RESEARCH AND ANALYSIS

A level maths: Maintenance of Standards Investigation: Technical Report

Investigation of the appropriateness of grading standards in A level maths 2018

ofqual

Contents

1 Executive summary

The reformed versions of the A level maths qualifications were available for first teaching from September 2017. Unique amongst A levels, candidates were allowed to certificate at the end of the first year of teaching, in summer 2018. These arrangements were in place due to the way in which candidates wishing to enter maths and further maths have historically structured their learning and assessment. Typically, students studying further maths enter for maths at the end of Year 12 followed by certification in further maths in Year 13. Were candidates not allowed to certificate in maths at the end of the first year of teaching, those wishing to follow this pattern of entry would either have had to sit the out-going (legacy) version of A level maths at the end of Year 12 and the reformed version of further maths at the end of Year 13 or to have sat both together in 2019. Making certification available for candidates at the end of the first year of teaching avoided this obstacle for centres and candidates, however, it did cause complications for the first award of the reformed qualifications in summer 2018.

The summer 2018 awards of A level maths were challenging due to both of the main sources of evidence used to set grade boundaries –expert qualitative judgement and statistical predictions – being weaker than would typically be the case. Expert judgement is always weakened at a time of qualification reform due to the unknowable impact that structural, contextual and content changes (and the interaction of these factors) should have on what is deemed an appropriate level of performance of candidates. The statistical predictions were potentially less reliable than is typical due to the majority of the cohort being 17-year-old (Year 12) candidates. This meant that, rather than basing the predictions on 18-year-old (Year 13) candidates, as is typically the case for all A levels, the predictions in this first year were based on this majority 17-year-old group. Uncertainty over the composition of this cohort meant that there was less confidence in the strength of this evidence than is typical.

During the summer 2019 exam series – the first 'full' award of A level maths after two years of availability, with a predominantly 18-year-old cohort – Ofqual became concerned about the differences between the grade boundaries that exam boards had chosen compared to those that had been set a year earlier. The reason for these concerns was the systematic and, in some instances, large differences in grade boundaries between years, with those set in 2019 lower than in summer 2018. Given the assurances provided by the far stronger evidence available for the summer 2019 awards, Ofqual decided to investigate the matter to determine the appropriateness of the grade boundaries set in 2018.

The investigation was composed of three strands:

Strand 1) A statistical analysis of candidate results

Strand 2) Analysis of relative question paper difficulty

Strand 3) Analysis of candidate performance

Strand 1 considered the results data of candidates across 2017 (legacy version only), 2018 (reformed and legacy versions) and 2019 (reformed version only). The analysis showed that the approach taken to setting standards in summer 2018 was effective in maintaining the statistical standards from previous years (defined by the

mean GCSE to A level maths value-added relationship) for the 17-year-old candidates, who were used as the basis of prediction. The analysis also showed that the statistical standards were effectively maintained in summer 2019 relative to the legacy qualifications when setting the grade boundaries based on 18-year-old candidates. However, it was shown that the relationship between the attainment of 17-year-old candidates relative to 18-year-olds is different on the reformed version of the qualifications compared to the legacy version. Given that the 17-year-old candidates were the basis for the first awards in 2018, this change in relationship led to a difference in standards set in 2018 compared with 2019.

The primary quantifiable source of this change in relationship between 17 and 18-year-old candidates when transitioning to the new version of the qualifications was identified to be the removal of resitting – or, more specifically, the removal of the opportunity for candidates to resit individual assessments in the new, linear, versions of the qualifications prior to certification. Overall statistical standards for 18-year-old candidates have been maintained in A level maths across the transition to the reformed versions. The protection provided by the use of statistical predictions during awarding ensures that structural changes were accounted for (including the change in availability of resitting). The attainment of 17–year-olds has, however, increased in the reformed version as boundaries were set which compensated for the removal of the opportunity to resit from which they did not typically benefit previously. The analysis shows that, in the A level maths qualification with the largest entry – offered by Pearson – up to 58% of the difference in grade boundaries between 2018 and 2019 was due to the compensation built in to account for the removal of resitting. Other, unquantifiable sources of change are also likely to have contributed to this change in relationship between age groups such as changes to the qualification content and curriculum.

In addition to the anticipated cohort of 17-year-old candidates sitting the reformed version of the qualification in 2018, a significant number of 17-year-old candidates chose to sit the legacy version of the qualifications that year. This investigation has demonstrated broad alignment between the standards set for these two sub-groups of 17-year-olds in 2018. Despite reservations regarding the reliability of the statistical evidence in 2018, the use of predictions is likely to have played an important role in this being the case.

The second strand of the investigation considered the contribution any change in difficulty between years may have had on the difference in grade boundary position. This consisted of capturing subject experts' judgements of the relative difficulty of the exam questions across 2018 and 2019. This information was combined with analysis of the operationally available question level candidate mark data. The analysis showed that there were differences in the difficulty of the assessments between 2018 and 2019 that will have contributed to the difference in grade boundaries between years. Overall, the difficulty of the Pearson question papers increased slightly between 2018 and 2019 partly contributing to the lower grade boundaries. The difficulty was more consistent across years for the two OCR qualifications with the modelling showing a necessary lowering of the grade boundaries for the OCR A qualification appearing to be appropriate due to a slight increase in difficulty. In contrast, overall, the AQA question papers were of slightly lower difficulty in 2019 compared to the previous year suggesting a slight increase in boundaries would have been appropriate to compensate purely for differences in difficulty.

Strand 3 considered the performance of candidates across 2018 and 2019. This analysis was based on comparative judgements made by subject experts of the performances of candidates on one component of each qualification across years. This analysis confirmed that, for the components selected, there was an identifiable difference in performance standard between years at grade A for the OCR A, OCR B (MEI) and Pearson qualifications and a difference at grade E for the AQA, OCR B (MEI) and Pearson qualifications. In these cases, these differences were shown to be greater than the uncertainty in the judgement process.

In summary, the difference in grade boundaries set in A level maths in 2018 and 2019 did lead to a discontinuity in grading standards between the first two years of the reformed qualifications. However, this discontinuity was inevitable and occurred at a point in the transition that appears the most equitable across the different sub-cohorts of candidates across years. This is deemed the most equitable due to the comparability of standards being achieved for the two groups of 17-year-old candidates in 2018: those sitting either the legacy or reformed versions of the qualification.

The cause of the discontinuity of standards was the change in relative relationship between the performance of 17-year-old and 18-year-old candidates, combined with the necessary use of 17-year-old candidates as the basis for the 2018 awards. The effect was not caused by the 2018 cohort of 17-year-olds sitting the reformed qualification being statistically atypical in comparison with other years. On this basis, it is reasonable to expect a broadly similar relationship to continue in the qualifications with the discontinuity being confined to the examination series scrutinised through this investigation.

The most significant contributor to the change in relationship between 17-year-old and 18-year-old candidates was the removal of the opportunity for candidates to resit assessments in the reformed version of the qualification. This effect, combined with a tendency for the assessments to be more difficult in 2019 compared with the previous year, accounts for the majority of the difference in boundary position between years. These two effects in isolation provide a satisfactory explanation of the differences in boundaries at grade A. At grade E, in addition to the effects of resitting and differences in difficulty, exam boards faced the challenge in 2018 of setting this grade with a very small number of candidates performing at that level. This additional uncertainty contributed to the difference in boundary marks at this grade.

2 Background

2.1 Specifications and availability

AS and A level qualifications in England have been through a phased period of reform in recent years. The first reformed AS and A levels were available to be taught from September 2015, with the first AS examinations in the summer of 2016 and the first A levels examinations in the summer of 2017. Two further phases followed the same pattern with the majority of the remaining subjects being introduced for teaching in 2016 and 2017[1].

The key changes to qualifications were:

1) assessment would be mainly by exam, with other types of assessment used only where essential to test assess the content in a valid way
2) the content for the new A levels was reviewed and updated with universities playing a greater role than was previously the case
3) AS and A levels would be assessed linearly at the end of the course. AS assessments would typically take place after one year's study and A levels after two. The courses would no longer be unitised to allow modular sitting
4) AS and A levels would be decoupled, meaning that AS results would no longer contribute towards the A level grade. AS qualifications could, however, be offered by exam boards and were typically designed with the intention of enabling teaching alongside the first year of the A level course

Reformed qualifications in AS and A level maths were available for teaching from September 2017. Uniquely among A levels, maths was available for examination after only one year of teaching. In the *Consultation on Conditions and Guidance for AS and A level Mathematics and AS and A level Further Mathematics*[2], Ofqual proposed that the first examination for the new A level maths should take place in summer 2018, at the end of the first year. This was to allow (but not require) students beginning their studies in 2017, and who were intending to take maths followed by further maths, to take their examinations in maths in one year and then in further maths a year later. This would be in-keeping with both existing practice and with opportunities that would be afforded to candidates in subsequent years.

In 2017 there were 1,952 certifications in A level maths by 17-year-olds and, in 2018, 966 of these candidates went on to certificate in A level further maths.

Over 80% (30 out of 36) of respondents to the consultation either agreed or strongly agreed with allowing this approach on the grounds of fairness to this first cohort of candidates. Two respondents were neural – one of which noted the technical challenges associated with awarding – with, four disagreeing or strongly disagreeing

[1] Some ancient languages and less commonly taught modern foreign languages were introduced in September 2018

[2] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/513672/as-and-A level-mathematics-and-further-mathematics-analysis-of-responses.pdf

on the basis of potential confusion for schools and a perceived lack of preparation time for teachers.

Three exam boards – AQA, OCR, and Pearson – are currently recognised to offer the reformed AS and A level maths qualifications in England. Across these boards, four accredited specifications are available, with AQA[3] and Pearson[4] each offering one and OCR offering two[5,6], including one developed in collaboration with Mathematics in Education and Industry (MEI).

2.2 Qualification structures

In most AS and A level subjects, the structural changes introduced by reform entailed moving from a four unit structure – AS qualifications were composed of two units with the A level composed of four (the two AS units plus two A2 units) – to typically, an AS qualification composed of two assessments and the A level three papers independent from the AS assessments. Maths, however, underwent a more radical change as previously communicated by Ofqual[7]. In the legacy qualifications, the AS level qualifications comprised three units and the A level, six, with complex rules for combination.

Table 1 illustrates the unitised structure of legacy AS and A level qualifications in maths outlining the combination of units permitted to contribute to a candidate's final grade. This is illustrated using the Pearson qualification as an example[8]. Not only were the AS level qualifications integral to the A level, but certain optional units – the Application units, Mechanics, Statistics, and Decision Mathematics – could be used towards different titles.

The reformed A level maths qualifications, which are the subject of this investigation, are far simpler in structure with no optionality at component level or within the question papers. These structures are summarised in Table 2.

2.3 Aggregation and optimisation

Aggregation of marks across assessments in AS and A level maths was simplified significantly through reform. In all the qualifications outlined in Table 2, the raw marks achieved by candidates in each component are summed, without scaling, to form the overall mark. This contrasts with the legacy qualifications, which required the use of the Uniform Mark Scale[9] (UMS) to account for differences in assessment difficulty across the different series in which candidates could accrue their marks.

As highlighted above, a key feature of the reforms was the decoupling of the AS qualification from the A level and the (re-)introduction of linear assessment in place of the unitised structure of the legacy qualifications. Previously, the unitised structure allowed candidates to resit units as often as desired prior to certification. When the legacy versions of the qualifications were originally introduced for 2008, candidates

[3] https://www.aqa.org.uk/subjects/mathematics/as-and-a-level/mathematics-7357
[4] https://qualifications.pearson.com/en/qualifications/edexcel-a-levels/mathematics-2017.html
[5] https://www.ocr.org.uk/qualifications/as-and-a-level/mathematics-a-h230-h240-from-2017/
[6] https://www.ocr.org.uk/qualifications/as-and-a-level/mathematics-b-mei-h630-h640-from-2017/
[7] https://ofqual.blog.gov.uk/2019/02/08/new-a-level-maths-in-2019/
[8] https://qualifications.pearson.com/en/qualifications/edexcel-a-levels/mathematics-2008.html
[9] https://store.aqa.org.uk/admin/results-days/AQA-UMS-GUIDE.PDF

could sit (and resit) units in either the January or June series. These opportunities were, however, reduced from January 2013 onwards due to the removal of the January series, with all A level exam assessment taking place in the summer series. However, in maths, candidates frequently resat AS units when certificating at A level; the effects of which are explored later in this report.

Table 1. The unitised structure of Pearson's legacy AS and A level qualifications in maths.

AS Mathematics		
Core Mathematics 1	Core Mathematics 2	Application unit M1, S1 or D1

AS Further Mathematics		
Further Pure Mathematics 1	Application or FP unit	Application or FP unit

AS Pure Mathematics		
Core Mathematics 1	Core Mathematics 2	Core Mathematics 3

AS Further Mathematics (Additional)		
Application or FP unit	Application or FP unit	Application unit

A Level Mathematics		
Core Mathematics 1	Core Mathematics 2	Application unit M1, S1 or D1
Core Mathematics 3	Core Mathematics 4	Application unit M1, S1 or D1 or M2, S2 or D2

A Level Further Mathematics		
Further Pure Mathematics 1	Application or FP unit	Application unit
Further Pure Mathematics 2 or 3	Application unit	Application unit

A Level Pure Mathematics		
Core Mathematics 1	Core Mathematics 2	Core Mathematics 3
Core Mathematics 4	Further Pure Mathematics 1	Further Pure Mathematics 2 or 3

A Level Further Mathematics (Additional)		
Application or FP unit	Application unit	Application unit
Application unit	Application unit	Application unit

Table 2. Structure of the reformed A level maths assessment frameworks.

	Component	Content	Marks (Weight)	Time (mins)
AQA (7357)	Paper 1	Pure Maths	100 (33.3%)	120
	Paper 2	Pure Maths & Mechanics	100 (33.3%)	120
	Paper 3	Pure Maths & Statistics	100 (33.3%)	120
OCR A (H240)	H240/01	Pure Maths	100 (33.3%)	120
	H240/02	Pure Maths & Statistics	100 (33.3%)	120
	H240/03	Pure Maths & Mechanics	100 (33.3%)	120
OCR B (MEI) (H640)	H640/01	Pure Maths & Mechanics	100 (36.4%)	120
	H640/02	Pure Maths & Statistics	100 (36.4%)	120
	H640/03	Pure Maths & Comprehension	75 (27.3%)	120
Pearson (9MA0)	9MA0/01	Pure Maths	100 (33.3%)	120
	9MA0/02	Pure Maths	100 (33.3%)	120
	9MA0/03	Statistics & Mechanics	100 (33.3%)	120

Given the interaction between units that could potentially contribute to the different AS and A level qualifications outlined in Table 1, rules were previously in place to optimise candidates' grades and ensure a consistency of approach across and within maths qualifications.

The Joint Council for Qualifications (JCQ) *GCE Mathematics Aggregation Rules*[10] – no longer applicable to reformed qualifications in England – were designed to ensure that "candidates receive the best possible set of unit grades and, where candidates have taken extra units, the best units are not left unused." The rules that underpinned implementation of this approach are provided in Annex A. In these rules for the legacy qualifications, the JCQ document advises candidates re-sitting one or more units "to re-enter for all relevant qualifications to make sure that all units are unlocked and can be re-combined in the best possible way". The significance of these rules for the 2018 and 2019 awards is that, previously, many 18-year-olds taking further maths and certificating – or indeed re-certificating – in maths would

[10] https://www.jcq.org.uk/Download/exams-office/entries/gce-maths-information/gce-maths-rules---guidance-for-centres

have had their best unit results counted toward A level maths. In the reformed qualifications, the two titles are independent of one another.

2.4 Approach to maintenance of grading standards

The transition from one version of a qualification to another, which occurs at the time of reform, poses challenges to the processes of setting and maintaining grading standards. Maintaining performance standards (the level of performance required from candidates to achieve equivalent grades between years) across the transition may seem the intuitively correct approach. However, it is frequently inappropriate for three reasons.

First, structural changes to the qualification may mean that candidates of the same ability are more or less able to demonstrate equivalent levels of performance either side of the transition. This may be due to changes to the aggregation of candidates' marks across assessments, the weighting of assessments and/or the assessment opportunities afforded to candidates. Not allowing for this may lead to candidates being unfairly advantaged or disadvantaged either side the transition if levels of performance were matched across the transition as it may be appropriate for the expectations to be modified.

Second, demands of the content, the mode of assessment and the interaction between the two may also impact on the appropriateness of expectations of performance either side of the transition. More or less demanding subject content, sampled and assessed in different ways will likely impact both on candidates' performance and the perceptions of those performances. Maintaining standards on the basis of performance alone would fail to account for these differences.

Third, unfamiliarity with the format of the assessment, the content or lesser availability of supporting materials (such as past papers and revision resources) by students and teachers may lead to poorer exam performance, but does not necessarily reflect lower ability in the subject itself. These effects can lead to a dip in the overall level of performance of candidates in the assessment at the point of transition to a new qualification. This 'saw tooth effect' has recently been explored in the context of GCSE and AS/A level assessments[11] suggesting it may take approximately three years for students and teachers to become familiar with the nature and requirements of new assessments.

Given these challenges to making judgments of candidates' performances around the time of qualification change, to ensure the fair treatment of candidates, exam boards should not necessarily be looking to match candidate performance from the final year of a legacy qualification with the first year of the new version. To mitigate these issues, the standard setting process in new qualifications is guided by statistical predictions. These predictions model the value-added relationship between national assessments at different stages of candidates' schooling – in the case of A levels, the relationship between candidates' GCSE attainment and the A level subject in question. These relationships are then carried forward to ensure that a cohort of a given ability profile would achieve the same grade distribution in the legacy and reformed A level qualification. The rationale for this approach is that

[11] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/549686/an-investigation-into-the-sawtooth-effect-in-gcse-as-and-A level-assessments.pdf

these predictions control for the structural, contextual and content changes that hamper maintenance of standards of the basis of performance, as outlined above, and that candidates have no control over the qualification changes, so should not be disadvantaged (or advantaged) by the happenstance of their year of birth.

The procedures used in awarding do, however, employ both statistical predictions and the judgements of the senior examining team about the performance of candidates on and around the grade boundary marks indicated by the statistics. Due to the limitations highlighted above, there is greater emphasis on the statistical evidence during the early awards of new qualifications with senior examiners using their experience and expertise to verify whether or not the performance standards indicated by the statistics represent reasonable expectations. This approach has been widely communicated by Ofqual throughout the reform process[12].

When the new qualifications have become more established, the sources of evidence remain unchanged, however, the motivation for the use of statistics changes and the question put to senior examiners recommending grade boundaries is refreshed to support a steady state maintenance of an overall level of performance.

2.4.1 Awarding A level maths in 2018

As highlighted above, statistical predictions are a critical source of evidence for awarding, particularly when transitioning from one version of a qualification to another. For the first award of the new A level maths qualifications in 2018 the statistical evidence available was atypical. Typically, the statistical predictions used to inform the award of A level qualifications are based on the value-added relationship demonstrated by previous 18-year-old candidates[13]. However, as described in Section 2.1, the circumstances surrounding maths differ from other subjects meaning the candidates seeking certification at the end of the first year of teaching in 2018 were predominantly 17-year-olds.

This group of 17-year-old candidates were, therefore, used as the basis for the statistical predictions, modelling the anticipated value-added relationship on the attainment demonstrated by 17-year-old candidates who certificated in A level maths across 2014-2017[14]. These predictions were suspected to be less reliable than would typically be the case due to the size of the 17-year-old entry and uncertainty over the impact the changes implemented through reform may have had on the nature of the entry and, therefore, the validity of the predictive model[15]. An additional source of uncertainty in the basis of these predictions was the lack of direct statistical control over this group across the years on which the model is based. The reason for this is that statistical predictions typically focus on the majority cohort; as highlighted

[12] For example: https://ofqual.blog.gov.uk/2018/01/19/gcse-and-a-level-awarding-in-2018/

[13] All ages referred to throughout this report are defined by candidates' age at 31 August in the year in question, i.e. typical year 13 students are referred to as *18-year-olds*, throughout.

[14] The motivation for selecting this range of years was to average any variations in statistical standard for 17-year-olds in the period following removal of the January exam series.

[15] Consideration of these factors was supported by survey information from centres in the lead up to the first award: https://ofqual.blog.gov.uk/2018/03/16/setting-standards-in-the-new-a-level-maths-qualifications/

above, this is usually 18-year-olds for A level qualifications. This means that the outcomes for other groups of candidates, outside of the basis of prediction, are not routinely used to inform the awarding process. Generally, in steady state conditions, these intra-age group relationships and, more generally, relationships between sub-cohorts, remain unchanged over time. The consequence of this is that, ensuring the standard of the majority cohort is maintained, ensures by default appropriate standards for all other candidates. Even if the performance of these sub-cohorts are very different to one another, if they are stable relative to one another, this ensures constancy.

Despite the uncertainty over the reliability of the predictions, the statistics provided what was believed to be the strongest evidence available given the limitations on examiner judgement as outlined above. Examiners were, nonetheless required to evaluate the reasonableness of the performance standard as an important stage of the process.

2.4.2 Awarding A level maths in 2019

The statistical evidence to support the award of A level maths in 2019 was believed to be far more robust than that available in summer 2018. Awards for A levels are typically guided by value-added predictions for 18-year-olds who have completed two years of study; in this case, 18-year-olds who had studied from September 2017 to June 2019. In line with first awards of other reformed A levels, the basis of predictions was the relationship between candidates' mean GCSE and A level attainment for the 18-year-old cohorts in 2010 and 2011. These years were selected as a basis as they represent the first two years of the previously reformed qualifications (the first time all exam boards used a common national predictive model) avoiding any impact from potential inflationary or deflationary effects across the intermediate years.

The only additional measure taken when formulating the predictions for maths was to account for the interaction with further maths. To isolate typical maths candidates sitting the course over two years, candidates opting to certificate simultaneously in maths and further maths were excluded from the basis of predictions. While dual certification across both subjects may seem counterintuitive given the learning pattern of students highlighted in Section 2.1, dual certification was encouraged in the JCQ guidance to ensure the maximisation of UMS scores across titles from all available units.

As discussed above, due to this series being early in the life of the qualification and being the first main award, the use of statistical evidence was particularly prominent in the standard setting process.

2.5 Basis for the investigation

During the summer 2019 series, as part of Ofqual's routine engagement with exam boards, it became apparent that there were notable differences in the grade boundaries which were being recommended for A level maths compared to those that were set in 2018. Grade boundaries generally differ from one exam series to the next, predominantly to account for variations in the difficulty of the assessments making up the qualification. The changes in boundaries observed between 2018 and 2019 in A level maths were notable. This was due to the systematic nature of the differences across exam boards – with the vast majority of boundaries being lower

than in the previous year – and, in some instances, the size of those differences. The boundary marks set at the judgmental grade boundaries[16] in 2018 and 2019 are shown in Table 3.

These grade boundary differences, combined with the relatively weak evidence available for conducting the awards in 2018 raised concerns regarding the security of that initial standard.

At Ofqual's request, following the 2019 awarding meetings but prior to the issuing of results, representatives from the exam boards carried out an additional script scrutiny across 2018 and 2019. Through this exercise, exam boards were asked to compare the relative difficulty of the assessment and the performances of candidates at equivalent grade boundaries across years before meeting with Ofqual to share their findings.

In the meeting with Ofqual, the Chair of Examiners for AQA believed that that there were differences in performance between candidates at the grade E boundaries across years – particularly on papers one and two – such that some ungraded scripts from 2018 were comparable with grade E scripts from 2019. The Chief Examiner – who was involved in setting the papers – recalled a conscious decision to ease the demand of the 2019 papers and gave examples of this, including attempting to write more accessible multiple-choice questions to ease candidates into the papers at the start and providing additional scaffolding. The AQA representatives also recalled the very scant evidence on which to base the 2018 award and that there was insufficient judgemental evidence to justify deviation from predictions (despite their known limitations).

OCR, represented by members of their Assessment Standards team, reported that their senior examiners had considered the demand of the 2018 and 2019 papers to be similar across years on all their assessments. Through the additional scrutiny process, they also identified that lower boundaries for both A level maths qualifications they offer would likely have led to a more comparable level of performance at the grade boundaries.

During the meeting with Ofqual, the Chair of Examiners for Pearson said that, through this post-award scrutiny activity, the Pearson examiners found performances at grade boundaries to be broadly comparable between the two years. Examiners were, however, split over whether the quality of candidates' performances on the pure maths paper at grade A were equivalent, with half believing the 2018 performances were stronger. They also felt the papers were of comparable demand overall, noting that, although Paper 2 (9MA02) had received feedback for being difficult in 2019, Paper 1 (9MA01) was possibly a little easier in 2019 than it had been the previous year.

In light of systematic differences in boundaries, the magnitude of the differences between years and the representations provided by the exam boards, the current investigation was commissioned to understand the likely source of differences of grade boundaries across the two years and to examine the appropriateness of the grading standards set in 2018.

[16] Judgemental grades are those at which exam boards consider statistical evidence and senior examiners scrutinise candidates' work through the process of awarding in order to set the grading standards. For AS and A levels, the judgemental grade boundaries are at grades A and E and are therefore used as the common reference points throughout this investigation.

Table 3. A level maths grade boundaries set in 2018 and 2019 including the relative differences.

		2018		2019		Difference	
		A	E	A	E	A	E
AQA (7357)	Paper 1	56	32	53	15	-3	-17
	Paper 2	65	30	62	16	-3	-14
	Paper 3	60	28	70	21	+10	-7
	Qualification level	**181**	**90**	**185**	**52**	**+4**	**-38**
OCR A (H240)	Paper 1	67	21	54	13	-13	-8
	Paper 2	61	18	58	15	-3	-3
	Paper 3	69	21	49	12	-20	-9
	Qualification level	**197**	**60**	**161**	**40**	**-36**	**-20**
OCR B (MEI) (H640)	Paper 1	74	45	70	23	-4	-22
	Paper 2	68	40	59	17	-9	-23
	Paper 3	55	30	49	12	-6	-18
	Qualification level	**197**	**115**	**178**	**52**	**-19**	**-63**
Pearson (9MA0)	Paper 1	70	26	56	15	-14	-11
	Paper 2	62	24	52	13	-10	-11
	Paper 3	52	20	57	15	+5	-5
	Qualification level	**184**	**70**	**165**	**43**	**-19**	**-27**

It is acknowledged that exam boards conducted these awards in good faith at the time and they were conducted with knowledge of the limitations of the available evidence.

From an initial consideration of the issues, it appears that a number of different scenarios may have contributed to the difference in grade boundaries between years.

Scenario A) A high proportion of the question papers in 2019 were more demanding in 2019 compared with the equivalent paper 2018 (in some instances significantly so) and the grade boundaries set in both years reflected an actual difference in difficulty.

Scenario B) 17-year-olds' performances in 2018 were under-rewarded as they should have been allowed to demonstrate a greater value-added than those in 2017, however, there was insufficient evidence for exam boards to make this judgement and allow this greater value-added to be reflected in their outcomes.

Scenario C) 18-year-olds in 2019 demonstrated less value-added than those in 2017 and were over rewarded meaning that the 2018 grade boundaries were appropriate, and 2019 grade boundaries were inappropriately lenient.

Scenario D) The relative difference between the value-added relationships for certificating 17-year-olds and certificating 18-year-olds changed across the

transition to the reformed qualifications. This would mean that both 2018 and 2019 grade boundaries appeared appropriate at the time of award based on the statistical evidence available, but a discontinuity in standards occurred for one of these sub-groups of candidates.

It should be noted that these scenarios are not mutually exclusive and may co-exist, combining to give rise to the observed outcomes and grade boundary differences.

It appears unlikely that Scenario A is the sole contributor to the difference in grade boundaries unless there was a concerted and co-ordinated effort to increase the difficulty across the majority of papers. While some differences in grade boundaries between years were relatively modest and might plausibly be due to differences in assessment difficulty, other changes were far greater suggesting an obvious and inappropriate difference in difficulty between years. Differences in difficulty may, however, have played some role in the differences in grade boundaries and are, therefore, explored through Strand 2 of this investigation.

Scenario B would require either a wholesale improvement in the preparation of 17-year-olds studying the new content and sitting unfamiliar style assessments in 2018 compared with previous cohorts on the legacy qualification or for there to be a significant difference in the nature of the 17-year-old cohort that meant the statistical predictions were no longer valid.

Scenario C would require a large-scale collapse in the preparation of 18-year-old candidates for the new examinations beyond the effects of any structural changes. This would be greater than the dip in performance standards typically observed when qualifications change and would have affected the 18-year-old candidates in a way it failed to affect the 17-year-olds a year earlier.

Were there to be no significant change in the relative preparation or nature of the 17 or 18-year-old cohorts, as required for Scenarios B and C, it may be that structural and/or content changes have impacted on the relative performances of different age groups leading to Scenario D. As described in Section 2.4.1, outcomes for 17-year-old candidates are usually set indirectly as a consequence of setting standards for 18-year-old candidates. The 17-year-old standards, therefore, remain broadly constant as long as the relative relationship with 18-year-olds remains unchanged.

Given the necessary role of statistics in supporting the awards in 2018 and 2019 and the reliance on the two different ages groups in the statistical models that acted as their basis, a change in the relationship across the transition has the potential to lead to distinct performance standards being set for these two sub-groups of candidates and, therefore, across the two years.

3 Approach

The investigation reported here was formed of three strands:

Strand 1) Statistical analysis of candidate results
Strand 2) Analysis of relative question paper difficulty
Strand 3) Analysis of candidate performance

These strands of investigation have distinct aims, as outlined in the relevant sections below, however, they are designed to be mutually complementary to support investigation of the scenarios described above.

3.1 Strand 1: Statistical analysis of candidate results

The focus of Strand 1 is the analysis of the results in A level maths achieved by candidates across the summer 2017, 2018 and 2019 exam series. This analysis draws on additional contextual data, such as candidates' prior attainment in other qualifications.

As highlighted in Section 2.5, the primary focus for the investigation is the appropriateness of the standards set on the reformed qualifications in 2018, relative to the standards set based on more reliable evidence in 2019. Consideration of the attainment of candidates in 2017 is, however, critical to support the understanding of any differences in behaviour and/or results across the transition to the reformed qualifications. Analysis of changes at this transition may be informative in explaining the cause of the differences in grade boundaries that have occurred post-reform.

This strand of work explores the extent to which the *statistical* standards have been maintained and to understand the potential source of any difference. There are obvious limitations in considering this statistical strand in isolation. For example, these analyses alone do not tell us anything direct about the quality of candidates' responses across years and how the difficulty of the assessments may have impacted on those performances. While statistical approaches to the maintenance of standards have been shown to be operationally highly effective, for the purposes of an investigation such as this, a more direct evaluation of these factors is desirable, as described below.

3.2 Strand 2: Analysis of relative question paper difficulty

The primary reason for grade boundaries on different versions of an assessment to be different between exam series is to account for variations in their difficulty. Lower grade boundaries are set on more difficult versions of an exam and higher boundaries set on less difficult assessments to ensure fairness for the groups of candidates sitting each version. Given the genesis of this investigation, and the need to explore the potential contribution of Scenario A to the observed effects, it is important to understand the extent to which differences in assessment difficulty might have impacted on the position of grade boundaries between years.

To enable the relative difficulty to be evaluated, two different approaches have been taken:

1) judgements of expected difficulty – a judgement of how difficult subject experts anticipate individual exam questions to be, therefore, providing a measure of expected assessment difficulty independent of the ability of the candidates who sit the assessments.
2) evaluation of actual difficulty – analysis of operationally available candidate data to identify the relative difficulty of assessments in practice.

3.3 Strand 3: Analysis of candidate performance

When assessing comparability, it is also important to consider the performance of candidates across the years in question. Evaluation of the results data in Strand 1 may indicate a statistical difference in standards, and Strand 2 may be able to identify the contribution to the change in boundaries due to differences in difficulty, however, without the evaluation of candidates' work, it is not possible to know whether any of these effects lead to a qualitatively meaningful difference in performance. This strand of work seeks to address this point by drawing comparisons between the quality of work produced by candidates at equivalent grades across 2018 and 2019. This will allow identification of the degree of similarity or difference in performance standards across years and, based on an evaluation of the uncertainty in the judgements, identify whether any differences can be reliably identified.

As discussed in Section 2.4, the evaluation of performance standards early on in the life of a qualification can be problematic due to the 'saw tooth' effect – the potential for candidate performance to dip in the early years of availability due to a reduction in the availability of practice and support materials and the experience of teachers with the new qualifications. It should be noted, however, that while this effect may be present within the reformed A level maths qualifications and may, therefore, impact on the continuity of performance standards across years, the difference in grade boundaries between 2018 and 2019 cannot be explained by this effect. Compensation for the saw tooth effect would typically require grade boundaries to rise over the early years of a qualification rather than lower as is the case here.

4 Strand 1: Statistical analyses of candidate results

As described in Section 2.4, statistical analyses provide a key source of evidence when setting grade boundaries, operationally. However, building an understanding of the statistical relationships between cohorts (and sub-cohorts) based on the grades actually achieved can also be informative for post-hoc exploration of standards issues, such as those of interest here. This strand of the investigation considers the proximity of the outcomes from awarding to the statistical predictions, analyses the relationships between different cohorts/sub-cohorts sitting A level maths comparing these relationships with expectations, and explores the data with a view to understanding any matters of note.

4.1 Data preparation

The data used in these analyses are largely the candidate level data supplied by the awarding organisations following each summer series. Exam boards routinely provide candidate result data at qualification and unit/component level to facilitate on-going monitoring. Exam boards were notified of its use in this investigative work. These data are provided by exam boards immediately following the summer series and, therefore, do not reflect any mark/grade changes that take place through any post-results review of marking or appeals process. Given that these data are collected at the same point each year, this has minimal impact on the legitimacy of the analyses. In addition to these data, exam boards provided candidates' GCSE results from the corresponding year for the A level cohorts to enable matching with their prior attainment. These prior attainment data are used as the basis for the majority of operational awarding activity by exam boards and underlie much of the analysis presented here. The candidate match rate for these data sets was 86% across the 17 and 18-year-old A level maths candidates across 2017, 2018, and 2019.

Only candidates sitting the four main legacy and four main reformed qualifications were included; for example, the small number of candidates sitting the legacy pure maths titles were removed from the data. Candidates entering maths and further maths in the same series were also excluded from the value-added analyses for consistency with the data used to guide the awards (see Section 2.4.2) and to better enable like with like comparisons across series.

4.2 Candidate entry behaviour

Table 4 shows the entry figures for candidates certificating in A level maths in 2017, 2018, and 2019. These demonstrate a relatively consistent level of entry across the three years with the largest proportional change being the increase in 17-year-old candidates in 2018 compared to 2017.

The first point to note is the significant proportion of 17-year-old candidates opting to certificate on the legacy qualification in 2018 (42.5% of the 17-year-old entry). While this is a legitimate and permitted choice for candidates, it is counter to expectation. Candidates may be taking this route for a range of reasons: they may have advanced an academic year at some stage in their schooling (and are therefore in

year 13), they may have decided to sit the legacy qualification at the age of 17 and go on to do further maths at 18 on the reformed qualification[17] or they may have decided to sit maths early at the age of 17 without the intention of going on to do further mathematical study.

While this is a legitimate choice for candidates, this splitting of the 17-year-old cohort does raise questions about the reliability of the prediction used to guide the 2018 award. Were the decision of 17-year-olds to sit the legacy rather than reformed qualification to be non-random (in terms of their representativeness of the overall value-added relationship), this may compromise the statistical models used to guide the award. This potential issue is considered below in Section 4.5.1.

Table 4 Certificating candidate entries in A level maths from 2017, 2018 and 2019

	2017			2018			2019		
	17 yo	**18 yo**	**All**	**17 yo**	**18 yo**	**All**	**17 yo**	**18 yo**	**All**
AQA (6361)	348	13,724	16,546	187	13,891	16,151			
AQA (7357)				141	27	256	197	10,034	11,270
OCR (7890)	302	10,353	12,104	91	10,696	12,069			
OCR A (H240)				100	<10	115	233	6,464	7,170
OCR (MEI) (7895)	153	8,687	10,182	62	9,258	10,447			
OCR B (MEI) (H640)				33	<10	36	27	5,907	6,260
Pearson (9371)	1,163	38,209	47,899	852	41,918	51,227			
Pearson (9MA0)				1,336	207	1,691	1,813	50,469	57,693
Total	**1,966**	**70,973**	**86,731**	**2,802**	**76,005**	**91,992**	**2,270**	**72,874**	**82,393**

4.3 Awarding outcomes

When using statistical predictions to support awarding, Ofqual has different expectations dependent on the number of candidates of the appropriate age that can be matched to their prior attainment. Candidates that can be matched to their prior attainment are those that are used to form the statistical predictions, and when this number gets too small, these predictions provide an increasingly unreliable guide as

[17] This route was taken by 143 of the 1,192 candidates across the four qualifications

to the appropriate standard. As reflected in the published Ofqual data exchange procedures, exam boards are not expected to raise as an exception any level of deviation from prediction for qualifications with matched candidate entries below 500. In the current context, however, the statistical predictions provide a useful start-point for script scrutiny. While it may be appropriate to deviate (significantly in some cases) from this initial position on the basis of quality of work, this common start point provides some protection against an inconsistency of approach across exam boards.

Table 5 and Table 6 show the awarding outcomes relative to prediction from 2018 and 2019.[18]

It is evident from the closeness of these matched candidates outcomes to the statistical predictions for AQA and Pearson in 2018 and for all qualifications in 2019 that the grade boundaries suggested by the statistics were, in these instances, deemed reasonable performance standards by senior examiners. On this basis, it appears reasonable to dismiss Scenario C, introduced in Section 2.5, as a credible rationale for the observed differences due to acceptability of the suggested statistical standards in 2019.

Where the OCR A qualification deviated from prediction in 2018 it was in the direction of leniency, therefore, recommending boundaries lower than those suggested by the statistical evidence. For OCR B (MEI), the number of matched candidates was insufficient to provide a meaningful guide on which awarders could have confidence.

Table 5 Matched candidate outcomes relative to prediction at time of award from summer 2018

Specification	Description	Outcome (cum %)		17 yo Matched candidates
		A	E	
AQA (7357)	Predicted outcome	59.1	97.8	98
	Matched outcome	59.2	98.0	
	Difference	+0.1	+0.2	
OCR A (H240)	Predicted outcome	67.4	98.9	95
	Matched outcome	74.5	100.0	
	Difference	+7.1	+1.1	
OCR B (MEI) (H640)	Predicted outcome	72.7	100.0	33
	Matched outcome	54.6	93.9	
	Difference	-18.2	-6.1	
Pearson (9MA0)	Predicted outcome	63.4	98.0	1,008
	Matched outcome	63.8	98.0	
	Difference	+0.4	0.0	

[18] For reference, the 17-year-old, 18-year-old and all candidate outcomes across 2017, 2018 and 2019 are provided in Annex B.

Table 6. Matched candidate outcomes relative to prediction at time of award from summer 2019

Specification	Description	Outcome (cum %)		18 yo Matched candidates
		A	E	
AQA (7357)	Predicted outcome	28.3	96.7	6,086
	Matched outcome	28.1	96.7	
	Difference	-0.2	0.0	
OCR A (H240)	Predicted outcome	36.9	97.5	4,396
	Matched outcome	36.7	97.5	
	Difference	-0.2	0.0	
OCR B (MEI) (H640)	Predicted outcome	35.7	97.5	4,246
	Matched outcome	35.4	97.6	
	Difference	-0.3	0.1	
Pearson (9MA0)	Predicted outcome	32.7	97.1	37,037
	Matched outcome	32.8	97.0	
	Difference	+0.1	-0.1	

4.4 Effectiveness of maintenance of statistical standards

The process of forming statistical predictions is based on establishing the value-added relationship between prior attainment (which in the case of A levels is defined by candidate' mean GCSE grades) and candidates' results in the subject of interest from a previous year of choice. This value-added relationship is then carried forward to the current cohort. Operationally, this is achieved through the formation of prediction matrices,[19] which are typically 'national' prediction matrices combining data from across exam boards to promote inter-exam board comparability.

Here, to aid the evaluation and visualisation of the achieved value-added relationships, rather than using prediction matrix representations, a multiple linear regression was fitted using A level maths grade (scored 0 to 6) as the dependent variable and mean GCSE score (scored 0 to 10) as the independent variable. Dummy variables were used representing each relevant combination of qualification type (legacy or reformed), examination series (2017, 2018, or 2019), and age group (17 or 18).

