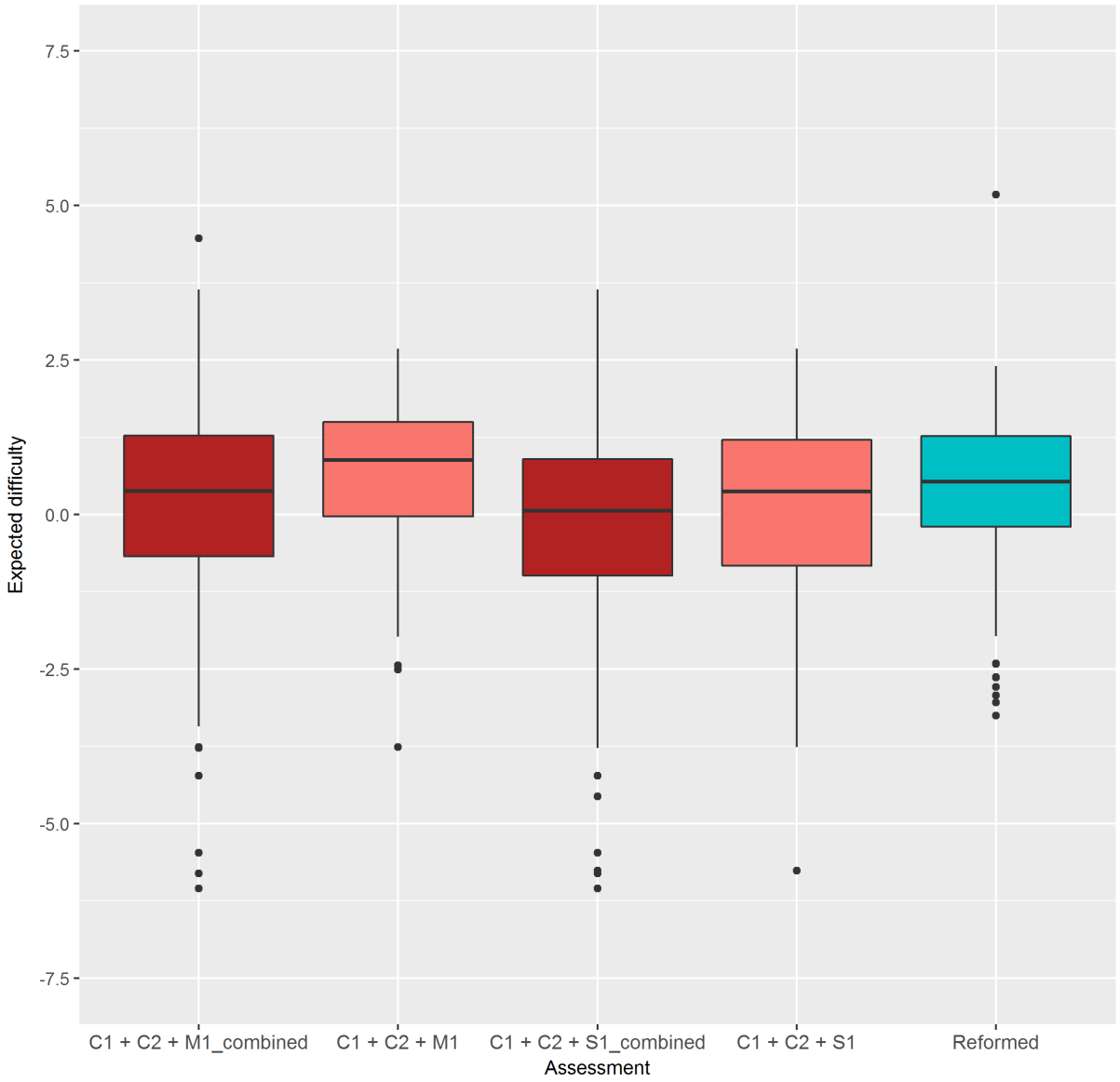

Figure 1: *AS data at whole assessment level for specification 1: Boxplots showing the median and interquartile ranges of expected item difficulty for all of the 2015 assessments combined, and specification 1's 2015 assessments and the final judged reformed sample assessments. The two routes through the 2015 assessments are C1 + C2 + M1 (core 1, core 2 and mechanics 1) and C1 + C2 +S1 (core 1, core 2 and statistics 1) respectively.*

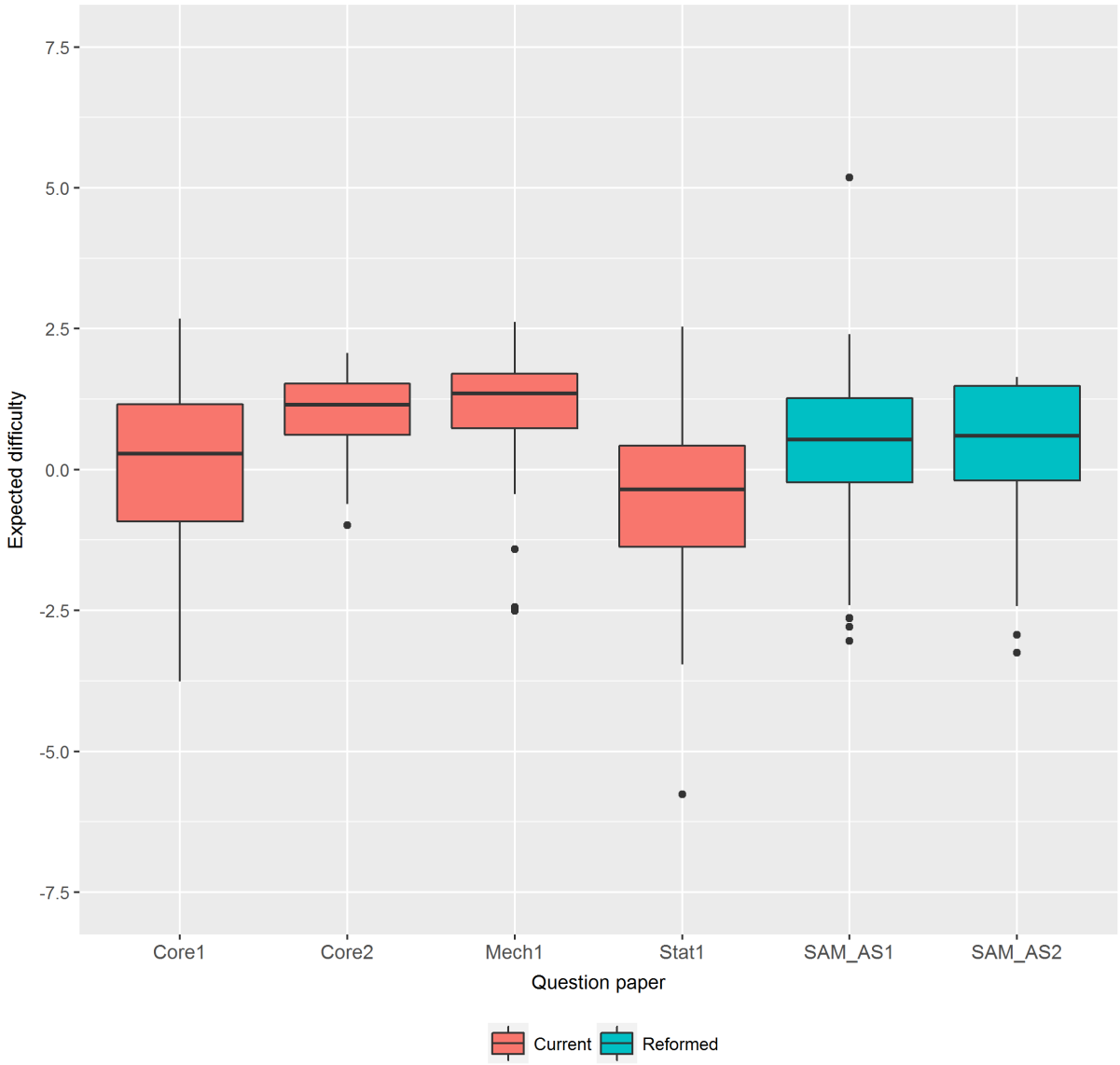

Figure 2: *AS data at paper level for specification 1: Boxplots showing the median and interquartile ranges of expected item difficulty for each paper in specification 1's 2015 assessments and the final judged reformed sample assessments.*

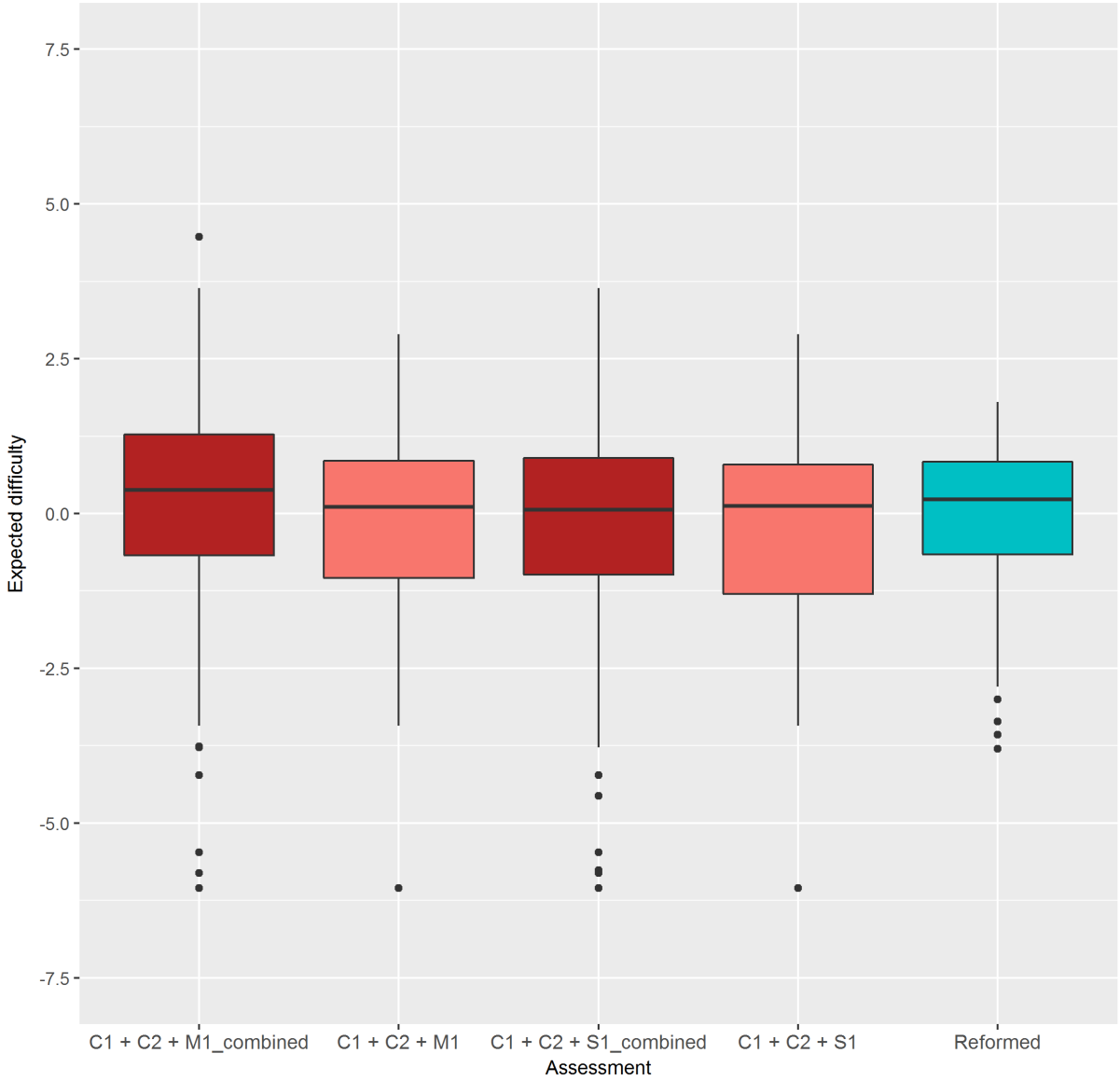

Figure 3: *AS data at whole assessment level for specification 2: Boxplots showing the median and interquartile ranges of expected item difficulty for all of the 2015 assessments combined, and specification 2's 2015 assessments and the final judged reformed sample assessments. The two routes through the 2015 assessments are C1 + C2 + M1 (core 1, core 2 and mechanics 1) and C1 + C2 +S1 (core 1, core 2 and statistics 1) respectively.*