This analysis treats the four legacy qualifications collectively and the four reformed qualifications collectively to identify national trends in the data. The analysis shows a

[19] https://ofqual.blog.gov.uk/2017/04/21/prediction-matrices-explained/

significant relationship between mean GCSE score and A level maths grade ($p < 0.001$).

Table 7 shows the beta coefficients for the regression model and the upper and lower bound estimates. All the coefficients were significant at the .001 level, meaning that there were statistically notable differences in results for each group at any given value of mean GCSE score when compared with the reference group used for the model – 18-year-olds sitting the legacy qualifications in 2017.

To support representation of the average value-added relationship across the grade range and improve stability and visualisation, the fixed gradient model was used for the majority of the analyses. The slope coefficient for mean GCSE score was B = 0.803; an increase in one point in the mean GCSE score was associated with a grade increase in A level maths of 0.8 grades. $R^2 = .332$.

Table 7. Beta coefficients for the regression model explaining variation in A level maths grade score using mean GCSE score, qualification type, examination series, and age group (n=162,289).

	Unstandardised Coefficients		95.0% Confidence Interval for B	
	B	Std. Error	Lower Bound	Upper Bound
(Constant)	-2.661***	0.024	-2.707	-2.615
Legacy 2017 18s (reference group)	0.803***	0.003	0.797	0.808
Legacy 2017 17s	0.508***	0.037	0.436	0.579
Legacy 2018 17s	0.479***	0.063	0.356	0.602
Legacy 2018 18s	0.027***	0.008	0.012	0.043
Reformed 2018 17s	0.450***	0.036	0.379	0.520
Reformed 2019 17s	0.761***	0.033	0.697	0.825
Reformed 2019 18s	-0.112***	0.008	-0.127	-0.097

(*** = significant at .001 level.)

To initially evaluate the effectiveness of the standard setting process, consideration should be given to the aims of the 2018 and 2019 awards. As described in Section 2.4.1, it was the value-added relationship for 17-year-olds from the legacy qualifications that was the basis for the statistical guidance in 2018. In 2019, the statistical basis for the awards was 18-year-old candidates from the legacy

qualifications as described in Section 2.4.2[20]. Figure 1 shows the relationship between mean GCSE score and attainment in A level maths for 17-year-olds in the legacy qualifications in 2017 and the reformed qualifications in 2018. Also included are the relationships for 18-year-olds in the legacy qualifications in 2017 and the reformed qualifications in 2019. These were the notional relationships on which the statistical guidance used for awarding were based.

Given the similarity of the relationships, the awards appear to have been successful in their aim of aligning the value-added relationship between legacy and reformed qualifications for the 17-year-olds in 2018 and the 18-year-olds in 2019. The small gap of approximately 0.14 grades between the lines for 18-year-old candidates is explored in Annex C.

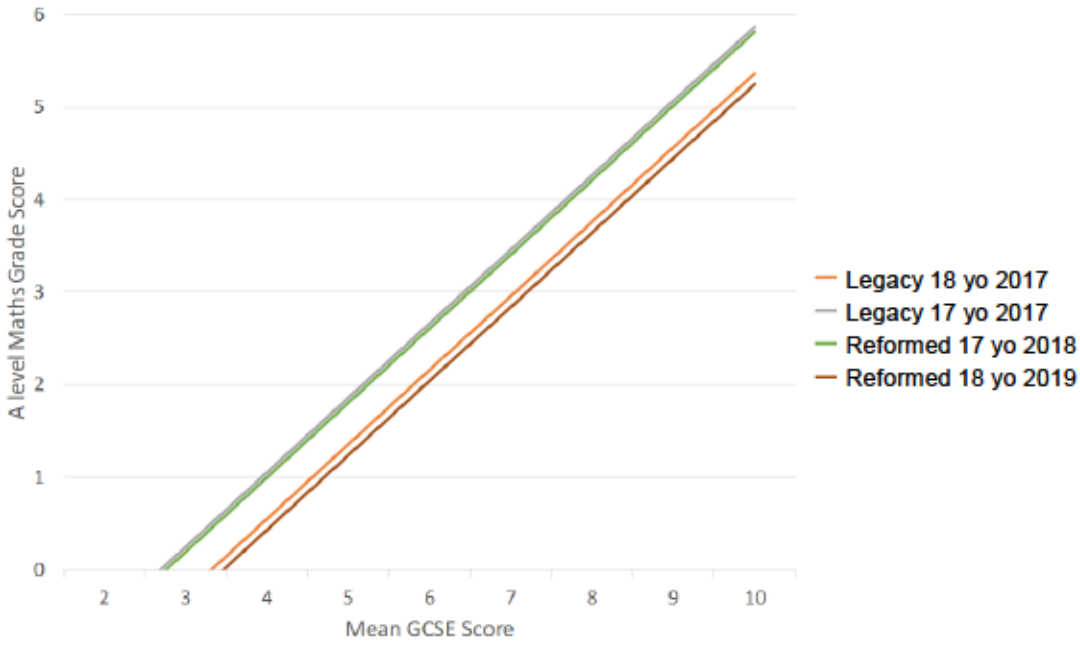


Figure 1. Relationship between mean GCSE score and attainment in A level maths for 17-year-olds in the legacy qualifications in 2017 and the reformed qualifications in 2018 and 18-year-olds in the legacy qualifications in 2018 and the reformed qualifications in 2019 (maths grades A – U reported as 6 – 0).*

Figure 2 reproduces Figure 1 and adds the relationship for the 17-year-olds in 2019. In both 2017 and 2019, the grade boundaries were set based on the majority 18-year-old cohorts; therefore, the 17-year-olds' grade distribution – and thus value-added relationship – were not controlled directly in the award, as described in Section 2.4.1.

As can be seen in Figure 2 and can be calculated from Table 7, the difference in value-added between 18-year-olds and 17-year-olds on the legacy qualification in 2017 is 0.51 A level grades for a given mean GCSE score. For the reformed qualifications awarded in 2019 this difference in average value-added had increased to 0.87 A level grades. As the 17-year-old standards were not set directly and were a

[20] Note that the basis for the awards was an aggregate of data across multiple years. Here, comparison is made with 2017 only as the most recent representation of the statistical standards on the legacy qualifications.

consequence of the statistical standard for 18-year-olds, this relative change in relationship has occurred naturally.

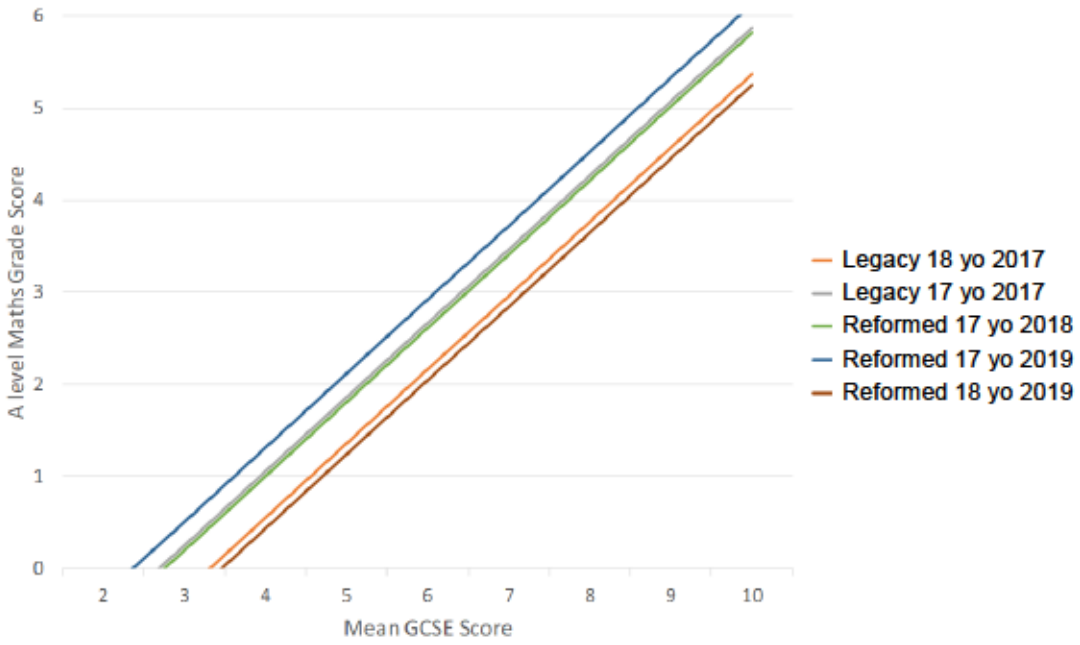


Figure 2. Relationship between mean GCSE score and attainment in A level Mathematics for 17-year-olds in the legacy qualifications in 2017 and the reformed qualifications in 2018 and 2019 and 18-year-olds in the legacy qualifications in 2018 and the reformed qualifications in 2019 (Maths Grades A – U reported as 6 – 0).*

For clarity, Figure 3 shows only the regression plots for 17-year-olds sitting reformed qualifications in 2018 and 2019 with 95% confidence intervals. The difference between intercepts of 0.31 grades appears reliable (Table 7 and Figure 3) demonstrating the difference in value-added for the two groups of 17-year-old candidates.

Given this change in relationship and the use of the legacy 17-year-old value-added relationship for the basis of the first year of the reformed qualifications, this could indicate a discontinuity in statistical standards between 2018 and 2019.

To help build an understanding of the impact this change in relationship might have had on standards over time and within year it is important to understand why this relationship might have changed and whether it reflects a genuine change in the value-added relationship that will likely be reflected in future awards of the reformed qualifications. These issues are discussed in the sections that follow.

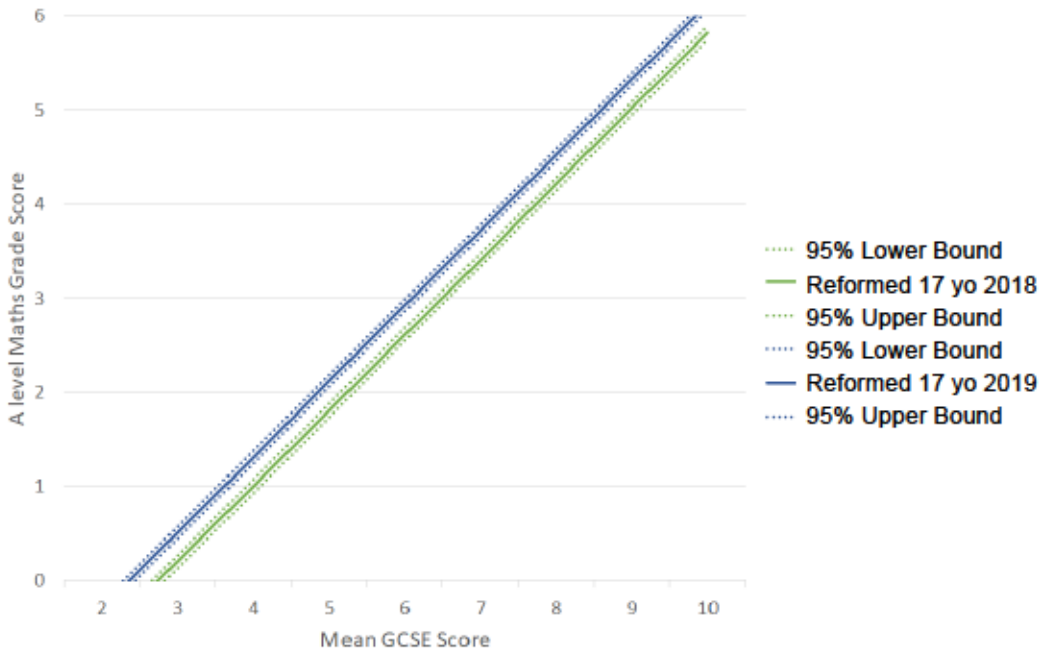


Figure 3. Relationship between mean GCSE score and attainment in A level maths for 17-year-olds in the reformed qualifications in 2018 and 2019 (Maths Grades A – U reported as 6 – 0; confidence intervals = 95%).*

4.5 17-year-old vs 18-year-old relationship

The evidence above supports the conclusion that, despite concerns about the reliability of the predictions used, the statistical standard for 17-year-old candidates was carried forward effectively between legacy and reformed qualifications in 2018, as intended and described in Section 2.4.1. The evidence also suggests that the standards were carried forward effectively between legacy and reformed qualifications in 2019 based on the predictions for 18-year-olds, as described in Section 2.4.2.

Despite the effectiveness of these two processes, there appears to be a discontinuity in statistical standards between 2018 and 2019 based on a comparison of the value-added relationships for 17-year-olds across the two years potentially suggesting that standards are not comparable between 2018 and 2019. As highlighted above, the difference in the relationship between ages needs to be better understood before conclusions can be drawn. To do so, the following four potential factors were identified, which may have impacted (legitimately or otherwise) on the relative value-added relationships for 17 and 18-year-old candidates either side of the transition to the reformed qualifications, and are considered in the sections that follow:

a) the presence of a significant 17-year-old cohort sitting the legacy qualification in 2018 meaning the 17-year-old cohort was split in an unexpected way
b) a change in aggregation and, in particular, the role of optimisation (see Section 2.3)
c) changes to the subject content/curriculum through the qualifications reform process
d) the change in opportunities for candidates to resit individual assessments

4.5.1 17-year-old split cohort in 2018

The original motivation for allowing candidates to certificate after the first year of teaching in the reformed versions of A level maths was primarily to afford 17-year-old candidates the opportunity to certificate in maths before going on to the reformed further maths a year later, should they so wish. It was, therefore, anticipated that those sitting A level maths in 2018 would be sitting the reformed version of the qualification. However, in 2018, the 17-year-old entries were unexpectedly split between the legacy (1,125 candidates) and the reformed qualifications (1,610 candidates). It seems likely this is a centre level decision rather candidate self-selection; however, that does not in itself preclude the possibility of there being a systematic difference between the two. This could mean candidates who perform better than average at A level, given their mean GCSE attainment, were differentially entered for either the legacy or reformed version of the qualifications. Were this to have happened it could lead to one of two scenarios:

1) the act of carrying forward the value-added relationship for 17-year-olds sitting the legacy qualification to those sitting the reformed qualification in 2018 (as was proved to be the case above) was inappropriate as this sub-group was not representative of the full 17 year-cohort, or
2) the combination of the value-added relationships for the two 17-year-old sub-cohorts in 2018 matches that of the value-added relationship for 17-year-olds in 2019. This would suggest that the combined standard across both groups in 2018 was appropriate for the reformed qualifications, or

For information, Table 8 below shows the outcomes for all 17-year-olds sitting reformed and legacy maths qualifications in summer 2018[21]. These outcomes do not control for the ability profile of the entries; therefore, differences between outcomes for the legacy and reformed groups do not necessarily reflect differences in the value-added relationship and, therefore, the grading standards.

Table 8. Cumulative percentage at grade for all 17-year-olds sitting reformed and legacy maths qualifications in June 2018. Outcomes for legacy qualifications are shown excluding and including candidates who certificated further maths in the same series.

	A*	A	B	C	D	E	U	Cands
Reformed	31.5	64.1	79.5	88.7	93.9	97.6	100.0	1,582
Legacy (excl. further maths)	17.1	49.5	71.1	86.5	92.6	96.4	100.0	895
Legacy (incl. further maths)	28.0	58.6	76.4	88.9	94.0	97.1	100.0	1,122

To explore the relative value-added relationship between the two groups of 17-year-olds in 2018, Figure 4 reproduces Figure 2 and adds the remaining relationships for the 17 and 18-year-olds sitting the legacy qualifications in 2018. Were the 17-year-

[21] Some of the candidates entered for the legacy maths qualifications also entered for further maths and the results are shown excluding and including these candidates as indicated above.

old cohort split differentially, in terms of value-added, between legacy and reformed qualifications in 2018, we would expect to see the value-added for 17-year-old candidates sitting the legacy qualification in 2018 to differ from the relationship for candidates entering the reformed version. It is clear from Figure 4 that this not the case. The inclusion of the 18-year-old relationship for the legacy qualification in 2018 was to confirm that there was no material difference for this group across years, which is indeed confirmed by the relationship shown.

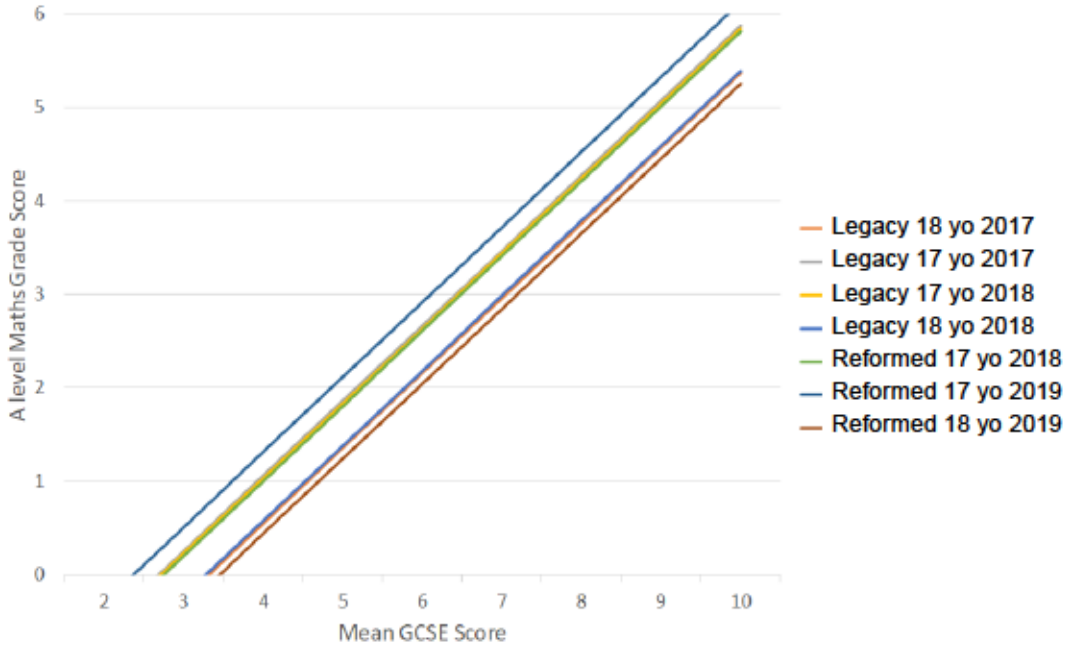


Figure 4. Relationship between mean GCSE score and attainment in A level maths for 17 & 18-year-olds in the legacy qualifications in 2017 and 2018 and the reformed qualifications in 2018 (17s only) and 2019 (maths grades A – U reported as 6 – 0).*

This finding therefore suggests that the split in the 2018 17-year-old cohort was not systematic in terms of the value-added relationship. This would appear to suggest that neither of the two potential scenarios above occurred. However, to confirm whether or not this value-added relationship for the two groups appeared appropriate relative to previous performance, a common centres analysis focussing on these 17-year-old candidates was performed.

In this process, the entry patterns of centres are analysed across the two years of interest with only the outcomes for centres with entries (in the sub-group of interest) in both years being retained. The outcomes for these common centres can then be compared across years. The principle on which this analysis is based is that, overall, the aggregated outcomes for common centres are unlikely to shift significantly between adjacent years (if standards have been successfully maintained). If similarity of value-added relationships demonstrated above were inappropriate this may be exposed through these analyses.

Table 9 shows the outcomes for candidates from centres that entered 17-year-olds for legacy qualifications in 2017 and legacy and/or reformed qualifications in 2018. Table 10 shows the outcomes for candidates from centres that entered 17-year-olds for legacy qualifications in 2017 and either legacy *or* reformed qualifications in 2018,

but not both[22]. These common centres analyses should be interpreted tentatively due to the small numbers of candidates on which they are based and the known limitations of the methodology. However, the cells reliant on particularly small numbers have been greyed out to indicate as they are not particularly informative, but have been included for completeness.

Table 9 suggests that the candidates in 2018 performed slightly less well than candidates from the same centres in 2017, whether they sat the legacy or reformed qualification. This is show by differences of -3.34%p at grade A and -0.09%p at grade E for the legacy qualification and -2.54%p at grade A and -1.12%p at grade E for the reformed qualification. However, given that it is the differences in relationship that are of particular interest here, it is fair to conclude that there is no material difference between groups.

Table 10 overcomes the issue of centres splitting their entries between legacy and reformed qualifications by excluding candidates from centres that entered some 17-year-olds for legacy qualification and some for the reformed version in 2018. The outcomes for candidates at centres that remained with the legacy qualification were slightly higher in 2018 (+1.05%p at grade A and +0.03%p at grade E) as was the case for those switching completely to the reformed qualification with a more notable +6.03%p increase in outcomes at grade A and a modest decline of -1.30%p at grade E.

Taken together, these common centres analyses do not present compelling evidence that 17-year-old candidates entered for the reformed qualifications in 2018 received collectively different grades than they might have received had they entered for the legacy qualifications instead.

In summary, the splitting of the cohort between legacy and reformed qualifications did not have a negative impact on the maintenance of standards for 17-year-olds from the legacy to reformed qualifications in 2018. Nor did these analyses suggest that, when the two groups of 17-year-olds from 2018 are combined, their value-added relationship reflected that for 17-year-olds candidates who certificated in 2019. This therefore, means that the split in entry for 17-year-old candidates in 2018 can be discounted as a potential source of change in relative value-added relationship between 17 and 18-year-old candidates.

[22] The 'combined' totals of candidates include those from centres that changed qualifications between years, whereas the qualification level count excludes them, hence the combined totals are typically greater than the sum of the specification totals. Note that a centre that has changed exam board in addition to splitting its entries – for example, that has entered AQA legacy in 2017 and AQA legacy *and* Pearson reformed in 2018 – will appear for AQA under the legacy-legacy analysis, but not in the combined row, so the combined totals can also be lower than the sum of the exam board totals.

Table 9. Centres splitting their entry across legacy and reformed allowed to contribute to both legacy and reformed CC analysis.

	Legacy								Reformed							
	Legacy 2017			Legacy 2018			Difference		Legacy 2017			Reformed 2018			Difference	
	A	E	n	A	E	n	A	E	A	E	n	A	E	n	A	E
AQA	69.02	96.74	184	63.24	97.06	136	-5.79	0.32	56.60	96.23	53	50.00	93.94	66	-6.60	-2.29
OCR A	71.84	99.03	103	73.21	100.00	56	1.37	0.97	45.45	96.97	33	69.09	100.00	55	23.64	3.03
OCR B (MEI)	80.00	97.50	80	66.67	100.00	39	-13.33	2.50	77.78	88.89	9	62.50	75.00	8	-15.28	-13.89
Pearson	69.09	98.32	537	67.12	97.97	295	-1.97	-0.36	66.67	97.79	543	64.30	97.26	731	-2.37	-0.53
Combined	70.25	98.02	911	66.91	97.93	532	-3.34	-0.09	65.38	97.88	754	62.84	96.76	958	-2.54	-1.12

Table 10. Centres splitting their entry across legacy and reformed excluded from the analysis.

	Legacy								Reformed							
	Legacy 2017			Legacy 2018			Difference		Legacy 2017			Reformed 2018			Difference	
	A	E	n	A	E	n	A	E	A	E	n	A	E	n	A	E
AQA	69.18	96.86	159	62.90	96.77	124	-6.28	-0.08	46.43	96.43	28	48.28	96.55	29	1.85	0.12
OCR A	74.71	100.00	87	73.08	100.00	52	-1.64	0.00	35.29	100.00	17	48.28	100.00	29	12.98	0.00
OCR B (MEI)	80.00	97.50	80	66.67	100.00	39	-13.33	2.50	77.78	88.89	9	62.50	75.00	8	-15.28	-13.89
Pearson	59.86	98.21	279	66.51	98.09	209	6.65	-0.12	55.44	97.19	285	62.47	96.91	421	7.03	-0.28
Combined	65.70	97.83	554	66.75	97.86	421	1.05	0.03	54.66	97.48	397	60.69	96.18	524	6.03	-1.30

4.5.2 Changes in aggregation

The next factor to be explored as a potential source of the change in the value-added relationship between age groups are changes to the aggregation rules.

As highlighted in Section 2.3, due to the interaction between the qualifications available in the GCE maths suite (AS and A level qualifications in different version of maths, further maths and additional further maths), the legacy maths qualifications required complex aggregation rules and had in place arrangements to optimise candidates results. In contrast, the simpler structure of the reformed qualifications means assessment results are contained within a single qualification and aggregation is performed independently. It should be noted that the removal of this optimisation of candidate grades on the legacy qualification does not represent a disadvantage to candidates on the reformed qualification. As discussed in Section 2.4, one motivation for using statistical predictions to guide awards at qualification level, particularly around a time of reform, is to not differentially advantage/disadvantage candidates either side of the transition due to this kind of structural change. Optimisation in and of itself, therefore, should not impact on the statistical standard set in 2019 relative to the reformed qualification. The motivation for its consideration here is whether or not it might have positively impacted on the relative outcomes for 17-year-olds candidates compared to 18-year-olds.

Intuitively, this change in aggregation approach should differentially affect the relationship between 17 and 18-year-olds, as candidates sitting at the age of 18 will have previously had more units on which to draw and also have had greater opportunity to sit units which could be combined in different ways. This combined with the optimisation rules giving preference to A level maths outcomes over further maths could, therefore, provide a differential benefit for 18-year-old candidates. While this change is likely to have had an effect, it cannot, however be the cause of the differences in the relative value-added relationship discussed in Section 4.4 and summarised in Figure 2. This is due to the approach taken to formulating the statistical predictions that has been mirrored in the data analysis used here. The candidates benefiting from optimisation are those certificating in A level maths and further maths simultaneously as highlighted through the JCQ guidance. As these candidates are excluded from the comparison, this factor cannot explain the difference identified above and can, therefore, be discounted from the considerations.

4.5.3 Curriculum effects

Changes to the content and curriculum as a consequence of the reform process would not, overall, impact on outcomes due to the protection provided by the statistical predictions as described in Section 2.4. However, were these changes to have a differential effect across age groups, this could impact on the relative performance of candidates.

The main curricular changes to the new A levels in maths are that:

(i) the applied content – statistics and mechanics – is now compulsory, sampling topics from previously optional units of the legacy qualifications, plus additional material for statistics;

(ii) there is no optional content - all candidates for a qualification sit the same question papers;

(iii) decision maths content has been removed from (AS and) A level maths.

Beyond this high level summary, provided in Annex D, is a summary of more detailed changes to A level maths, adapted from an AQA publication[23].

It is not possible to say anything conclusive about the effects of these changes on subgroups of students without a thorough analysis of the choice of optional content and patterns of performance in the legacy qualifications by 17 and 18-year-old candidates. Such an analysis may still be speculative given performance data would clearly not be available for options candidates chose not to take. That said, due to the removal of optionality as part of the structural changes made, the better achieving candidates post-reform will be, by definition, the better all-round performers. Those candidates who can tackle content that might not be their strength. It seems likely that many of those taking maths aged 17 and continuing to study further mathematics aged 18 will fit this description, whereas the 18-year-old, non-further maths cohort will, disproportionately, contain those that benefited from optionality and, therefore, may be more likely to demonstrate overall weaker performance post-reform. These candidates are likely to have demonstrated the greatest saw-tooth effect (see Section 2.4), which has been counteracted by the comparable outcomes approach to maintaining standards.

While this is speculative, it appears plausible that these changes to the content (and related assessment requirements) may have impacted on the relative relationship between 17 and 18-year-old candidates for reasons other than their age.

Counter arguments to this position are that 17-year-old candidates may have been afforded less time to prepare for this broader range of content. It is, however, unlikely that these early certificating candidates have not started covering at least some of the A level content prior to the year of their maths certification.

In conclusion, it is not possible to draw strong conclusions regarding the impact of the curriculum changes on different age groups of candidates and these effects cannot be readily quantified. However, it appears entirely plausible for the reasons given that this might have a positive impact on the performance of 17-year-old candidates relative to 18-year-olds in the reformed versions of the qualifications.

4.5.4 Resitting opportunities

The fourth and final identified source of a potential change in the value-added relationship between 17 and 18-year-old candidates on the reformed qualifications is the removal of opportunities for candidates to resit individual assessments prior to certification. As described in Section 2.2, two structural changes to the A level qualifications impact on the practicalities and effectiveness of resitting: the move from modular to linear qualifications and the decoupling of AS and A level qualifications. Similar to the case of optimisation as described above, the use of statistical predictions through awarding seeks to prevent any overall advantage/disadvantage for the majority cohort due to this effect, however, it may have a differential effect between age groups.

In the legacy qualifications where candidates' AS unit marks contributed to their A level grade, candidates would frequently resit AS units at the age of 18. This

[23] https://www.aqa.org.uk/resources/mathematics/as-and-a-level/mathematics/plan/summary-of-changes

approach is something maths, in particular, lends itself to given the cumulative nature of learning in the subject.

To model the magnitude of this effect, the attainment of candidates certificating at A level in 2017 were analysed. Those certificating candidates were matched with the AS maths certification data from 2016. Once candidates were matched, their unit results used for certification at AS in 2016 and A level in 2017 were analysed. Where candidates had increased their UMS mark between the two certification events, this benefit was due to unit level resitting. To determine the cohort level effect of resitting, each candidate had the number of additional marks accumulated through resitting removed from their UMS total mark and were regraded. These modelled results were then aggregated to determine revised outcomes separately for 17 and 18-year-old candidates. The magnitude of the effect is shown in Table 11 with the impact of this resitting effect on the value-added relationship for these candidates. This effect, aggregated across qualifications, is shown in Figure 5 with the orange line representing candidates' actual grades and the blue line representing the grades they would have obtained were their unit resit results discounted. The figures in Table 11 show a very similar effect of resitting on cumulative percentage outcomes across the legacy qualifications. From comparison of Figure 5 and Table 11 it is interesting to note that, despite resitting having a far greater impact on candidates at lower levels of mean GCSE prior attainment, the effect on cumulative percentage outcomes is greater at grade A than grade E.

The aggregated impact across the legacy qualifications was an increase of 6.5% at grade A and 1.8% at grade E due to the resitting improvement. A similar sized effect was also observed, however, for 17-year-old candidates. The critical difference, however, is that a far smaller proportion of 17-year-old candidates followed this route of certification and, in doing so benefitted from resitting between AS and A level.

To determine the potential impact of this effect on the 2019 grade boundaries, national prediction matrices for 18-year-old prior attainment matched candidates were built based on the 2017 outcomes; one version including the effects of resitting and one with the effect removed. The prior attainment profiles for matched 18-year-old candidates entering for each qualification in 2019 were then applied to each matrix to form a prediction with and without resitting. These two sets of predictions were then subtracted to establish the difference in predicted outcomes due to the resitting effect for 18-year-old candidates as shown in Table 12.

To translate these adjustments into mark differences, qualification and component level mark distributions were generated for prior-attainment matched 18-year-old candidates. To establish the difference in marks due to the resitting effect, the outcome at the operationally set grade boundary was adjusted by the percentages quoted in Table 12 minus the resitting benefit seen for 17-year-old candidates shown in Table 11, with this value scaled by the proportion of 17-year-old candidates benefitting from resitting[24]. The results of this process are presented in Table 13.

[24] Ideally, it would have been desirable to replicate the process of outcome matrix and prediction generation carried out for 18-year-olds for the group of 17-year-old candidates. However, due to the small number of candidates this approach did not provide usable results. The proportionate weighting of the 17-year-old effect and adjustment relative to the 18-year-old mark distribution was deemed an appropriate approximation.

Table 11. Modelled impact of cumulative percentage outcomes of unit level resitting between candidates certificating to AS maths in 2016 and A level maths in 2017 weighted by proportion.

	Age	Cands	As prop'n of age	Original outcomes		Modified Outcomes		Difference	
				A	E	A	E	A	E
AQA	18	12,836	0.94	39.8	97.2	34.0	95.5	+5.8	+1.7
(6361)	17	73	0.21	63.0	98.6	58.9	97.3	+4.1	+1.3
OCR	18	8,822	0.85	49.4	98.3	42.7	96.7	+6.7	+1.6
(7890)	17	85	0.28	76.5	97.6	69.4	96.5	+7.1	+1.1
OCR (MEI)	18	7,164	0.82	47.2	97.7	41.1	96.1	+6.1	+1.6
(7895)	17	39	0.25	71.8	94.9	69.2	92.3	+2.6	+2.6
Pearson	18	31,698	0.83	41.1	97.5	34.4	95.6	+6.7	+1.9
(9371)	17	194	0.17	66.0	99.5	58.8	99.0	+7.2	+0.5
Combined	18	60,520	0.85	42.8	97.6	36.3	95.8	+6.5	+1.8
	17	391	0.20	68.3	98.5	62.1	97.4	+6.2	+1.1

Table 12 Impact of the removal of the resitting effect on the 2019 18-year-old matched candidate predicted outcomes at grades A and E

	Prediction incl. resitting		Prediction excl. resitting		Resitting effect	
	A	E	A	E	A	E
AQA (7357)	30.89	96.47	24.17	93.97	+6.72	+2.50
OCR A (H240)	34.31	97.13	27.16	95.05	+7.15	+2.08
OCR B (MEI) (H640)	34.33	97.22	27.17	95.16	+7.16	+2.06
Pearson (9MA0)	32.03	96.68	25.17	94.32	+6.86	+2.36

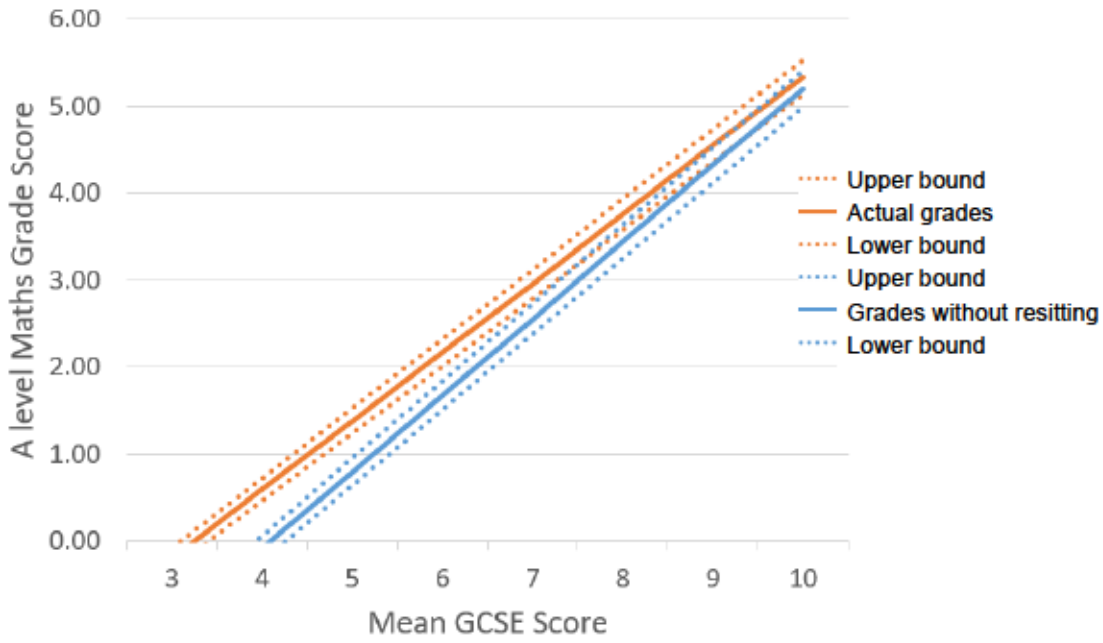


Figure 5. Relationship between mean GCSE score and attainment in A level maths for 18-year-olds in the legacy qualifications in 2018 – Actual grades vs grades without the benefit of resitting (maths grades A – U reported as 6 – 0; confidence intervals = 95%).*

The qualification grade boundary adjustments to account for resitting are summarised in Table 14. This shows the greatest proportion of the difference in grade boundaries between 2018 and 2019 being accounted for at grade A in the Pearson qualification (58%) and the smallest being the OCR B (MEI) qualification at grade E (17%).

Table 13 Impact of the resitting effect on 2019 grade boundary position based on the resitting benefit demonstrated by candidates between certification at AS maths in 2016 and certification to A level maths in 2017. All outcomes are for 18 year-olds.

		Paper 1		Paper 2		Paper 3		Overall	
		A	E	A	E	A	E	A	E
AQA	2019 Boundary	53	15	62	16	70	21	**185**	**52**
	Outcome at original (%)	31.08	96.68	31.04	96.60	33.41	96.69	30.31	97.15
	Outcome at adjusted (%)	24.64	94.11	24.75	93.96	27.68	94.52	24.68	94.84
	Adjusted boundary	57	19	66	20	73	26	**194**	**65**
	2018 Boundary	56	32	65	30	60	28	**181**	**90**
OCR A	2019 Boundary	54	13	58	15	49	12	**161**	**40**
	Outcome at original (%)	38.68	96.76	37.07	96.40	36.33	97.22	36.23	97.39
	Outcome at adjusted (%)	33.47	94.72	31.99	94.77	29.59	95.43	30.32	95.02
	Adjusted boundary	57	16	61	17	53	14	**171**	**49**
	2018 Boundary	67	21	61	18	69	21	**197**	**60**
OCR B (MEI)	2019 Boundary	70	23	59	17	49	12	**178**	**52**
	Outcome at original (%)	38.35	96.97	38.53	97.05	35.50	96.50	36.13	97.35
	Outcome at adjusted (%)	31.55	95.22	32.20	95.30	28.52	96.64	30.10	94.25
	Adjusted boundary	73	27	62	20	52	15	**187**	**63**
	2018 Boundary	74	45	68	40	55	30	**197**	**115**
Pearson	2019 Boundary	56	15	52	13	57	15	**165**	**43**
	Outcome at original (%)	33.93	96.27	32.03	95.10	35.36	95.73	29.92	89.29
	Outcome at adjusted (%)	28.26	94.26	26.17	92.41	29.46	93.54	24.33	87.00
	Adjusted boundary	59	18	56	16	61	18	**176**	**53**
	2018 Boundary	70	26	62	24	52	20	**184**	**70**

Table 14 Estimated qualification level boundary adjustment to account for resitting as a proportion of the 2018 to 2019 grade boundary difference

	Qualification level boundary adjustment for resitting (in marks)		Proportion of 2018 to 2019 change (%)	
	Grade A	Grade E	Grade A	Grade E
AQA	9	13	N/A	34
OCR A	10	9	28	45
OCR B (MEI)	9	11	47	17
Pearson	11	10	58	37

4.6 Intermediate findings from Strand 1

This strand of the investigation has considered the extent to which the statistical standards were effectively maintained between the legacy A level maths qualifications and the reformed versions awarded in 2018 and 2019. This has provided an effective mechanism to evaluate Scenario B (the potential under-reward of 17-year-old performances in 2018) and Scenario D (the potential change in value-added relationship between 17 and 18-year-old candidates).

In summary, based on the analyses presented here, the following intermediate findings can be drawn:

i. statistical predictions were effective in maintaining grading standards for 17-year-old candidates between 2017 and 2018, as was the intention of the models applied during awarding. The concerns regarding the reliability of the predictions for this purpose appear, in hindsight, to have been unnecessary. The continuity of these standards with the legacy qualification is supported by an analysis of outcomes for common centres and demonstrates why these outcomes were not unexpected for centres. These standards do not, therefore, represent disadvantage to these candidates in 2018 relative to those certificating in earlier years

ii. statistical predictions were also effective in maintaining grading standards for 18-year-old candidates between 2017 and 2019, as was the intention of the models applied during awarding

iii. a change in value-added relationship between 17-year-old and 18-year-old candidates has occurred between the legacy and reformed qualifications. This could be observed for the first time in summer 2019. In combination with points i and ii, this suggests that performance standards are unlikely to have been maintained between 2018 and 2019

iv. there was no systematic difference, from a value-added perspective, between those candidates choosing to sit the legacy qualification at the age of 17 in 2018 compared to those sitting the reformed version in the same year. The effects described in points i to iv are represented graphically in Figure 6 with statistical standards that were directly set for the majority cohort being represented by bold lines and those that are for a sub-cohort and therefore set consequentially by fainter lines

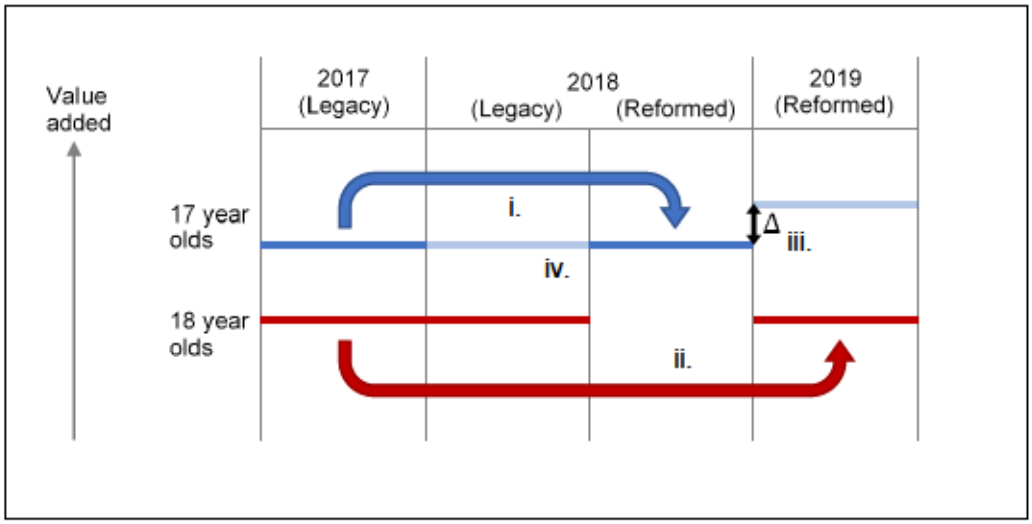


Figure 6 Summary of relationship between value-added rates across the reform transition for 17 and 18-year-old candidates

v. while 17-year-old candidates in 2018 do not appear to have been disadvantaged relative to previous years, they have not been afforded the value-added benefit occurring the year later to 17-year-old candidates (and that is likely to continue for the lifetime of the qualification) due to the change in relationship with 18-year-olds

vi. the predominant isolatable and quantifiable cause of this change in relationship appears to be the removal of the benefit of resitting in the reformed version of the qualifications, which previously had greatest benefit across the 18-year-old cohort. The analysis presented here estimated this effect to be equivalent to between 9 and 13 marks based on the 2019 mark scale

vii. the other likely contributor to the change in relationship between 17 and 18-year-old candidates are the changes to the curriculum that took place across the transition to the reformed qualifications. This effect cannot be effectively quantified, however, it is plausible that such changes would impact on the different sub-cohorts of candidates, potentially distinguished by age

This change in relationship between 17 and 18-year-old candidates has been shown to be key in explaining the difference in grade boundaries between 2018 and 2019 (Scenario B). It is the role of Strands 2 and 3 presented here to explore further the extent to which the difference in grade boundaries between years may also reflect differences in assessment difficulty over time or might represent further discontinuity in performance standards.

5 Strand 2: Analysis of question paper difficulty

The previous section considered the potential difference in grading standards that might exist between the 2018 and 2019 awards of A level maths and identified resitting as one likely cause of a difference in grade boundaries between the two years. This section explores the potential for differences in the grade boundaries to have emanated from differences in assessment difficulty.

The central purpose of the awarding process is to achieve fairness for candidates by setting grade boundaries that account for differences in the demands put on them. A key contributing factor to differences in grade boundaries between years is, therefore, differences in difficulty of the assessments; lower grade boundaries are set on more difficult assessments and higher boundaries set on less difficult assessments. It is possible that at least some of the changes in grade boundaries between 2018 and 2019 are due to differences in assessment difficulty. This strand of work is focussed on identifying, any differences in difficulty across the two years and quantifying the impact.

Two approaches were taken to evidence these differences: i) an item level comparative judgement of difficulty to quantify the expected difficulty and ii) an analysis of the operationally available item level candidate mark data. Evidence from these two analyses are explained below and brought together in Section 0 to quantify the impact of any differences.

5.1 Expected difficulty

Evaluating the difficulty of assessments across versions can be problematic when relying on operationally available data as effective interpretation relies on some form of linkage – either through having common items or common candidates across versions. In the context of the current comparison of interest – A level maths across 2018 and 2019 – there are no common items across years on which to rely and there are, typically, relatively few candidates who sit both versions of the assessment across years. Those who do cannot be considered *common*, however, due to the likely progression of their learning between the two sittings. The challenge is heightened in the current context due to significant differences, not just in the ability profile of the cohort across years, but also differences in composition and value-added for the two groups. This means that commonly used covariates such as mean GCSE grade do not provide adequate controls.

To evaluate differences in difficulty directly, independent from these effects, a comparative judgement exercise of item difficulty was performed. This provides an efficient approach to identifying differences in expected assessment difficulty across years for all components in scope for the investigation. Through the design process, consideration was given to alternative methods of gathering equivalent difficulty information independent of the confounding factors discussed above. One alternative would have been to ask an experimental set of participants – such as current year 13 or first year undergraduate students – to sit assessments from across both years in a counterbalanced cross-over design. Such approaches have been shown to generate results of comparable accuracy to the comparative judgement methodology through

the testing of a modest number of students. This may be valuable in instances where expected difficulty is to be used purely as a proxy for actual difficulty. Due to a combination of the logistical limitations, potential disruption to participating schools, limitations on the meaningfulness of the results due to a lack of preparedness of students and the magnitude of the assessment and marking challenge to generate data for all 24 assessment versions that are the subject of this investigation, the comparative judgement methodology was favoured. In addition, the proposed approach provides potentially valuable information regarding reasonable expectations of item difficulty for comparison with the expectations of difficulty articulated by exam boards. This information is not available from actual or experimental difficulty data which has the 'benefit' of hindsight.

5.1.1 Method

The comparative judgement method broadly mirrors the approach used in similar research into the comparison of assessment difficulty in recent years[25]. Briefly, the current study involved a number of A level teachers and maths experts using an online system to remotely judge the relative difficulty of individual items. Through the on line system, judges were asked select the more difficult item for students to answer from pairs of questions presented side by side on screen. The methodology relies on each judge seeing a random selection of questions with each question being judged against many other questions by many judges. In this study, the items were presented with their mark schemes, as it was possible that differences in their design or the approach to allocating marks could have an effect on item difficulty. Previous work, related to a study of AS and A level maths questions demonstrated the value of including the mark scheme in improving the correspondence between the judged difficulty and actual difficulty (quantified by the item facility)[26].

To construct a single scale of expected difficulty, a statistical model is then fitted to the judgement data which gave an estimate of difficulty for each item which best explains the pattern of judgements made.

5.1.1.1 Procedure

Comparisons were conducted using the online comparative judgement platform, No More Marking[27]. 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

[25] For example:
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
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/668587/Summer_2017_GCSE_Maths_assessments_review.pdf
https://www.gov.uk/government/publications/an-evaluation-of-the-item-difficulty-in-as-and-a-level-maths

[26] See appendix A of: https://www.gov.uk/government/publications/an-evaluation-of-the-item-difficulty-in-as-and-a-level-maths

[27] www.nomoremarking.com

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 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).'*

This approach is common with other recent studies. 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.

5.1.1.2 Materials and participants

All items from the assessments comprising the summer 2018 and 2019 A level maths assessments were included in the comparative judgement exercise[28]. This corresponds to 832 live items. In addition, 100 items were included from the sample assessment materials (SAMs) that were submitted by exam boards to Ofqual as part of the qualification accreditation process. The inclusion of these anchor items was to enable the constructed scale to be linked back to previous work, should it be necessary for the purposes of analysis, and to potentially support future work outside the scope of this investigation. These anchor items have been judged as part of previous work[29] and were selected on the basis of being spread across the difficulty scale constructed in that study.

Provided in Annex E is the breakdown of the items included for each paper with the corresponding tariffs. This information is summarised below in Table 15.

[28] Question 12 (parts a and b) from OCR B (MEI) paper 2 from 2019 was excluded from the exercise due to an error in a question that rendered it unanswerable. This was identified during the live series leading to all candidates being awarded maximum marks for the item. The absence of this item from the comparative judgement study and maximum marks being awarded operationally has been factored into the remaining analysis presented here.

[29] https://www.gov.uk/government/publications/an-evaluation-of-the-item-difficulty-in-as-and-a-level-maths

Table 15 Breakdown of items included in the comparative judgement exercise

		Items included		
		2018	2019	SAMs
AQA	Paper 1	36	39	9
	Paper 2	33	32	9
	Paper 3	38	39	8
OCR A	Paper 1	29	33	8
	Paper 2	42	38	10
	Paper 3	29	30	7
OCR B (MEI)	Paper 1	37	32	10
	Paper 2	43	42	6
	Paper 3	26	27	3
Pearson	Paper 1	31	35	11
	Paper 2	33	35	7
	Paper 3	39	34	12
Overall	-	416	416	100

The formatting of the items presented to judges matched that used in the live question papers. The items were presented at the top of each judging window followed by the relevant section of the mark scheme. To ensure that the difficulty of the items could be judged in the context of the question as a whole, including any prompt/source materials, judges were presented with the complete question with the item subject to the current judging decision highlighted for attention. A similar approach was taken to the presentation of the mark scheme with the relevant section highlighted on screen. An example of the item presentation can be seen in Figure 7.

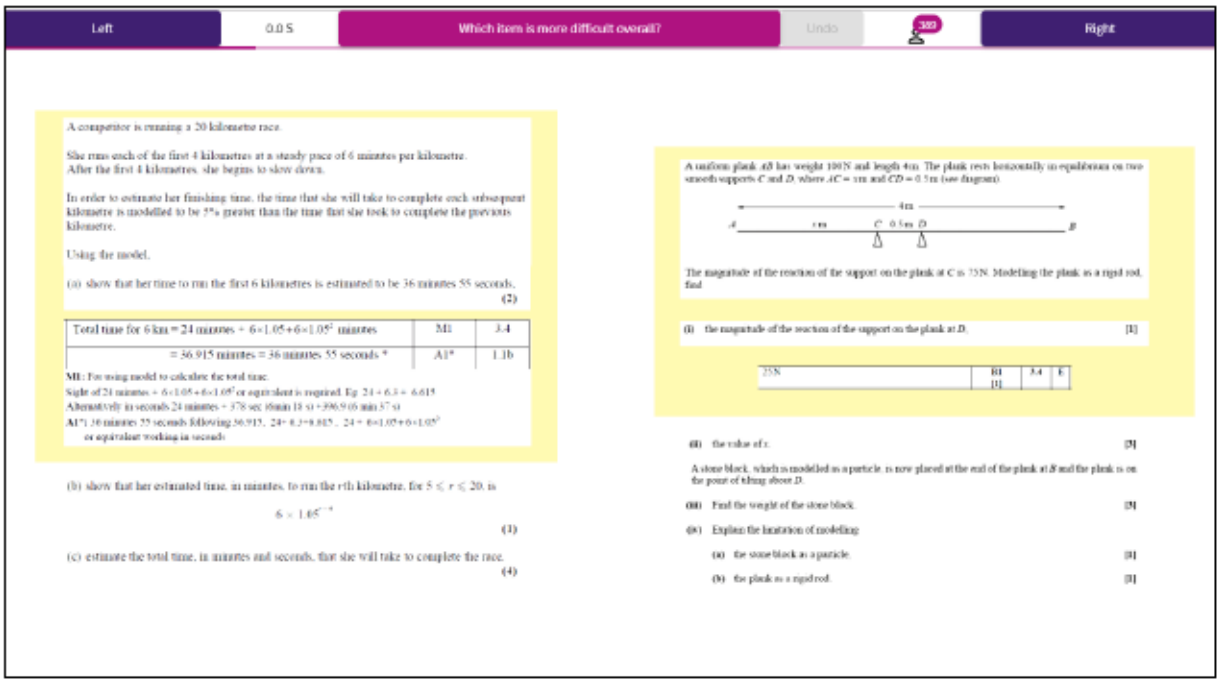


Figure 7 Screenshot of item difficulty comparative judgement exercise through the No More Marking system

A panel of 42 A level maths teachers and subject experts were recruited as judges. In order to maximise recruitment efficiency, these teachers were identified by contacting the top 100 centres in terms of entry size for A level maths in summer 2019. Judges were randomly presented with items from across all qualifications, papers and years, and were asked to complete 500 judgements across a week long window at their own convenience. 40 of the judges completed a full allocation with one judge completing 111 judgements and, another, 30. The judges were paid for their involvement in the study.

5.1.2 Analysis

The R package *sirt*[30] 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 in-fit, scale-separation reliability (SSR) and split-half reliability.

5.1.2.1 Judge consistency and exclusions

After the initial model was fitted to the set of judgements, judge in-fit was checked. In-fit is a measure of the consistency of the judgements made by a judge compared to the overall model. A high in-fit indicates that the judge was either inconsistent within their own judgements, or was applying different criteria from the consensus. Two outlying judges were identified and excluded using the criteria of an in-fit more than two standard deviations above the mean in-fit value for all judges. The

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

distribution of median judging times is shown in Figure 8. These median judging times were comparable to the time taken by judges in previous similar studies[31].

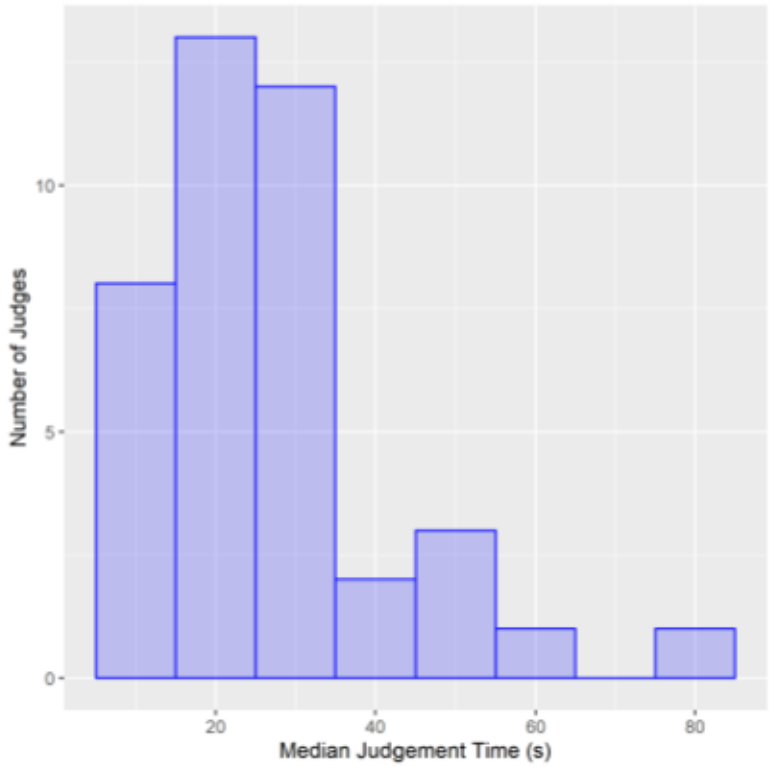


Figure 8 Median judging times for the 40 judges taken forward into the main analyses

Following the exclusion of judges producing misfitting data, the model was refitted and all other statistics are based on this final model fit.

Reliability is quantified in comparative judgement studies by the scale separation reliability (SSR) statistic that is derived in the same way as the person separation reliability index in Rasch analyses. It is interpreted as the proportion of 'true' variance in the estimated scale values. The SSR was 0.932 which shows a good level of reliability.

To represent the results of this analysis, each assessment is shown in the figures in the following sub-sections as a box plot displaying the median (solid bar), inter-quartile range (height of the box) and mean (diamond) of the expected item difficulties on a logit scale on the y-axis. This probabilistic scale can be interpreted as describing the log odds of one item being judged more difficult than another item. The absolute value is arbitrary and, in this case, is set in relation to the SAM items used as anchors during the item calibration. 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.

The outputs for each qualification are provided in the sub-sections that follow, with plots of the underlying item level expected difficulty estimates provided in Annex F.

31 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/676730/A_level_and_AS_mathematics_An_evaluation_of_the_expected_item_difficulty.pdf

5.1.2.2 AQA

Shown in Figure 9a is the aggregate expected item difficulty for the individual AQA assessments along with the aggregate across the qualification in Figure 9b. The statistics summarised in these plots are provided in Table 16.

These outputs show that, on the basis of the judgement data collected, the composite expected difficulty of Paper 1 has slightly reduced in 2019 compared to 2018 along with an increase in spread of the item difficulties. For Paper 2, the aggregate differences in expected difficulty appear to be small between years. For Paper 3, the expected difficulty of Paper 3 was lower in 2019 compared with the previous year.

When aggregated together across the qualification, this corresponds to an overall reduction in the expected difficulty of the combined assessments in 2019 compared with 2018. It is worth noting that this is in-line with the intention indicated by senior examiners in Section 2.5. As highlighted above, the units of this scale are arbitrary and, therefore, the materiality of this difference in expected difficulty is still to be established as is the case for the summary of differences in expected difficulty that follow for the other qualifications.

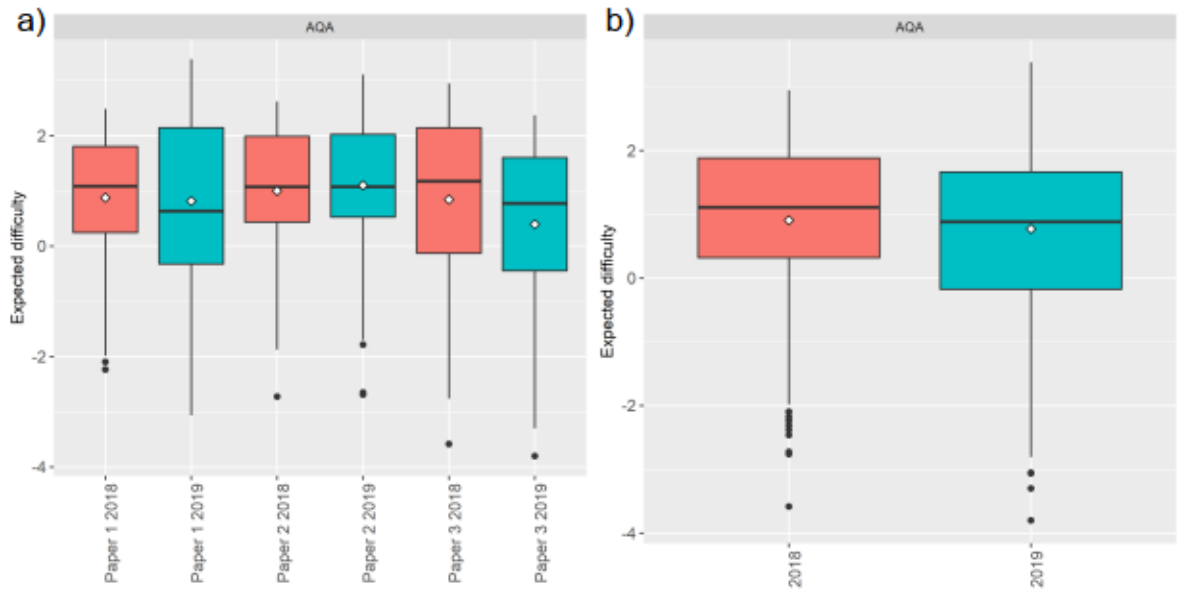


Figure 9 Summary boxplots of a) question paper expected difficulty and b) aggregated expected difficulty across the qualification for AQA question papers

Table 16 Summary statistics of expected difficulty for AQA question papers

	2018			2019		
	Mean	Median	IQ-range	Mean	Median	IQ-range
Paper 1	0.88	1.08	1.56	0.82	0.63	2.47
Paper 2	1.00	1.07	1.55	1.10	1.07	1.50
Paper 3	0.84	1.17	2.27	0.40	0.77	2.05
Combined	0.91	1.11	1.56	0.77	0.88	1.84

5.1.2.3 OCR A

Shown in Figure 10a is the aggregate expected item difficulty for the individual OCR A assessments along with the aggregate across the qualification in Figure 10b. The statistics summarised in these plots are provided in Table 17.

These data show that, for Paper 1, the expected difficulty between years is similar with a slight reduction in the spread of items. The overall picture for, Paper 2, suggested a different distribution of expected item difficulties within the assessment across the two years with the median item difficulty increasing in 2019 and the mean expected difficulty reducing. The results for Paper 3 suggest a slight increase in difficulty in 2019 compared to the previous year with a reduction in the spread of item difficulties.

When aggregated across all papers, as shown in Figure 10b, this indicates a similar level of difficulty between years. This appears broadly in line with the views of senior examiners expressed through the awarding process.

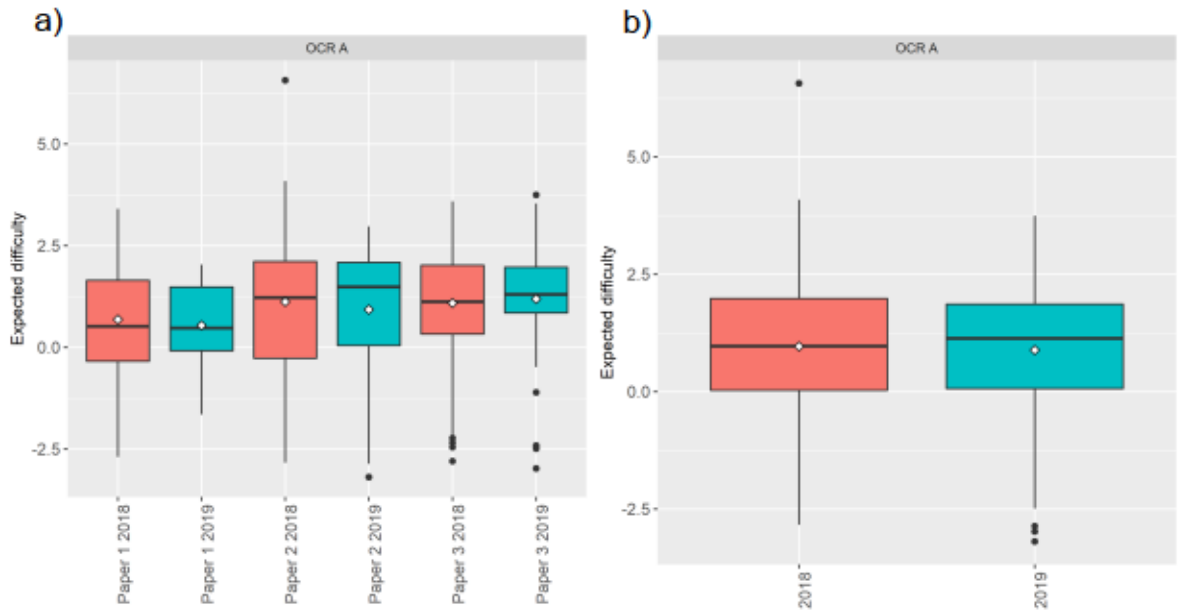


Figure 10 Summary boxplots of a) question paper expected difficulty and b) aggregated expected difficulty across the qualification for OCR A question papers

Table 17 Summary statistics of expected difficulty for OCR A question papers

	2018			2019		
	Mean	Median	IQ-range	Mean	Median	IQ-range
Paper 1	0.68	0.52	1.99	0.54	0.47	1.57
Paper 2	1.12	1.22	2.38	0.93	1.49	2.04
Paper 3	1.09	1.12	1.69	1.19	1.30	1.12
Combined	0.96	0.97	1.95	0.89	1.13	1.80

5.1.2.4 OCR B (MEI)

Shown in Figure 11a is the aggregate expected item difficulty for the individual OCR B (MEI) assessments along with the aggregate across the qualification in Figure 11b. The statistics summarised in these plots are provided in Table 18.

Figure 11a shows that, for Paper 1, the expected difficulty appears similar between the two years. For Paper 2, there appears to have been an increase in the spread of expected item difficulty with a greater number of more difficult items increasing both the spread and the overall difficulty of the assessment. In contrast, Paper 3, shows a notable increase in the spread of expected difficulty in 2019 compared with 2018 resulting in a decrease in the difficulty of the paper.

When aggregated up to qualification level, this suggests a slight increase in the overall difficult of the assessment in 2019, with an increased spread of expected difficulty. The different maximum mark of Paper 3 is taken into account in this qualification level aggregate.

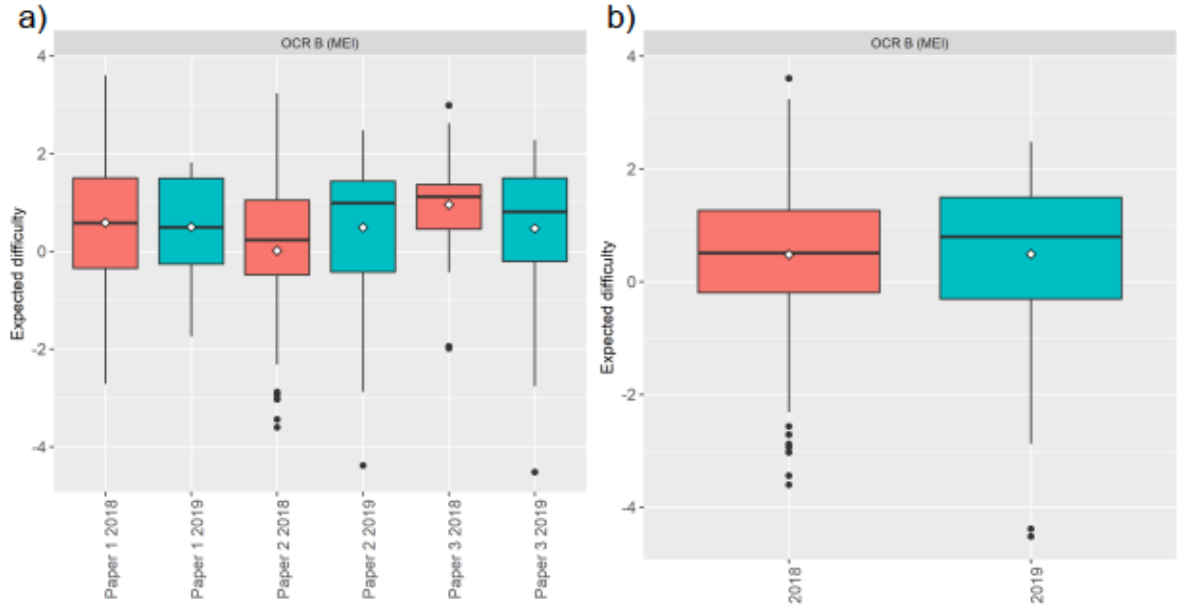


Figure 11 Summary boxplots of a) question paper expected difficulty and b) aggregated expected difficulty across the qualification for OCR B (MEI) question papers

Table 18 Summary statistics of expected difficulty for OCR B (MEI) question papers

	2018			2019		
	Mean	Median	IQ-range	Mean	Median	IQ-range
Paper 1	0.59	0.58	1.85	0.50	0.50	1.75
Paper 2	0.01	0.23	1.53	0.49	0.99	1.86
Paper 3	0.96	1.12	0.91	0.47	0.81	1.70
Combined	0.48	0.51	1.46	0.49	0.80	1.80

5.1.2.5 Pearson

Shown in Figure 12a is the aggregate expected item difficulty for the individual Pearson assessments along with the aggregate across the qualification in Figure 12b. The statistics summarised in these plots are provided in Table 19.

These outputs show that, on the basis of the judgement data collected, the expected difficulty of Paper 1 was similar across 2018 and 2019, with a suggestion of a different distribution of expected item difficulties within the assessment with the median item difficulty increasing in 2019 and the mean expected difficulty reducing. For Paper 2, the aggregate differences in expected difficulty appear to indicate an increase in difficulty of the assessment in 2019 compared to 2018 with a reduction in the spread of expected difficulties. This increase in difficulty appears in line with the feedback received from centres as noted in the feedback provided by the Pearson senior examiners. For Paper 3, the spread of expected difficulty is also reduced in 2019 compared to the previous year with a similar mean expected difficulty but increase in median.

When aggregated together across the qualification, this corresponds to an overall increase in the expected difficulty of the combined assessments in 2019 compared with 2018.

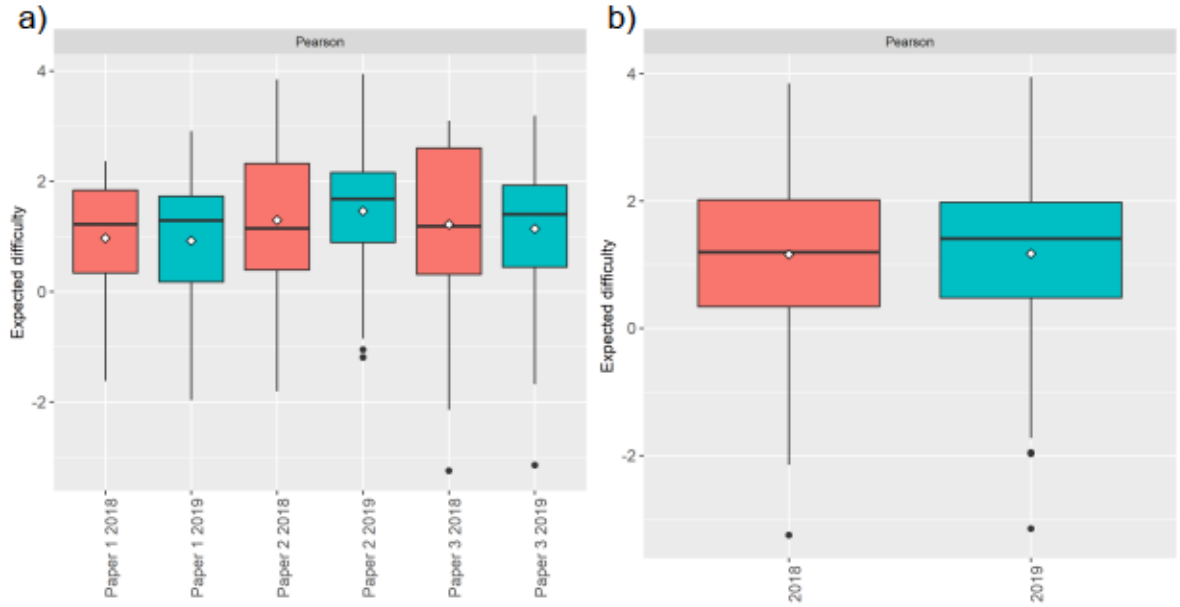


Figure 12 Summary boxplots of a) question paper expected difficulty and b) aggregated expected difficulty across the qualification for Pearson question papers

Table 19 Summary statistics of expected difficulty for Pearson question papers

	2018			2019		
	Mean	Median	IQ-range	Mean	Median	IQ-range
Paper 1	0.97	1.22	1.50	0.92	1.29	1.55
Paper 2	1.29	1.15	1.93	1.46	1.68	1.27
Paper 3	1.22	1.19	2.29	1.14	1.40	1.48
Combined	1.16	1.20	1.68	1.17	1.41	1.50

5.1.3 Expected difficulty summary

In isolation, consideration of the practical implications of the measures of expected difficulty presented above is problematic. As described, the units of these measurements are arbitrary so quantification of the differences in terms of the potential impact on marks achieved by candidates. Therefore, the extent that these differences should be compensated by grade boundary positioning is not clear. The use of these measures, in combination with measures of actual difficulty is revisited in Section 0.

5.2 Actual difficulty

The classic measure of item difficulty is the item facility index – a measure of the mean number of marks, expressed as a percentage, achieved by candidates sitting an item. This can provide a useful measure of the relative difficulty of items within an assessment or across assessments sat by the same candidates. In contrast to the measures of expected difficulty discussed above, comparison of these measures is problematic across versions of an assessment that have been sat by different candidates due to potential differences in ability. This is particularly challenging in the current instance where the most valuable comparisons are across 2018 and 2019 and it is known that the cohort sitting the reformed qualification is significantly different across years. However, these operational data provide a valuable basis for validating the item level expected difficulty estimates in addition to being used to translate the expected difficulty measures into estimated mark differences.

To support the later interpretation, the operational item level data have been fitted using the partial credit Rasch model[32] – a psychometric model which enables the situation and analysis of candidate ability and polytomous item difficulty on the same scale. The partial credit model (PCM) was fitted using R package *sirt*. All items from across a single qualification could be fitted to the model given the commonality of candidates sitting all papers facilitating construction of a single scale for each qualification in each year. Before linking, comparisons cannot be made across models as a separate model was fitted for each year.

In the following sub-sections, for each qualification are presented box plots of item facility and tables with the corresponding summary statistics. Item facility histograms for each paper are provided in Annex G, and test characteristic curves. These curves represent the expected test score of candidates with different abilities as defined by the model. When the curves for two tests on the same ability scale have the same shape, a test shifted to the left will be easier than tests on the right, since for the same ability the expected score on the test will be higher than those on the other tests.

Due to insufficient data, it was not possible to fit the PCM to OCR A and OCR B (MEI) data from 2018.

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

5.2.1 AQA

Figure 13 and Table 20 show the distribution of item facilities for the AQA question papers. These figures show a lower than typical spread of item facilities for Paper 2 in 2018 and a tendency for candidates to find items on Paper 3 less difficult than Paper 2 and, in turn, those less difficult than Paper 1. It should be noted that this information does not account for the weighting of items and, therefore, does not directly indicate the relative difficulty of the papers as a whole once the item weightings have been taken into account.

The test characteristic curves do, however, provide a better indication of difficulty of the whole paper across the ability range. These are provided in Figure 14 and show that, in 2018, for lower ability candidates, Papers 1 and 2 were the most difficult. For higher ability candidates, however, Paper 1 was the least difficult of the three papers. In 2019, Paper 3 was the least difficult for candidates of all abilities. At high ability, Paper 2 was the most difficult with Papers 1 and 2 having similar levels of difficulty for those of lower ability.

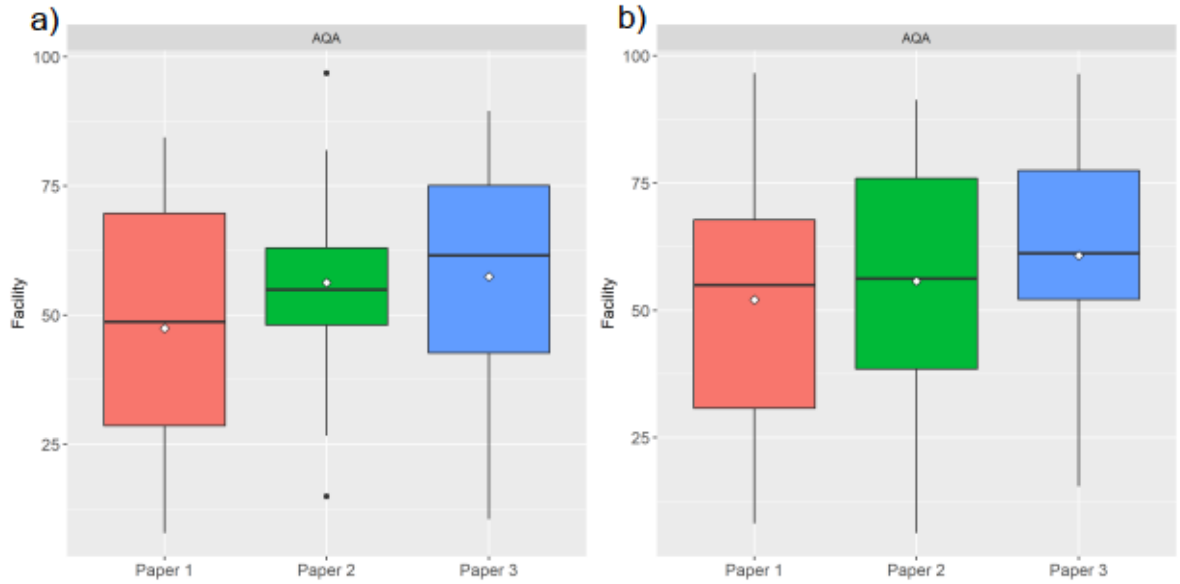


Figure 13 Boxplots of item facility index for AQA papers from a) 2018 and b) 2019

Table 20 Summary statistics of item facility indices for the AQA papers

	2018			2019		
	Median	Mean	S.D.	Median	Mean	S.D.
Paper 1	48.7	47.48	23.3	55.0	52.1	23.7
Paper 2	54.9	56.30	16.3	56.2	55.8	24.3
Paper 3	61.5	57.47	21.1	61.2	60.8	20.8

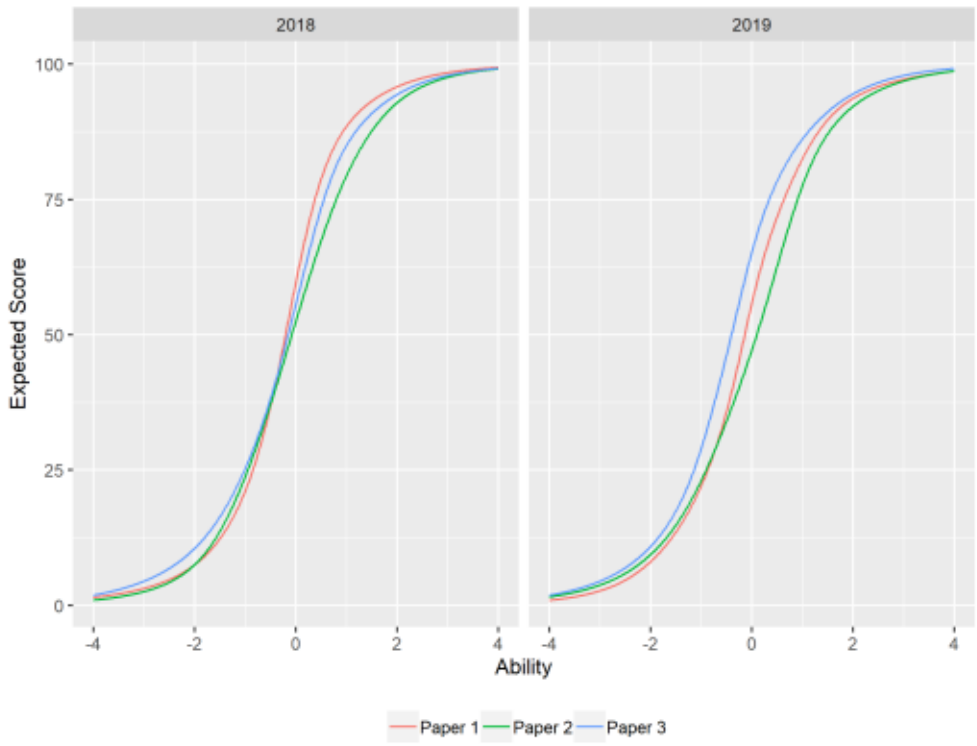


Figure 14 Test characteristic curves for the AQA 2018 and 2019 question papers from fitting the PCM

5.2.1 OCR A

Figure 15 and Table 21 show summaries of the item facilities for the OCR A question papers. The item facilities for all papers are higher in 2018 than in 2019, however, as highlighted above, this is likely to be due to cohort ability effects across the two years. The mean facility indices were broadly similar across all papers within each year with no obvious trend.

Figure 16 shows the test characteristic functions from 2019. These plots show that high ability candidates and lower ability candidates found all three papers of very similar levels of difficulty. Interestingly, for candidates of average ability, Paper 3 appears to have been distinctly more difficult that the other two assessments.