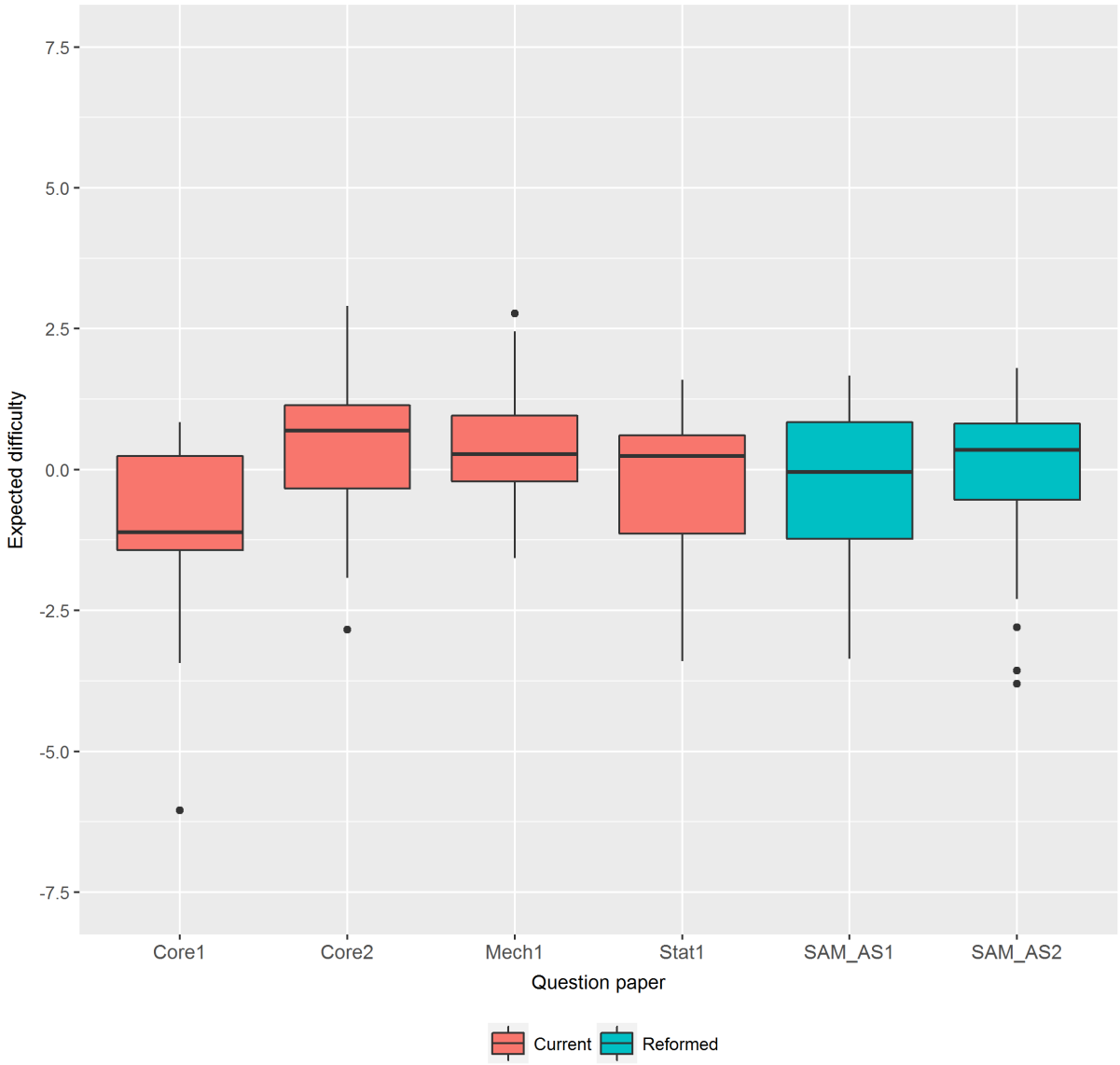

Figure 4: *AS data at paper level for specification 2: Boxplots showing the median and interquartile ranges of expected item difficulty for each paper in specification 2's 2015 assessments and the final judged reformed sample assessments.*

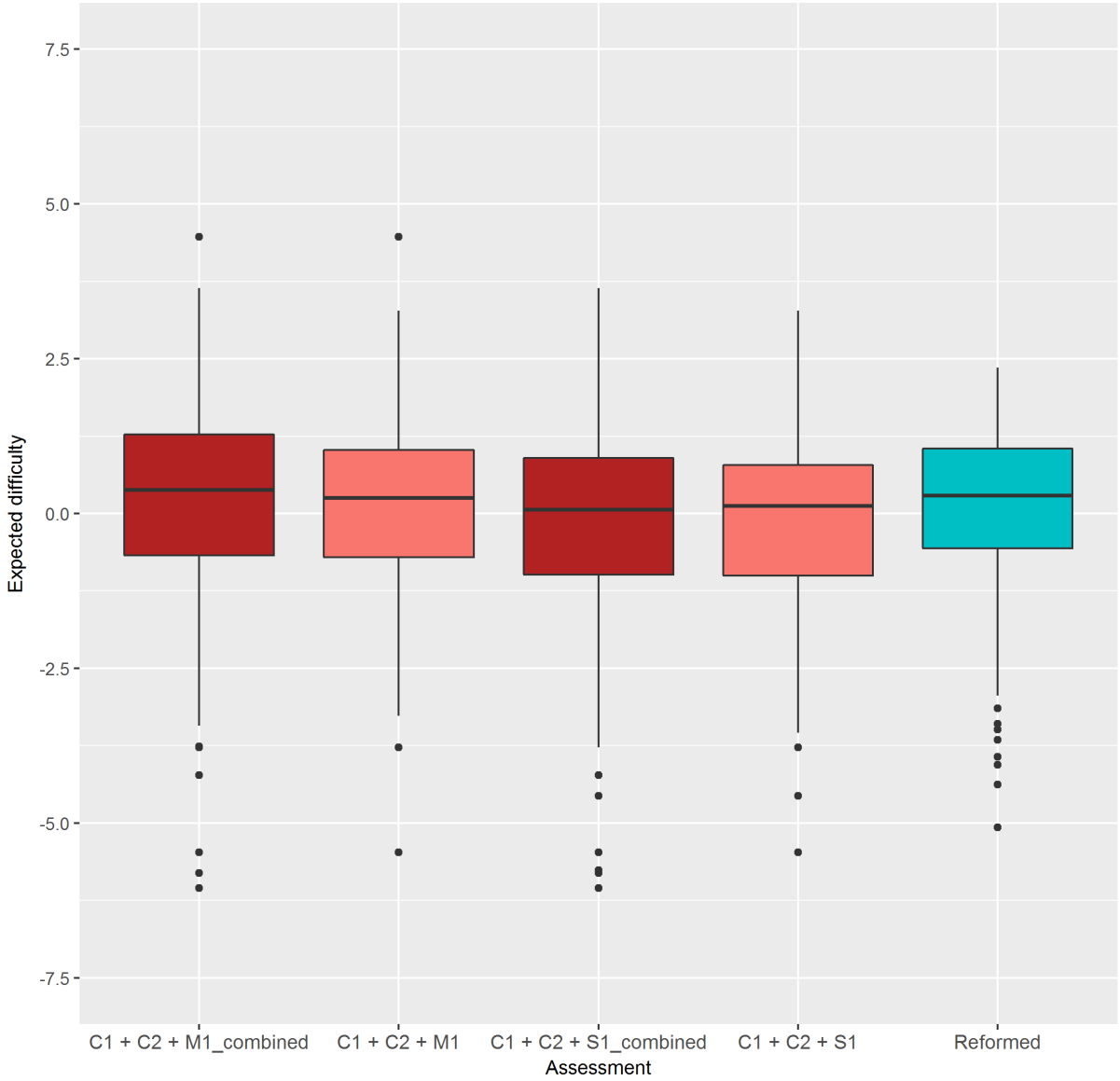

Figure 5: *AS data at whole assessment level for specification 3: Boxplots showing the median and interquartile ranges of expected item difficulty for all of the 2015 assessments combined, and specification 3's 2015 assessments and the final judged reformed sample assessments. The two routes through the 2015 assessments are C1 + C2 + M1 (core 1, core 2 and mechanics 1) and C1 + C2 +S1 (core 1, core 2 and statistics 1) respectively.*