Table 21 Summary statistics of item facility indices for the OCR A papers

	2018			2019		
	Median	Mean	S.D.	Median	Mean	S.D.
Paper 1	80.8	74.5	17.3	48.6	52.0	21.2
Paper 2	78.0	70.7	24.1	56.1	56.6	18.7
Paper 3	75.5	74.6	17.9	51.2	52.3	23.5

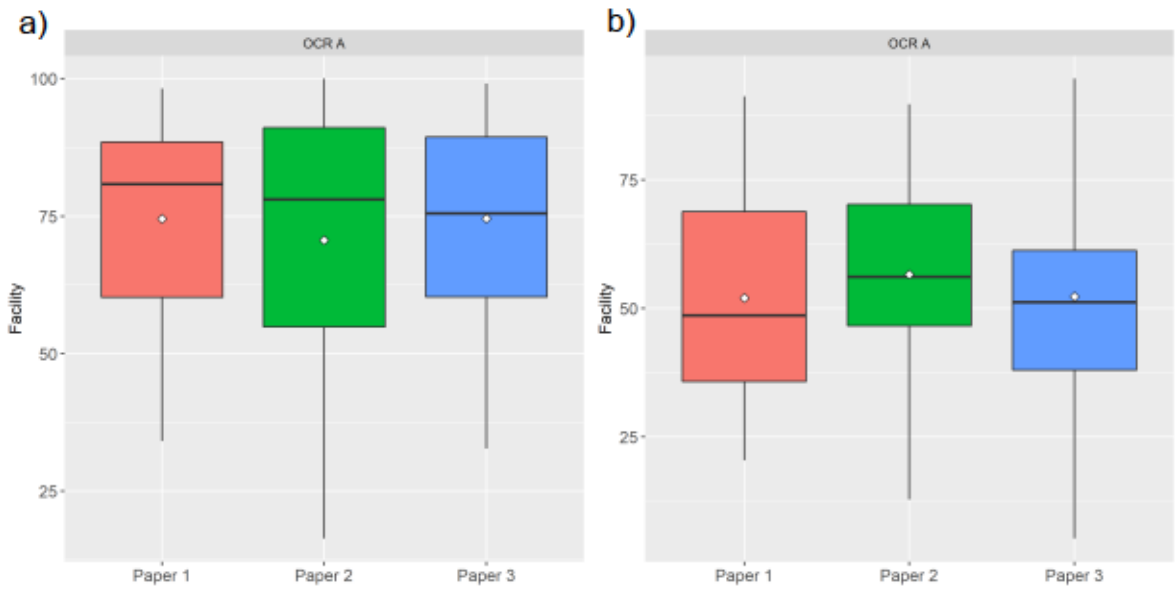


Figure 15 Boxplots of item facility index for OCR A papers from a) 2018 and b) 2019

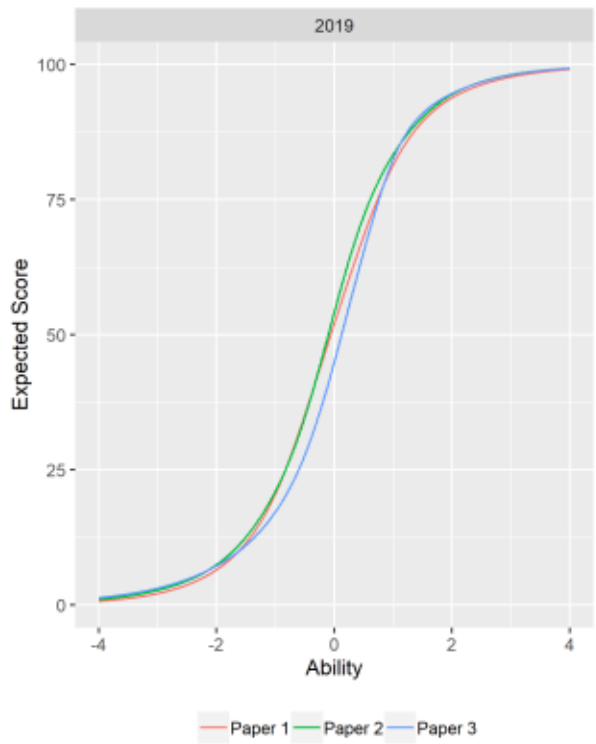


Figure 16 Test characteristic curves for the OCR A 2019 question papers from fitting the PCM

5.2.2 OCR B (MEI)

The item facility box plots for the OCR B (MEI) question papers are shown in Figure 17 along with the summary statistics in Table 22. These plots show that there appeared to be a greater number of more difficult questions on Paper 2 in 2018, however, there was a range of item difficulties indicated by the large spread of facility

indices. The item mean and spread of item difficulties across the 2019 papers appears, on average, to be broadly consistent.

Figure 17a shows the unscaled test characteristic curves for the 2019 question papers with Figure 17b showing the versions scaled for total mark available. These curves show that Paper 1 was the least difficult for candidates across the ability range and Paper 2 the most difficult for all abilities with the exception of the least able. Paper 3 was the most difficult for the weakest candidates.

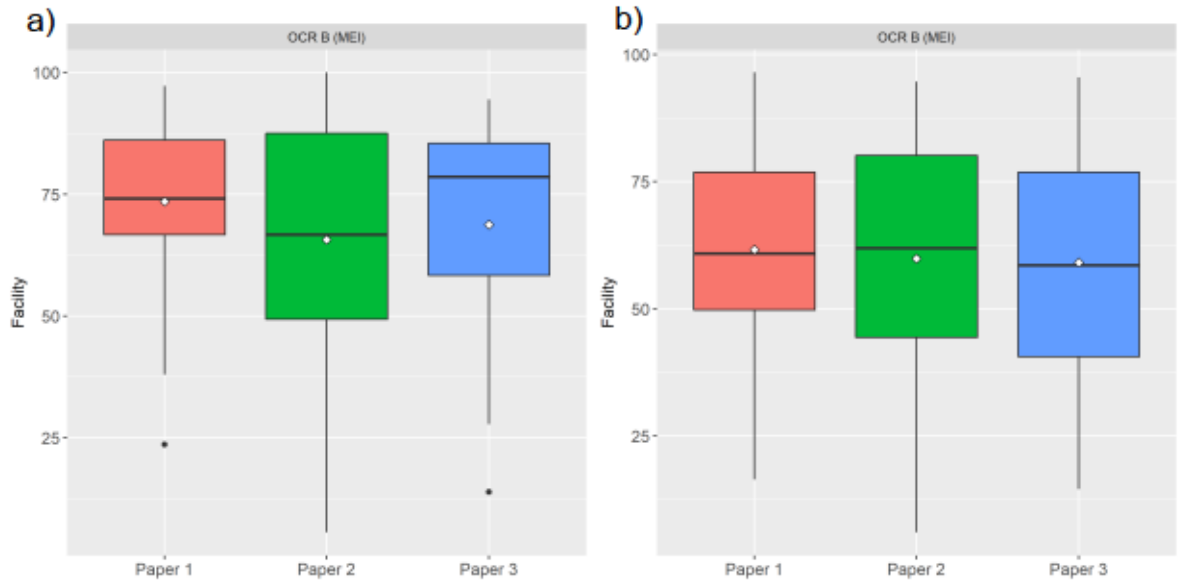


Figure 17 Boxplots of item facility index for OCR B (MEI) papers from a) 2018 and b) 2019

Table 22 Summary statistics of item facility indices for the OCR B (MEI) papers

	2018			2019		
	Median	Mean	S.D.	Median	Mean	S.D.
Paper 1	74.1	73.5	18.2	60.9	61.6	20.1
Paper 2	66.7	65.7	25.7	61.9	59.9	24.4
Paper 3	78.6	68.7	22.3	58.5	59.1	22.0

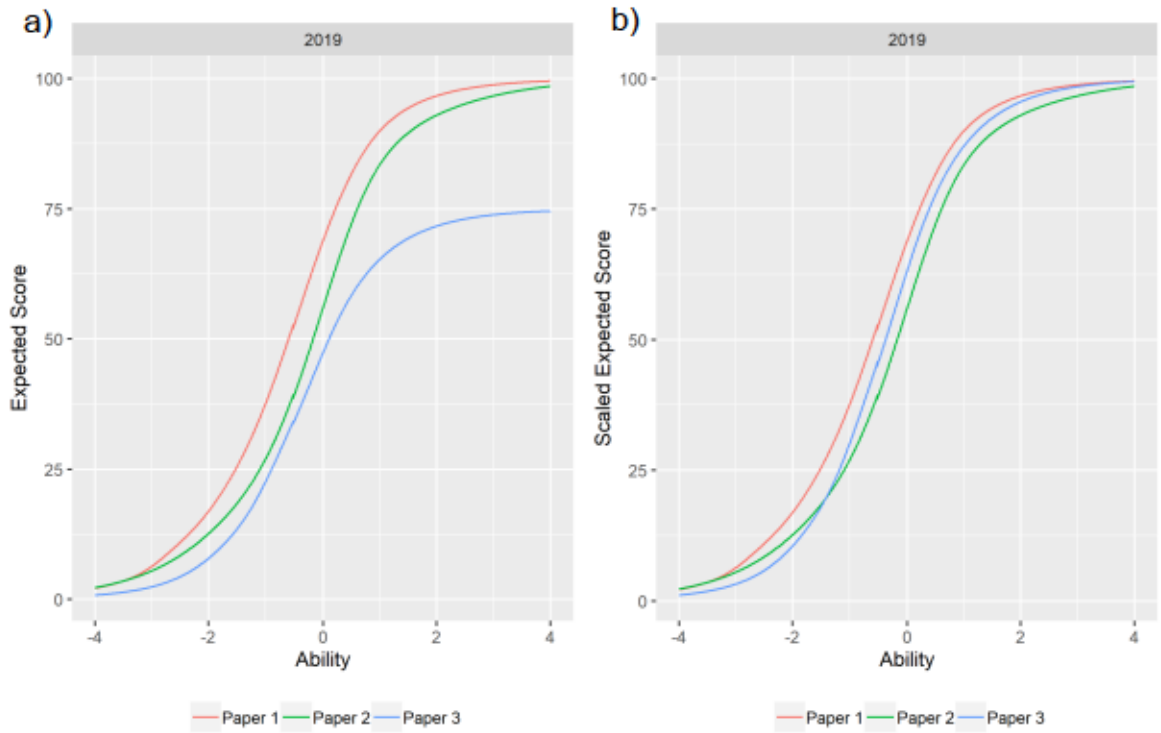


Figure 18 Test characteristic curves for the OCR B (MEI) 2019 question papers from fitting the PCM based on a) raw expected score and b) expected score scales for total mark

5.2.3 Pearson

Figure 19 and Table 23 summarise the item facility indices for the Pearson question papers across the two years. In 2018, there was a high proportion of items that candidates found difficult on Paper 3 compared with the other two papers indicated by the larger spread of facility indices and slightly lower average facility indices. In 2019, the average item facility was similar for Papers 1 and 2 with a notably higher spread of items on Paper 1. The average item facility on Paper 3 was higher than for the other papers, indicating that candidates found these, on average, less difficult.

Taking into account the relative weighting of items and the distribution of items across the difficulty range, the test characteristic curves are provided in Figure 20. These show that, in 2018, candidates across the ability range found Paper 3 the most difficult, followed by Paper 2 then Paper 1. Although the difference in difficulty is very low for lower ability candidates. In 2019, for high ability candidates, Paper 2 was the most difficult with similar difficulty to Paper 3. For lower ability candidates, the difficulty appears broadly similar across all three papers.

Table 23 Summary statistics of item facility indices for the Pearson papers

	2018			2019		
	Median	Mean	S.D.	Median	Mean	S.D.
Paper 1	72.4	67.2	21.1	47.2	46.3	23.3
Paper 2	61.1	63.4	22.1	45.8	47.7	19.7
Paper 3	56.0	53.2	25.0	54.7	52.9	24.2

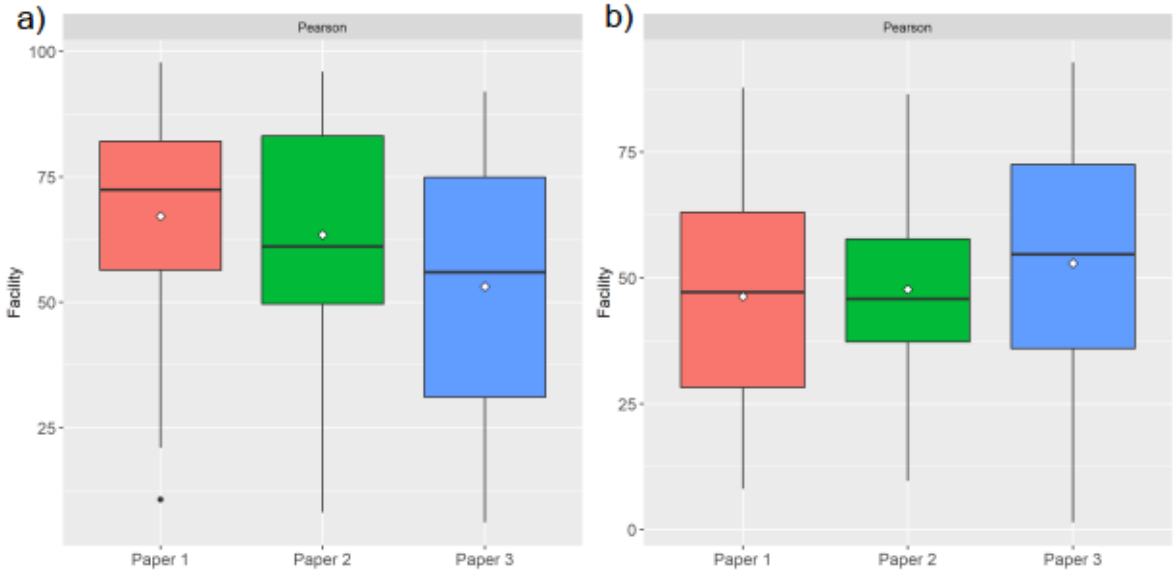


Figure 19 Boxplots of item facility index for Pearson papers from a) 2018 and b) 2019

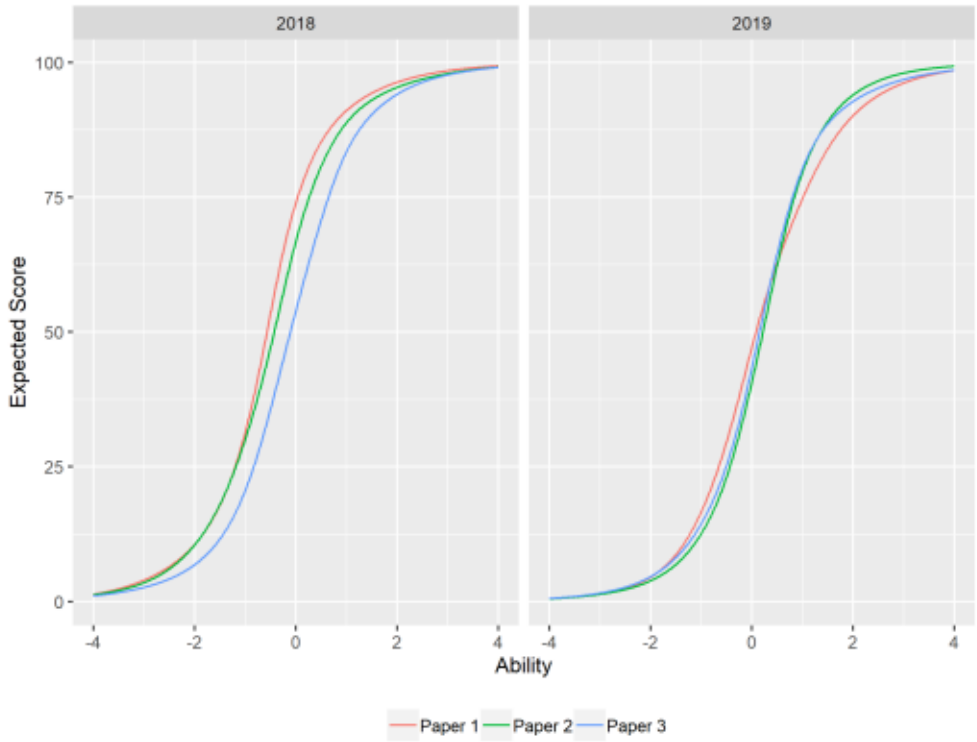


Figure 20 Test characteristic curves for the Pearson 2018 and 2019 question papers from fitting the PCM

5.3 Overall evaluation of difficulty

Practical interpretation of the two sets of data – expected difficulty and item facility indices – across 2018 and 2019 in isolation is difficult. The expected difficulty parameter estimates do not indicate the magnitude of the impact of any differences in terms of marks and the facility indices are subject to effects of cohort ability differences. To aid a more meaningful interpretation addressing these issues, the two measures can be combined.

Three approaches are used here to enable this translation:

1) mark scale linking via the expected difficulty distribution
2) mark change modelling based on 2019 relative difficulty
3) facility to difficulty parameter regression matching

These different approaches and their application are outlined in the following sections.

5.3.1 Approach 1: Mark scale linking via the expected difficulty distribution

Being able to link scales across different versions of an assessment requires a common element – typically items or candidates – across versions as a basis for the linkage. Due to the linear nature of the qualifications, common candidates do exist across papers within each qualification in each year and, therefore, as described in Section 5.2, the PCM has been fitted within year to put all candidates and items on a common scale. These scales across years and qualifications are, however, independent of one another and cannot be directly compared prior to linking (i.e. a candidate with an ability of θ=1.0 on one scale is not necessarily more able than a candidate of ability θ= 0.9 on another.)

In this work, the only scale available to us on which all items have been calibrated is the scale of expected difficulty constructed through the item comparative judgement exercise. Approach 1 uses this expected difficulty scale as a common basis through which the PCM item parameters, and therefore the test characteristic curves, are linked.

The procedure is outlined below and represented graphically in Figure 21. This procedure is carried out separately for each qualification.

Step 1: Fit the partial credit model separately to the 2018 and 2019 item level data across all assessments

Step 2: Determine the distribution and summary statistics for the item difficulties (β parameters) from the model for each year

Step 3: From the Rasch model fitted as part of the item comparative judgement exercise, determine the mean and standard deviation of the expected difficulty item parameters for all items in each year

Step 4: Separately, for each year, translate the item difficulties (β parameters) and corresponding category thresholds from the PCM to match the corresponding distribution of expected difficulties based on the following expression:

$$\beta'_x = \frac{(\beta_x - \overline{\beta_x})\sigma_{\beta x}}{\sigma_\delta} + \bar{\delta}$$

where:

- β'_x is the difficulty parameter from year x
- β_x is the original difficulty parameter from year x
- $\overline{\beta_x}$ is the mean of the difficulty parameters across all assessments in a given year x
- $\sigma_{\beta x}$ is the standard deviation of difficulty parameters across all assessments in year x
- $\bar{\delta}$ is the mean of the equivalent expected difficulty parameters
- σ_δ is the standard deviation of the equivalent expected difficulty parameters

Step 5: Based on the adjusted item difficulties, β'_x, construct test characteristic curves (expected score *v* ability) for the individual components and the aggregate of the components in each year

Step 6: Identify the values of ability (θ) relating to the 2019 grade A and E boundaries for the overall qualification

Step 7: Identify the equivalent 2018 grade boundaries for the identified values of θ to determine the adjustment required to account for the differences in assessment difficulty

Step 8: Bootstrap Steps 3 to 7 (100 iterations) sampling the expected difficulty parameters (δ) based on the standard error for the Rasch parameter estimates

A requirement of this approach is that a valid model can be fitted to the data from both 2018 and 2019. This was not possible for the OCR A and OCR B (MEI) qualifications due to insufficient candidate data from the 2018 series, hence the need to explore other methodologies. It is recognised that the scale used to perform the linking – the expected difficulty scale – has been defined on a different basis to the item parameters estimated using the PCM. However, as this intermediary scale is used purely to match the distributions of item parameters, no claim of equivalence of scales is made, nor is it explicitly necessary.

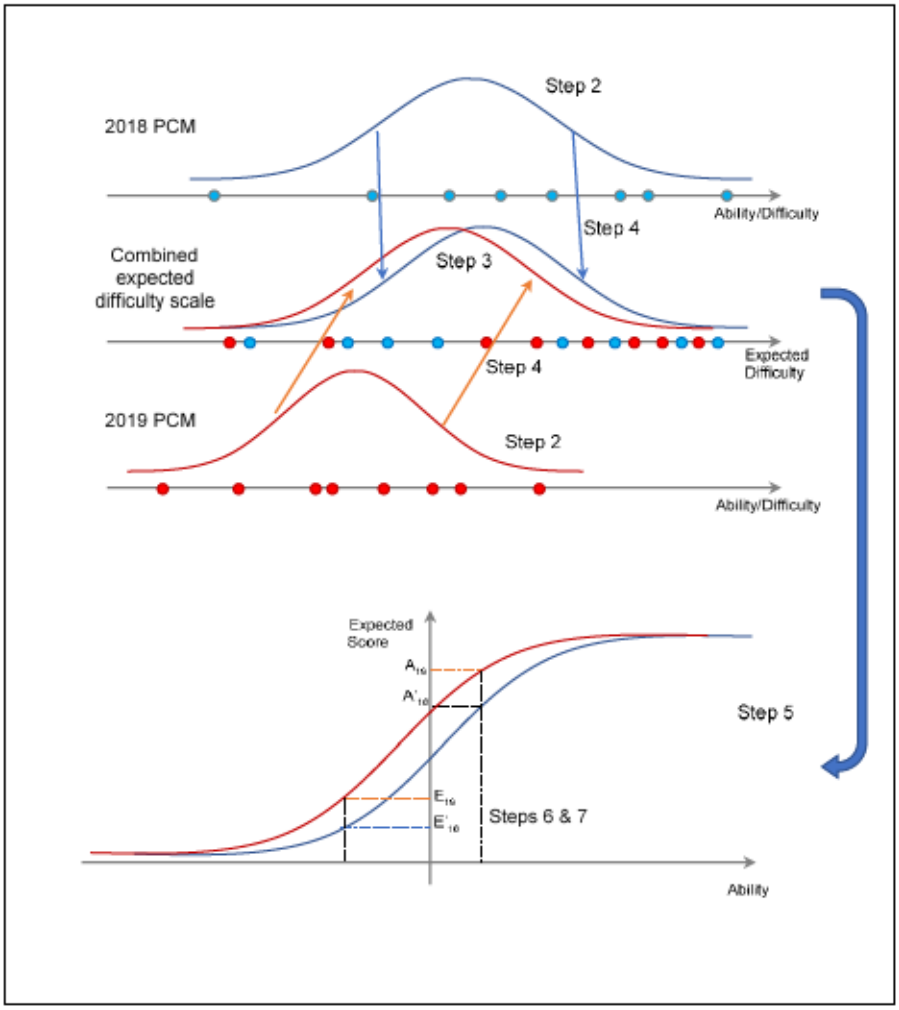


Figure 21 Graphical representation of the linking process via the expected difficulty scale

5.3.2 Approach 2: Mark change modelling based on 2019 relative difficulty

Where it has not been possible to fit the PCM for 2018 data there is no immediately available scale of actual difficulty to use as a basis for linking from that year. In these instances, an alternative approach has been applied that seeks to use the characteristics of the 2019 PCM link to the expected difficulty Rasch scale to approximate the impact of differences in difficulty on the 2018 marks. This approach constructs the relationship between differences in expected difficulty and differences in expected score based on the 2019 data and uses that relationship to model the impact of differences in difficulty across years.

The procedure to deliver this approach is outlined below and demonstrated graphically in Figure 22.

Step 1: Fit the partial credit model to the 2019 item level data for each qualification across all assessments

Step 2: Determine the distribution and summary statistics for the item difficulties (β parameters) from the model for each qualification

Step 3: From the Rasch model fitted as part of the item comparative judgement exercise, determine the mean and standard deviation of the

expected difficulty of the item parameters for all items for each qualification in 2019

Step 4: For 2019, translate the item difficulties (β parameters) and corresponding category thresholds from the PCM to match the corresponding distribution of expected difficulties based on expression given in Step 4 of Approach 1

Step 5: Based on the adjusted item difficulties, β'_x, construct test characteristic curves for the individual components and overall qualification from 2019

Step 6: Bootstrap Steps 3 to 5 (100 iterations), sampling the expected difficulty parameters (δ) based on the standard errors for the parameter estimates to create a number of test characteristic curves

Step 7: Using the AQA qualification level test characteristic curve as the arbitrary reference, build the relationship between difference in mean expected difficulty and expected score at the ability, by recording the difference in expected score and the difference in mean expected difficulty parameter for each qualification and iteration of the test characteristic curves produced through the bootstrapping process. This is performed separately at grades A and E to reflect the impact of different gradients of the test characteristic curves[33]. Centre the distribution about the origin to remove any differences in standard and regress the change in expected score (ΔX) on the difference in aggregate expected assessment difficulty ($\Delta\bar{\delta}$) to represent this relationship

Step 8: From the Rasch model fitted as part of the item comparative judgement exercise, determine the mean expected difficulty of the item parameters for all items for each qualification in 2018 and 2019. For each assessment, and for the qualification overall, determine the difference in mean expected item difficulty between 2019 and 2018 for each iteration of the bootstrapped parameters

Step 9: Using the differences in difficulty calculated in Step 8 and the relationship between the change in mean expected difficulty and mean mark change estimated in Step 7, estimate the difference in expected score at the grade A and E boundaries

Step 10: Bootstrap Steps 8 and 9 (1,000 iterations), sampling the expected difficulty parameters (δ) for both the 2018 and 2019 assessments based on the standard errors for the parameter estimates

[33] To ensure this process is not overly impacted by the properties of the assessment used as a reference the grade A and E boundaries are calculated simply as the mean of (scaled) grade boundaries from 2019.

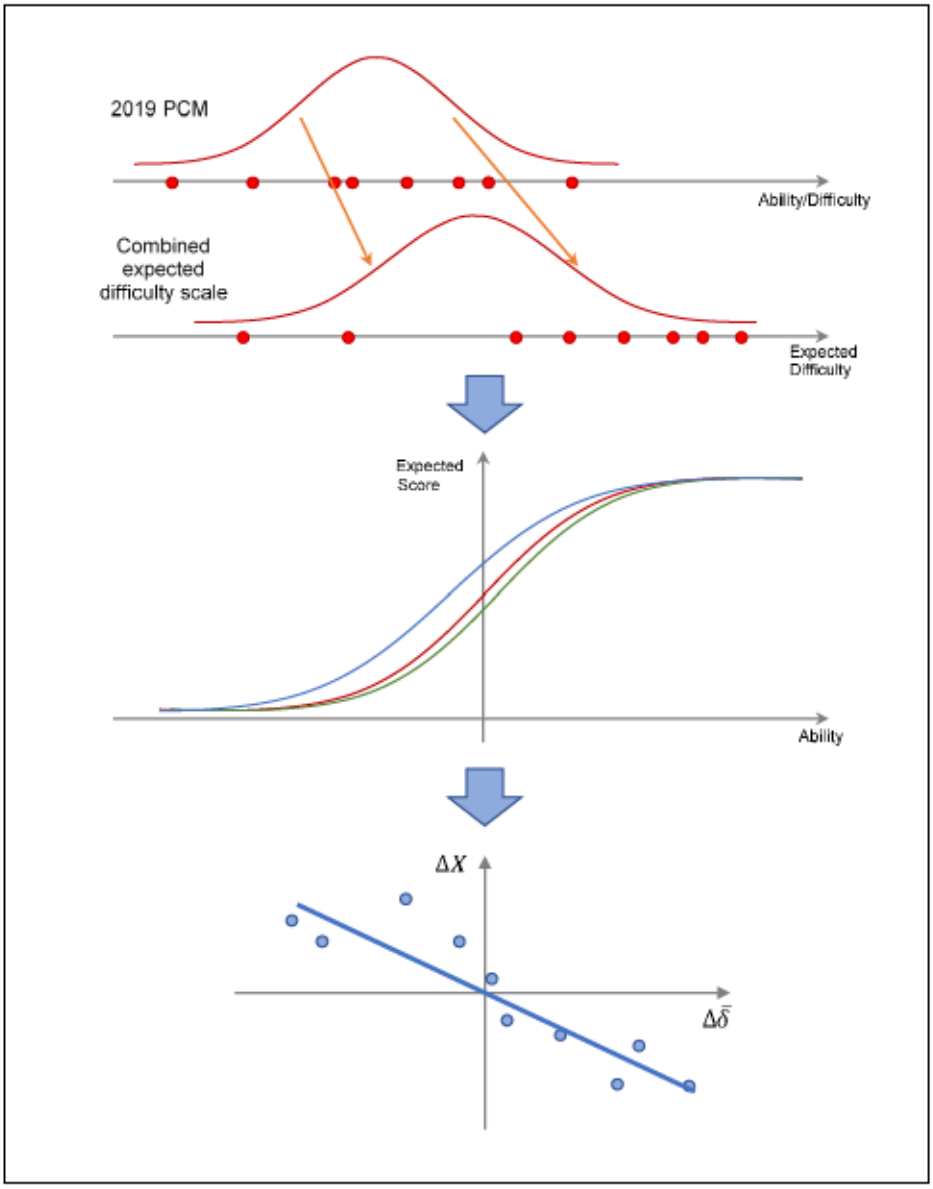


Figure 22 Graphical representation of the 2019 relative difficulty modelling approach

5.3.3 Approach 3: Facility to difficulty parameter regression matching

The final approach, to convert expected difficultly parameters into changes on the mark scale, is the most crude. The approach assumes adjustments at grades A and E match the changes in mean mark due to differences in the assessment difficulty, but overcomes the need to fit a psychometric model to the data.

The approach relies on the direct relationship between the expected difficulty parameters and the item facilities. A typical relationship between item facility and expected difficulty parameter is shown in Figure 23. As discussed above, it is not possible to interpret the relative item facility indices directly due to their dependence on the ability of the cohort sitting the item. In the example shown, all points on a vertical line (constant expected difficulty) have been judged to be at the same level of expected difficulty, independent of any cohort effects, through the comparative

judgement exercise. Any difference between the relationship in the vertical direction can therefore be considered to be cohort ability effects.

In the illustrative example given, the 2018 cohort was composed of better scoring candidates than in 2019 for items of a given difficulty. This approach attempts to remove this cohort effect to allow direct comparison of mean item facilities.

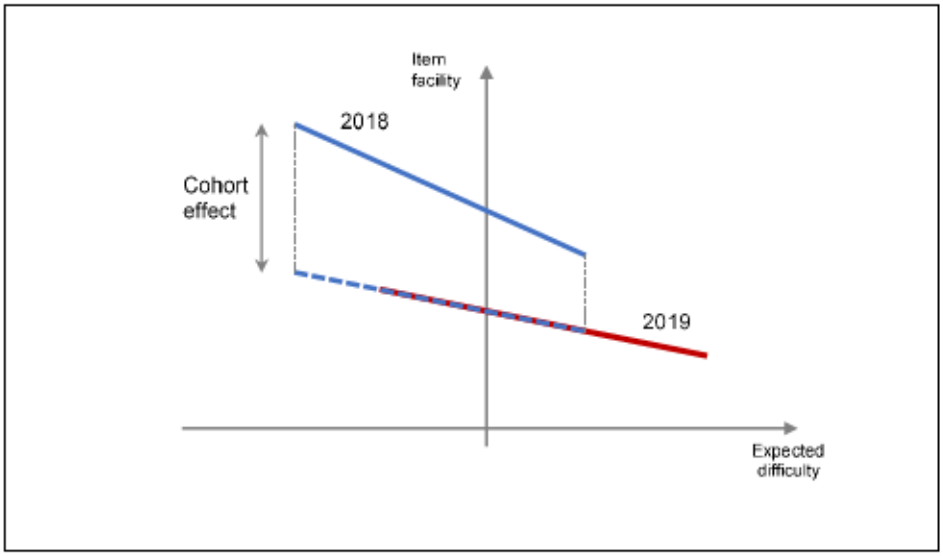


Figure 23 Typical relationship between expected difficulty and item facility index across years

The procedure is outlined below, along with the supporting graphical representation in Figure 24.

Step 1: Fit a linear regression of item facility index on expected difficulty parameter across all papers in a qualification, separately for 2018 and 2019. It is possible to consider all items from a qualification simultaneously for this purpose due to the linear nature of the qualifications and, therefore, the commonality of candidates across papers

Step 2: Determine the relationship required to remove the cohort effect and map the 2018 item facilities onto the 2019 relationship using the following expression:

$$X_j' = \delta_j(M_{19} - M_{18}) + C_{19} - C_{18}$$

where:

- X_j' is the cohort effect adjusted item facility for item j
- δ_j is the expected difficulty parameter for item j
- M_x is the gradient of the relationship between expected difficulty and item facility index in year x
- C_x is the intercept of the relationship between expected difficulty and item facility index in year x

This approach allows a different level of adjustment across the expected difficulty range reflecting the fact that the cohort effect might vary with item difficulty[34]

Step 3: For each component within each qualification, perform a weighted aggregate (based on tariff) of the adjusted 2018 item facility indices and the 2019 original facility indices. The difference in the aggregate provides an estimation of the mean change in marks for each component based on differences in expected difficulty of the items with the cohort effect removed

Step 4: Bootstrap steps 1 to 3 sampling the expected difficulty parameters (δ) based on the standard error for the parameter estimates

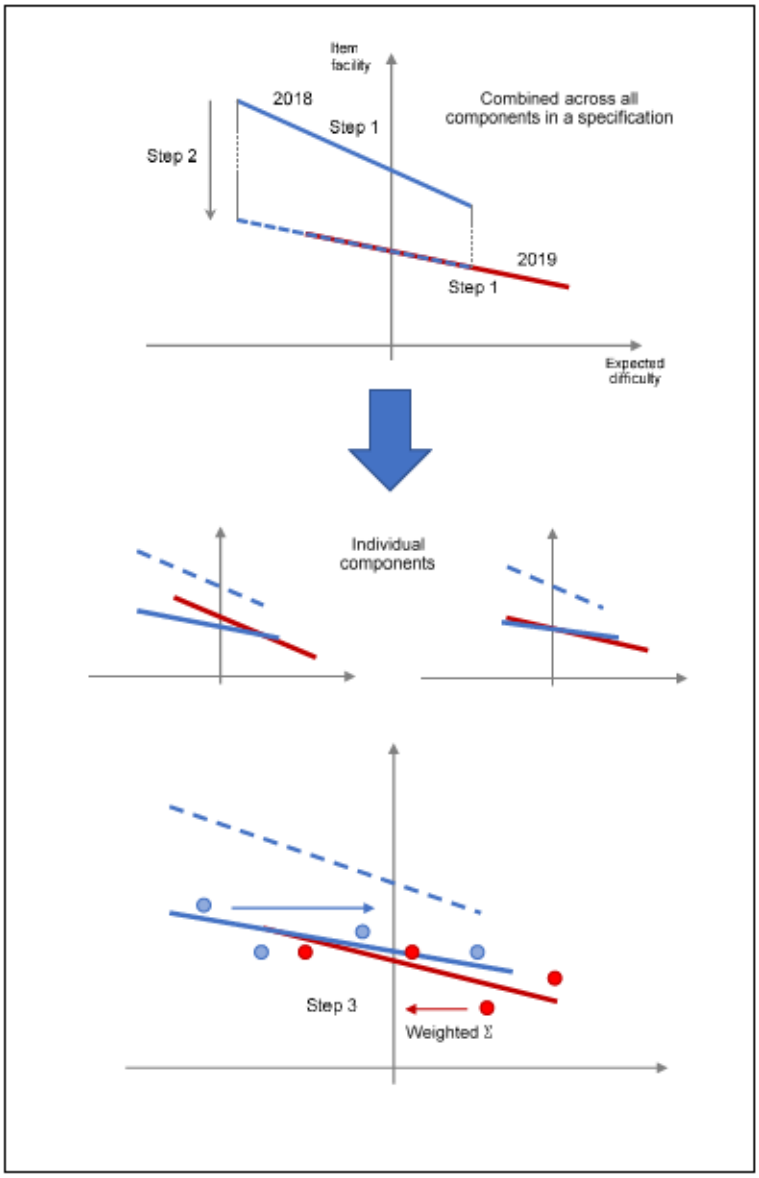


Figure 24 Graphical representation of the regression line matching approach

[34] It is acknowledged that this relationship might vary on different dimensions other than expected difficulty (such as topic), however, the approach outlined here reflected the pragmatic nature of the adjustment process.

This approach makes the distributional assumption that a change in mean mark for the paper overall corresponds to a similar change in mark at each grade boundary.

5.3.4 Analysis

5.3.4.1 Approach 1

Due to insufficient data available for fitting of the partial credit model for OCR A and OCR B (MEI) in 2018, Approach 1 can only be performed for AQA and Pearson. The test characteristic curves resultant from Step 5 for AQA and Pearson are shown in Figure 25 and Figure 27, respectively. The mean and standard deviation of the corresponding grade boundary adjustments, resultant from the bootstrapping process, are shown in Figure 26 and Figure 28 with the results summarised in Table 24.

The results in Section 5.1.2.2 suggest that, overall, the difficulty of the AQA assessments reduced slightly between 2018 and 2019. This modelling suggests that, to account purely for this reduction in difficulty, the overall grade A boundary would need to have been increased by between 10 and 24 marks and the grade E boundary by between 7 and 12 marks.

The results in Section 5.1.2.5 suggest that, overall, the difficulty of the Pearson assessment increased slightly between the two years. Based on this modelling, to account purely for this difference in difficulty between 2018 and 2019, the overall grade boundaries would have had to have been lowered by between 7 and 23 marks at grade A and by between that between 4 and 11 marks at grade E to account for the overall increase in difficulty.

It should be noted that the component level modelled adjustments should be interpreted with extreme caution and are likely to be unreliable in their current form. This does not, however, detract from the qualification level findings. The reason for the potential unreliability in the component level adjustments is the potential for inter-component differences in standard. In addition to the size of the adjustment being related to the magnitude difference in difficulty, it is also very sensitive to the gradient of the test characteristic curve at the point used to sample the expected score. As described in Step 6 in the approach described in Section 5.3.1, selection of this point is dictated by the qualification level boundary position. Different component level standards would, therefore, mean sampling the test characteristic curve at points other than where the component level boundary was located potentially leading to over or under sensitivity in terms of expected score differences. This can be interpreted in two different ways. Either, the model is an insufficient representation of the data as single values of ability (as defined by the grade boundaries) map to different points on the ability scale or the inter-component standards are misaligned. With linear qualifications, the implications of inter-component level misalignment of standards is not significant due to standards being set at the qualification level.

The focus in this investigation is on the overall differences in qualification level grade boundaries between years. Should it be the component level adjustments that are of interest, the approach outlined above could be modified accordingly.

Table 24 Modelled adjusted grade boundaries for AQA and Pearson derived using Approach 1

		2019 Boundaries		Modelled Adjustment (2018 to 2019)			
		A	E	A		E	
				Mean	SD	Mean	SD
AQA (7357)	Paper 1	53	15	1.88	2.05	3.44	0.74
	Paper 2	62	16	2.39	2.65	2.81	0.79
	Paper 3	70	21	12.73	2.30	3.94	0.87
	Combined	**185**	**52**	**17.44**	**7.04**	**9.78**	**2.43**
Pearson (9M0A)	Paper 1	56	15	-11.05	2.57	-2.82	1.37
	Paper 2	52	13	-8.59	2.65	-5.51	1.34
	Paper 3	57	15	4.58	2.66	1.03	1.07
	Combined	**165**	**43**	**-15.17**	**7.81**	**-7.46**	**3.81**

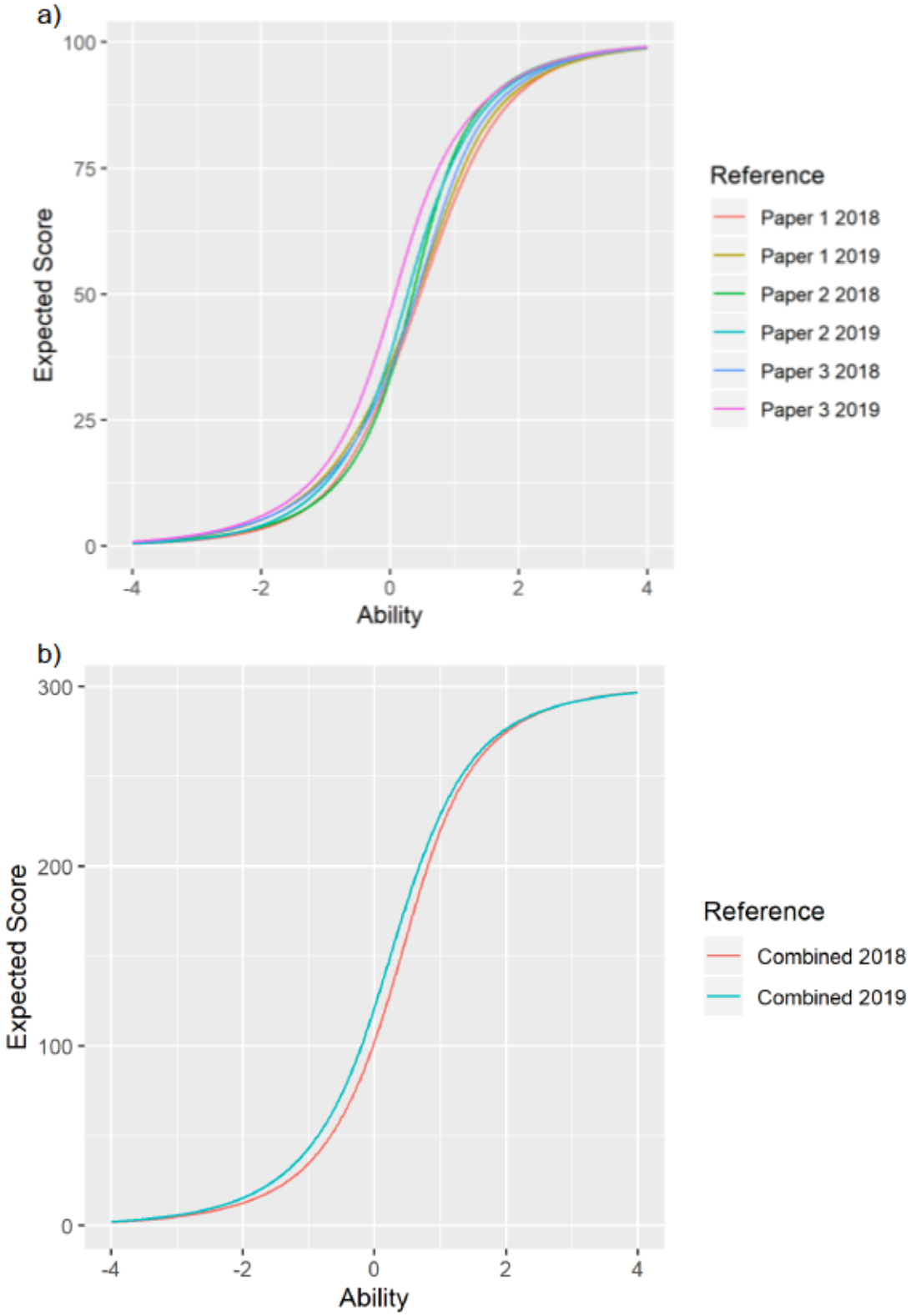
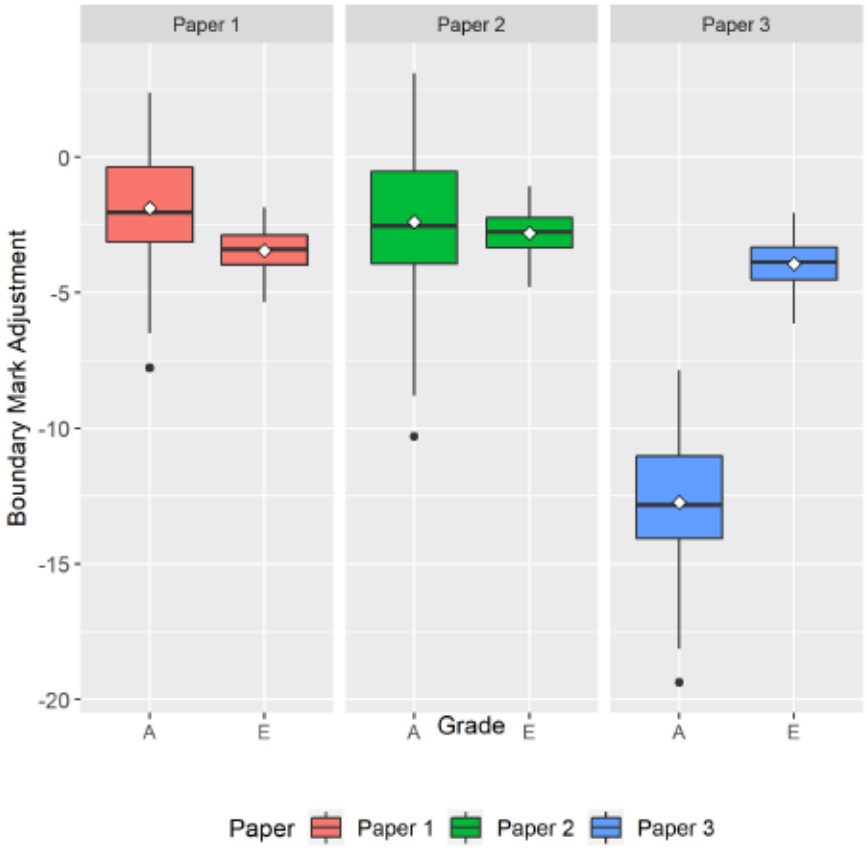


Figure 25 AQA linked test characteristic curves for a) each assessment and b) the combined assessments within each year



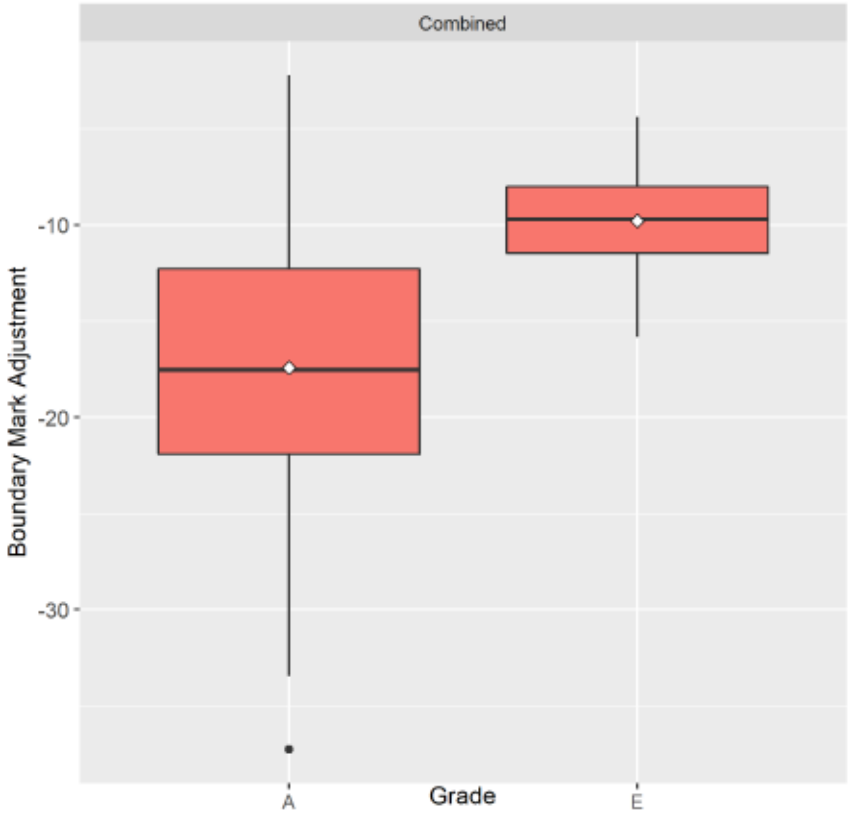


Figure 26 AQA grade boundary adjustments to account for differences in expected difficulty for a) each assessment and b) the combined assessments

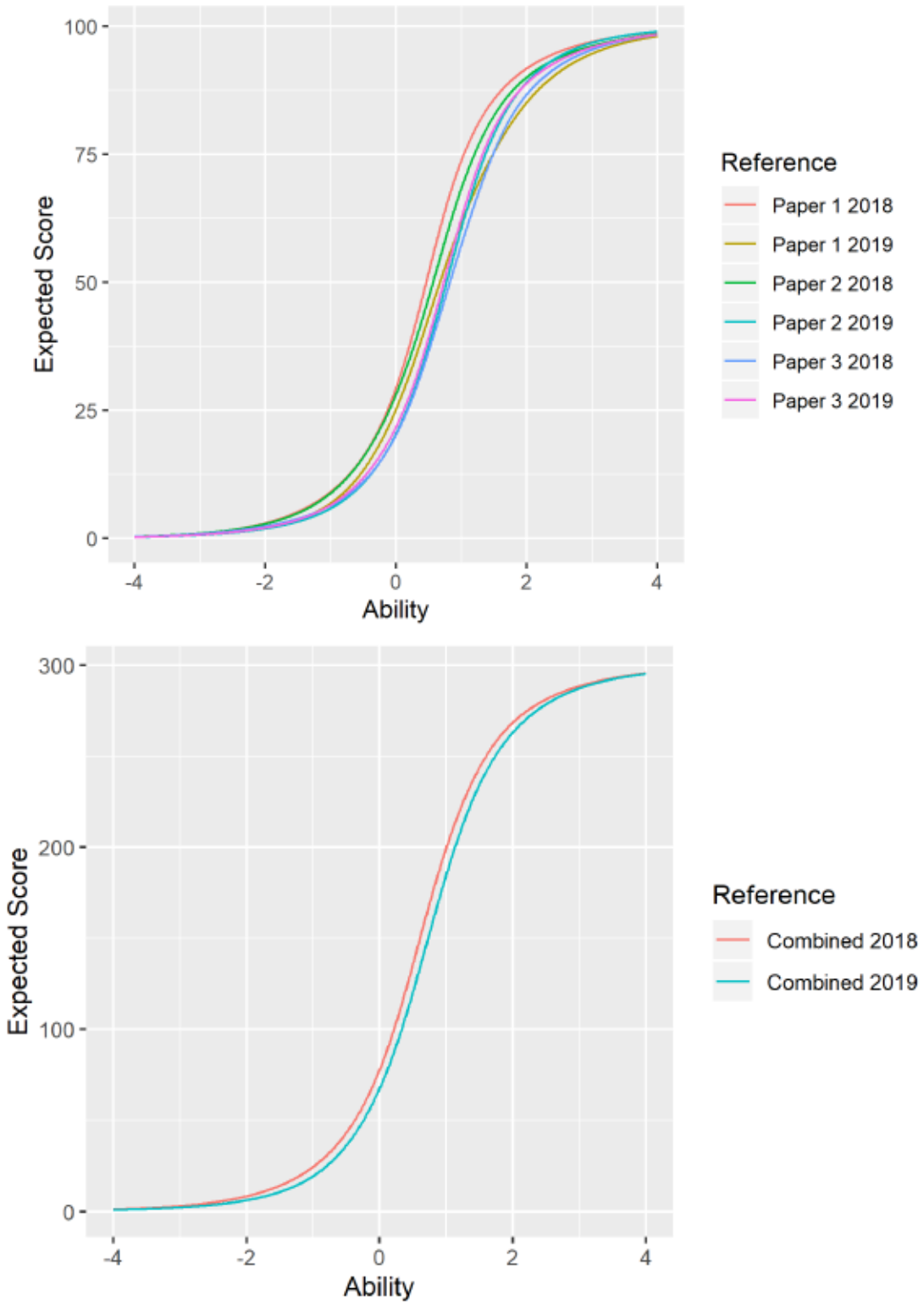


Figure 27 Pearson linked test characteristic curves for a) each assessment and b) the combined assessments within each year

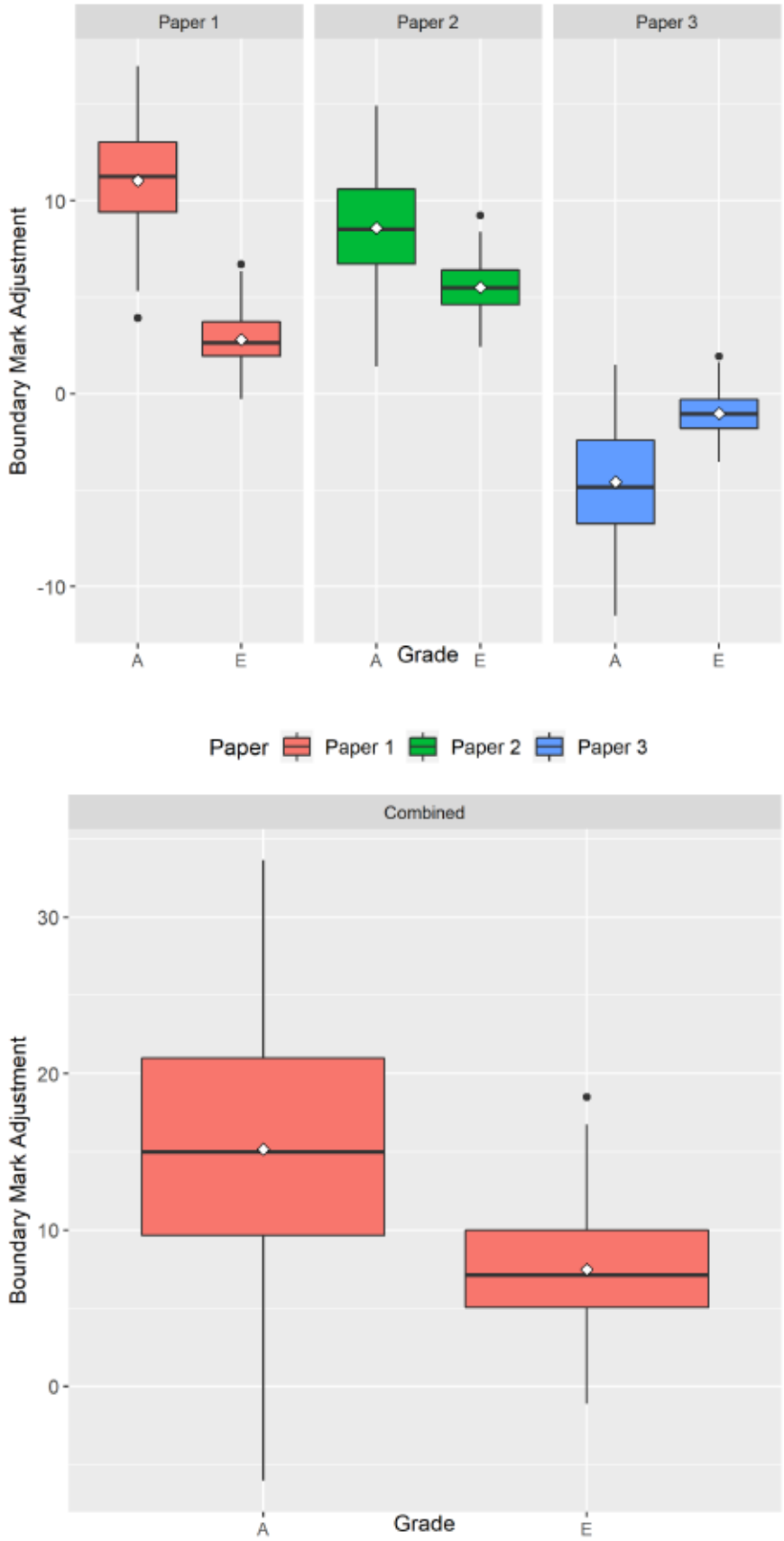


Figure 28 Pearson grade boundary adjustments to account for differences in expected difficulty for a) each assessment and b) the combined assessments

5.3.4.2 Approach 2

This approach was able to be implemented for all qualifications. Shown in Figure 29 are centred distributions of the change in expected score based on differences in expected item difficulty. The differences in these distributions as grades A and E are due to differences in gradient of the test characteristic curves for different levels of ability.

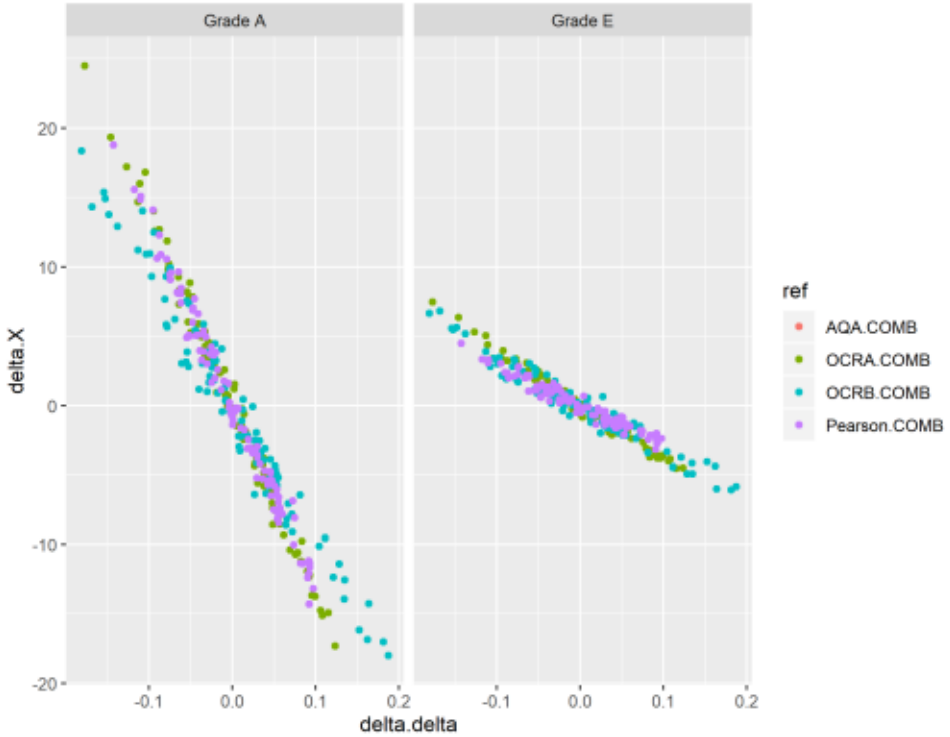


Figure 29 Modelled relationships between changes in expected difficulty and resultant changes in expected score

The results of applying the modelling described in Section 5.3.2 are provided in Table 25.

For AQA, this modelling suggests that, to account purely for differences in difficulty between 2018 and 2019, overall the grade A boundary would have needed to increase by between 12 and 25 marks and, at grade E, increase by between 3 and 7 marks. For OCR A the equivalent changes would have been a lowering between 3 and 17 marks at grade A and between 1 and 5 marks at grade E. For OCR B (MEI) difficulty was relatively constant across years, overall, meaning the required changes at grade A would have been between +4 and -11 marks and, at grade E, between +1 and -3. The overall grade boundaries for Pearson would have to have reduced between 13 and 26 marks at grade A and between 4 and 8 marks at grade E.

Table 25 Modelled adjusted grade boundaries derived using Approach 2

		2019 Boundaries		Modelled Adjustment (2018 to 2019)			
		A	E	A		E	
				Mean	SD	Mean	SD
AQA (7357)	Paper 1	53	15	12.37	3.79	3.56	1.09
	Paper 2	62	16	-1.98	4.12	-0.57	1.18
	Paper 3	70	21	7.45	4.15	2.14	1.19
	Combined	**185**	**52**	**18.61**	**6.97**	**5.36**	**2.01**
OCR A (H240)	Paper 1	54	13	-2.70	4.31	-0.78	1.24
	Paper 2	58	15	0.17	4.19	0.05	1.21
	Paper 3	49	12	-10.08	4.56	-2.90	1.31
	Combined	**161**	**40**	**-10.21**	**7.59**	**-2.94**	**2.18**
OCR B (MEI) (H640)	Paper 1	70	23	-6.39	3.96	-1.84	1.14
	Paper 2	59	17	-11.04	4.02	-3.18	1.16
	Paper 3	49	12	15.99	3.35	4.61	0.96
	Combined	**178**	**52**	**-3.71**	**7.40**	**-1.07**	**2.13**
Pearson (9MA0)	Paper 1	56	15	-3.86	3.97	-1.11	1.14
	Paper 2	52	13	-14.36	3.99	-4.13	1.15
	Paper 3	57	15	0.36	3.95	0.10	1.14
	Combined	**165**	**43**	**-19.62**	**6.69**	**-5.65**	**1.93**

5.3.4.3 Approach 3

Shown in Figure 30 are the assessment level relationships between expected item difficulty parameters and item facility values. As highlighted in Section 5.3.3 (and Figure 23) the vertical offset between the regression lines between 2018 and 2019 relates to the cohort effect impacting on the item facility index values. While there is some variation in the relationships between years, there is a reassuring similarity of relationship across assessments within the same qualification due to the commonality of candidates sitting all assessments within each year. With the exception of AQA, the facility indices for a given expected difficulty are notably higher in 2018 compared to 2019, fitting with the expectations of the higher ability of the 2018 cohort. The absence of an obvious cohort effect for AQA can be explained by consideration of the 2018 outcomes for the reformed qualification provided in Annex B. It can be seen in Table 36 in the Annex that almost half of the candidates were of an age other than 17-years-old and were significantly weaker that the 17-year-old cohort. This will have significantly reduced the facility indices for AQA in a way that was not the case for the other qualifications.

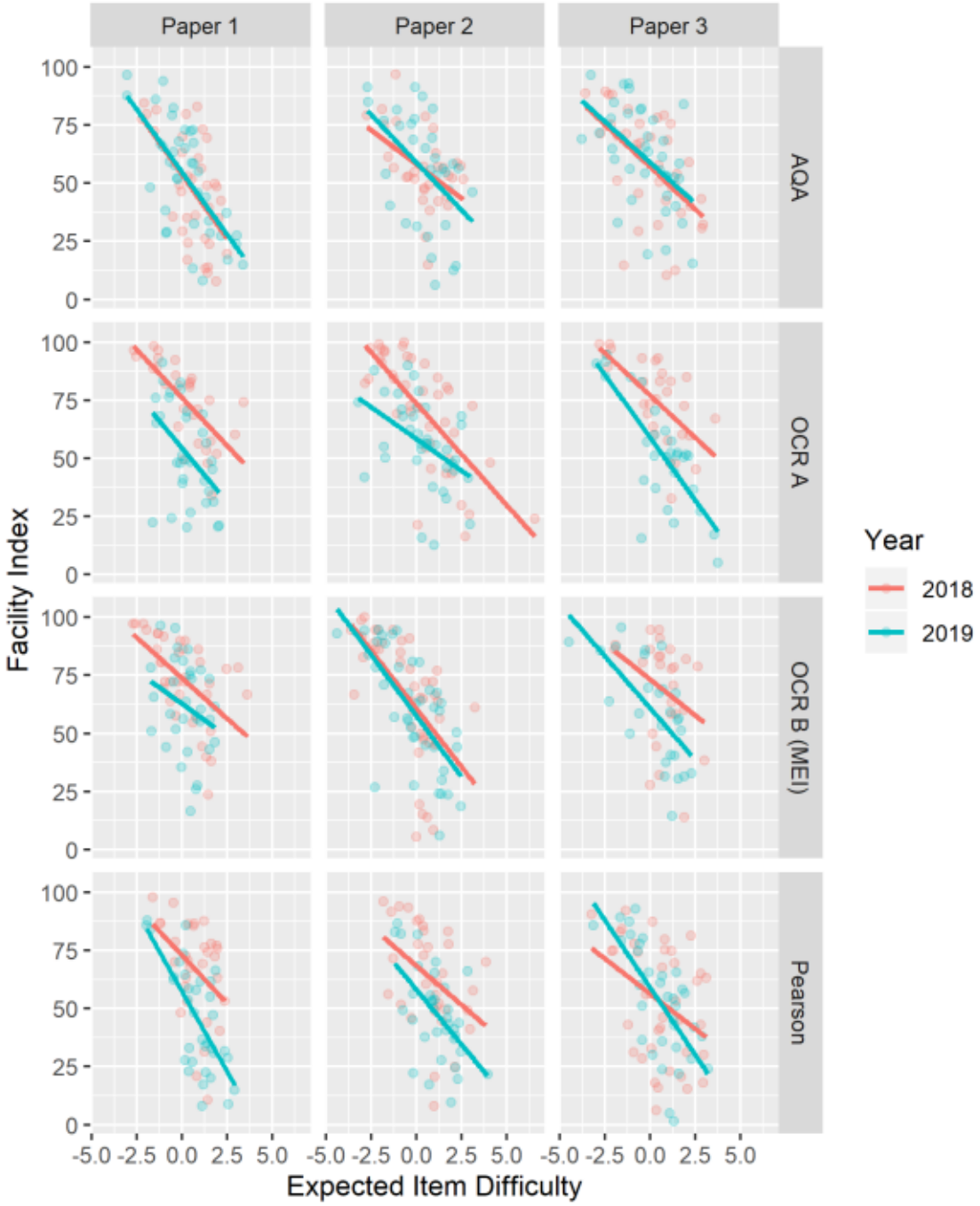


Figure 30 Plots of expected item difficulty against facility index by question paper across 2018 and 2019

Figure 31 shows the required adjustment to the item facilities accounting for the cohort effect and Figure 32 shows the pre- and post- adjusted item facilities having performed the procedure described in Section 5.3.3.

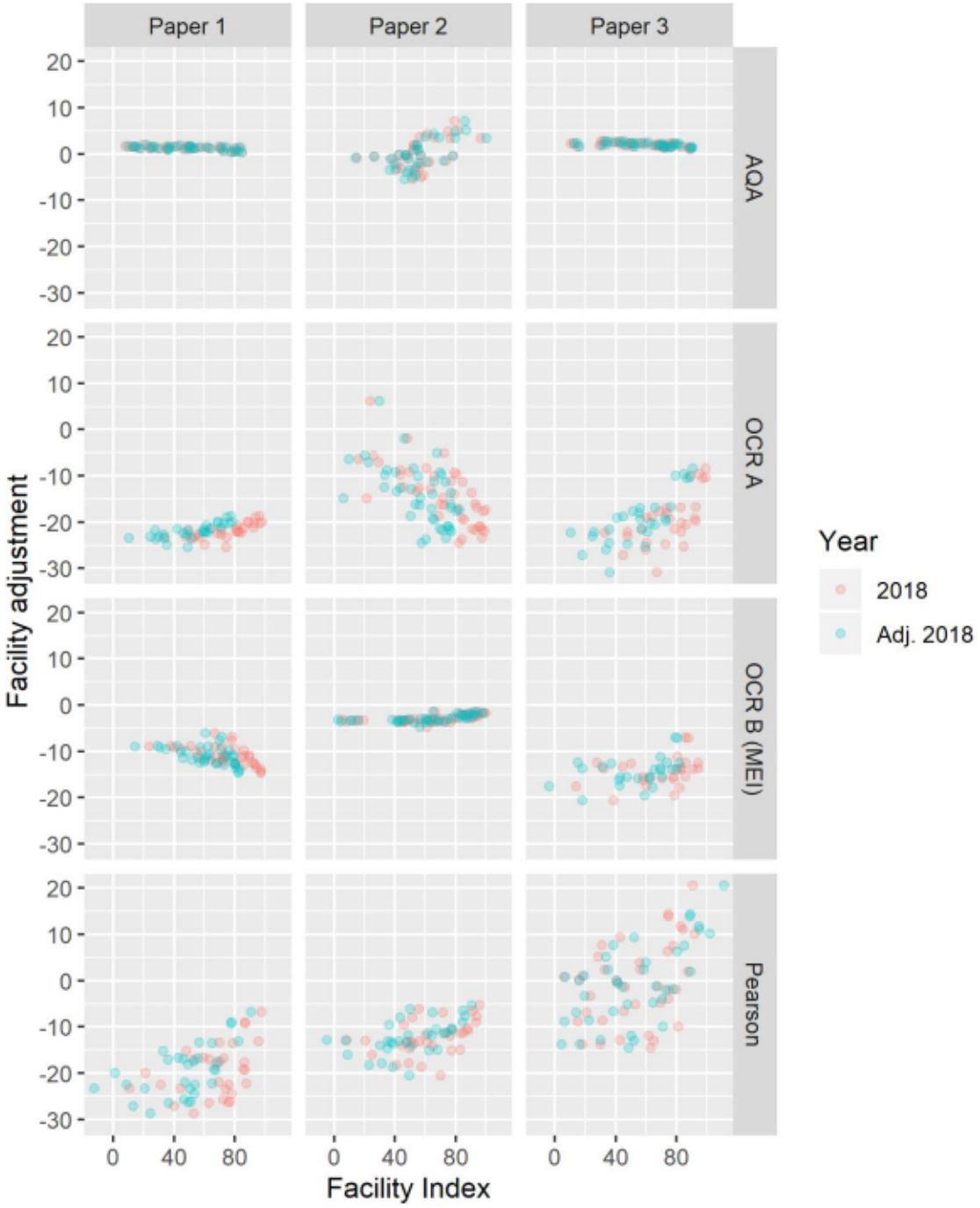


Figure 31 Plots of item facility adjustment by question paper to account for cohort effects

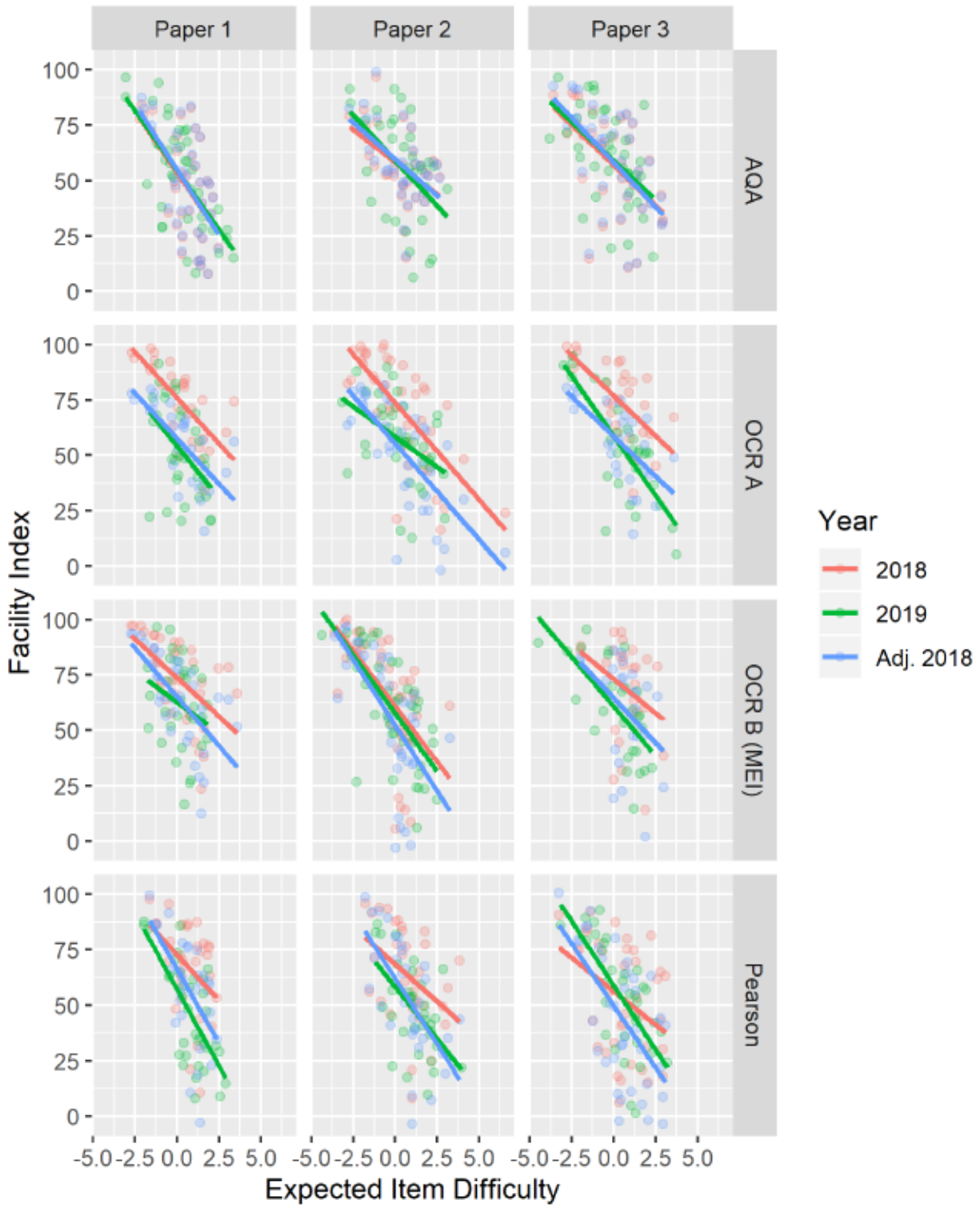


Figure 32 Plots of expected item difficulty against facility index by question paper including the adjusted 2018 facility relationships

The impact of these adjustments on mean mark at question paper and qualification level is provided in Table 26. Making the assumption that a difference in mean mark corresponds to the same difference in boundary mark between sittings, the adjustments suggested by this method are significantly lower than those suggested by Approaches 1 and 2. This analysis suggests that, to account for differences in difficulty, AQA's qualification level boundaries should have increased by between 0

and 4 marks, the OCR A qualification level boundaries should have reduced by between 1 and 4 marks, the OCR B (MEI) boundaries changed between +2 and -1 marks, and Pearson's qualification level boundaries should have been lowered by between 2 and 7 marks.

Table 26 Adjusted mean facility values removing the cohort effect

		Mean of weighted facilities				
	Paper	2018	2019	Modified 2018	Modelled Adjustment (2018 to 2019)	
					Mean	S.D.
AQA	1	47.90	47.26	48.67	-1.41	0.60
	2	54.64	52.44	55.31	-2.87	0.64
	3	53.10	60.21	53.89	6.32	0.59
	Total	**155.64**	**159.91**	**157.87**	**2.04**	**1.83**
OCR A	1	73.00	53.17	54.65	-1.48	0.54
	2	68.25	54.89	49.95	4.94	0.67
	3	79.98	48.84	54.69	-5.85	0.66
	Total	**221.23**	**156.90**	**159.29**	**-2.39**	**1.84**
OCR B (MEI)	1	70.58	60.09	60.96	-0.87	0.61
	2	64.03	55.38	55.38	0.00	0.54
	3	51.78	45.53	44.09	1.44	0.56
	Total	**186.39**	**161.00**	**160.43**	**0.57**	**1.54**
Pearson	1	69.88	50.26	58.34	-8.08	0.71
	2	65.10	46.98	52.09	-5.11	0.92
	3	54.73	50.84	42.05	8.79	0.81
	Total	**189.71**	**148.08**	**152.48**	**-4.40**	**2.20**

5.4 Intermediate findings from Strand 2

This Strand of work has considered the difficulty of the assessments making up each of the reformed A level maths qualifications from two perspectives – the expected difficulty of the items as judged by teachers and subject experts and the operationally available candidate result data. These two sources have then been brought together to translate the arbitrarily scaled differences in expected item difficulty into differences in marks. This was to estimate the appropriate degree of difference in grade boundaries required to account for question papers being of different levels of difficulty across years. In doing so, this analysis has been directly targeted at understanding Scenario A as defined in Section 2.5 – the extent to which changes in difficulty of the assessments may account for differences in grade boundary position between 2018 and 2019.

The key findings for this strand of work are as follows:

i. based on the initial evaluation of expected difficulty, differences in difficulty of the assessments (individually and collectively within qualifications) between years were identified. For the AQA assessment there appeared to be an,

overall, reduction in difficulty of the assessments between 2018 and 2019. For the OCR qualifications the overall difficulty appeared relatively stable across years, and for the Pearson assessments, the overall difficulty appeared to increase in 2019 compared to 2018

ii. on the basis of the summary statistics of the expected difficulty distributions, these characterisations of relative difficulty between years were broadly in line with the representations provided by exam boards during the summer 2019 series (see Section 2.5)

iii. modelling of the impact of these differences in terms of marks showed that, due purely to differences in difficulty across years, the AQA qualification level boundaries would need to increase notably at grade A with a slight increase at grade E. Despite appearing of similar difficulty based on the summary statistics, the grade boundaries for OCR A needed to be moderately reduced at grades A and E with this inconsistency in finding likely due to the change in distribution of item difficulty masked by the summary statistics. The consistent levels of difficulty across years for OCR B (MEI) led to only a marginal reduction in grade boundaries between 2018 and 2019. The greatest downwards adjustment of boundaries due to differences in difficulty between 2018 and 2019 was for Pearson with a notable downward adjustment (likely 15-20 marks) at grade A and an adjustment of around 6 marks at grade E. It is, however, important to note the levels of uncertainty associated with these estimates as reported in the tables above

iv. on the basis of this evidence, differences in difficulty of the assessments between years, accounts for a considerable proportion of the reduction in grade boundaries at grade A, particularly for the OCR A qualification (~30%) and the Pearson qualification (up to 100% depending on the analysis model applied).

6 Strand 3: Analysis of candidate performance

So far, the statistical relationship between cohorts within and across years has been considered as a potential indicator of differences in standard that could have resulted from the differences in grade boundaries between years. The relative difficulty of the assessments has also been considered to understand the extent to which changes in difficulty might have contributed to a necessary difference in grade boundaries. To this point, however, there has been no consideration of the materiality of any differences from the perspective of candidate performance. That issue is the focus of this final strand of the investigation. The work presented here seeks to determine the consequences of differences in boundaries set across the two years for the quality of work at those boundaries. This strand asks the question of whether subject experts are able to distinguish between the quality of this work across years, taking into account the inherent uncertainty of the judgement process.

The use of examiner judgement to precisely and reliably identify the quality of candidate work and, by inference, the ability of the candidate is a well-researched area and is widely recognised as facing significant challenges. Those challenges need to be considered as part of the design process for this strand of the investigation to ensure the results provide a sufficiently reliable and accurate basis for interpretation. The ability of human judges to make absolute judgements about the quality of a piece of work is limited and has been shown to be unreliable[35]. This builds on work by Thurstone[36] and Laming[37] with the latter claiming that there are no absolute judgements made by humans – all of our judgements are relative comparisons – and that humans can carry out these relative judgements much easier with far greater reliability than attempts to make judgements against absolute criteria.

Based on this thinking, methodologies that rely on comparisons of quality of candidates' work relative to one another have been increasingly applied in studies of comparability in the UK[38]. Similar approaches are adopted here.

These methodologies can either involve the repeated comparison of two artefacts in isolation about which a relative judgement on some property is made – similar to the comparative judgement methodology applied in Strand 2, where the property being

[35] Baird, J. and Dhillon, D. Qualitative Expert Judgements on Examination Standards: Valid, but Inexact. AQA research report RPA_05_JB_RP_077. Guildford: AQA. (2005); Cresswell, M. "Examining Judgements: Theory and Practice of Awarding Examination Grades." PhD thesis, University of London Institute of Education, London (1997)

[36] Thurstone, L.L., Psychological Review 3: 273-286 (1927)

[37] Laming, D., Human judgement: the eye of the beholder, London: Thomson (2004)

[38] For example: Bramley, T. In Techniques for monitoring the comparability of examination standards (pp. 246–300). London, U.K.: Qualifications and Curriculum Authority (2007); Bramley, T. and Gill, T. Research Papers in Education, 25(3), 293-317 (2010); Jones, Wheadon, Humphries, and Inglis British Educational Research Journal 42 (4), 543-560 (2016); Pollitt, A., & Elliott, G., Monitoring and investigating comparability: A proper role for human judgement. Cambridge: Research and Evaluation Division, University of Cambridge Local Examinations Syndicate (2003)

compared was item difficulty – or multiple repeated comparisons of a larger number of artefacts can be made; such as the rank ordering of a number of scripts on the basis of quality. These approaches have different benefits and limitations, however, both are based on a sequence of relative judgements.

As described in the following sections, both approaches were explored as methodologies for realising the purpose of this strand of the investigation.

Irrespective of the details of the design, there are two primary issues raised when comparative methodologies are applied for the purposes of analysing the quality of candidates' work. These are issues of bias and validity.

The primary concerns regarding bias are due to judgements of performance being influenced by the difficulty of the task (or assessment) that candidates have been asked to perform. Important work in this area[39] has demonstrated that, when forming judgements of the quality of candidates' work, experts systematically underestimate the performance of candidates asked to carry out a more difficult task in comparison to those asked to carry out an easier one. This means that two candidates of the same ability carrying out tasks that are of differing demand may be judged unfairly due to this bias.[40] This effect is not something that is resolved through the application of comparative judgement methodologies. In the current work, this bias is attempted to be mitigated in three ways:

1) where possible, the assessments identified for comparison were selected on the basis of comparability of difficulty with a view to minimising the effects of bias. This approach is applied in the later phases of judgement activity, as discussed below
2) when interpreting the results of the judgement exercise, the potential for bias is acknowledge and qualitatively accounted to ensure that conclusions are not drawn which might be equally explainable by the presence of bias
3) when judges were briefed to take part in the work they were asked to explicitly consider the differences in difficulty when making their judgements. It is acknowledged that this is likely to have a minimal effect due to it being an inherent property of the judgement process rather than something that can be consciously controlled by participants, however, it may prove of some practical benefit

The validity of the holistic judgements that characterise judging processes such as those considered here is the most contentious area when applying comparative judgement methodologies to qualifications such as A level and GCSEs. By definition, the qualities of candidates' work that are deemed to be valued when sitting these qualifications are defined by the expectations of the mark scheme and any weighting applied to the different elements of the assessment through the aggregation process.

---

[39] Good, F. and Cresswell, M., Educational Studies 14(3), 1988

[40] This effect has been considered further in a comparative judgement study of standards in A level maths (Jones, Wheadon, Humphries, and Inglis British Educational Research Journal 42 (4), 543-560 (2016)) which included a condition to measure this effect with their findings suggesting that the mathematicians' judgement of candidates' performances appeared not to be biased by differences in task difficulty. A cautious approach is, however, followed through this work to protect against over-interpretation of the findings.

Comparative judgement exercises, however, take a significantly different approach. Rather than focusing on detailed criteria, these approaches seek judgement of 'quality' at an holistic level – asking expert judges to form their opinions based on their own internalised set of values.

Champions of comparative judgement methodologies would argue that moving away from detailed and forensic approaches to considering candidate performance, centred around the use of mark schemes, avoids the false precision of such approaches that can actually undermine achieving a valid rank order of candidates. This could be due to anomalies in the marking criteria that might lead to 'better' candidates not necessarily achieving more marks. Such protagonists argue that the results of such an approach are inherently valid as they are based on value judgements of experts.