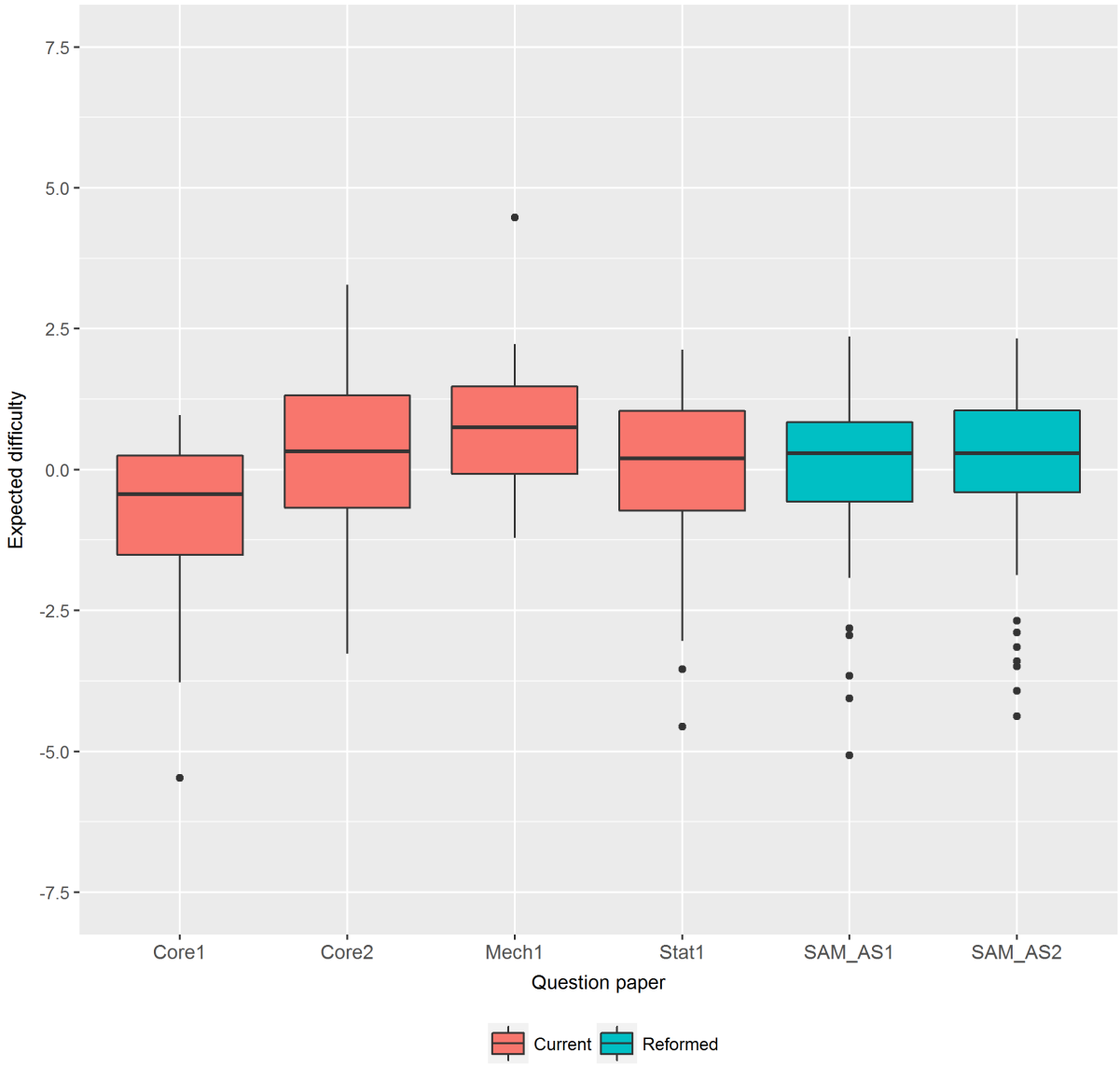

Figure 6: *AS data at paper level for specification 3: Boxplots showing the median and interquartile ranges of expected item difficulty for each paper in specification 3's 2015 assessments and the final judged reformed sample assessments.*

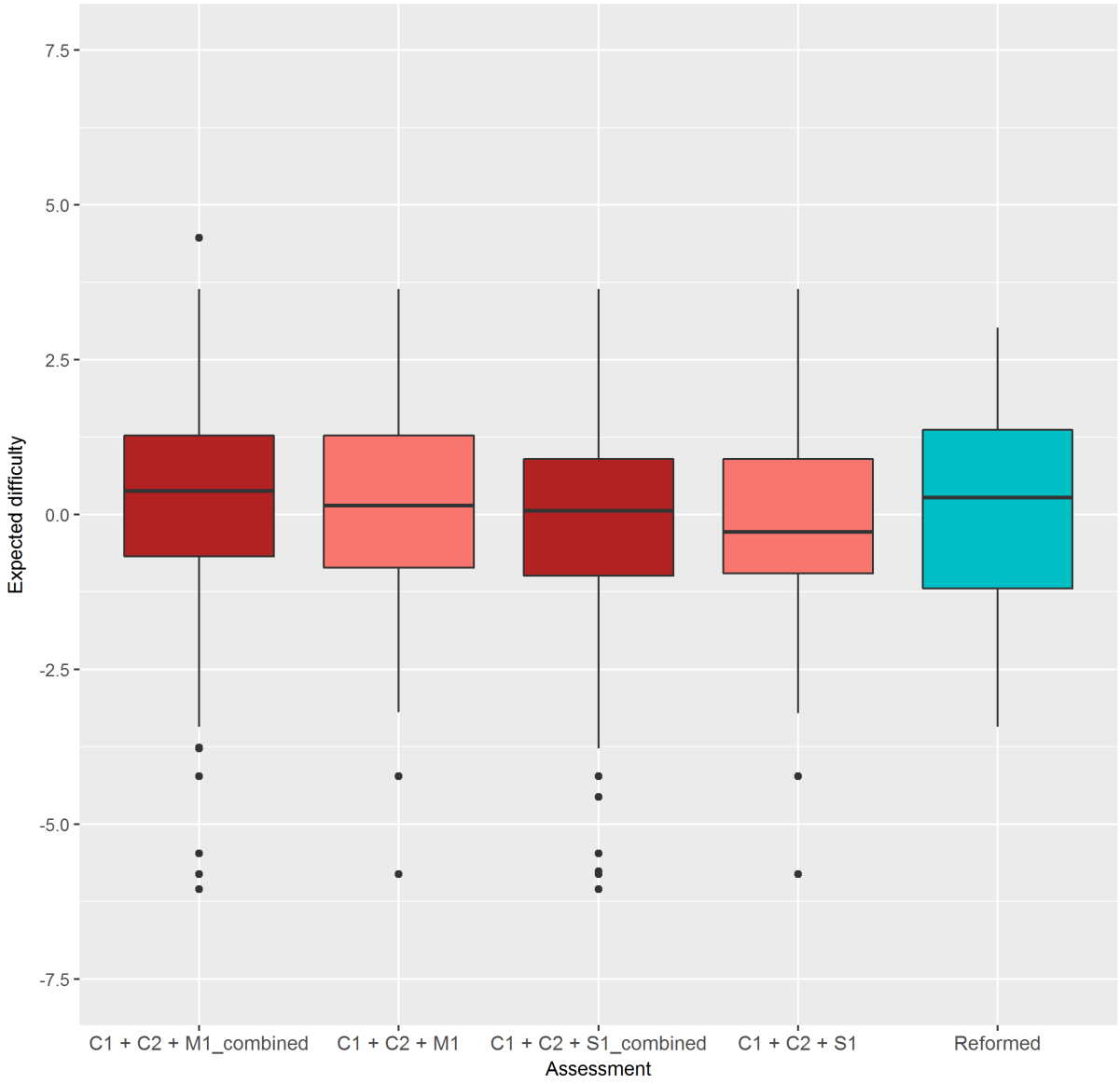

Figure 7: *AS data at whole assessment level for specification 4: Boxplots showing the median and interquartile ranges of expected item difficulty for all of the 2015 assessments combined, and specification 4's 2015 assessments and the final judged reformed sample assessments. The two routes through the 2015 assessments are C1 + C2 + M1 (core 1, core 2 and mechanics 1) and C1 + C2 +S1 (core 1, core 2 and statistics 1) respectively.*

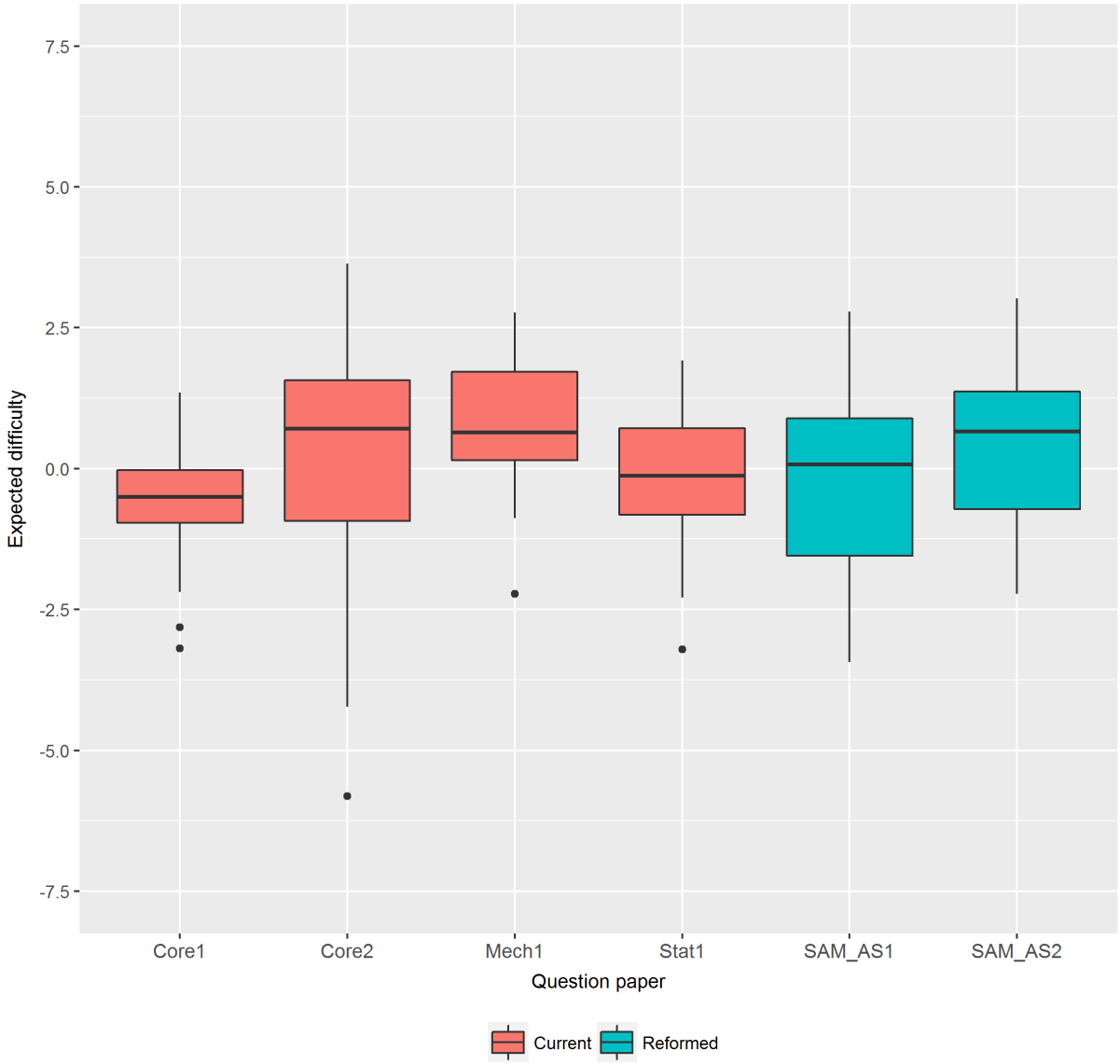

Figure 8: *AS data at paper level for specification 4: Boxplots showing the median and interquartile ranges of expected item difficulty for each paper in specification 4's 2015 assessments and the final judged reformed sample assessments.*

For the mechanics route through the 2015 AS assessments (C1 + C2 + M1) the median difficulty varied from 0.11 to 0.88 (overall median = 0.38), a range of 0.77 logits. For the statistics route through the 2015 AS assessments (C1 + C2 + S1), the median difficulty varied from -0.28 to 0.37 (overall median = 0.06), a range of 0.65 logits. For the final judged versions of the reformed AS sample assessments the median difficulty varied from 0.23 to 0.53 (overall median = 0.34), a narrower range of 0.30 logits.

All 4 reformed AS assessments therefore have median difficulties which are very close to that of the mechanics route combined across all of the legacy assessments. There are also no big differences between each specification's own legacy mechanics route and its reformed AS. Slightly larger differences do exist between the reformed assessments and each specification's legacy statistics route, due to the consistently lower expected difficulty of the statistics route.

The individual papers in the reformed assessments are closer together in median difficulty (and overall distribution) in every case than the papers in the corresponding legacy assessment (while noting that only 3 of the 4 legacy papers would be sat by each candidate). The individual reformed papers are therefore more representative of the overall assessment difficulty than were the legacy papers. The legacy core 1 paper was almost always of very low difficulty, and in some cases so was the statistics 1 paper.

Overall, for the reformed AS sample assessments, relative to the 2015 assessments there appears to be a slight increase in difficulty overall and the spread of the assessment medians are smaller for the reformed sample assessments than the 2015 assessments.

5.2 A level

Figures 9 to 16 show the distributions of expected item difficulty aggregated by assessment and paper respectively, for the four A level specifications in turn. Figures 9, 11, 13 and 15 combine the data across papers into a whole assessment distribution. For the 2015 assessments, the corresponding specification is shown, together with the combined distribution of all 4 specifications to give a picture of overall qualification difficulty. Figures 10, 12, 14 and 16 plot the individual paper distributions.

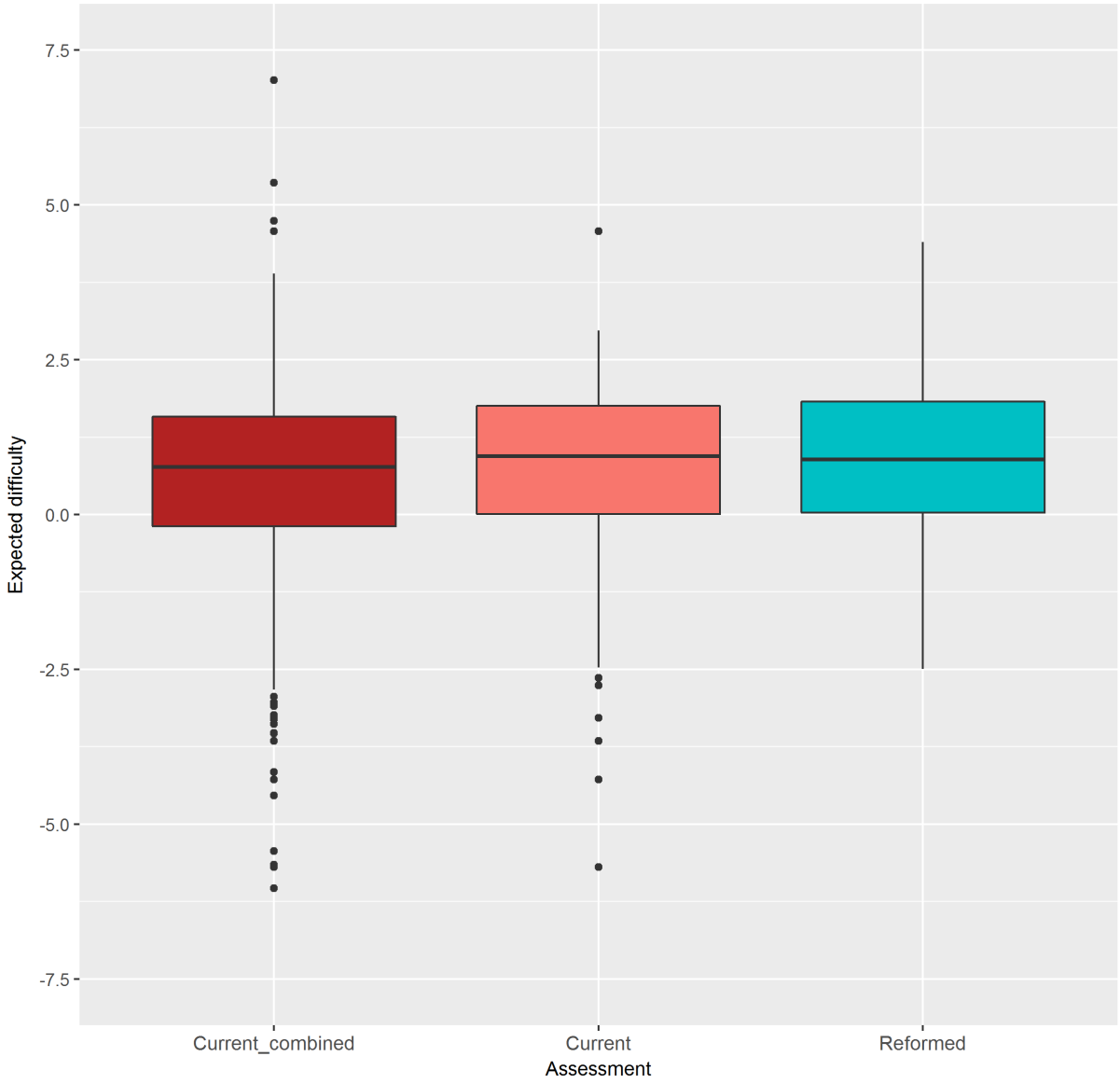

Figure 9: *A level data at whole assessment level for specification 1: Boxplots showing the median and interquartile ranges of expected item difficulty for all of the 2015 assessments combined, and specification 1's 2015 assessments and the final judged reformed sample assessments.*

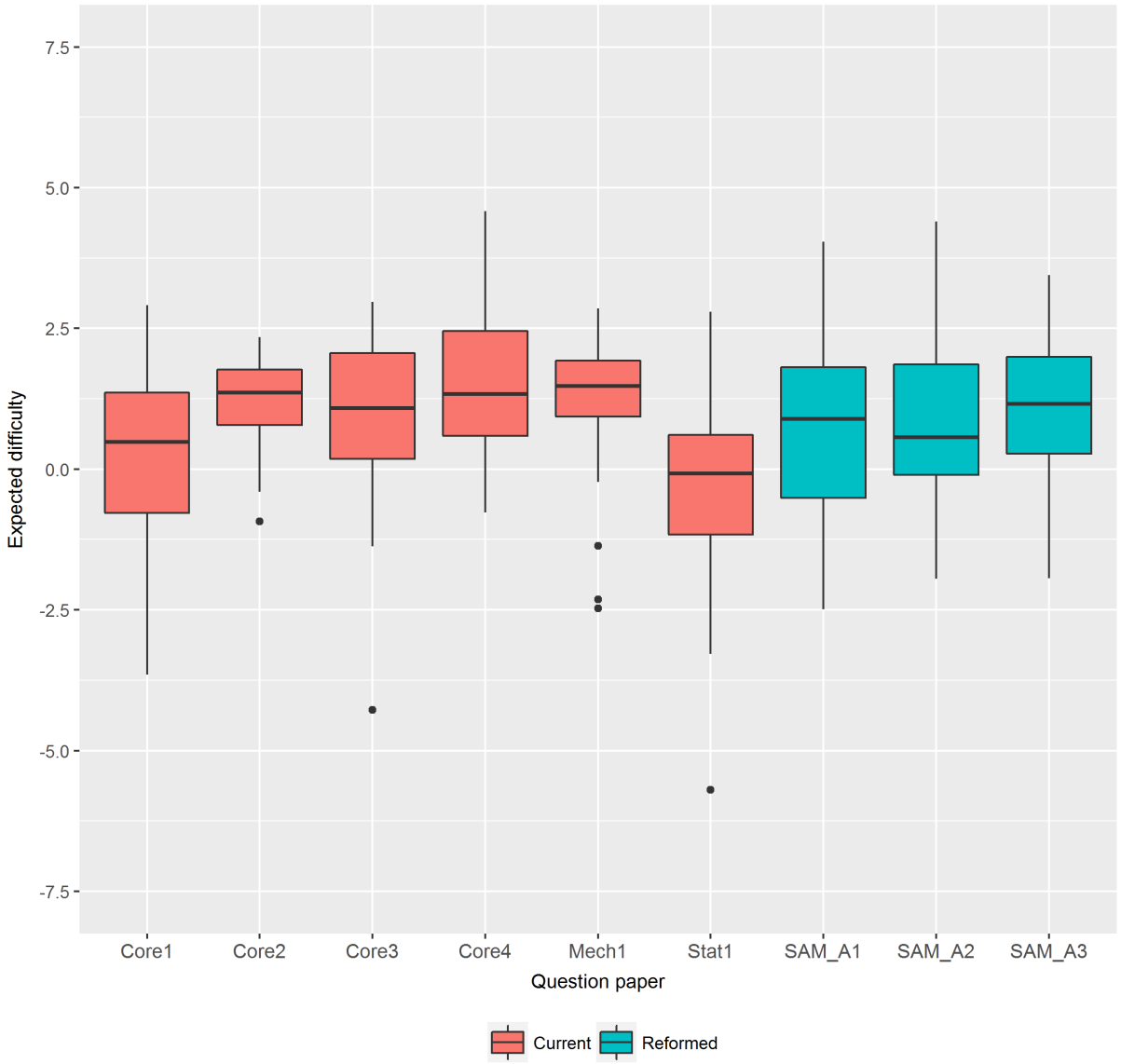

Figure 10: *A level data at paper level for specification 1: Boxplots showing the median and interquartile ranges of expected item difficulty for each paper in specification 1's 2015 assessments and the final judged reformed sample assessments.*