Critics of applying the approach for this purpose would argue that the results are not inherently valid as they do not reflect the intentional design features of the assessment. Also, for results of the exercise to be credible and defensible, they would argue it is necessary to understand (and be able to communicate) the basis on which those making judgements are doing so. This is extremely difficult to effectively capture and, where it can, may differ from the criteria for success that were explicitly defined as being of value through the design of the assessment, mark schemes and aggregation approach.

A further factor when considering the appropriateness of the approach is the properties of the assessments candidates have been asked to complete. Maths assessments typically contain items that are of relatively low tariff compared with subjects such as history and English and tend towards employing 'differentiation-by-task' (i.e. candidates of different ability are separated on the mark scale based on their ability to respond correctly to questions of different difficulty as opposed to differentiation-by-outcome where candidates are separated based on varying qualities of the response to questions of largely non-specific difficulty). Intuitively, candidates' responses to assessments that require extended or essay based evidence are more appropriate for comparison compared to the atomised, short response, questions that may be typical in subjects like maths[41]. However, several recent studies have been conducted evaluating the performance of candidates in maths assessment using this approach with promising findings[42]. Also, the nature of the questions in the reformed A level maths qualifications are more likely to extract richer responses from candidates than may traditionally be the case for maths assessments due to the role of skills such as problem solving.

These differing views on the appropriateness of comparative judgement methodologies come down to value judgements themselves and consideration of the acceptability of the limitations given the context of application. Based on the aims of the exercise performed here, the experience of the subject experts used for the judgement task (who will largely have had experience of working closely with the assessments and mark schemes under consideration), the ability to evaluate the

[41] Subjects requiring extended responses are also often selected as being the subject of experimental studies of comparative judgement as a replacement for marking due to the greater risk of subjectivity in the application of marking criteria in those circumstances.

[42] For example: Jones I. and Inglis M. Educational Studies in Mathematics, 89, 337–355 (2015); Jones, I., Swan,M., & Pollitt, A. International Journal of Science and Mathematics Education, 13, 151–177 (2014).

strength of relationship between the judgements of performance and actual marks awarded and for that to be factored into the interpretation of the results, the use of comparative judgement of candidates' work was deemed appropriate for the purposes of this investigation.

6.1 Pilot activity

When designing the approach to this strand of work, consideration was given to the efficiency with which expert judgements of performance could be captured and a pragmatic approach to capturing suitable evidence was sought. The initial proposed design was seeking to directly link the scales between 2018 and 2019 on the basis of expert judgement. The approach was based on making relative judgements between packs of scripts, as discussed above. The first stage of the proposed process was a rank ordering activity through which subject experts would sort a pack of 2018 scripts into order based on quality of candidates' work. Each script in this rank ordered list would then be compared with an already rank ordered range of scripts from the same component in 2019. Based on the comparison of quality across years, the judge would then insert the 2018 script into the pile at the point of equivalent performance. Performing this procedure would then provide a direct link between components from across years. Given uncertainty over the effectiveness and manageability of the task a pilot was undertaken with an experienced subject expert with significant experience of maths education.

Through this pilot, this pragmatic approach to capturing comparisons of performance, proved to be ineffective and it was deemed not possible to take this approach to make a meaningful link between the scales. There were two key barriers to the approach being effective.

First, was the confidence with which the 2018 scripts could be rank ordered. The initial proposal was for a pack of 12 scripts from 2018 to be sorted. It proved the case that any more than five scripts was unmanageable, which limited the information available through the approach.

Second, and more significantly, the exercise of selecting a point in the 2019 mark distribution that represented equivalent levels of performance was deemed not possible. This was due to the relatively small differences in marks between the 2019 scripts provided (typically every other mark), meaning there was insufficient qualitative difference in performance between adjacent scripts for the subject expert to deem the judgements to be meaningful. Again, a far smaller number of 2019 scripts with lower resolution on the mark scale was considered, however, this would have significantly reduced the information produced through the process and, therefore, the usefulness of the data.

Adaptations of the approach were considered such as the use of pre-ordered packs of scripts, however, they risked masking the unreliability of the process that could lead to misleading results.

On the basis of this pilot activity, it was, therefore, decided that a more conventional comparative judgement methodology would be a more effective and efficient approach to the collection of raw judgemental data.

6.2 Comparative judgements of performance

The remainder of the data collection for this strand was based on the pairwise comparisons of candidates' scripts. The data collection for this strand was divided into three phases:

1) Phase 1: A face-to-face judging activity using six maths experts who made comparisons of the relative quality of candidates' scripts within each component but with each judgement being made across years
2) Phase 2: An on-screen judgement activity performed by A level maths teachers who judged the relative performance of candidates work, randomly selected from across years, based on scripts that included the marks achieved on the individual questions
3) Phase 3: A repeat of Phase 2 with candidates' marks removed

The methods applied and the rationale for the different approaches are outlined in the following sections.

6.3 Method

6.3.1 Inclusion of candidate mark data

With all three phases of judging there were common design decisions to be made. One such consideration was the role that the marks awarded to candidates might play in supporting (or potentially undermining) judges to make valid judging decisions. The case for the removal of script level marks from candidate generated materials for the purposes of judging is relatively straight forward. The inclusion of scripts level marks would risk biasing judgements in two ways:

i. When comparing scripts produced in response to the same version of an assessment, knowledge of the overall marks achieved by two candidates would likely have a significant impact on judgments of which script was deemed of better quality. This is likely to bias the strength of relationship between marks awarded and script quality and underestimate the unreliability in the judgement process. In simple terms, judges may naturally favour one script over the other having been pre-conditioned by knowledge of the marks achieved. Were the judge to deem the lower scoring script to be of better quality, they would (implicitly or explicitly) be either questioning the accuracy of the marking or the validity of the approach to crediting the work of candidates defined by the mark scheme (see the discussion of validity above). Either way, knowledge of the total marks would be distracting and undermine in the aims of this work.
ii. When comparing scripts generated in response to different versions of the same assessment, knowledge of the overall mark could undermine the judges' attempts to account for the difference in difficulty of the two assessments. Were a judge presented with one script awarded a high mark on an assessment of low difficulty and another script awarded a lower mark in response to a more difficulty assessment this may inappropriately influence the judgement made, reinforcing the issues of bias discussed above. At the very least these marks would provide an unhelpful distraction for judges.

All script level marks were, therefore, removed for all phases of judging.

The case for the removal of item level marks is, however, less obvious. This is a manifestation of the tension between the holistic approach taken to forming judgements of quality in a comparative judgement methodology and the importance of understanding the correctness of candidates' responses where the assessments have been designed to differentiate-by-task when judging qualify.

From one perspective, the inclusion of candidate marks on the scripts supports judges in forming their holistic view. Their inclusion means judges do not need to engage with the detailed workings out within each candidates' responses and can focus more readily on formulating a high level picture of a candidates areas of strength and weakness in order to form his or her judgements. However, similar to the discussion of bias from the inclusion of script level marks, presented above, knowledge of the marks achieved (on either the same or different versions of assessments) may inappropriately influence the judgements made, moving them away from holistic judgements of quality towards a simple comparison of marks accumulated. Also, the inclusion of marks on scripts violate the assumptions of independence necessary for fitting the statistical models that follow.

The removal of marks from candidate scripts helpfully removes this potential source of bias and, therefore, eliminates this risk. However, one risk introduced by the removal of item level marks is that judges spend time evaluating the correctness of individual responses (and in extreme cases remarking them) while simultaneously attempting to form a high-level picture of the candidates' overall abilities. This approach would greatly increase the cognitive (and time) demands of the judgement task. Without item level marks, there is also the risk that judges routinely make incorrect evaluations of the quality of candidates' responses as they fail to detect incorrect responses. This risk reemphasises the validity risks raised by opponents of the methodology as judges may be forming their judgements on an illegitimate basis (incorrect interpretations of responses). Also, they may be inadvertently influenced by 'surface features' of the scripts (such as the amount of written response or the clarity of presentation) which, if positively correlated with candidates' marks, could mask the reduced validity of the exercise.

Given the complexity of these issues, and the need to ensure a perspective on relative performance standards could be formed through this investigation, a staged approach was taken to mitigate the risk. Phases 1 and 2 of the judging were performed based on scripts including item level marks in order to ensure a viewpoint (albeit a caveated one) could be formed. The item level marks were then removed from the scripts for Phase 3 with the intention of these judgements being used if the judging process proved to be successful.

6.3.2 Script selection

A further design consideration is the range of scripts that candidates are asked to consider. As the motivation for this strand of the investigation is the identification (or otherwise) of a difference in performance standard across 2018 and 2019, one approach could be to perform judgements of scripts from only on and immediately around the grade boundaries in each year. There are a number of issues with this approach, however. The first practical issue is the availability of work. For all four qualifications being considered through this investigation, there was a sparsity of candidate work at lower marks on the range and, therefore, around the grade E

boundary mark in 2018. Therefore, there would be insufficient work for this to be a meaningful exercise. From a technical perspective, there are issues making these judgements within or across years. Judgements within years would involve the comparison of scripts of very similar quality (based on the operational marks awarded). Given the known limitations of judging script quality, these judgements are unlikely to add value and result in a coherent scale. When making judgements across years other issues emerge. In one scenario, there is no discernible difference in performance standards and, therefore, judgements of which script is of better quality is essentially a random process providing no more useful evidence other than that the performance standards overlap. Alternatively, if the performance standards are distinct, judgements across years are likely to result in the scripts from one year being continuously selected as being of the highest quality. Setting aside the human factors and issues of bias associated with participants repeatedly making such a judgement, no information about the degree of separation/overlap would be available as a result of this process.

The approach taken here is therefore, to use scripts from across the mark range barring the very top and bottom of the mark distribution with the range broadly covering the operationally set grade boundaries. Scripts were typically separated by two or three marks in order to balance range and resolution of the selection, availability permitting.

Twenty scripts were selected for each component in each year with the exception of OCR B (MEI) papers 1 and 3 from 2018, due to the lack of work on appropriate marks. The mark ranges used are shown in Table 27. Candidates' personal identifiers were redacted from all script materials.

Table 27 Script selection for all phases of the comparative judgements of performance

	Component	Year	Sampled Scripts by Component Mark Total
AQA	7357/1	2018	17, 19, 20, 22, 24, 26, 28, 29, 32, 35, 39, 40, 44, 48, 52, 53, 56, 57, 60, 62
		2019	15, 19, 23, 26, 28, 30, 33, 36, 40, 43, 46, 48, 50, 53, 56, 59, 61, 63, 65, 69
	7357/2	2018	19, 20, 21, 22, 26, 30, 31, 33, 36, 38, 41, 47, 48, 52, 59, 65, 67, 70, 71, 76
		2019	16, 18, 22, 26, 28, 31, 35, 37, 41, 45, 49, 51, 55, 56, 59, 61, 65, 69, 71, 73,
	7357/3	2018	18, 20, 22, 24, 25, 27, 29, 33, 34, 39, 44, 44, 50, 54, 58, 59, 60, 61, 65, 68
		2019	21, 24, 26, 28, 32, 34, 38, 40, 42, 45, 48, 52, 56, 58, 60, 64, 66, 68, 70, 73
OCR (A)	H240/1	2018	24, 28, 37, 40, 43, 45, 50, 52, 54, 56, 59, 61, 62, 64, 66, 67, 69, 71, 74, 75
		2019	13, 17, 23, 25, 29, 31, 34, 37, 39, 41, 43, 45, 47, 51, 54, 57, 59, 61, 65, 67
	H240/2	2018	25, 29, 32, 39, 39, 45, 49, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 66, 66, 68

		2019	15, 20, 22, 25, 28, 32, 35, 37, 41, 43, 45, 47, 50, 54, 56, 58, 61, 63, 67, 69
	H240/3	2018	22, 23, 28, 34, 36, 37, 44, 46, 49, 51, 53, 54, 62, 63, 64, 66, 68, 71, 75, 77
		2019	12, 18, 20, 24, 26, 29, 31, 33, 37, 39, 42, 44, 46, 49, 51, 55, 57, 61, 63, 65
OCR B (MEI)	H640/1	2018	5, 31, 50, 53, 58, 59, 63, 65, 66, 69, 71, 74, 74, 75, 76, 78, 79
		2019	23, 30, 32, 35, 38, 40, 42, 46, 48, 52, 54, 56, 59, 62, 64, 66, 68, 70, 73, 77
	H640/2	2018	4, 26, 28, 33, 46, 48, 55, 56, 57, 58, 60, 63, 63, 64, 65, 67, 70, 71, 74, 77
		2019	17, 22, 26, 28, 30, 32, 34, 37, 40, 44, 46, 48, 51, 53, 57, 59, 62, 64, 66, 68
	H640/3	2018	13, 25, 35, 38, 40, 43, 49, 49, 51, 52, 53, 54, 54, 55, 55, 58, 59
		2019	12, 16, 18, 20, 22, 24, 26, 29, 31, 35, 37, 40, 42, 44, 46, 48, 50, 53, 57, 60
Pearson	9MA0/1	2018	19, 22, 24, 26, 29, 31, 33, 36, 40, 48, 54, 55, 59, 62, 63, 66, 68, 70, 73, 77
		2019	15, 18, 20, 24, 28, 30, 34, 36, 40, 42, 45, 48, 50, 52, 54, 57, 60, 64, 66, 68
	9MA0/2	2018	18, 21, 24, 24, 30, 33, 36, 42, 43, 48, 51, 55, 56, 60, 62, 63, 66, 68, 69, 73
		2019	13, 17, 19, 22, 25, 29, 32, 35, 37, 39, 43, 45, 47, 51, 53, 56, 58, 60, 62, 64
	9MA0/3	2018	10, 12, 14, 15, 17, 20, 25, 30, 36, 39, 41, 43, 46, 48, 49, 50, 52, 53, 54, 57
		2019	15, 18, 22, 25, 28, 32, 34, 36, 38, 42, 44, 47, 49, 51, 55, 57, 60, 62, 66, 68

6.3.3 Phase 1: Face-to-face judging activity

Six mathematicians, all of whom were either academic experts with an understanding of maths education or senior examiners responsible for A level maths assessments, participated in the study. Ofqual recruited the three independent experts and one senior examiner was nominated by their respective exam board at Ofqual's request. The only restriction on nomination was that they were not to have been involved in the additional scrutiny work that took place in summer 2019 as referenced in Section 2.5. Judges were paid for their participation.

Participants were sent all twenty-four question papers and mark schemes to familiarise themselves with them ahead of the judging exercise. Some participants were however, already familiar with one or more question papers, through their work for exam boards or through other regulatory activity commissioned by Ofqual.

In judging sessions lasting 90 to 120 minutes, participants were presented with pairs of scripts (in printed form) from a single qualification component. These scripts were selected such that always one was from the 2018 series and one from 2019. Participants were then asked to judge which of the two candidates was the better mathematician on the basis of the script evidence.

There were twelve components/judging tasks in total across the four qualifications and participants completed as many tasks as possible over several days. Each participant judged the components in a unique order over the course of several days in an attempt to eliminate ordering effects.

Participants were asked to re-familiarise themselves with the question papers at the start of each judging session and to review any notes they may have made in preparation. Participants were also asked, to the best of their abilities, to factor the relative demand of the question papers into their judgement of candidates' performances. It was stressed that, although the item-level marks and annotations were left on the scripts to support judgements, because they were comparing two different papers, judgements should concern the overall quality of candidates' responses not the number of marks awarded.

When a participant made a judgement, a facilitator recorded the response. As is typical for comparative judgement exercises, ties were not permitted. For each judgement, participants retained one of their current scripts and the facilitator replaced the second script with another selected randomly (with replacement[43]) from the same series. For the next judgement, the facilitator replaced the script previously retained.[44] An example is shown in Table 28

Table 28 Example script selection pattern for Phase 1 of the comparative judgement process

Pair Number	2018 Script	2019 Script
1	Candidate 10 (retain)	Candidate 34 (replace)
2	Candidate 10 (replace)	Candidate 27 (retain)
3	Candidate 76 (retain)	Candidate 27 (replace)
4	Candidate 76 (replace)	Candidate 19 (retain)
5	Candidate 42 (retain)	Candidate 19 (replace)

Participants were asked to keep in mind that the random selection of script pairs mean that some pairs of scripts might be more of less apparent that the difference between others. This may mean that not all judgements would necessarily require the same amount of time. Judges were also reminded that particular comparisons may be

[43] Note that, to prevent a participant judging the same script pair in succession, the script being replaced could *not* replace itself.

[44] One participant was unable to attend in person for a proportion of the time available for the judging exercise and, therefore, a significant proportion of the activity was performed remotely. In order to facilitate this approach a pre-prepared sheet for script pairings was provided to the participant for the judgements to be recorded manually.

incredibly difficulty to make a decision and, in those circumstances, a choice that may have felt arbitrary was acceptable.

As a follow-up to the judging exercise, participants were asked to capture their views on the exercise in order to better understand any factors that may have impacted on their judgements. The template used for this exercise is included in Annex H[45].

6.3.4 Phase 2: On-screen judging (window 1)

Through the face-to-face judging activity described in the previous section, 952 judgements were captured across the 12 components under consideration. While these provide a useful basis for initial comparison they represented an insufficient number of judgements to construct a reliable quality scale for each component. In order to supplement these judgements an additional phase of judging was required, as described here.

To facilitate collection of a sufficient number of judgements, in the most efficient way, Phase 2 was performed on-screen and focused on four components – one for each qualification. The selection of component for each qualification was made on the basis of the expected difficulty characterised through Strand 2 of the investigation. The component with the most similar median expected difficulty across 2018 and 2019 for each qualification was selected in order to minimise the effects of bias introduced due to differences in difficulty. Shown in Table 29 are the components rank ordered by absolute difference in median expected difficulty. By chance, each qualification has a component in the top four based on similarity. This led to the selection of paper 2 for AQA, and paper 1 for the other three qualifications, as indicated in the table. It should be noted that this approach meant that different topics were the subject of the papers for different qualifications. AQA paper 2 and OCR B (MEI) paper 1 assess Pure Maths and Mechanics, whereas OCR A paper 1 and Pearson paper 1 assess Pure Maths. Given the comparisons of performance were being made within components, these relative judgements of performance should not be systematically influenced by the difference in topics.

The recruitment process for judges built on that performed to recruit judges for the comparative judgement of item difficulty performed in Strand 2 (see Section 5.1.1.2). Through a combination of judges involved in the judging of item difficulty and those on the reserve list for that activity, 43 A level maths teachers were recruited and contracted as judges for the Phase 2 activity. Following recruitment, participants were asked which qualifications they were most familiar with, either through their current teaching or other exam board engagement. Based on this information, participants were allocated to one of the four components on which all of their judging activity would be based. Where possible, participants were matched to one of the components with which they were most familiar, however, due to an imbalance in profile, some participants familiar with the Pearson qualification were asked to act as judges on the AQA and OCR qualifications. Eleven judges were allocated to the AQA, OCR A and Pearson qualifications with 10 judges being allocated to OCR B (MEI). This allocation to the most familiar qualification was motivated by practical convenience rather than methodological issues which were managed through the preparation activities outlined below.

[45] Additional evidence was also captured verbally from one participant following the judging activity.

Table 29 Absolute difference of component level aggregated expected difficulty from Strand 2

Component	Absolute difference in median expected difficulty
AQA Paper 2	0.001
OCR A Paper 1	0.042
Pearson Paper 1	0.068
OCR B (MEI) Paper 1	0.088
OCR A Paper 2	0.181
Pearson Paper 3	0.218
OCR A Paper 2	0.263
OCR B (MEI) Paper 3	0.308
AQA Paper 3	0.401
AQA Paper 1	0.451
Pearson Paper 2	0.535
OCR B (MEI) Paper 2	0.755

To prepare for the task, all participants attended briefing sessions involving up to 11 participants either via video call or teleconference lasting between 20 and 30 minutes. In these briefings, participants were made aware of the aims of the investigation, the methodology being applied and the logistical arrangements. Through discussion of the methodology, the intended use of the item level marks, purely as a basis for interpreting the quality of responses to individual items to avoid the need for marking, was emphasised, as was the importance of taking into account any differences in perceived assessment difficulty.

Following the briefings, participants were sent the question papers and mark schemes for the 2018 and 2019 assessments for the purposes of familiarisation before beginning their judgements. Judges were requested to make 28 judgements each at their own convenience across a 6 day judging window.

To facilitate the judgement process, the No More Marking on-line platform, as described in Section 5.1.1.1, was again used; this time with participants being presented with candidate scripts randomly selected from the script lists in Table 27 for the component to which they had been allocated. To perform their judgements, participants were again, asked to identify 'which candidate was the better mathematician' on the basis of the script evidence presented with judges registering their decision by clicking either 'left' or 'right'.

All 43 judges completed their full allocation. This means AQA, OCR A and Pearson scripts were judged an average of 15.4 times and during the window and OCR B (MEI) scripts an average of 15.1 times. For the purposes of analysis the judgements from the first two phases of judging were combined.

6.3.5 Phase 3: On-screen judging (window 2)

Following on from the issues discussed in Section 6.3.1, the most desirable approach was deemed to be the judging of candidates scripts cleaned of item level marks. This would provide judgements free from the bias of marks, but risked the ability of participants to be able to perform the task effectively.

To test this condition, the on-screen judging procedure described above for judging window 1 was repeated with the same scripts, however, in window 2, the item level marks were removed.

Participants involved in Phase 2 were invited to continue their involvement for Phase 3. For this phase, 9 participants judged the AQA component, 10 judged the OCR A component, 9 the OCR B (MEI) and 10 the Pearson component. Due to four judges choosing not to continue their involvement, those continuing to participate were offered the opportunity to judge on multiple components with four judges carrying out this additional activity.

Judges were allocated 33 judgements each with the exception of those judging on multiple components, who, for contractual reasons, were allocated 21 judgements on the additional component to which they were allocated.

Given participants' previous involvement in the activity, there was no need to brief judges further beyond the basic arrangements and signalling the change in task. However, where participants were required to change the component they were judging or to judge on multiple components, additional familiarisation time and corresponding payment was made available.

All judges completed their full allocations resulting in AQA scripts being judged an average of 14.9 times, OCR A scripts an average of 15.9 times, 14.1 times for OCR B (MEI) scripts and 16.5 times for Pearson.

6.4 Analysis

6.4.1 Judge consistency and exclusions

Similar to the analysis of comparative judgement data in Section 5.1, the R package *sirt* was used to fit the data. This provided estimates of quality for each script. In contrast to the analysis of item difficulty where a single scale was calibrated for all items across all components, four independent models – one for each component – were constructed here. This was necessary due to judges making comparisons exclusively *within* components. These models were also fitted separately for the judging exercises with item marks included and excluded from the scripts.

After the initial model fit the judge in-fit was checked for each model to test the consistency of judgements made by individual participants. One judge was removed from the AQA judging in both judging windows due to having an in-fit outside of two standard deviations from the mean population in-fit. Following the exclusion of the judge producing misfitting data, the models were refitted with all other analyses based on these final model fits.

The scale separation reliability (SSR) was calculated for each model following the same procedure as outlined in Section 5.1.2.1. These figures are provided in Table 30.

Table 30 Reliability coefficients for the performance judgement model fits

		SSR
Phase 1 & 2	AQA[46]	0.882
	OCR A	0.916
	OCR B (MEI)	0.879
	Pearson	0.901
Phase 3	AQA	0.847
	OCR A	0.830
	OCR B (MEI)	0.881
	Pearson	0.886

Unsurprisingly, the levels of reliability tend to be slightly lower (or extremely similar) in Phase 3 where judges did not have access to candidates' marks. However, with the lowest SSR of 0.830 still reflecting a high proportion of the true score variance in the estimated scale values, all of these levels of reliability were deemed sufficiently high for the analysis to proceed.

Shown in Table 31 are the ranges of median judgement times for each component across the two on-screen judging windows[47]. These data do not show a consistent pattern across the different stages of the activity with some judging times shortening in Phase 3 (arguably due to increased familiarity) and some lengthening (arguably due to an increase in the demands of the judging task because of the absence of marks). Given these variations it would be inappropriate to draw any strong conclusion from these average judging times other than they appear to have been sufficient for participants to perform the task reliably[48].

The analysis above reported the reliability of the constructed models, however, as discussed in the introduction to Section 6, a further consideration when performing any comparative judgement task is the validity of the judgements being made. While judges may collectively be making judgements in a consistent way (resulting in high reliability) they may not be making judgements based on factors that are a meaningful or useful basis for analysis.

[46] For simplicity, throughout the analyses in this section, the components will be referred to by the qualification from which they have been drawn rather than full reference to each paper number. Caution should be taken, however, not to make inferences about the trends across the qualification beyond the specific components considered.

[47] Only timing data for those conducting the activity on-screen can be included in these data due to the mode of delivery and reliable data capture.

[48] For those judges with relatively short median judging times, their fit statistics were revisited and there were no grounds for their removal from the data.

Table 31 Range of median judgement times across both on-screen judging windows

	Range of media judging times in min:sec (mean in brackets)	
	Phase 2	Phase 3
AQA	0:52 - 8:04 (4:05)	1:27 - 9:38 (4:04)
OCR A	1:43 - 11:48 (5:29)	0:33 – 9:13 (4:14)
OCR B (MEI)	1:34 - 7:11 (4:01)	0:48 – 6:48 (3:02)
Pearson	2:03 - 4:44 (3:45)	2:14 – 6:47 (4:53)

The most relevant and readily available comparison that is of interest here is the relationship between the judgements made and the marks awarded to candidates. This comparison makes assumptions about the quality of marking, however, this is unlikely to be a significant factor given the historically high marking reliability in maths[49].

Figure 33 and Figure 34 show the relationships between marks awarded and estimated script parameters for each component Phases 1 & 2 and Phase 3, respectively. Also, provided in Table 32, are the awarded mark-to-script parameter correlations.

For the OCR and Pearson qualifications the correlation coefficients are high, however, those for AQA, although relatively strong, are lower than might be expected. The purpose of this analysis was, predominantly, to evaluate the appropriateness of the script parameter estimates, however, while the assumption has been made that the quality of marking is high, lower than expected marking quality could also reduce these correlations. To explore this, the Spearman rank correlations between the script parameter estimates across judging windows were calculated. Were the source of the reduced correlation to be related to marking, rather than some other factor within the judgement process it would be expected that the Spearman correlation coefficients of the script parameters would be no different for AQA than for the other qualifications. These coefficients are also shown in Table 32. Given that this coefficient is also lower for AQA, it appears that the slightly reduced strength of relationship between marks and script parameters is not linked to the marking.

As noted above, the difference in reliability coefficients for the Phases of judging was relatively small, however, while likely, it would be inappropriate to assume that the basis for the constructed scale in Phase 1 & 2 and Phase 3 was the same. The relatively high Spearman correlations (particularly for OCR A, OCR B (MEI) and Pearson) suggest that these scales are broadly consistent and therefore, likely to be formed on a similar basis. For consideration when designing other future studies, it is worth pointing out that it is unclear the extent to which participants' experience through Phase 2 acted as training for the additional judgment task, strengthening their familiarity with the assessment and the expectations of candidates.

[49] For example: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/759207/Marking_consistency_metrics_-_an_update_-_FINAL64492.pdf

Given improved validity of the judgement process being performed without the marks on the scripts and the reduction in assumptions that are violated by this approach, the data from Phase 3 will be taken forward for the evaluation of performance standards in the next section.

Table 32 Pearson and Spearman correlation coefficients between marks awarded and script parameter

		Pearson correlation		Spearman
		2018	2019	
Phase 1 & 2	AQA	0.897	0.917	0.843
	OCR A	0.971	0.977	0.903
	OCR B (MEI)	0.924	0.958	0.913
	Pearson	0.969	0.959	0.947
Phase 3	AQA	0.808	0.797	
	OCR A	0.895	0.909	
	OCR B (MEI)	0.975	0.851	
	Pearson	0.957	0.938	

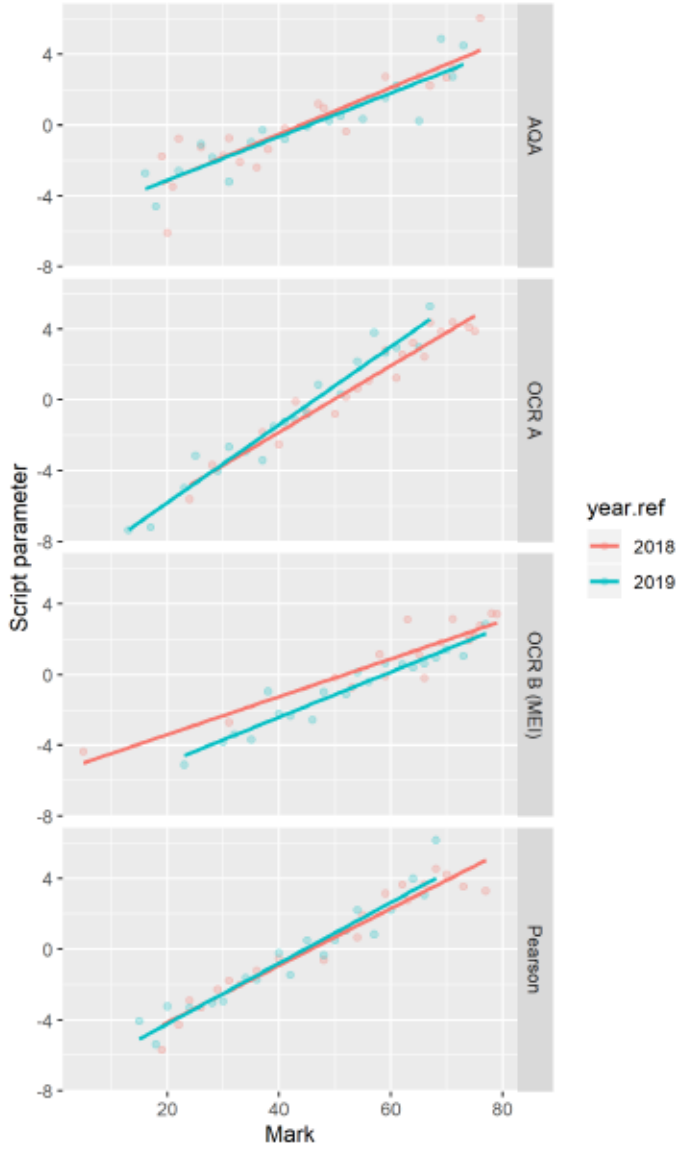


Figure 33 Relationship between mark awarded and script parameter for each component defined through judging Phases 1 & 2

6.4.2 Analysis of relative performance standards

The previous section established that the data collected and the models fitted represent a sufficient basis on which to compare performance standards. The purposes of these analyses are to identify whether or not the quality of work at the operationally set grade boundaries is discernibly different between the two years.

In performing this evaluation, it is important to reflect the uncertainty in the judgement process as any differences within the limits of that uncertainty cannot be argued to represent a meaningful difference.

By fitting the data to a single scale representing candidates' work from both years (Figure 33 and Figure 34) this has already accounted for any difference in difficulty (under the assumption that the difference in minimal and/or the judges have been effective in eliminating this source of bias – see discussion below).

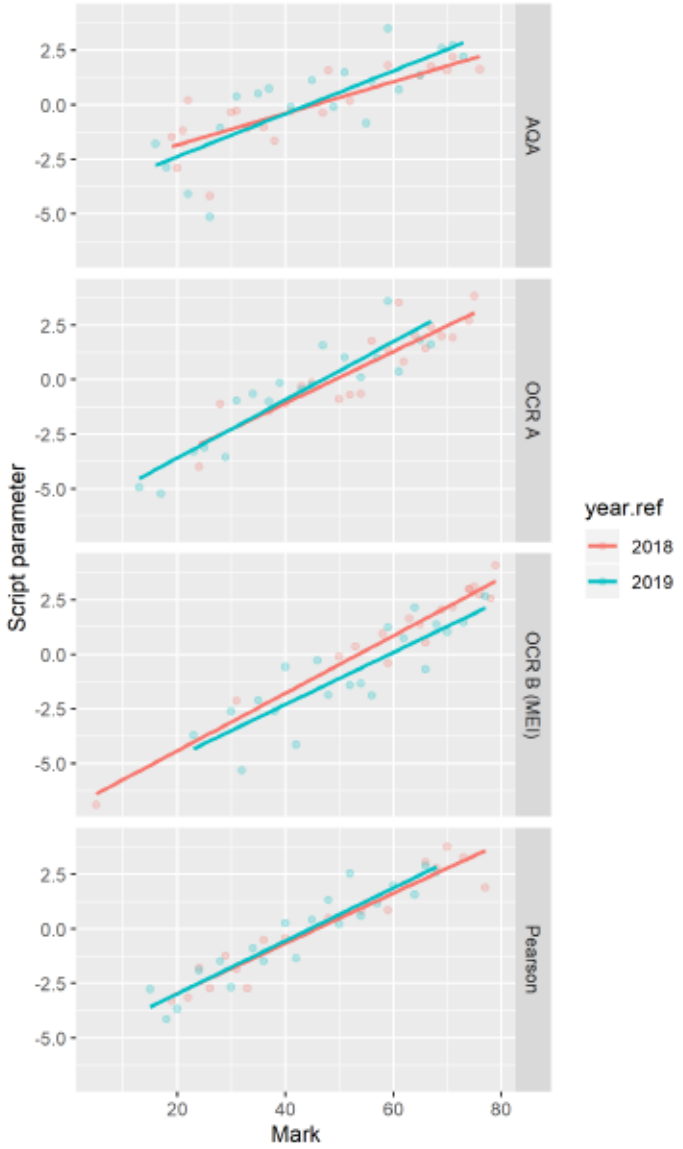


Figure 34 Relationship between mark awarded and script parameter for each component defined through judging Phase 3

Therefore, to establish whether or not the differences between the relationships in these figures represents a meaningful one in terms of script quality the following procedure was followed separately for each component:

Step 1: Regress script parameter estimates on awarded mark separately for the 2018 and 2019 data.

Step 2: For each year, based on the operationally set A and E grade boundaries, use the relationships established in Step 1 to look-up the script parameter and record the parameter value as a measure of script quality.

Step 3: Bootstrap Steps 1 and 2 (1,000 iterations) sampling the script parameter estimates based on the associated standard error for each script.

The results of applying this process are shown in Figure 35. The distributions on the right hand side of these plots represent the script quality at grade A. The distributions to the left represent grade E. It should be noted that these scales have been fitted independently from one another and therefore no attempt should be made to make

comparisons between these plots for the different qualifications as the relationship between them is arbitrary.

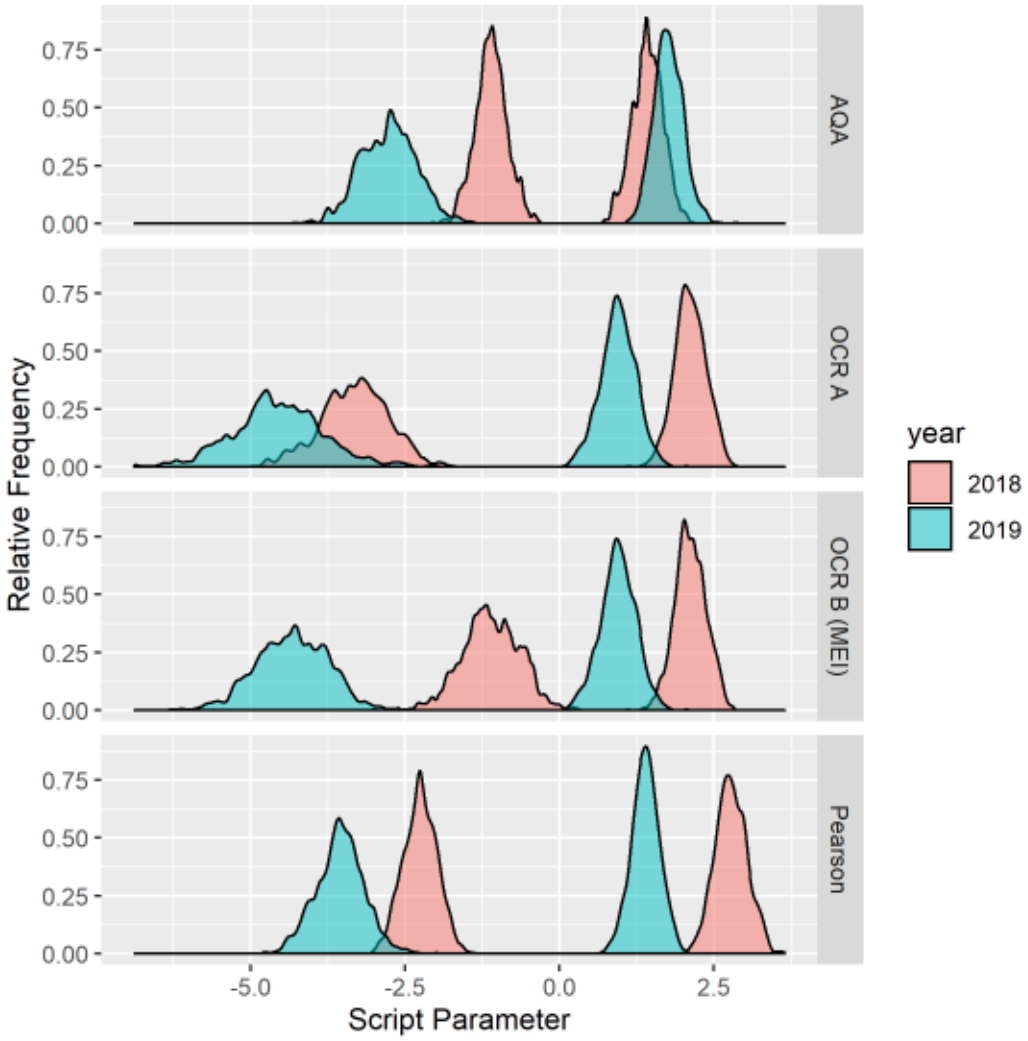


Figure 35 Distributions of script parameter at grades A (right of plots) and E (left of plots) in 2018 and 2019 based on the judgements of script quality performed during judging Phase 3

Assuming a normal distribution, based on the mean and standard deviations, the proportional overlaps of these distributions are shown in Table 33. Taking these as distributions of judgemental uncertainty, these values of overlap can be interpreted as the probability that two scripts (one in each year), selected at random from the respective boundary mark, *cannot* be distinguished from one another in terms of quality (the more intuitive reverse of this measure is included in the table).

Based on these analyses, these figures suggest that, at grade A, performance standards are largely indistinguishable for the AQA component, can be mostly distinguished for the OCR qualifications and are completely separable for Pearson. At the grade E boundary, for OCR A, the performances between years are reasonably common, for AQA and Pearson, are largely distinguishable and for OCR B (MEI) performances are completely distinct.

6.4.3 Subject expert views of the judgement process

As indicated in Section 6.3.3, following completion of the comparative judgement exercise, the participants involved in the Phase 1 judging activity were asked to

provide their views on the aspects of the question papers and candidates responses that made the judging process easy or difficult. These comments were scrutinised to support understanding of the data plotted in Figure 35 and are also captured to provide a potentially useful record when considering further studies of this kind.

One theme that was raised by several participants that hampered their ability to make what they perceived to be reliable judgements was when they encountered significantly imbalanced scripts and scripts with a sparsity of evidence. With imbalanced scripts, two candidates may provide similar amounts of correct material, however, the distribution of the quality of their responses are very different, potentially in different topic areas. This issue has a tendency to be more prevalent for more weakly performing candidates. Making judgements in that context are more reliant on the relative value that judges attribute to the topic areas or type of questions on which candidates provide better responses. This links to the challenge raised around judging the relative performance of candidates providing very little correct work, or indeed relatively few responses at all. The need to make comparisons of candidates who had provided a large number of weak responses compared to those providing few responses was identified as a challenge to judging.

One participant commented that it appeared candidates were being highly tactical in the items/sections of the question paper they decided to tackle or apparently spend more time completing. This made the judgement exercise particularly difficulty as some of scripts were reported as containing high quality responses to challenging questions with weaker responses elsewhere.

The issues raised above, while potentially effecting all candidates, tend to be relevant to candidates who are likely to have scored relatively few marks and, therefore, have been awarded a low grade. It is apparent from Figure 35 that there appears to be greater uncertainty in the judgements at grade E compared to grade A which is consistent with these views.