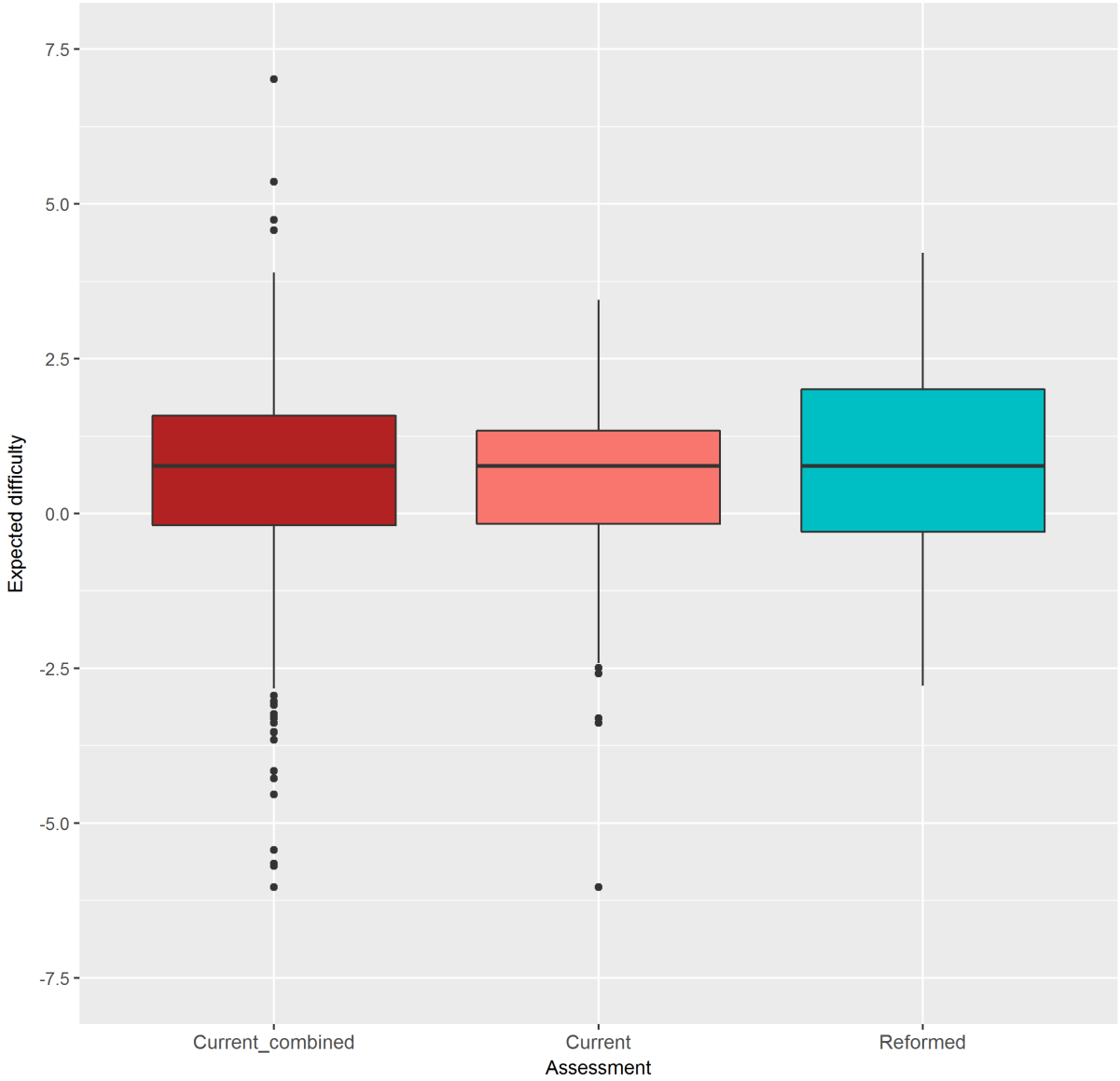

Figure 11: *A level data at whole assessment level for specification 2: Boxplots showing the median and interquartile ranges of expected item difficulty for all of the 2015 assessments combined, and specification 2's 2015 assessments and the final judged reformed sample assessments.*

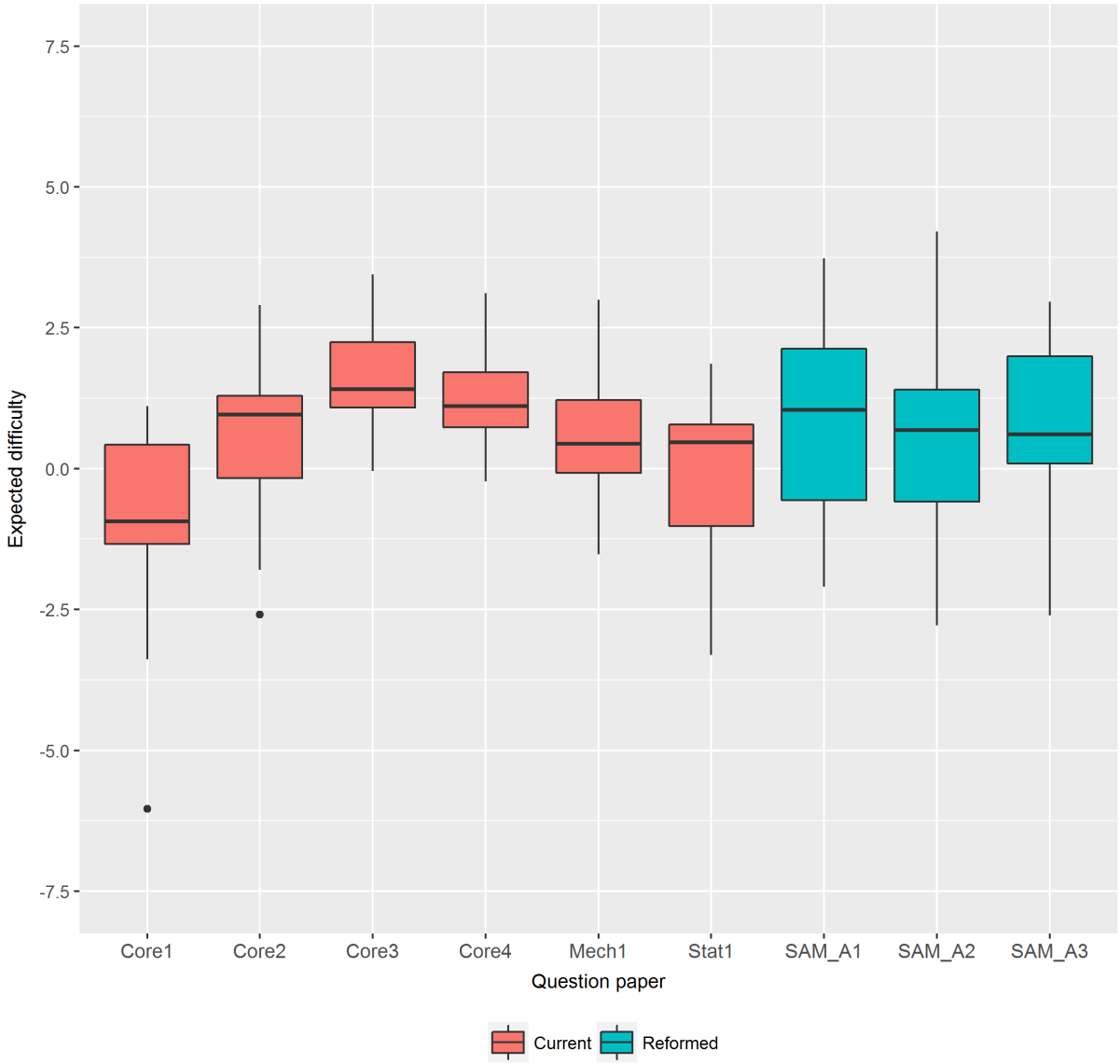

Figure 12: *A level data at paper level for specification 2: Boxplots showing the median and interquartile ranges of expected item difficulty for each paper in specification 2's 2015 assessments and the final judged reformed sample assessments.*

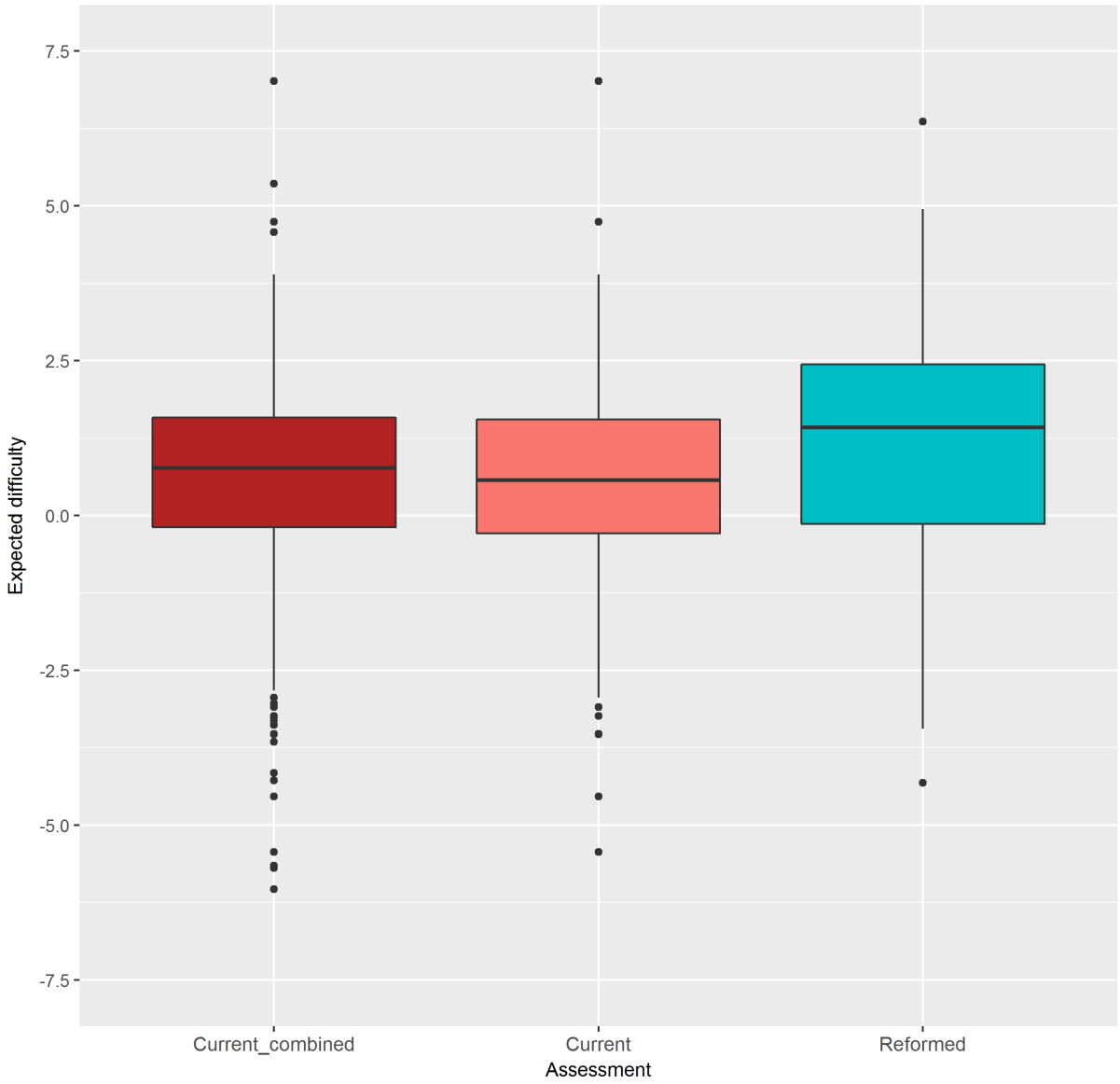

Figure 13: *A level data at whole assessment level for specification 3: Boxplots showing the median and interquartile ranges of expected item difficulty for all of the 2015 assessments combined, and specification 3's 2015 assessments and the final judged reformed sample assessments.*

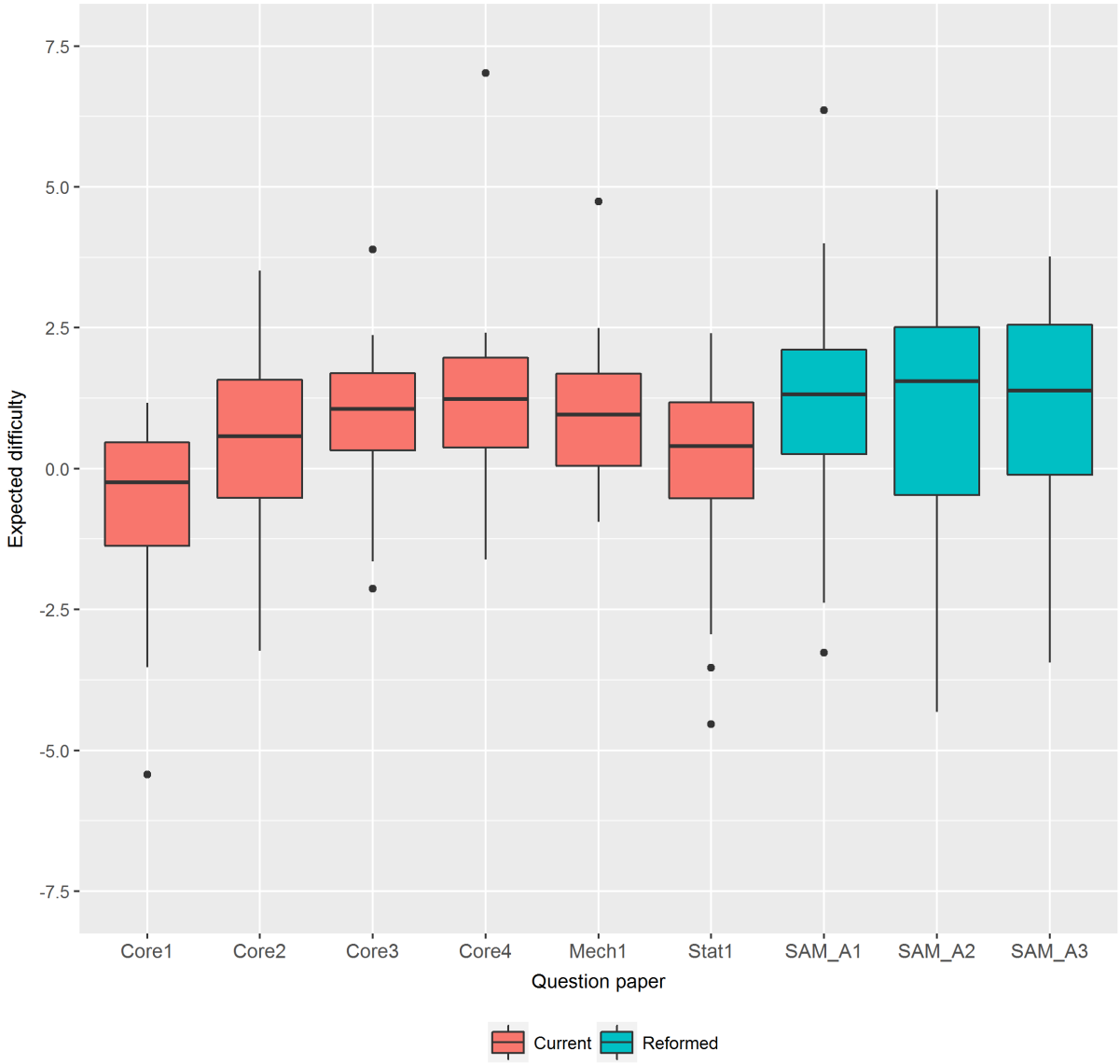

Figure 14: *A level data at paper level for specification 3: Boxplots showing the median and interquartile ranges of expected item difficulty for each paper in specification 3's 2015 assessments and the final judged reformed sample assessments.*

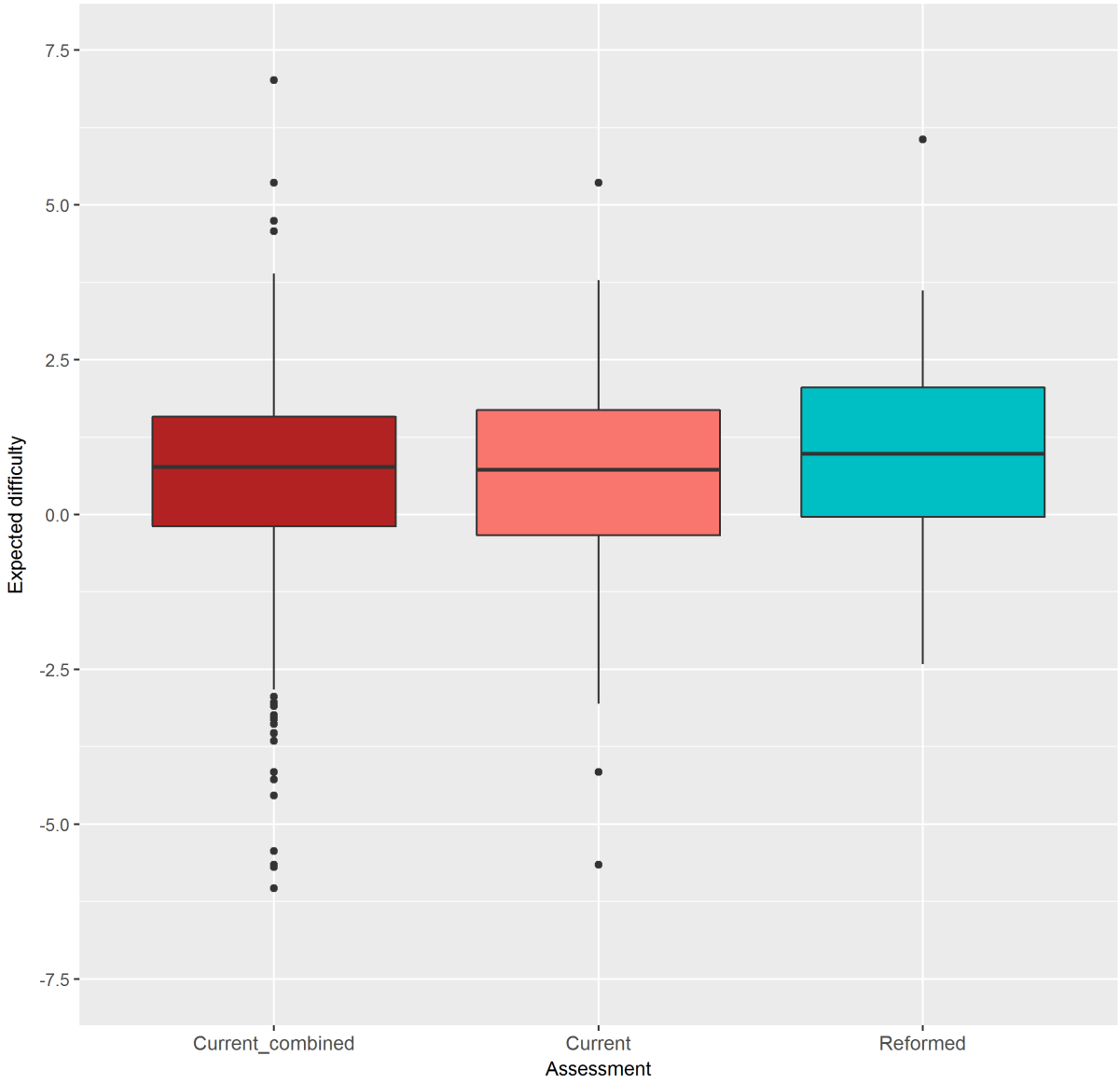

Figure 15: *A level data at whole assessment level for specification 4: Boxplots showing the median and interquartile ranges of expected item difficulty for all of the 2015 assessments combined, and specification 4's 2015 assessments and the final judged reformed sample assessments.*

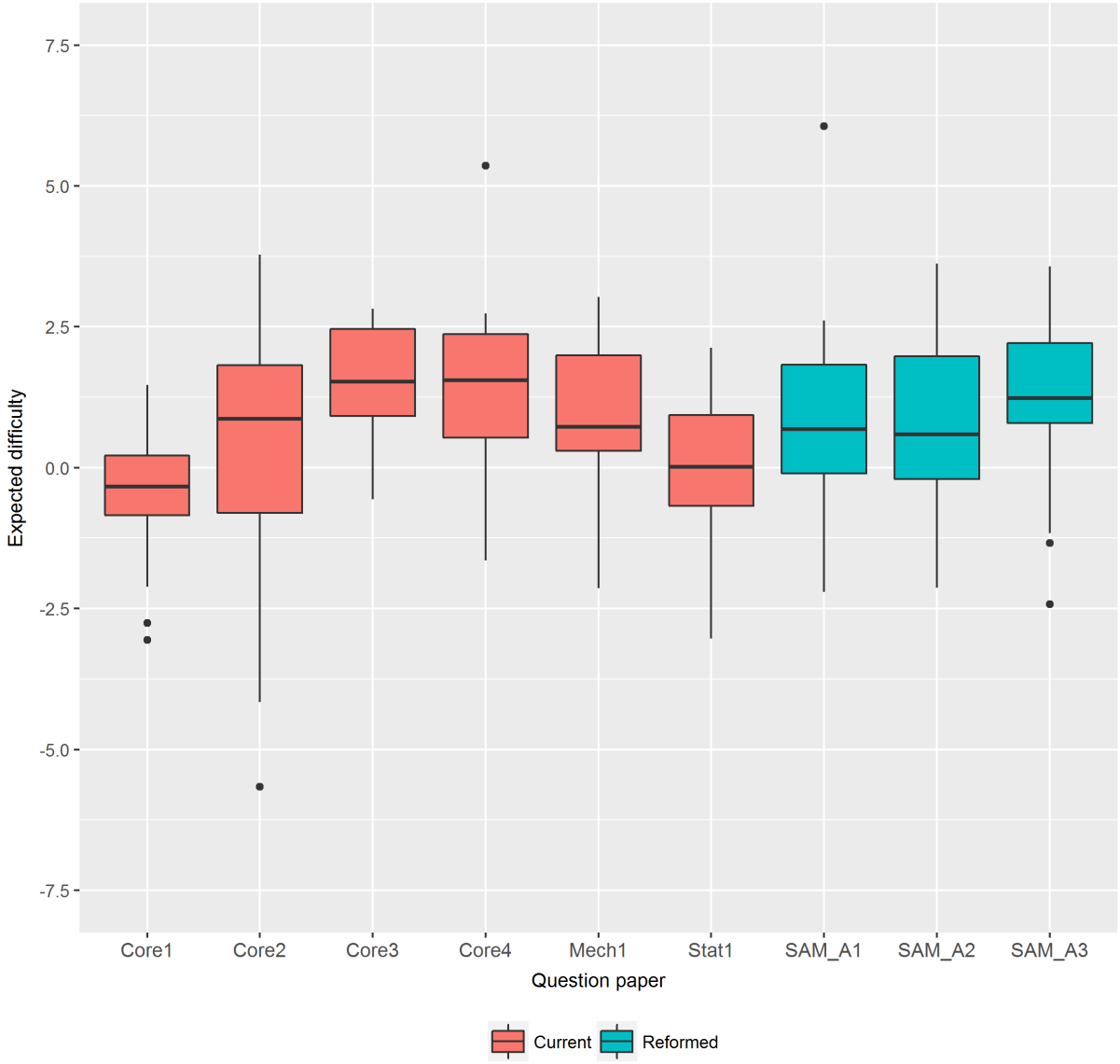

Figure 16: *A level data at paper level for specification 4: Boxplots showing the median and interquartile ranges of expected item difficulty for each paper in specification 4's 2015 assessments and the final judged reformed sample assessments.*

For the 2015 A level assessments, the median difficulty varied from 0.57 to 0.94 (overall median = 0.77), a range of 0.37 logits. For the final judged versions of the reformed A level sample assessments the median difficulty varied from 0.77 to 1.45 (overall median = 1.04), a range of 0.68 logits. This variation was largely caused by a higher median for the specification 3 assessment. Taking all aspects of the specification 3 assessment into consideration the accreditation panel considered this was of appropriate demand. Note that this range (0.68 logits) is similar to the range

of the 2015 AS assessments and will not lead to substantive differences between assessments.

Three of the 4 reformed AS assessments therefore have median difficulties which are close to that of the legacy assessments combined across all specifications. The same is true when comparing each legacy assessment and its corresponding reformed assessment. For specification 3 there is a larger difference, particularly since that specification's corresponding legacy assessment was judged to have lower difficulty than the other legacy assessments.