An issue over which there were split views was the helpfulness of the assessment design in terms of the arrangement of candidates' responses relative to the questions. Although different in style, the AQA and Pearson assessments have response areas integrated into the question booklets meaning candidates responses are presented close to the corresponding question. For the OCR assessments, this was not the case with the questions being provided in a separate booklet to those in which candidates respond. Some judges deemed this separation of questions from candidates' responses as a significant inconvenience and source of confusion due to the need to keep cross-referring to the two sets of materials disrupting the judging process. One of these participants contrasted this with the integrated answer booklet which they felt supported their judgements more efficiently.

In contrast, some participants found the need to constantly refer across pages in the same document as unhelpful and, particularly with the Pearson assessment, found themselves searching across a number of pages to find the relevant information they were seeking. One participant commented that the OCR approach assisted them in forming an holistic judgement as they could access all of the candidates' responses with ease in a compact format. It is inappropriate to compare scales across components and, therefore, there is insufficient experimental evidence to explore this effect confidently.

The final prominent theme in the responses related to the variations in the design of the different versions of the assessments between years. This was in terms of the balance of marks allocated to different topics/disciplines (such as pure, mechanics and statistics) and differences in item types and tariffs across years. These differences hampered the comparison process as it risked biasing the views of judges not just because it may provide a more or less demanding assessment, but one version may facilitate candidates to respond in a way the judge particularly values when formulating their perception of the ability of the candidate's work. A degree of variation in assessment design is expected to ensure assessments are not predictable, however, variations to the extent of those described by judges risks causing unnecessary confusion for candidates (let alone judgements of the quality considered here). This issue is out-of-scope for the current investigation so will be considered separately.

There was some evidence that participants were judging in a manner that was counter to the intentions of the assessment design. For example, one judge commented that they felt the level of penalty applied to candidates who failed to be as accurate in their responses as might be expected was overly punitive given the context of candidates sitting the exam. In forming their judgements they considered this to be less of an issue. This, is a practical example of the potential limitations of the approach and raises questions over the validity of the comparative judgment approach.

Table 33 Distributional overlap of script quality distributions across 2018 and 2019

	Proportion of overlap		Probability of detectable difference in quality	
	Grade A	Grade E	Grade A	Grade E
AQA	0.48	0.02	0.52	0.98
OCR A	0.03	0.33	0.97	0.67
OCR B (MEI)	0.03	<0.01	0.97	>0.99
Pearson	<0.01	0.05	>0.99	0.95

6.4.4 Judgement bias

As highlighted in Sections 6.3.3 and 6.3.4, judges participating in the exercise were asked to factor in any differences in difficulty of the assessment into their judgements of performance. Also, the components selected for focus in this strand of work were those that were the most similar in terms of difficulty across years, on the basis of the average expected item difficulty parameters estimated in Strand 2. The motivation for this was to reduce the known effects of bias when making judgements such as those considered here; where candidates' performances on more difficult assessments are undervalued compared to less difficult assessments. While the selected question papers appeared to be the most similar across years, it is, however, important to consider this potential effect as some differences in assessment difficulty are evident. This can be observed from the fact that the relationships between script parameter and awarded mark (Figure 34) are not

identical across 2018 and 2019, suggesting a difference in difficulty (i.e. candidates of the same ability have different expected scores in the two years).

Summarised in Table 34 is the direction of any difference in difficulty (based on Figure 35), including an indication of whether this would have led to an increased or decreased separation of script parameter distributions if the bias were to be present.

Table 34 Estimated impact of potential bias on judgements introduced due to differences in judged question paper difficulty

	More difficult based on Figure 35				Tendency of potential bias	
	At grade A		At grade E		Grade A	Grade E
	2018	2019	2018	2019		
AQA		✓	✓		Reduced separation	Reduced separation
OCR A		✓	✓		Increased separation	Reduced separation
OCR B (MEI)	✓		✓		Reduced separation	Reduced separation
Pearson		✓	=	=	Increased separation	NA

This table shows that, at grade E, any separation of distributions cannot be attributed to bias. If anything, elimination of bias would increase the likelihood of an identifiable difference in quality. At grade A, there is a mixed picture with the separation for OCR A and Pearson being, if anything, potentially accentuated in the presence of bias.

6.5 Intermediate findings from Strand 3

Through this strand of the investigation, consideration was given whether or not there was a difference in performance standards between the grading standards set in A level maths in 2018 compared with 2019. These analyses were based around a comparative judgement approach conducted in a number of phases. The main findings are drawn from the judgemental exercise conducted by judges based on scripts which had the marks candidates had been rewarded removed. Using these data allowed fewer of the assumptions of the approach to be violated while still resulting in a useable scale of script quality. This phase of judging focused on four components – one from each qualification – selected on the basis of expected similarity of difficulty across years to minimise the effects of bias.

Key findings from this strand are:

i. Through the comparison of script quality parameters at the operationally set grade boundaries, differences in performance standards were identified at the majority of grade boundaries for the components considered. Exceptions to this were AQA Paper 2 at grade A and OCR A Paper 1 at grade E. There was

also, a low level of overlap at grade A for the OCR components and grade E for AQA and Pearson.

ii. From consideration of the potential bias in the judging process due to differences in assessment difficulty across years, it appears unlikely that it can explain the instances of separation of performance standards.

iii. Consistent with the experiences of those conducting the judgemental process, there was greater judgemental uncertainty at grade E compared with grade A, particularly for the OCR qualifications, reducing the ability to reliably distinguish performances of different quality.

7 Summary and findings

This investigation has explored the potential sources and causes of the differences in grade boundaries in the reformed version of A level maths between 2018 and 2019 with a view to establishing the appropriateness of the grade boundaries set in 2018. The sources of the effects are largely common across the qualifications offered by each exam board, however, there are differences between the qualifications in terms of the size and direction of those effects and the context in which the awards took place.

Qualification level commentaries covering the four qualifications are presented in the following sub-sections, followed by the overall findings of the investigation and wider relevant considerations.

7.1 AQA summary

AQA had a modest entry size in 2018 with an entry of around 250 candidates with 98 17-year-old prior-attainment matched candidates on which to form a statistical prediction. Despite this being a small number of matched candidates, for the reasons described in Section 2.4 regarding the challenges to professional judgement at the point of reform, the boundaries suggested by the statistics were adjudged to be acceptable by the awarding committee. In 2019, AQA's entry was over 11,000 with over 6,000 matched 18-year-old candidates providing a strong basis on which to form a statistical prediction. Again, the boundaries suggested by the statistics were deemed an acceptable standard by the awarding committee.

AQA's grade A boundary was distinct amongst judgemental grade boundaries in this investigation insofar as it increased rather than decreased between 2018 and 2019. This does not, however, mean that it was not subject to the individual effects that may have acted to decrease the grade boundaries between years. For instance, the resitting effect at grade A was estimated to be 9 marks suggesting that, in order to regulate the impact of removing resitting, it was necessary to set boundaries that were around 3% of the total mark for the qualification lower to compensate for this effect than would otherwise have been the case. The reason for the increase rather than decrease in boundaries can, however, be accounted for by the apparent overall reduction in difficulty of the question papers between 2018 and 2019. The analysis suggests an increase in boundary of between 17 and 19 marks would have been necessary to account solely for this effect at grade A.

This demonstrates that the differences in boundaries at grade A between 2018 and 2019 can be fully explained by the difference in difficulty of the assessments and the need to compensate for the removal of resitting. It is worthy of note that the performance of candidates was not seen to be clearly separable at this grade through the analysis performed in Strand 3 of the investigation. However, the analysis indicated that, for the AQA component considered in this part of the work, the difficulty for more able candidates appeared higher in 2019 compared with 2018 and, therefore, judgemental differences in candidate quality may have been suppressed by the known effects of judgemental bias known to relatively under-reward performances on more difficult assessments.

At grade E, the size of the resitting effect with reference to the 2019 grade boundaries was estimated to be 13 marks at grade E. This equates to around one

third of the difference in grade boundaries between years. Based on the analysis in Strand 2, this difference was then also slightly increased rather than reduced due to differences in the difficulty of the assessments between years. However, uncertainty around the positioning of the grade E boundary in 2018 is unsurprising and understandable given the low matched candidate entry and the extremely small number of candidates towards the bottom end of the mark distribution on whom to form both a statistical and judgemental recommendation of grade boundary position.

Two other interesting points of note were evident through the analysis of candidates' performance in the AQA qualification. One was the apparent lack of a cohort effect between the 2018 and 2019 through the analysis of differences in difficulty explored using Approach 3 (as outlined in Section 5.3.3). AQA was atypical in there appearing to be little or no difference in the overall ability of candidates across the two years. This fact is, however, supported by the apparent presence of a relatively large number of weaker (non-prior attainment matched) candidates entering with AQA in 2018. This is evidenced by the large difference in the 17-year-old grade E outcome (95.6%) compared with the all candidate grade E outcome for the qualification (75.9%) as shown in Annex C. This can be put down to these unexpected properties of the AQA entry to the reformed qualification in 2018 rather than indicating any issue with either the analysis presented here or any operational processes.

The other point of note is the weaker relationship between judgements of script quality and candidates marks explored in Strand 3 compared to the other components. An initial consideration of this difference does not appear to suggest an issue with the underlying marking as a source of the difference. This is, therefore, not necessarily a matter for concern and may be as much a limitation of the analytical methods rather than the functioning of the assessment. However, this point may benefit from further follow-up consideration, as a disconnect between the differences in perceived quality of candidates' overall performance and the marks achieved could impact on the confidence in the results or suggest a broader validity challenge.

7.2 OCR A summary

The OCR A reformed qualification had an entry of only 115 candidates in 2018 with 95 of those able to be matched to appropriate prior-attainment data for us in forming the statistical predictions. Despite the limitations of both the judgemental evidence available for awarding at the time of transition to the reformed qualifications, awarders in 2018 were not satisfied that the quality of work at the boundaries suggested by the predictions reflected an appropriate standard, recommending a lower boundary at grade A than that suggested by the statistical evidence. This led to the award being 7% above prediction at grade A. A similar situation occurred at grade E with all matched candidates being deemed to have achieved this grade at least. Both of these deviations from predictions will have acted to have reduced the difference in grade boundaries between years which were 36 and 20 marks lower in 2019 at grades A and E respectively.

Based on the findings of Strand 3 of the investigation, the performances of candidates at grade E in the OCR A qualification were not reliably separable in terms of the quality of work. While there was a relatively large difference in the mean script parameter estimates across the two years, there was a large spread of these parameters across iterations of the analysis reflecting the difficulty subject experts

had in judging the relative quality of candidates work at the lower end of the mark scale as reflected upon in Section 6.4.3. In contrast, the judgements of script quality were largely separable at grade A with a small level of overlap in the script parameter distributions. However, if anything, this separation may have been slightly overestimated in the presence of bias introduced due to differences between years in difficulty of the assessments considered.

The analysis of resitting and the consequential impact on the 2019 grade boundaries was shown to account for 10 marks of difference between years at grade A and 9 marks at grade E. This corresponds to around 26% and 45% of the difference in boundaries at grades A and E, respectively.

The analysis of the expected difficulty of OCR A assessments identified an unclear picture in terms of whether the difficulty had increased or decreased overall with the mean expected item difficulty being lower in 2019 but the median being higher suggesting a different distribution of item difficulties across years. Based on the modelling to transform these expected difficulty scales into estimated necessary differences in grade boundary position, the analysis (based on Approach 2) suggested a mean difference of 10 marks at grade A and 3 marks at grade E purely due to the overall increase in difficulty identified by the analysis.

In combination, taking into account the uncertainty in the adjustments for assessment difficulty, the resitting effect and difference in assessment difficulty accounted for the majority of the difference between grade boundary marks at grade A between years. At grade E, the combination of the compensation required due to the removal of resitting and slightly increased assessment difficulty explained the majority of the difference in boundary marks between years. However, even neglecting the resitting effect, subject experts were unable to reliably identify a qualitative difference in level of performance across years.

7.3 OCR B (MEI) summary

The OCR B (MEI) qualification had the lowest entry of all of the reformed A level maths qualifications in 2018 exacerbating the challenges of awarding in the first series. The challenges to the role of expert judgement at the point of transition to a newly reformed qualification has been outlined in Section 2.4 due to the structural and content changes that limit its effectiveness. This fact, allied with the highly limited statistical evidence with only 33 prior-attainment matched 17-year-old candidates in 2018, led to increased uncertainty regarding the surety of the boundary setting process. This was a particular challenge at grade E where there were only three candidates scoring fewer than 100 marks.

The difference in grade boundary marks between 2018 and 2019 for OCR B (MEI) was 19 marks at grade A and 63 marks at grade E. The analysis of script evidence performed through Strand 3 of the investigation identified a separation in identifiable script quality at grade E (taking into account potential bias and a relatively large judgemental uncertainty). There was some overlap in quality of work at grade A suggesting it may be difficult to effectively distinguish between the quality of some work at the grade A boundary across years.

The resitting effect accounts for around 9 marks of the difference at grade A and around 11 marks of the difference at grade E. This corresponds to 47% of the

difference at grade A and 17% of the difference at grade E. However, this is due to the large difference between years at this grade rather than any effect of the cohort.

The analysis of assessment difficulty suggested the overall difficulty of the OCR B (MEI) assessment increased only very slightly between 2018 and 2019 with the level of difficulty being very similar overall. When seeking to quantify this effect in terms of mark only Approach 2 provide a meaningful estimate given the small number of candidates in 2018. This estimated a necessary reduction of around 4 marks at grade A and 1 mark at grade E (both based on the mean of the modelled distributions).

Taking into account the cumulative effect of the removal of resitting and the difference in difficulty of the assessment within years, the difference in grade boundaries at grade A is fully explained when accounting for the uncertainty in these estimates. This is not the case at grade E with a notable proportion of the difference likely being due to the uncertainty of awarding grade E in 2018 as discussed above.

7.4 Pearson summary

Given the proportional size of the Pearson entry, much of the analysis considered in this investigation looking at the national level effects were heavily dominated by the effects from this qualification. In 2018, Pearson had an entry of around 1,700 candidates in the reformed qualification with a matched 17-year-old candidate entry of over 1,000. This size of matched entry provided a stronger basis for greater confidence in the statistical predictions than was the case for the other qualifications.

The difference in the grade A boundary between 2018 and 2019 for the Pearson qualification was 19 marks, with a difference of 27 at grade E. Based on the analysis presented in Section 6.4.2 the difference in grade boundaries between years did lead to a judgementally distinguishable difference in performance standards. It is, however, possible that some of the difference at grade A might be attributable to bias in the judgement process due to the differences in difficulty of the assessments across years.

The analysis presented in Section 4.5.2 showed the magnitude of the resitting effect relative to the 2019 mark scale was estimated to be 11 marks at grade A and 10 marks at grade E. This corresponds to 58% of the boundary difference at grade A and 37% of the difference at grade E. Given the size of entry, this provides the best indicator of the impact of the effect and the compensation necessary in terms of grade boundary position, nationally.

Aggregated across the Pearson question papers, it was shown that the expected difficulty was higher in 2019 compared with 2018. The need to account for this difference was, therefore, an additional contributing factor to the need to set lower grade boundaries in 2019. The analysis in Section 5.3.4, suggested that the grade A boundary difference between years, necessary to account purely for the increase in difficulty, was 15 to 20 marks (based on the mean of the modelled boundary distributions using Approaches 1 and 2). At grade E, the equivalent boundary difference was 6 or 7 marks using the two difference approaches.

This analysis shows that the differences in grade boundaries between 2018 and 2019 for the Pearson qualification can be fully explained by the resitting effect and the change in difficulty between the assessments between the two years. At grade E, the difference can again be largely accounted for in these two effects taking into

account the uncertainty in the modelled adjustments for increased difficulty. A small proportion of the difference in boundaries at grade E does, however, remain unexplained. Despite the large entry size for the Pearson qualification relative to the other A level maths qualifications, there remains issues with the relative sparseness of the distribution at lower abilities that will have contributed to statistical and judgemental uncertainty over the placement of grade E boundaries. This cohort effect can be seen from the all candidate grade A outcome (see Annex C) and predicted outcome for matched 17-year-olds candidates being over 60% for the Pearson qualification. These high outcomes are not a matter of concern but merely indicate more about the tendency for early sitting 17-year-old candidates to be more able.

7.5 Overall findings

The key overall findings from this investigation are outlined below:

i. The difference in grade boundaries set in A level maths in 2018 and 2019 did lead to a discontinuity in grading standards between the first two years of the reformed qualifications. However, this discontinuity was inevitable and occurred at a point in the transition that appears the most equitable across the different sub-cohorts of candidates across years.

ii. Taking into account the uncertainty in professional judgement, the differences in standards set in 2018 compared with 2019 were qualitatively separable in the majority of cases based on the evidence generated through this investigation.

iii. The cause of the discontinuity was the change in the relative relationship between the performance of 17-year-old and 18-year-old candidates, combined with the necessary use of 17-year-old candidates as the basis for the 2018 awards. Allowing candidates the opportunity to certificate at the end of the first year of availability gave visibility to this change in relationship in a way that, would otherwise, not have been the case.

iv. The effect seen was not caused by the 2018 cohort of 17-year-olds sitting the reformed qualification being statistically atypical in comparison with other years. On this basis, it is reasonable to expect a broadly similar relationship to continue in the qualifications with the discontinuity being confined to the exam series scrutinised through this investigation.

v. A significant contributor to the change in relationship between 17-year-old and 18-year-old candidates was the removal of the opportunity for candidates to resit assessments in the reformed qualifications. On the legacy qualifications, 18-year-old candidates had greater opportunity to resit assessments and therefore received a relative benefit from resitting when compared with 17-year-old candidates. A proportion of the difference in grade boundaries between 2018 and 2019 was necessary to ensure that 18-year-old candidates were not disadvantaged due to the resitting provision being removed.

vi. The removal of resitting has impacted on the rank order of candidates relative to that on the legacy qualifications (evidenced by the change in relationship between 17 and 18-year-old candidates) as typically occurs at a time of change in qualifications. Typically, such a change in the rank order of candidates is confined to the point of transition to the reformed version of qualifications. However, the circumstances surrounding the early certification

opportunity made available for A level maths qualifications in 2018 led to this change being evident in the grade boundaries across the two years.

vii. A significant sub-cohort of 17-year-old candidates in 2018 certificated in the legacy version of the qualification rather than the reformed version which had not been anticipated. This investigation has demonstrated broad alignment between the standards set for the two groups of 17-year-old candidates in 2018 – those certificating to the legacy and reformed versions of the qualification. The operationally set grade boundaries, therefore, retain this within-year fairness. Despite reservations regarding the reliability of the statistical evidence in 2018, the use of predictions is likely to have played an important role in this being the case.

viii. As evidenced through the investigation, differences in the difficulty of assessments between years also contributed to a necessary lowering of grade boundaries in most cases. For three of the four qualifications considered, evidence suggested that, in the absence of any other effects, lower grade boundaries in 2019 were appropriate to allow for this effect.

ix. At grade A across all qualifications considered, the difference in grade boundaries between 2018 and 2019 can be largely explained by compensation for the resitting effect and the estimated difference in assessment difficulty between years.

x. At grade E, the resitting effect and differences in assessment difficultly explain the majority of difference in boundary marks. In most cases, however, there remains a proportion of unexplained difference. These differences in boundaries likely arise from the high levels of uncertainty when awarding at this grade in 2018 due to the sparseness of candidates, particularly at the lower end of the mark distribution, which further weakened the statistical and judgemental evidence. Other potential sources of difference between 17 and 18-year-olds are changes to the content/curriculum for which it is not possible to effectively quantify their impact.

7.6 Wider considerations

7.6.1 Generalisability of the age group effect

At the heart of the findings of this investigation is the change in the value-added relationship between 17-year-old candidates relative to 18-year-olds as the structural changes have been introduced through the process of qualifications reform. It should be noted that this change in relationship is a reflection on the historically observed relationship between 17-year-olds and 18-year-olds, and is a commentary on candidate attainment to date. This finding should not be read to imply that it is in the best interests of candidates to enter at the age of 17 rather than 18 for individual candidates. It is important that the decision over when to enter a qualification is based on the specific circumstances of each candidate which reflects their preparedness for assessment and certification. While it is anticipated that this recalibrated – and arguably more equitable – relationship between age groups will broadly continue for the lifespan of the reformed qualifications, it may not continue to be the case were changes to entry strategy to occur and inappropriately prepared candidates be entered early.

7.6.2 Implications for future candidate preparation

As highlighted in the findings above, the grade boundaries set in 2018 and 2019 realised a broadly different performance standard across the two years. It is important that exam boards reflect on the potential unintended consequences of these differences for the preparation of future cohorts. The use of past papers for exam practice, progress checking and mock examinations is widespread in preparing candidates for upcoming exams. In doing so, teachers, future candidates and their parents may choose to use historical grade boundaries as a mechanism to (formally or informally) predict their likely outcome. Inaccuracy in the boundaries used for this purpose could therefore be misleading. The findings of this investigation would suggest that use of the operational grade boundaries from the 2018 reformed qualifications would underestimate the grades that candidates may achieve.

Ofqual have previously signalled caution regarding over reliance on the grade boundaries set in the first series[50]. However, exam boards should consider how they best support, and communicate that support, to centres and candidates in order to avoid any such unintended consequences of the use of the 2018 assessments and grade boundaries for that purpose.

7.6.3 Considerations for future reforms to A level maths

We are not aware of any current plans to reform general qualifications. However, it is important that the findings of this investigation are reflected upon at the point that the qualifications are next revisited. Specifically, consideration should be given to the appropriateness of allowing candidates to certificate after the first year of availability of the new qualifications as an intermediary step before going on to sit further maths. It is inevitable that the relative rank order and therefore grading standards of different sub-cohorts of candidates are changed when the structure or content of a qualification is modified. While the magnitude, direction, predictability and desirability of such a change is dependent on the nature of the reforms, differences are, however, inevitable. If it is deemed essential to afford candidates the opportunity to sit their A level maths assessment before their A level further maths then two potential approaches could be considered:

a. The transition phase from one qualification to another (extended by one year relative to the most recent arrangements) with the first certification of the reformed version taking place after two years of teaching could be explored. This would afford candidates seeking to certificate in maths at the age of 17, followed by further maths a year later the opportunity to complete this route on the legacy qualification. Their peers, seeking to enter maths at the age of 18 would, however, be able to transition to the new qualification at the start of their course of study. This may have intolerable and undeliverable implications for centres depending on the co-teachability of the two versions which may preclude the approach. It would also mean challenges to the quality of evidence into awarding, however, it may provide increased visibility and separation of the effects compared to the recent arrangements.
b. An alternative approach might be to allow candidates to enter the A level assessments at the age of 17 on the reformed qualifications after

[50] https://ofqual.blog.gov.uk/2018/03/16/setting-standards-in-the-new-a-level-maths-qualifications/

a year of availability, as was the case through the recent reforms. However, the award of these qualifications could be deferred by a year to ensure greater control over the relative standards across years and for sub-cohorts within them. This approach would allow centres and candidates to spread the assessment burden across two years, as is currently the case for candidates following this path, and would give visibility of the marks achieved. However, grade outcomes would not be available until a year later. This would impact of the use of candidates' grades for formative purposes including decisions over progression. Depending on the nature of the changes, this approach (and indeed any approach) is unlikely to avoid a discontinuity in standards for some sub-cohorts of candidates which, as described through this investigation, is inevitable at times of change. It may, however, afford greater visibility and control over those differences.

7.6.4 Implications beyond maths

Given the issues highlighted through this investigation regarding the change in relationship between 17 and 18-year-old candidates, and the implications that has and for A level maths, it is worth considering if these findings should be an indication of issues for other subjects. While the certification arrangements for A level maths gave visibility to the issue, thought should be given to whether such an effect may be present in other subjects where the effects may have been masked due to early certification opportunities.

There are a number of reasons to suggest that the issue is largely localised to maths.

A level maths is unique in terms of its relationship with another A level subject: further maths. No other subjects have the built in incentive and consequential benefit to certificate early as a mechanism for either preparation or selection into another A level subject. In other subjects, while candidates may have sat units early in the legacy versions of qualifications, they tended not to certificate early. Far fewer candidates, chose to certificate before the age of 18 in other subjects meaning that a smaller number of candidates would be affected by the changes under consideration here.

In addition, there are factors relating to the nature of maths as a subject and the aggregation rules on the legacy qualification that suggest resitting in maths was particularly beneficial. As a subject, the acquisition of knowledge and skills is particularly accumulative in nature. In legacy structures, the barriers to re-sitting and benefits of doing so appear reduced. For example, a decision on the legacy qualification to resit an 'early' pure maths unit at the point of certification would seem attractive due to the recency of learning more advance pure maths content which has built directly on the prior learning necessary for the unit being resat. This reduces the preparation burden and increases the likely chance of benefit. Another consideration is the extreme flexibility (and with it, complexity) of the aggregation arrangements for the legacy qualifications. The range of combinations of units available for candidates to use towards certification and the smaller size of those units due to the six rather than four unit structure of maths made it more feasible for

candidates to be able to optionally draw on a larger number of unit results in a way that is not possible for other subjects.

This combination of significantly lower numbers of certifications from 17-year-olds in other subjects allied with likely reduced benefit of resitting on the legacy version due to content and structural differences make it unlikely that similar effects are material elsewhere.

8 Annexes

ANNEX A – Legacy qualification optimisation rules

ANNEX B – Qualification level outcomes for 17 and 18-year-old candidates in 2017, 2018 and 2019

ANNEX C – Explanation of minor difference in value-added relationship for 18-year-old candidates between 2017 and 2019.

ANNEX D – Summary of changes to qualification content through reform

ANNEX E – Breakdown of items included in the item difficulty comparative judgment exercise

ANNEX F – Item level estimates of expected difficulty

ANNEX G – Item facility index histograms

ANNEX H – Template for the capture of views on the comparative judgement exercise

ANNEX A –Legacy qualification optimisation rules

Rule 1

Grading of qualifications is determined as follows:

Step (i) maximisation of the qualification **grades** (including A*).

Step (ii) for the qualification grades determined under step (i), the maximisation of the **uniform mark totals** for each qualification.

The maximisation of grades and uniform mark totals for qualification titles is determined using the sequence:

Mathematics; Further Mathematics; Additional Further Mathematics

The highest possible grade is awarded for the first qualification title requested in the above sequence, followed by the highest possible grade for the second qualification title requested in the above sequence (if the candidate has entered for two titles), followed by the highest possible grade for the third qualification title requested in the above sequence (if the candidate has entered for three titles). Only one qualification (AS or A level) is maximised for each title.

For example, if a candidate has entered for AS and A level Mathematics and AS and A level Further Mathematics (i.e. two titles), the highest possible grade is awarded for A level Mathematics followed by the highest possible grade for A level Further Mathematics. The uniform mark totals for A level Mathematics and Further Mathematics (in that order) are maximised before the AS qualification grades are considered.

Rule 2

Once grades have been issued, units used towards a qualification award will become 'locked' to that qualification's group. This means that these units can only subsequently be used towards qualification awards in the same 'qualification group' (as defined below); the units cannot be used towards a qualification in a different qualification group.

The groups and levels within the groups are as follows:

Group	Level	Qualification
A: Mathematics	1	AS Mathematics / AS Pure Mathematics†
	2	A level Mathematics
B: Further Mathematics	1	AS Further Mathematics
	2	A level Further Mathematics
C: Additional Further Mathematics*	1	AS Further Mathematics (Additional)
	2	A level Further Mathematics (Additional)

*not all awarding bodies offer this group.

†Pure mathematics units (i.e. the C units and FP units) will not be locked when used to certificate any AS award.

A unit may have been 'single-locked' by being used towards the award of only one of the qualifications in the group.

OR

A unit may have been 'double-locked' by being used for the awards of both the AS and the A level qualifications in the group.

A unit that has been 'single-locked' will become 'unlocked' from a qualification group by re-entering the qualification for which it was used.

A unit that has been 'double-locked' will become 'unlocked' by re-entering the A level qualification only.

If a unit is not unlocked by re-entering the appropriate qualification, it is only available for re-use towards qualifications within the group to which it is locked.

Rule 4

Entitlement to re-enter qualifications

An entitlement to re-cash-in is achieved if:

a unit has been taken or re-sat since the qualification award was last made,

or

there are units in the results bank that have not been locked to a qualification group.

Once satisfied, an entitlement to re-cash-in places no restrictions on the number of cash-in entries that can be made. For example, a candidate with awards for both A level Mathematics and A level Further Mathematics can re-enter both when an entry for just one unit is made.

ANNEX B – Qualification level outcomes for 17 and 18-year-old candidates in 2017, 2018 and 2019

Table 35. A level maths outcomes in 2017 for 17 & 18-year-olds and candidates of all ages by qualification (legacy).

		Grade							
	Academic Age	A*	A	B	C	D	E	U	Total
AQA 6361	17	35.3	65.9	80.8	88.9	93.1	96.7	100.0	334
	18	18.5	40.4	61.2	78.0	90.0	97.0	100.0	13,692
	All ages	17.7	39.1	59.7	76.8	89.2	96.5	100.0	16,606
OCR 7890	17	39.1	70.5	88.4	94.4	96.4	98.7	100.0	302
	18	22.5	51.0	72.7	86.1	93.8	98.3	100.0	10,334
	All ages	22.0	49.7	71.6	85.5	93.4	98.1	100.0	12,098
OCR 7895	17	47.7	80.1	89.4	93.4	95.4	97.4	100.0	151
	18	26.7	48.8	69.2	83.8	93.0	97.8	100.0	8,670
	All ages	25.5	47.2	68.1	83.2	92.8	97.9	100.0	10,177
Pearson 9371	17	29.4	62.4	80.7	89.3	94.2	97.6	100.0	1,163
	18	16.8	43.4	66.0	81.8	92.0	97.4	100.0	38,205
	All ages	16.4	41.9	64.2	80.2	90.8	96.7	100.0	48,842
All	17	33.3	65.6	82.6	90.4	94.5	97.6	100.0	1,950
	18	19.2	44.6	66.4	81.9	92.0	97.5	100.0	70,901
	All ages	18.5	43.1	64.8	80.6	91.1	97.0	100.0	87,723

Table 36. A level maths outcomes in 2018 for 17 & 18-year-olds and candidates of all ages by qualification (legacy & reformed).

	Grade								
	Academic Age	A*	A	B	C	1D	E	U	Total
AQA 6361	17	22.3	58.7	76.1	87.0	95.1	96.7	100.0	184
	18	14.1	40.1	61.7	78.8	90.6	96.8	100.0	13,873
	All ages	13.5	38.5	60.2	77.5	89.7	96.2	100.0	16,289

AQA 7357	17	24.8	59.1	73.0	82.5	89.8	95.6	100.0	137
	18	11.1	29.6	55.6	63.0	74.1	77.8	100.0	27
	All ages	17.2	42.2	54.3	61.6	69.4	75.9	100.0	232
OCR 7890	17	41.8	70.3	83.5	96.7	97.8	100.0	100.0	91
	18	21.8	48.2	70.7	85.0	93.5	97.8	100.0	10,693
	All ages	21.5	47.3	69.6	84.3	93.1	97.7	100.0	12,107
OCR 7895	17	46.8	62.9	79.0	93.5	98.4	100.0	100.0	62
	18	23.3	48.8	69.6	82.9	92.0	97.5	100.0	9,255
	All ages	22.5	47.3	68.2	82.2	91.6	97.4	100.0	10,504
OCR H240	17	44.0	73.0	82.0	93.0	98.0	100.0	100.0	100
	18	0.0	57.1	71.4	85.7	85.7	100.0	100.0	7
	All ages	39.8	70.8	80.5	92.9	97.4	100.0	100.0	113
OCR H640	17	27.3	54.5	75.8	84.8	90.9	93.9	100.0	33
	18	0.0	0.0	0.0	0.0	100.0	100.0	100.0	1
	All ages	30.6	52.8	72.2	80.6	88.9	91.7	100.0	36
Pearson 9371	17	26.2	56.8	75.4	88.2	93.0	96.6	100.0	785
	18	14.9	43.7	67.3	82.9	92.1	96.7	100.0	41,813
	All ages	14.3	42.0	65.3	81.2	90.8	96.1	100.0	49,808
Pearson 9MA0	17	31.4	64.2	80.1	89.2	94.1	97.7	100.0	1,312
	18	24.8	52.5	64.9	75.2	82.2	86.6	100.0	202
	All ages	30.6	61.1	76.1	85.4	91.1	94.8	100.0	1,626
All	17	30.1	61.8	78.2	88.8	94.0	97.4	100.0	2,704
	18	16.8	44.3	67.0	82.4	91.9	97.0	100.0	75,871
	All ages	16.4	43.1	65.5	81.1	91.0	96.4	100.0	90,715

Table 37. A level maths outcomes in 2019 for 17 & 18-year-olds and candidates of all ages by qualification (legacy[51] & reformed).

		Grade							
	Academic Age	A*	A	B	C	D	E	U	Total
AQA 6361	17	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1
	18	31.7	75.0	93.3	100.0	100.0	100.0	100.0	60
	All ages	8.9	35.7	64.5	82.3	94.1	98.4	100.0	740
AQA 7357	17	35.2	65.3	81.3	90.2	94.8	96.9	100.0	193
	18	13.2	37.5	58.1	76.0	90.3	97.5	100.0	10,018
	All ages	13.1	36.6	56.6	74.5	89.0	96.6	100.0	11,300
OCR 7890	17	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1
	18	50.0	83.3	100.0	100.0	100.0	100.0	100.0	6
	All ages	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
OCR 7895	17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
	18	77.8	77.8	88.9	100.0	100.0	100.0	100.0	9
	All ages	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
OCR H240	17	29.4	69.3	80.1	90.9	96.1	99.1	100.0	229
	18	20.6	48.4	65.2	79.3	90.8	98.0	100.0	6,331
	All ages	20.6	48.5	64.9	78.8	90.4	97.8	100.0	7,149
H640	17	33.3	55.6	77.8	88.9	92.6	100.0	100.0	27
	18	21.4	47.2	67.8	82.4	92.4	98.1	100.0	5,791
	All ages	21.1	46.4	67.0	81.6	91.9	97.9	100.0	6,253
Pearson 9371	17	52.4	66.7	81.0	90.5	95.2	100.0	100.0	21
	18	51.4	73.1	88.0	93.7	96.0	99.4	100.0	175
	All ages	15.9	39.4	64.9	82.5	91.8	96.8	100.0	2,903
Pearson 9MA0	17	41.5	71.5	83.2	90.7	96.9	99.0	100.0	1,781
	18	16.1	40.6	57.7	74.7	88.6	97.2	100.0	50,137

[51] The reformed specifications were available in 2019 primarily as a final resit opportunity, hence the low entries, particularly from 17 and 18-year-olds.

	All ages	16.3	40.2	56.8	73.6	87.6	96.6	100.0	55,755
All	17	39.8	70.5	82.7	90.6	96.6	98.8	100.0	2,253
	18	16.6	41.5	59.3	76.0	89.4	97.4	100.0	72,527
	All ages	16.5	40.8	58.6	75.1	88.6	96.8	100.0	84,100

ANNEX C - Explanation of minor difference in value-added relationship for 18-year-old candidates between 2017 and 2019.

As shown in Figure 1. there is a small gap of approximately 0.14 grades between the regression lines for 18-year-olds, despite the awards using the same combined 2010 and 2011 outcome matrices. In the data exchange procedure for summer 2019, explicit reference was made to the removal from the predictions of candidates certificating both maths and further maths in the same series. Although previous versions of this document did not make explicit reference to this, such candidates have always been removed from the predictions, so their absence in 2019 does not explain this difference.

The likely cause of this gap is the change in 2017 from GCSEs in English, English literature, and maths that were graded A* – G to those grade 9 – 1. Until 2018, mean GCSE was reported on a 0 – 8 scale, 0 being U and 8 being A*. From 2018, both A* – G and 9 – 1 grades have been reported on a 0 – 10 scale. For the purposes of this analysis, mean GCSE scores were calculated for all candidates in all years using the new conversion:

Score	*0*	*1*	*1.25*	*2*	*2.5*	*3*	*3.75*	*3.9*	*5*	*5.1*	*6*	*6.8*	*7*	*8*	*8.5*	*9*	*9.5*	*10*
9 to 1	U		1		2		3		4		5		6	7		8		9
A to G*	U	G		F		E		D		C		B			A		A*	

Candidates taking new GCSEs in England have the potential to score slightly higher on mean GCSE than those taking GCSEs grade A* – G. The 2019 (18-year-old) cohort would have been the first to do 9 – 1 GCSEs in English, maths, and sciences, so their scores may have been a little higher overall. Whilst this should have little to no impact when using prediction matrices (where the distribution is divided into deciles) it might have a small effect when using a regression model. The apparent lower value-added in the model is possibly a result of 18-year-olds in 2019 having slightly higher mean GCSE, but being in the same mean GCSE deciles and receiving the same A level grades they always would have done. This may also apply to the 17-year-olds in 2018, whose value-added is fractionally lower than that of 17-year-olds in 2017 (if their GCSE mix were similar to the 18-year-olds' the following year – it may not be because they are more likely to have done more GCSEs early than the 18-year olds did). 17-year-olds in 2019 will have taken a greater proportion of 9 – 1 GCSEs than the previous cohort did; however, in 2019 the regression line is above those from previous years. It is possible that the model slightly underestimates the extent of their increase in value-added.

ANNEX D – Summary of changes to qualification content through reform

Table 38. Changes to pure maths Content – adapted from AQA (2017).

What's new
A greater overarching emphasis on modelling and problem solving (AS and A-level).
Specific methods of proof eg disproof by counter example (AS and A-level) and proof by contradiction, including irrationality of $\sqrt{2}$ and the infinity of primes (A-level only).
Use of functions, parametric equations, sequences and series in modelling (A-level only).
Use of logarithmic graphs for estimating parameters in exponential relationships (AS and A-level).
Trigonometric exact values, small angle approximations, trigonometric functions, geometric proofs of formulae (A-level only).
Gradient functions of a curve (AS and A-level).
Differentiation from first principles for polynomials (AS and A-level), sin and cos (A-level only).
Connected rates of change (A-level only).
Integration as the limit of a sum (A-level only).
Use of second derivatives for determining convexity, concavity and points of inflection (A-level only).
The Newton-Raphson method (A-level only).
What's gone
Remainder Theorem.
Volumes of revolution.
Mid-ordinate and Simpson's rule.
Vector equations of lines.
Scalar product (of vectors).
What's moved from A-level to AS
Use of exponential and logarithmic models using base e.
What's moved from AS to A-level
Sequences given by a formula for the n th term; increasing, decreasing and periodic sequences; sigma notation; arithmetic sequences and series; geometric sequences and series.
Radian measure, arc length, area of sector, area between two curves.
Trapezium rule.

Table 39. Changes to Mechanics Content – adapted from AQA (2017).

What's new
Derivation of formulae for constant acceleration for motion in a straight line (AS and A-level).
What's moved from A-level to AS
Use of vectors in two dimensions, magnitude and direction of a vector, position vectors, vector addition and multiplication by scalars.
Use of calculus in kinematics for motion in a straight line.
What's moved from AS to A-level
Derivation of formulae for constant acceleration for motion in two dimensions (using vectors).
Resolution of forces using Newton's second law.
Forces and dynamics for motion in a plane.
Use of the $F \leq \mu R$ model.

Table 40. Changes to Statistics Content – adapted from AQA (2017).

What's new
Selection and critique of sampling methods and data presentation techniques (AS and A-level).
Greater emphasis on making connections when calculating probability (AS and A-level).
Correlation coefficients (A-level only).
Statistical hypothesis testing (AS and A-level).
What's moved from A-level to AS
Application of the language of statistical hypothesis testing.
What's moved from AS to A-level
Conditional probability.
Normal and binomial distribution models.