The individual papers in the reformed assessments are much closer in median difficulty (and overall distribution) in every case than the papers in the corresponding legacy assessment. The individual reformed papers are therefore more representative of the overall assessment difficulty than were the legacy papers.

For the reformed A level sample assessments, relative to the 2015 assessments, there appears to be a slight increase in difficulty overall and the spread of medians is larger but within an acceptable range.

6 Discussion

Overall this comparative judgement analysis shows slightly higher levels of expected difficulty for items from the sample assessments relative to the 2015 assessments (see Tables C1 and C2, Appendix C). The reform of AS and A level mathematics did not require an increase in the difficulty of the items on the assessments. The increase seen here is small, and effectively meets this requirement, considering that there is always some degree of variability in assessment difficulties. The range of assessment median difficulties for the 2015 AS and A level assessments and the reformed AS and A level specifications is relatively small, indicating there will not be substantial differences in overall difficulty between assessments. Such small differences can easily be accommodated by the setting of grade boundaries at awarding. The choice of specifications to teach should be based more on content and style as there is little appreciable difference in difficulty.

Finally, note that this data only covers the reformed sample assessments up to the final submission for accreditation where it was considered necessary to judge item difficulty. Some of the specifications went through additional submissions which included some very minor changes to their sample assessments. However, no significant changes to item difficulty were requested and so the final phase study for each specification will represent quite closely the final expected difficulty distributions of the accredited sample assessments.

Appendix A – Pilot study

Eighty-nine items from the 2015 core 1 papers from the 4 specifications were used in a group of 8 studies which tested various combinations of the following 3 factors:

- type of expert judge – A level / AS maths teacher or PhD mathematics students
- judging prompt – either the difficulty of achieving full marks, or an estimate of the 'overall' difficulty of the item
- inclusion of mark schemes or not

All 8 combinations of these 3 factors were trialled. Two outcomes were of concern:

1. The robustness of the statistical model fit to the judgement data

2. The correlation between estimated difficulty and item facility (actual difficulty - the average proportion of the maximum mark for the item which the student cohort achieved) in the 2015 summer series

Judge type

PhD mathematics student had proven consistent judges of mathematical difficulty in our previous GCSE maths comparative judgement studies. However, in our GCSE science work, GCSE science teachers were asked to judge the items, as they may have had slightly greater insight into what students may find difficult. We decided to directly compare the 2 types of judges.

Judging prompt

In all of our previous work we asked judges to think about the difficulty of giving a fully correct, or full mark, answer. In the GCSE maths context, this was quite representative of the (often binary) way students perform on the items – either completely wrong or fully correct. For GCSE science items this criterion provided a clearly defined standard to judge against, and facilitated planned modelling of the performance of papers, where the maximum mark difficulty could be used to estimate the intermediate mark difficulty. It was also thought that a clearly defined judgement criterion would promote more consistent judgements, and therefore a better model fit to the data.

However, full mark difficulty is not the same measure as item facility (average actual difficulty), and so the correlation between the two measures of difficulty could be compromised. We wanted to trial a judgement of 'overall difficulty' – something more comparable to item facility. One concern we also wanted to test was that this kind of judgement against a potentially less clearly defined criterion could lead to different interpretations/applications of the criterion by each judge and a less robust model fit.

Therefore we compared 'full mark difficulty' to 'overall difficulty' judging prompts, in order to compare the correlations to item facility and robustness of model fit for both judging prompts.

The instructions the judges received for the two prompts are given below.

"Which item is the more difficult to answer fully?"

This refers to the difficulty of giving a complete answer that would achieve full marks. Where multiple solution paths are possible, only the easiest one would be required. This means that you may find yourself comparing a 1-mark question against an 8-mark question. You must think about the difficulty a student would experience in getting 1/1 for the first question, and 8/8 for the second question, and decide which case is harder.

"Which item is more difficult overall?"

This refers to the average difficulty for students. So thinking about students across the whole ability range, for which question do you think that on average students will achieve the lower proportion of the total marks available. You can think about how a whole range of students might perform on the two questions. Alternatively, you might want to consider a single 'average' student, and how that one student would perform on the two questions. Your benchmark measure for both is the proportion of full marks that would be achieved.

Example: For an 8 mark question you might expect, on average, students to earn around 3 of the marks available. The other question is worth 3 marks, and you might expect students, on average, to earn 2 marks. Therefore, the 8 mark question is more difficult – even though students might be getting more marks, they are earning a smaller proportion (0.375) of the maximum mark available compared to the other question (0.667).

Mark scheme inclusion

It had previously been thought that including a copy of the mark scheme with the question risked mental overload for the judges (they are making a large number of judgements in a relatively short time) while adding little to the accuracy of their judgements. However, in this instance where changes to the mark scheme design may have impacted on difficulty, it was desirable to include the mark schemes. We compared the judging with and without mark schemes to determine the effect of this manipulation.

PhD maths student judges were likely to be less familiar with the mark schemes than the teachers. In order to minimise the effect of any difficulty understanding the mark scheme, all participants were provided with guidance on the common abbreviations

used in the mark schemes and how to interpret them, alongside the general task instructions.

Design

We recruited 24 PhD students and 23 maths teachers (the target was 24 to allow full counterbalancing). Each judge took part in 4 studies (2 judging criterion by 2 mark scheme levels). Judges completed the studies in a counterbalanced order. To avoid confusion over which criterion to apply, the studies with the same judging criterion were completed together. Half the judges started with the full mark difficulty criterion, half with the overall difficulty criterion. For each criterion pair, half the judges started on the study with mark schemes, and half on the study without.

Each judge was assigned 45 judgements per study. Due to some judges not completing their allocation, and some items needing to be removed due to errors on the item, there were between 835 and 930 judgements per study, equalling around 20 judgements on average for each of the 89 items (each judgement includes 2 items).

Results

No judges were excluded from the studies, as the use of a repeated measures design minimises the effect of any poorly-fitting judges. The same model fitting took place as described in Section 4 the main body of the report, giving expected difficulty values for all items. Facility values for each item were obtained from the exam boards and were correlated with the expected item difficulties.

Table A1 below summarises the various measures obtained across the 8 studies (see Section 4.1 of the main report for a description of scale separation reliability (SSR) and split-half reliability). There were only minor differences in the reliability of the statistical model fitted to the data. The SSRs, indicating the robustness of the statistical model fit, varied very little, from 0.76 to 0.83 (the number of judgements per item was relatively low, and SSR tends to increase with more judgements which is why these are a little on the low side). The split-half reliability was a little more variable, and was slightly higher when the mark scheme was included in the judging and when the criterion was the full mark difficulty.

The length of time taken to complete each judgement, as measured by the median judging time (the mean would be affected by the occasional extreme time when, e.g. the computer was left unattended) was longer when the mark scheme was included than when it was not. A median time of around 23 seconds with no mark scheme increased to 31 seconds when the mark scheme was present, indicating that consideration of the mark scheme did take up additional time.

Table A1: *Statistical model fit to paired judgement data, and correlation of model parameters to live item facility for each of the 8 pilot studies.*

	No mark scheme		Mark scheme	
	Full mark difficulty	Overall difficulty	Full mark difficulty	Overall difficulty
Maths Teachers				
SSR	0.81	0.81	0.85	0.79
Split-half reliability	0.65	0.58	0.74	0.60
Median judging time	22 s	21 s	31 s	31 s
Correlation with facility	0.47	0.53	0.60	0.58
PhD Students				
SSR	0.76	0.82	0.83	0.78
Split-half reliability	0.53	0.70	0.71	0.58
Median judging time	24 s	24 s	31 s	30 s
Correlation with facility	0.36	0.38	0.37	0.45

The correlations between the ranked expected difficulty values from the comparative judgement exercise and the ranked item facilities are shown in Figure A1.

For all 4 of the mark scheme/criterion combinations, teacher judgements were more highly correlated to the item facilities than PhD student judgements. In general, inclusion of the mark schemes led to higher correlations, particularly for the teacher judges. Although the very highest correlation was seen for teachers judging the maximum mark difficulty with the mark scheme present, on balance, the overall difficulty criterion produced higher correlation than the full mark difficulty. For the teacher judgements, the fractionally higher correlation for full mark difficulty with the mark scheme (0.60) than the overall difficulty with the mark scheme (0.58) is not a substantive difference.

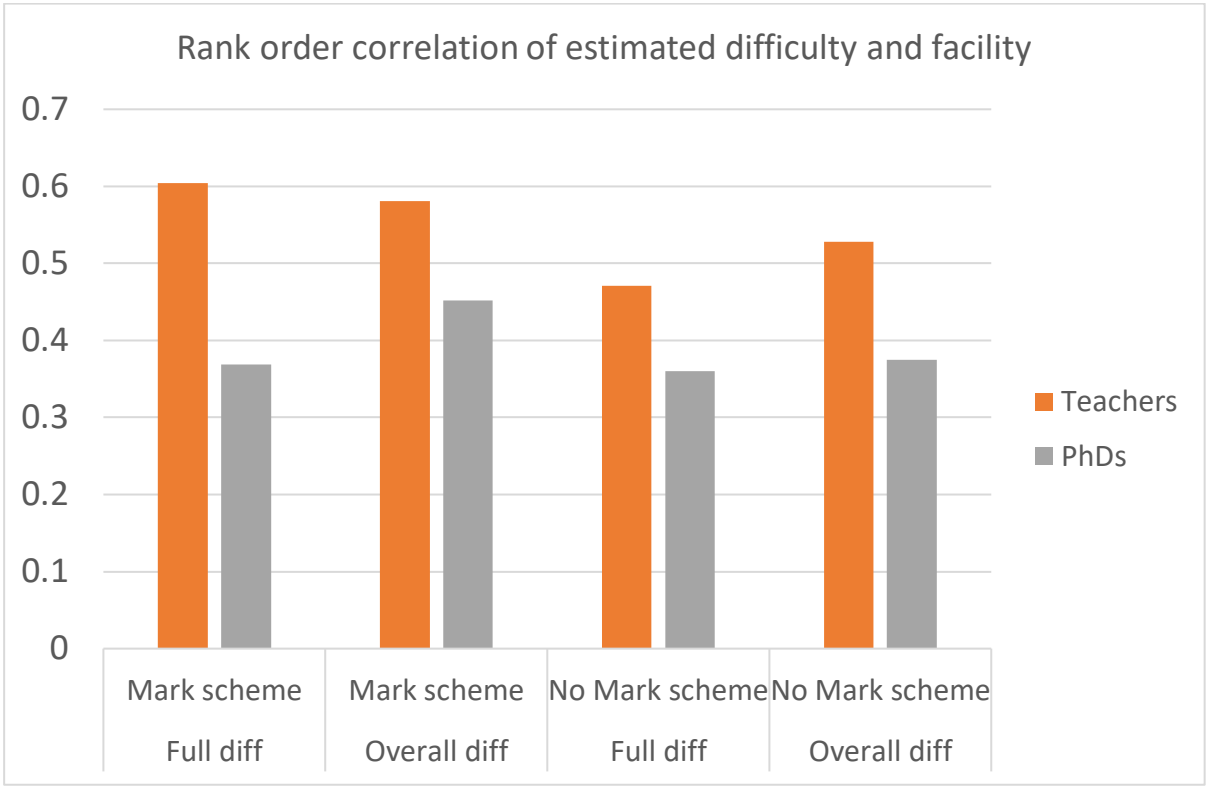

Figure A1: *Correlation of model parameters to live item facility for each of the 8 pilot studies*

In a post-study debrief, most judges said that overall the mark schemes were moderately useful, and that they tended to ignore the mark scheme if the questions were of appreciably different difficulty, but use the mark schemes in those judgements where the questions were not obviously different in difficulty.

Following these results, the following design decisions were made for the main study:

- current A level / AS maths teachers would judge the items;
- mark schemes would be included for all items;
- item difficulty would be judged against the idea of overall difficulty.

Appendix B – Adjustment of AS item expected difficulty within the legacy A level

AS items are targeted at year 12 students. They were judged by teachers in the context of A level questions targeted at year 13 students and, as a result, the AS items were judged as slightly easier than they would be for the cohort they were primarily intended for. It was necessary to introduce a correction that would slightly increase the difficulty for the AS items within the A level (core 1, core 2, mechanics 1, and statistics 1).

Initially, the facilities for year 12 and year 13 candidates for all items on the legacy AQA, MEI, OCR and Pearson AS papers were calculated. The ability (mean GCSE score) for each item was higher for year 12 candidates; the ability between the 2 year groups were matched by the removal of the lowest ability candidates from year 13. Once ability was matched, the facility was typically higher for year 13 candidates: a result of the extra year of teaching.

The difference in the overall facility for each item between year 12 and year 13 is shown in figure B1, where a positive difference corresponds to a higher facility for year 13 candidates and the reverse is true for a negative difference. The difference was modelled by fitting a curve fixed at zero difference at a facility score of one. For each item the year 13 facility can be adjusted based on the curve; bringing the facility of the two cohorts into agreement with one another (figure B2). No correction is applied to items with a year 12 facility of less than 0.25 due to the small number of items.

The adjusted facility is converted to an adjusted expected difference by multiplying the adjusted facility with the gradient obtained from the regression line of expected difficulty on facility (figure B3). On average, the expected difficulty for AS items within the A level is adjusted upwards by approximately 0.2.

The Spearman rank order correlation for the data in Figure B3 was 0.49. This is lower than that obtained in the pilot studies, probably because the items, including both mechanics and statistics items, were more diverse in topic and question size than those found on the core 1 pure maths papers used in the pilot studies and probably more difficult for the judges to evaluate. This suggestion is supported by a larger Spearman correlation of 0.63 between the judged difficulty and facility when the items are restricted to just the core1 and core 2 papers.

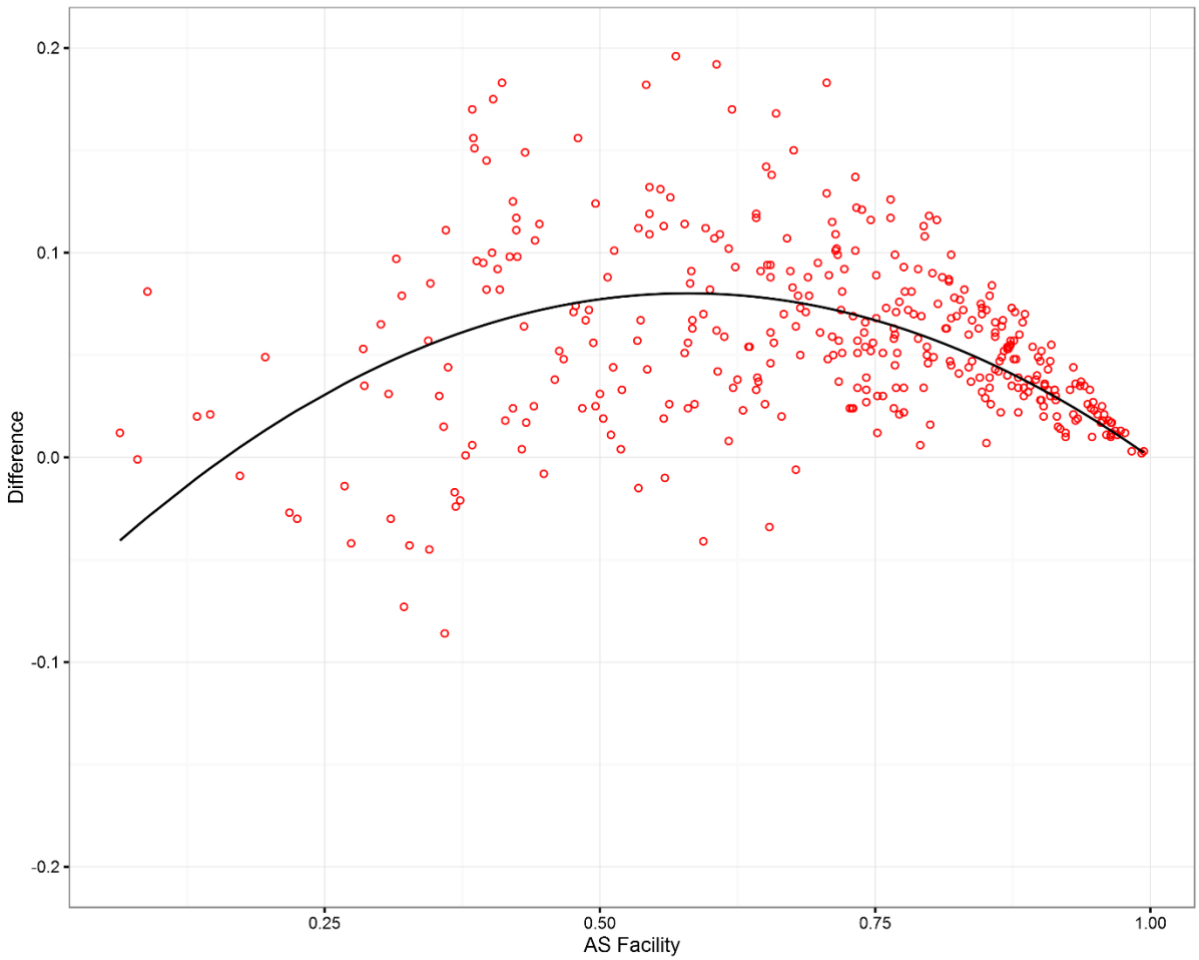

Figure B1. Difference in facility between year 12 and year 13. The modelled difference is illustrated by the black curve.

Figure B2. Adjustment of the year 13 facility.

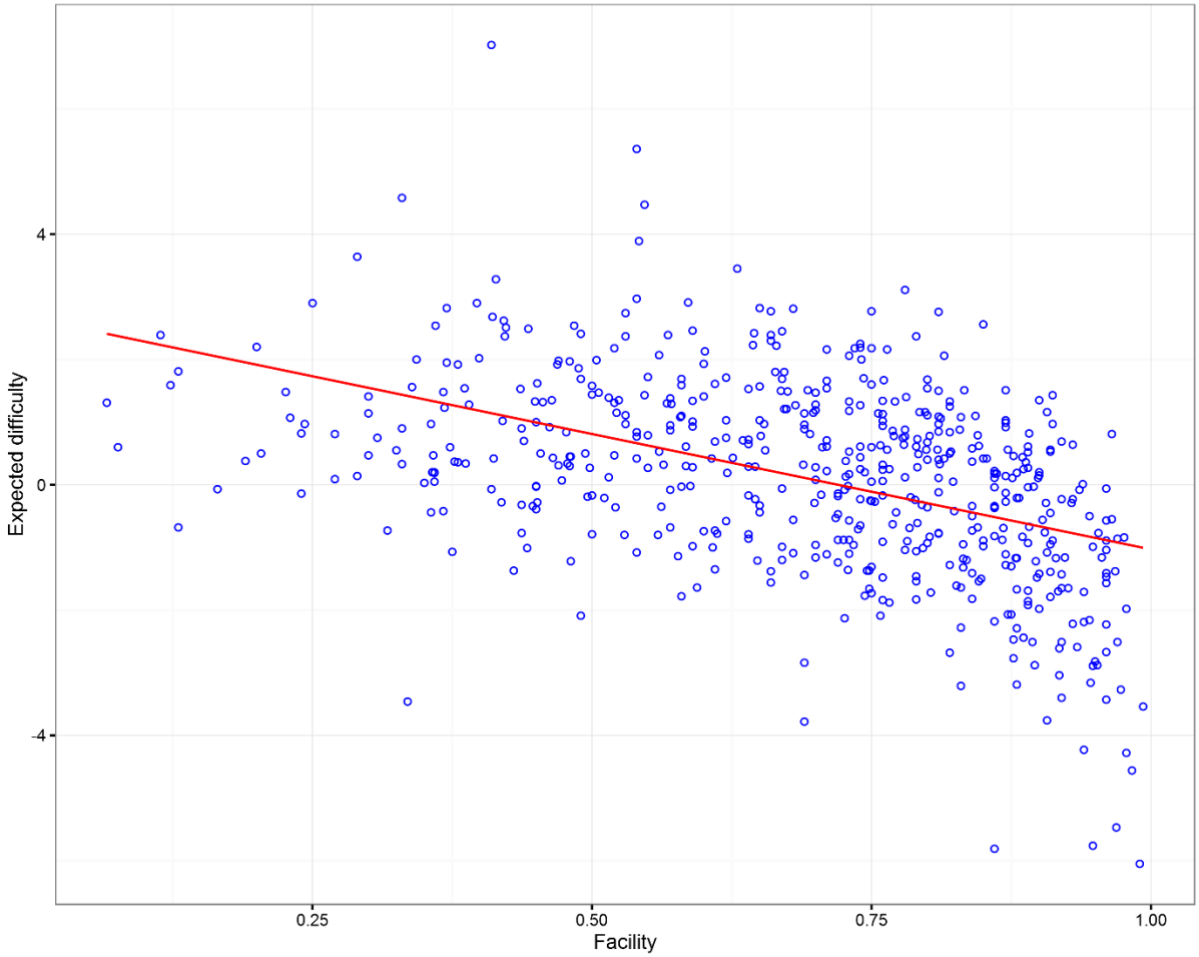

Figure B3. Expected difficulty as function of facility for all AS items

Appendix C – Additional data tables

Table C1: *Median expected item difficulty for all items from the 2015 AS assessments and the final judged version of all the reformed AS assessments.*

	2015 assessment		Final version of reformed sample assessments
	C1 + C2 + M1	C1 + C2 + S1	
AS	0.38	0.06	0.34

Table C2: *Median expected item difficulty for all items from the 2015 A level assessments and the final judged version of all the reformed A level assessments.*

	2015 assessment	Final version of reformed sample assessments
A level	0.77	1.04

Table C3: *Median of expected item difficulty for all AS specifications from the 2015 and reformed AS sample assessments.*

Specification	2015 assessments		Final version of reformed sample assessments
	C1 + C2 + M1	C1 + C2 + S1	
1	0.88	0.37	0.53
2	0.11	0.12	0.23
3	0.25	0.12	0.29
4	0.15	-0.28	0.25

Table C4: *Median of expected item difficulty for all A level specifications from the 2015 and reformed A level sample assessments.*

Specification	2015 assessments	Final version of reformed sample assessments
1	0.94	0.89
2	0.77	0.77
3	0.57	1.45
4	0.72	0.98

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