ANNEX E - Breakdown of items included in the item difficulty comparative judgment exercise

Table 41 Tariff counts for the items from 2018 and 2019 assessments included in the comparative judgement of expected difficulty

			2018	2019
		Tariff	Count	Count
AQA	Paper 1	1	9	10
		2	7	12
		3	11	8
		4	6	5
		5	2	3
		6	0	0
		7	0	1
		8	0	0
		9	0	0
		10	1	0
	Paper 2	1	10	8
		2	5	8
		3	7	4
		4	4	5
		5	3	2
		6	1	2
		7	2	2
		8	1	1
		9	0	0
		10	0	0
	Paper 3	1	15	12
		2	8	13
		3	5	5
		4	4	3
		5	1	3
		6	2	1
		7	3	2
		8	0	0
		9	0	0
		10	0	0

			2018	2019
		Tariff	Count	Count
OCR A	Paper 1	1	3	5
		2	6	9
		3	7	7
		4	10	8
		5	0	2
		6	1	1
		7	0	0
		8	1	1
		9	0	0
		10	1	0
	Paper 2	1	18	13
		2	9	8
		3	6	6
		4	5	5
		5	1	4
		6	0	1
		7	3	1
		8	0	0
		9	0	0
		10	0	0
	Paper 3	1	6	6
		2	5	6
		3	8	4
		4	1	8
		5	3	3
		6	3	1
		7	2	1
		8	0	0
		9	1	0
		10	0	1

		2018		
		Tariff	Count	Count
OCR B (MEI)	Paper 1	1	7	7
		2	9	5
		3	15	9
		4	3	8
		5	2	1
		6	0	1
		7	0	1
		8	1	0
		9	0	0
		10	0	0
	Paper 2	1	15	18
		2	16	7
		3	6	9
		4	1	4
		5	3	1
		6	0	1
		7	1	2
		8	0	0
		9	1	0
		10	0	0
	Paper 3	1	6	5
		2	6	9
		3	7	6
		4	3	4
		5	2	0
		6	1	3
		7	0	0
		8	1	0
		9	0	0
		10	0	0

		2018		
		Tariff	Count	Count
Pearson	Paper 1	1	4	6
		2	7	10
		3	7	8
		4	7	9
		5	4	0
		6	1	1
		7	1	0
		8	0	1
		9	0	0
		10	0	0
	Paper 2	1	7	5
		2	6	11
		3	10	11
		4	5	4
		5	3	1
		6	1	2
		7	0	1
		8	0	0
		9	0	0
		10	1	0
	Paper 3	1	15	9
		2	8	10
		3	7	5
		4	3	3
		5	3	3
		6	2	2
		7	0	0
		8	0	1
		9	1	1
		10	0	0

Annex F – Item level estimates of expected difficulty

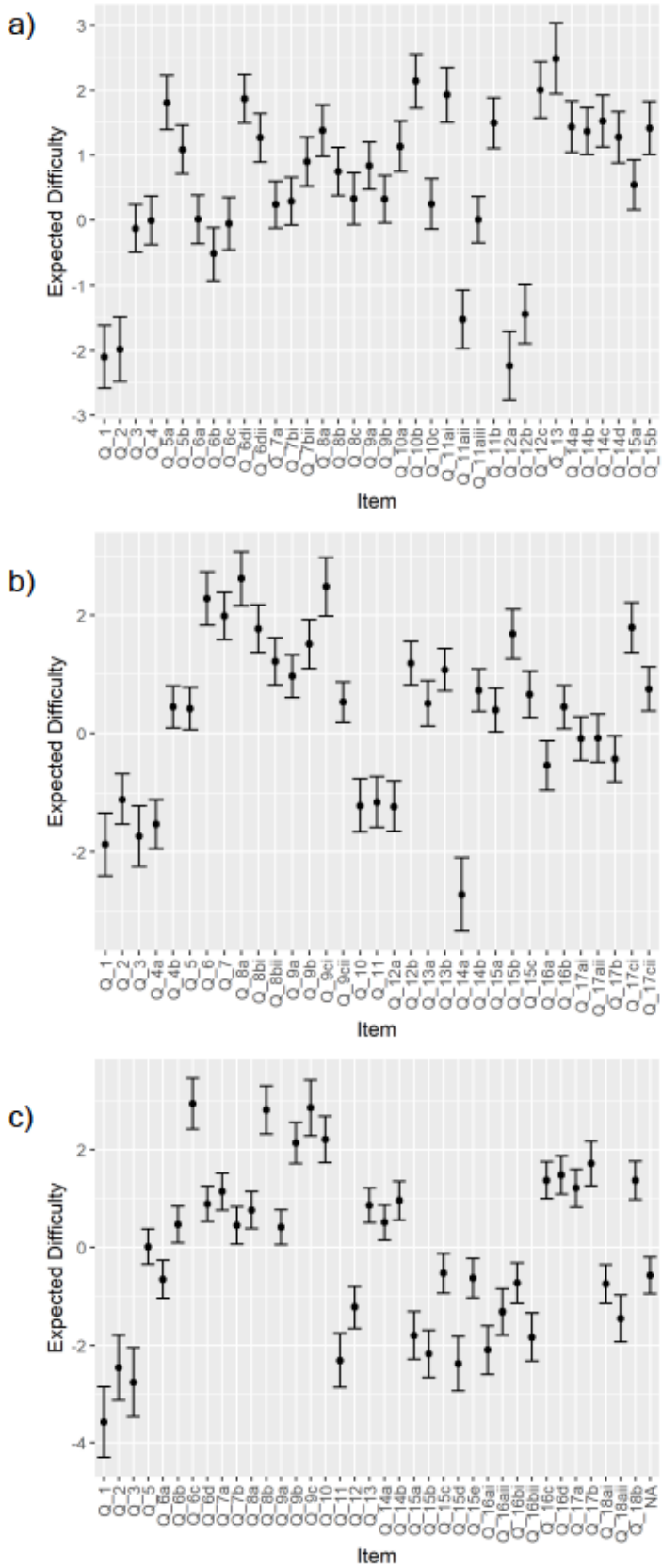


Figure 36 Item estimates of expected difficulty for AQA 2018 a) Paper 1, b) Paper 2 and c) Paper 3

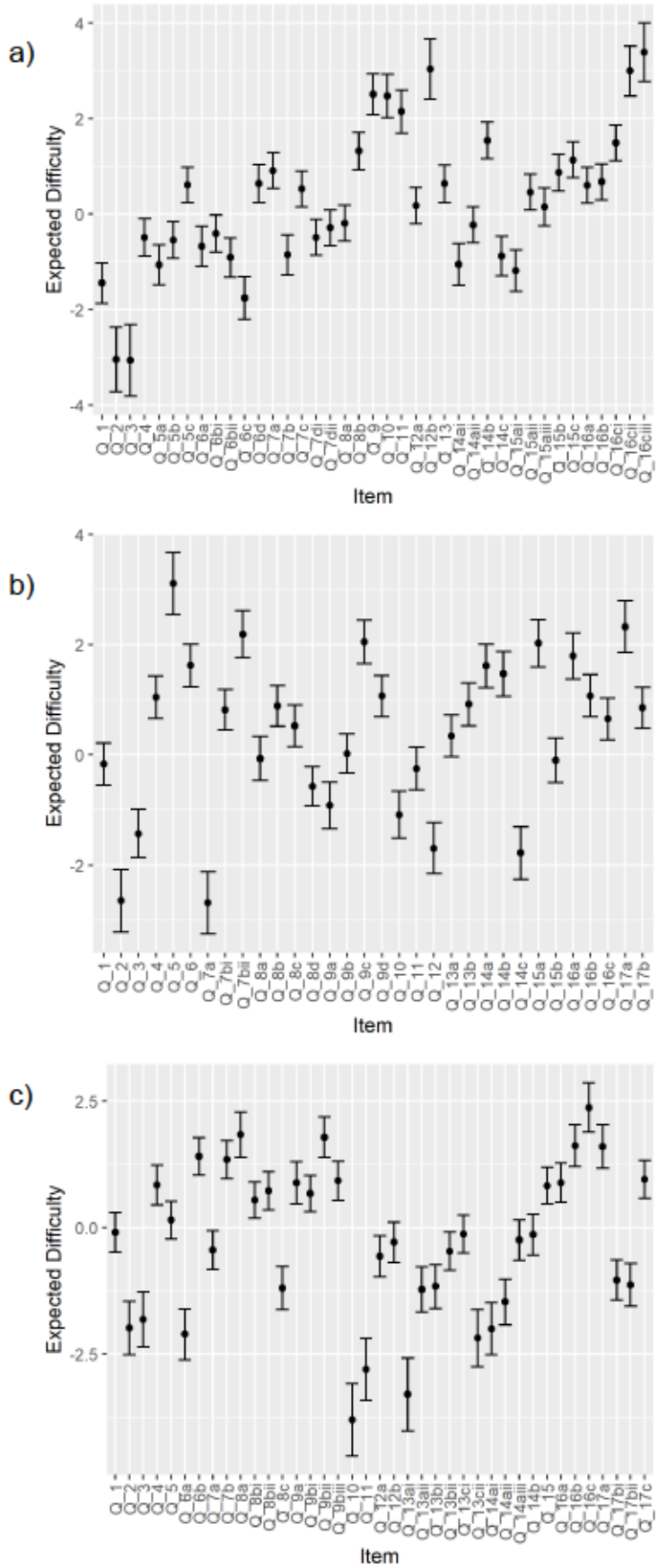


Figure 37 Item estimates of expected difficulty for AQA 2019 a) Paper 1, b) Paper 2 and c) Paper 3

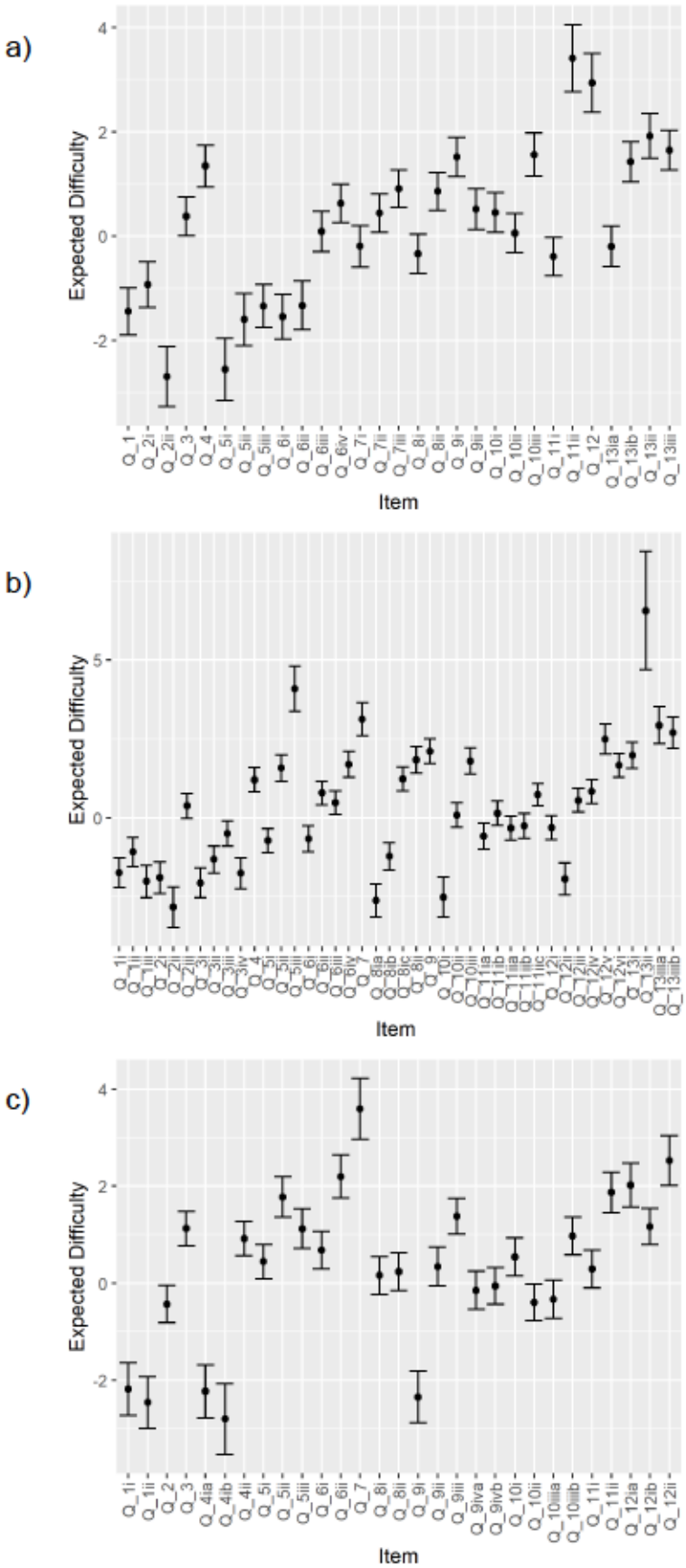


Figure 38 Item estimates of expected difficulty for OCR A 2018 a) Paper 1, b) Paper 2 and c) Paper 3

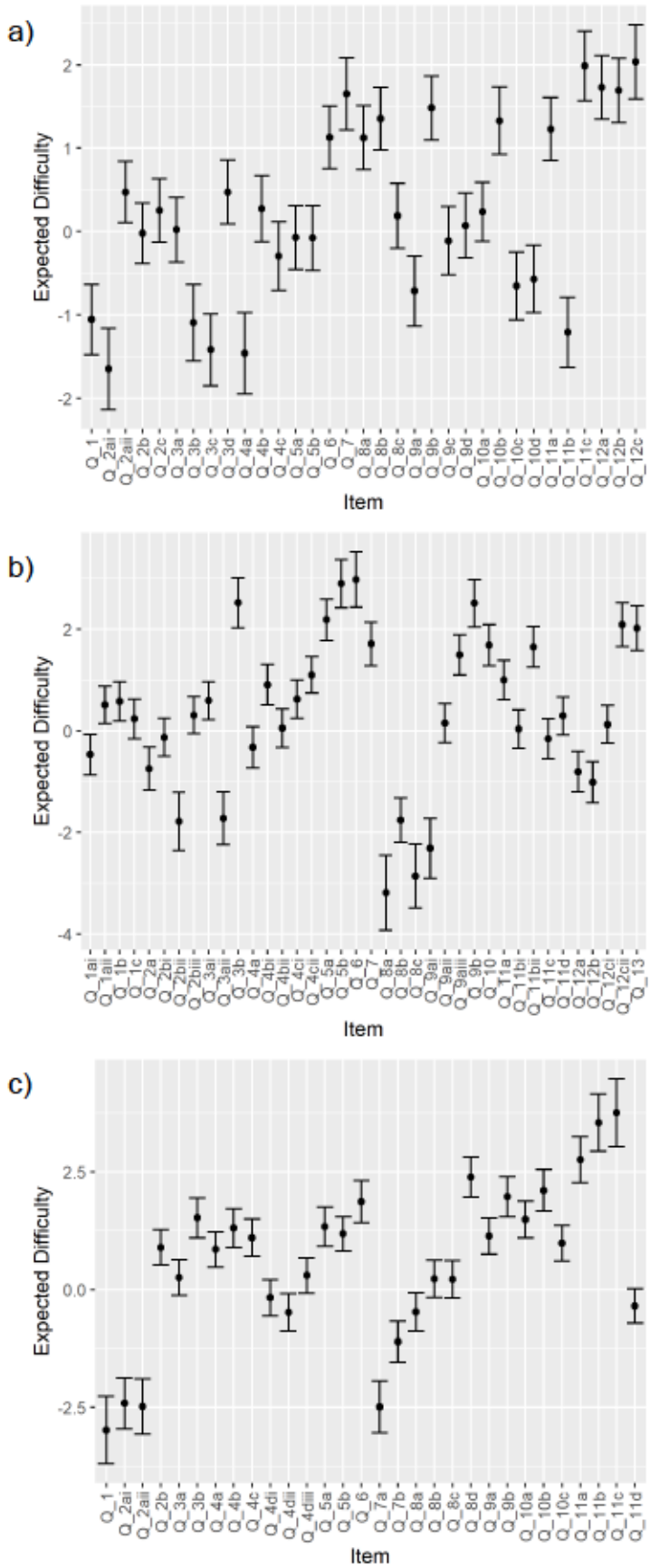


Figure 39 Item estimates of expected difficulty for OCR A 2019 a) Paper 1, b) Paper 2 and c) Paper 3

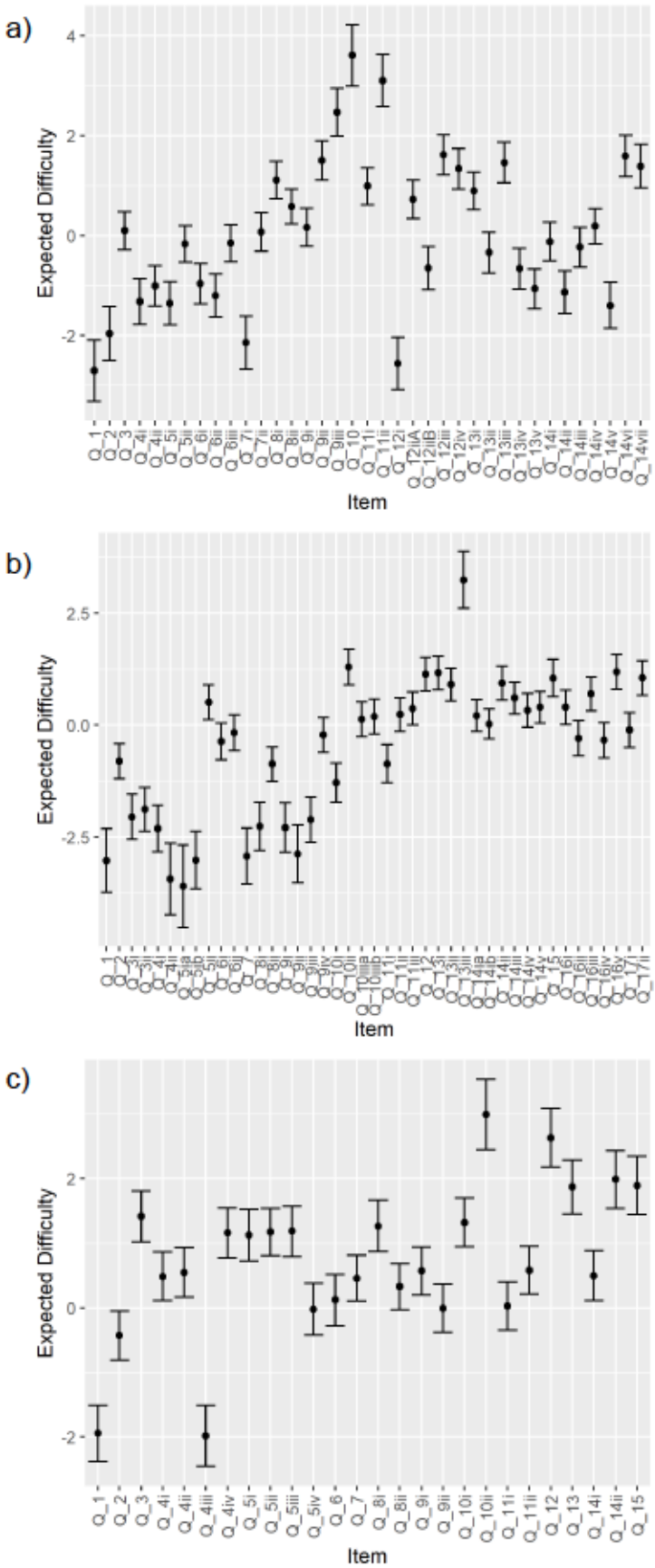


Figure 40 Item estimates of expected difficulty for OCR B (MEI) 2018 a) Paper 1, b) Paper 2 and c) Paper 3

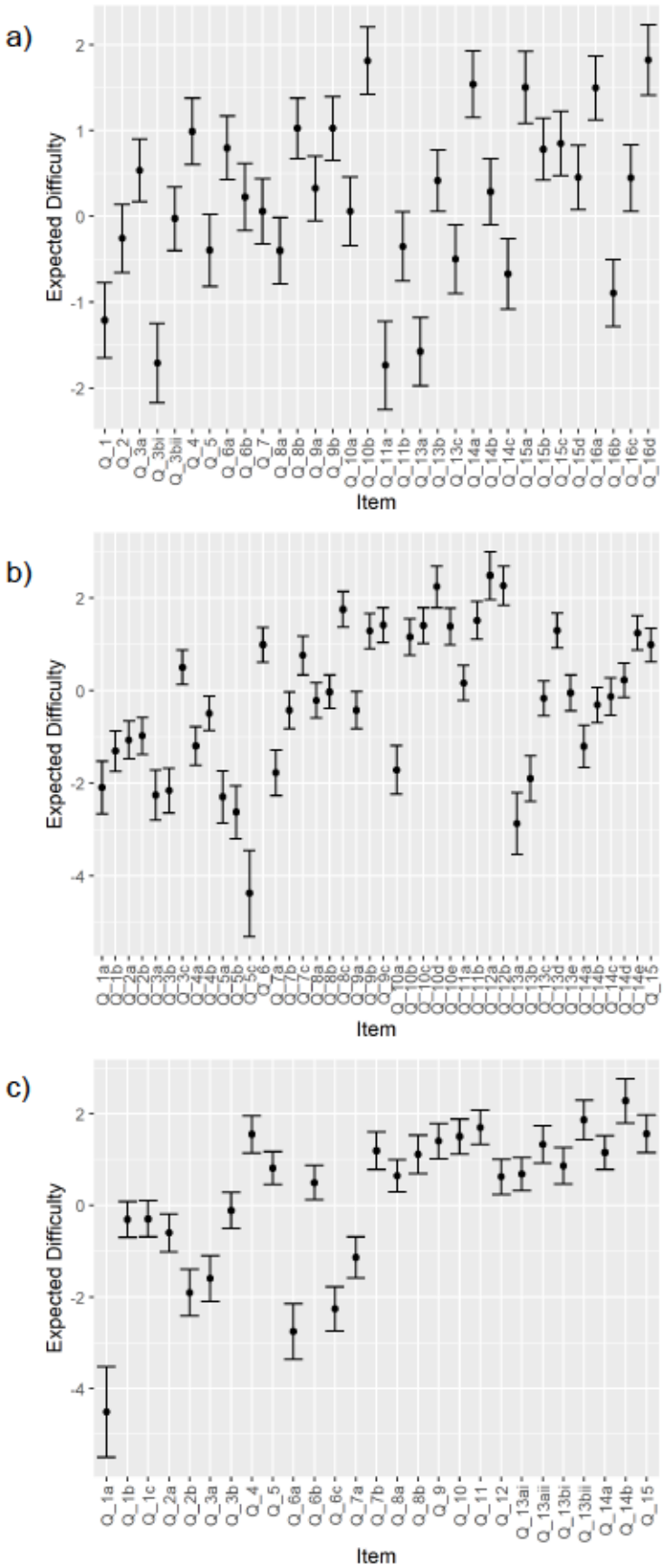


Figure 41 Item estimates of expected difficulty for OCR B (MEI) 2019 a) Paper 1, b) Paper 2 and c) Paper 3

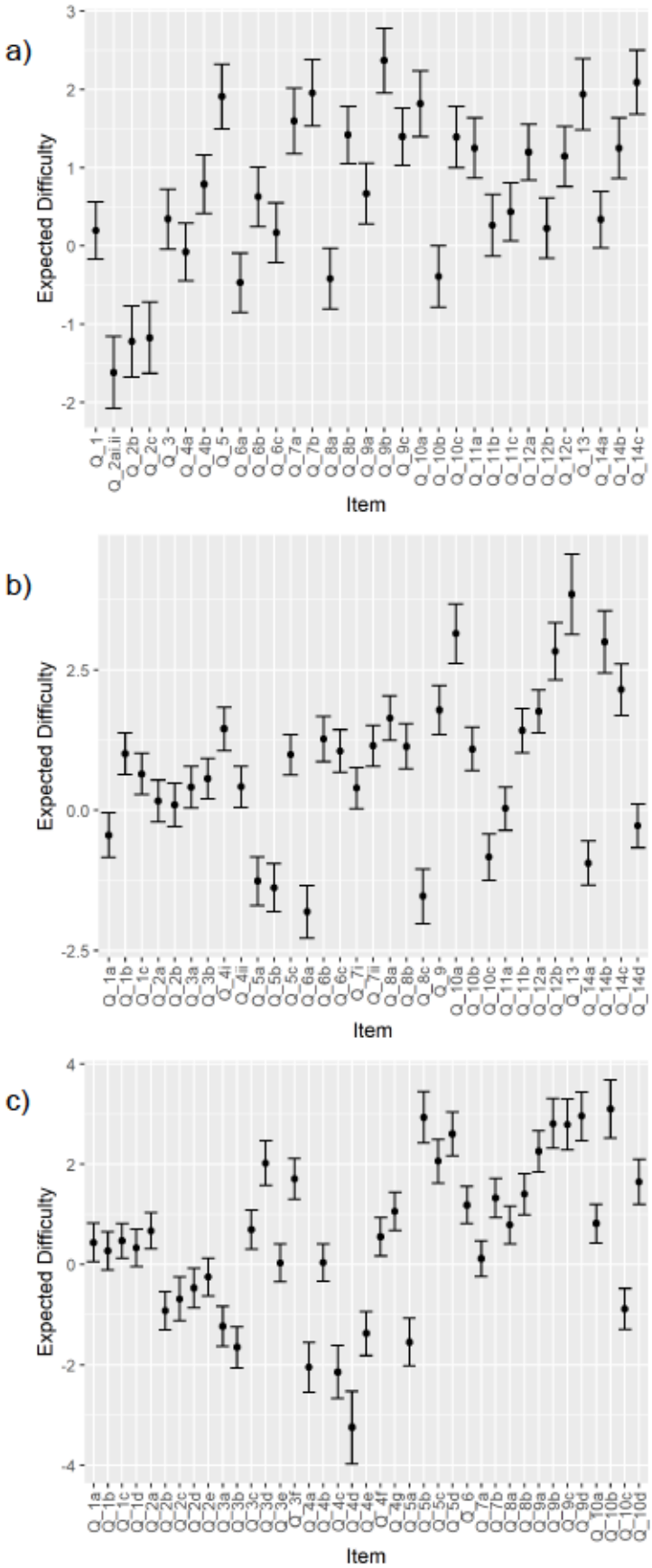


Figure 42 Item estimates of expected difficulty for Pearson 2018 a) Paper 1, b) Paper 2 and c) Paper 3

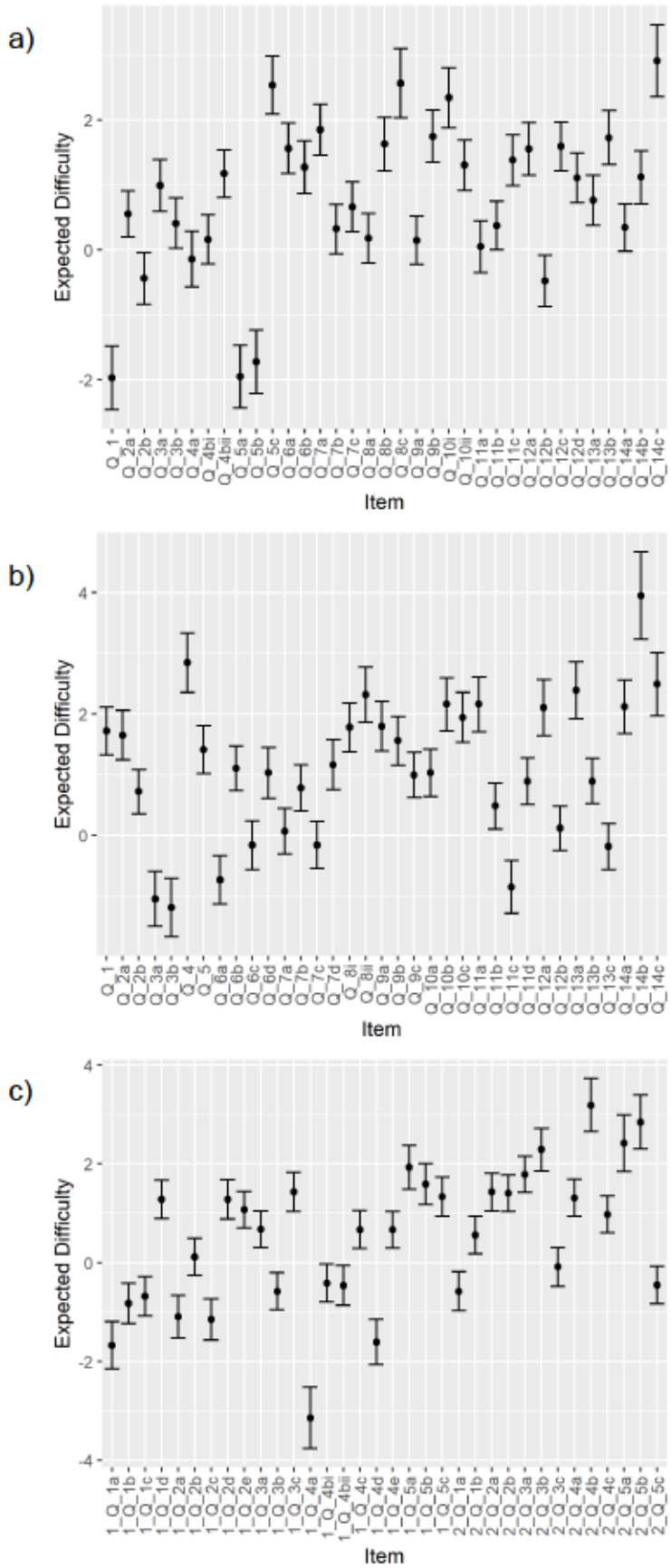


Figure 43 Item estimates of expected difficulty for Pearson 2019 a) Paper 1, b) Paper 2 and c) Paper

ANNEX G – Item facility index histograms

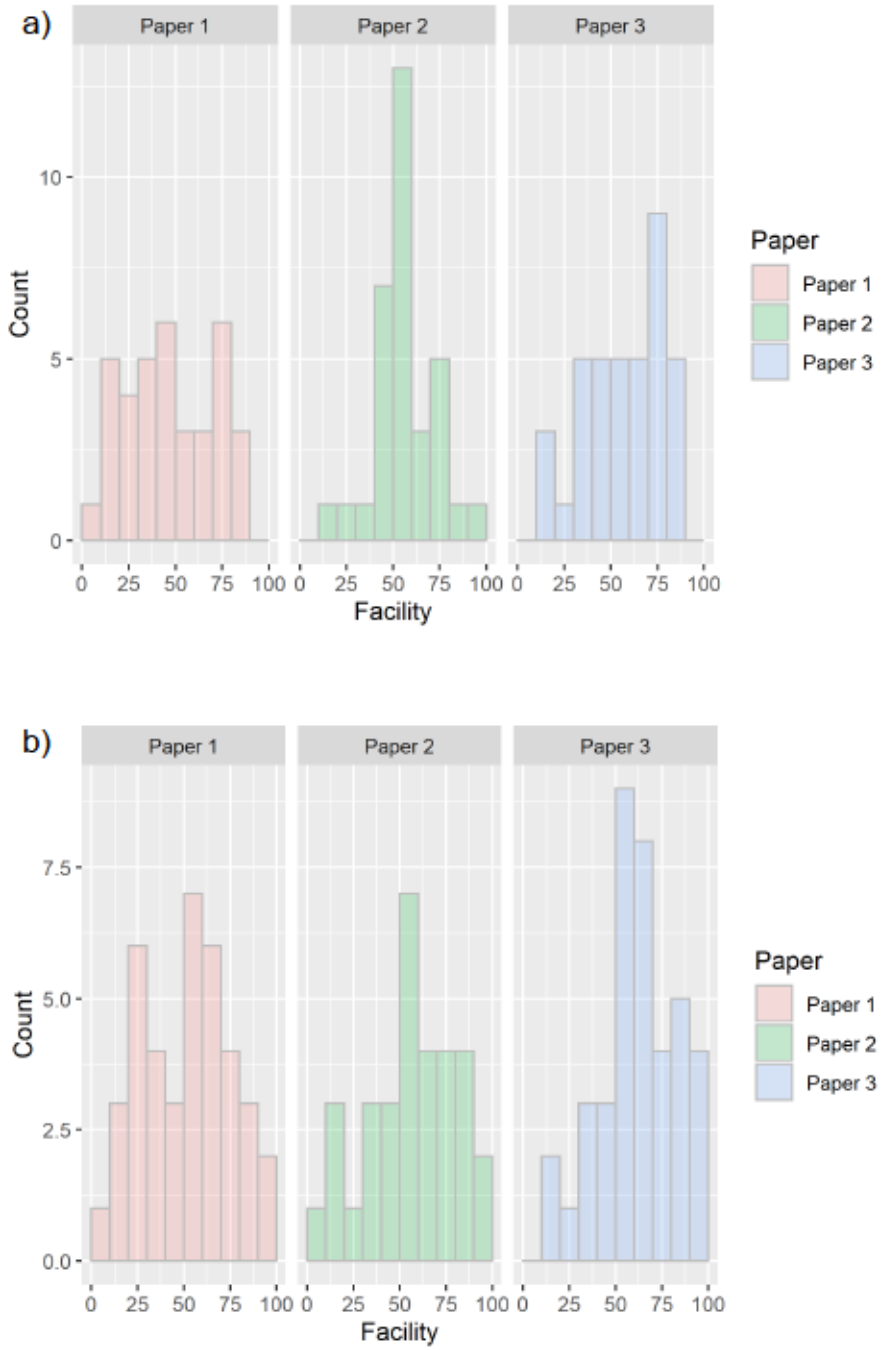


Figure 44 Item facility indices for the AQA question papers from a) 2018 and b) 2019

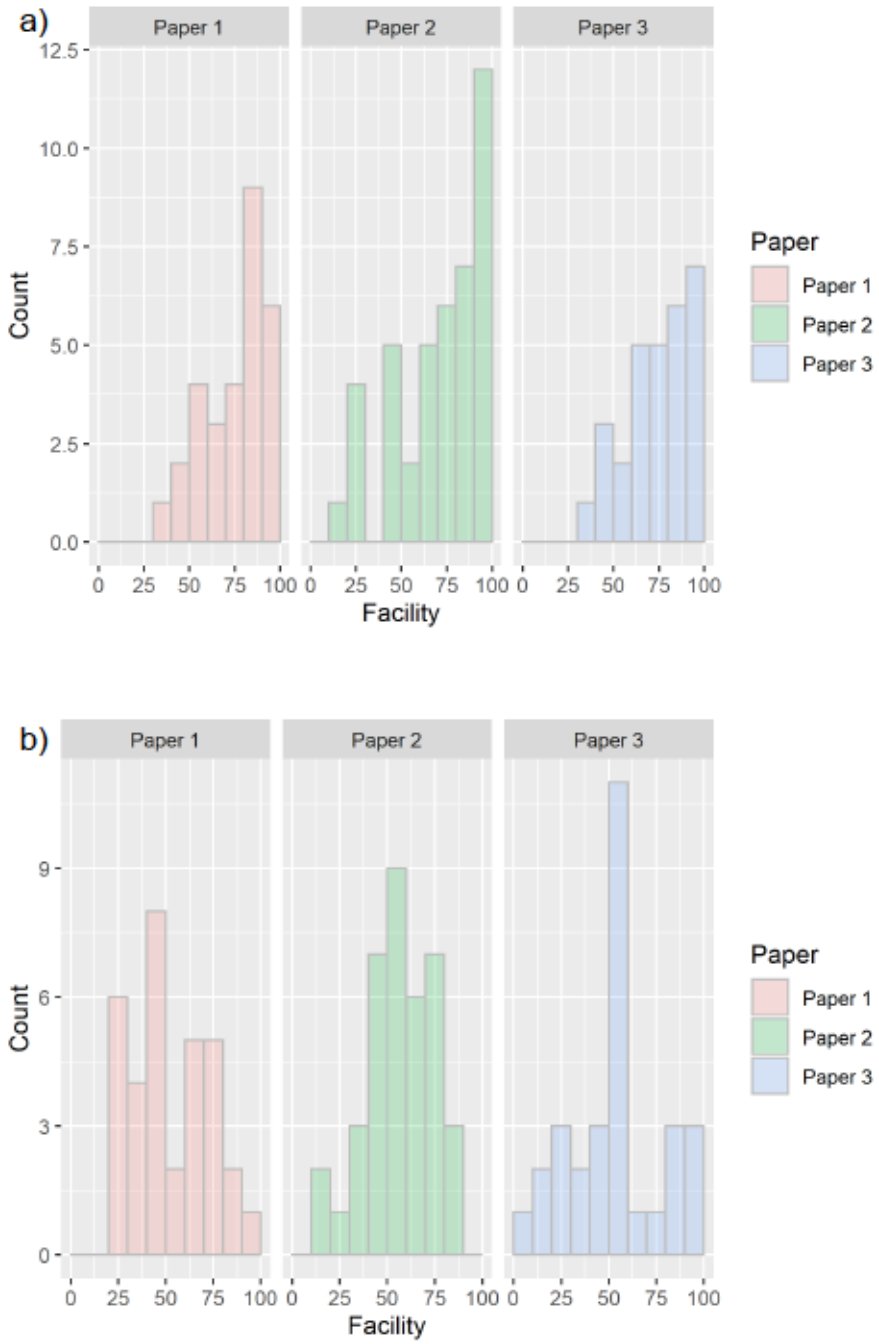


Figure 45 Item facility indices for the OCR A question papers from a) 2018 and b) 2019

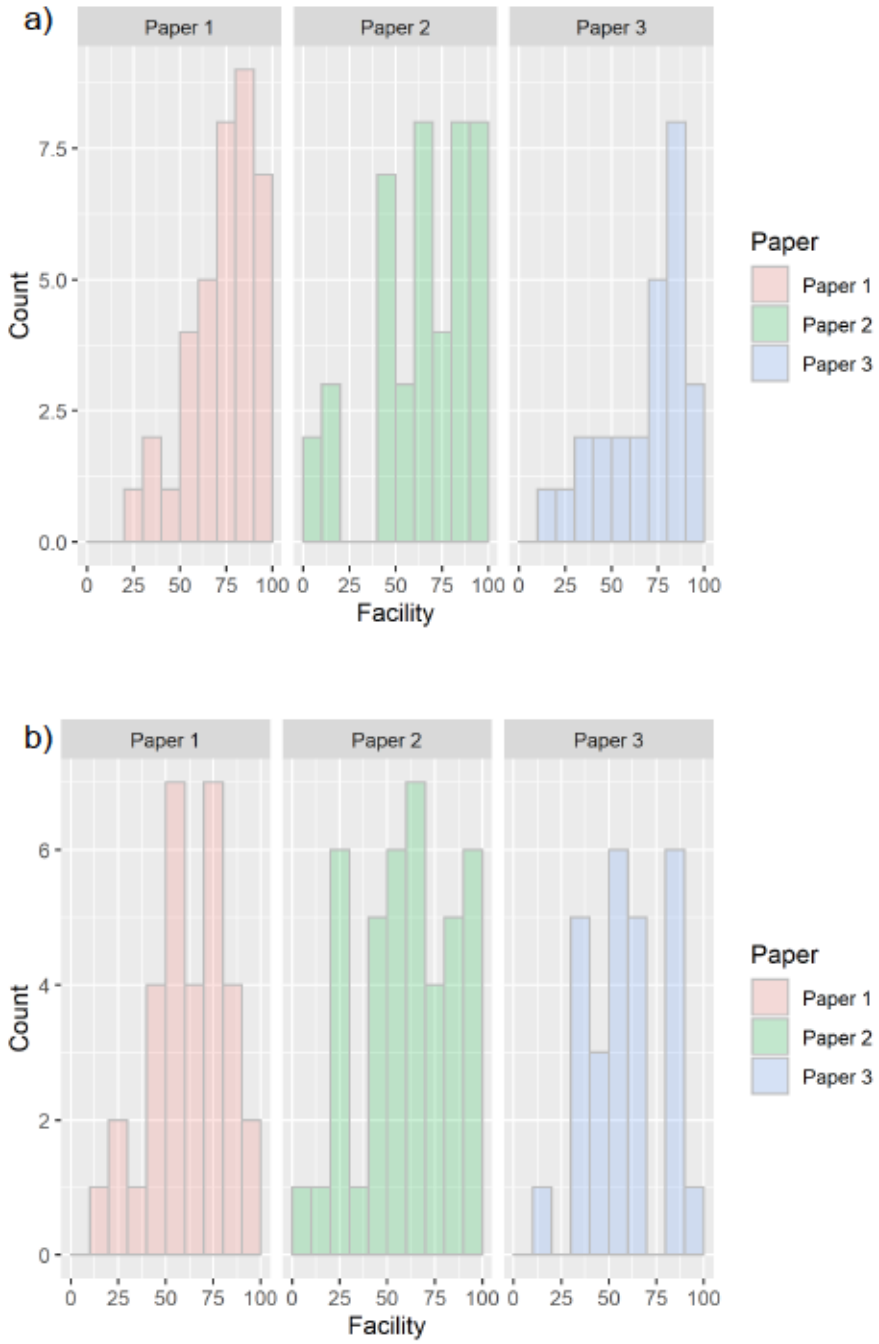


Figure 46 Item facility indices for the OCR B (MEI) question papers from a) 2018 and b) 2019

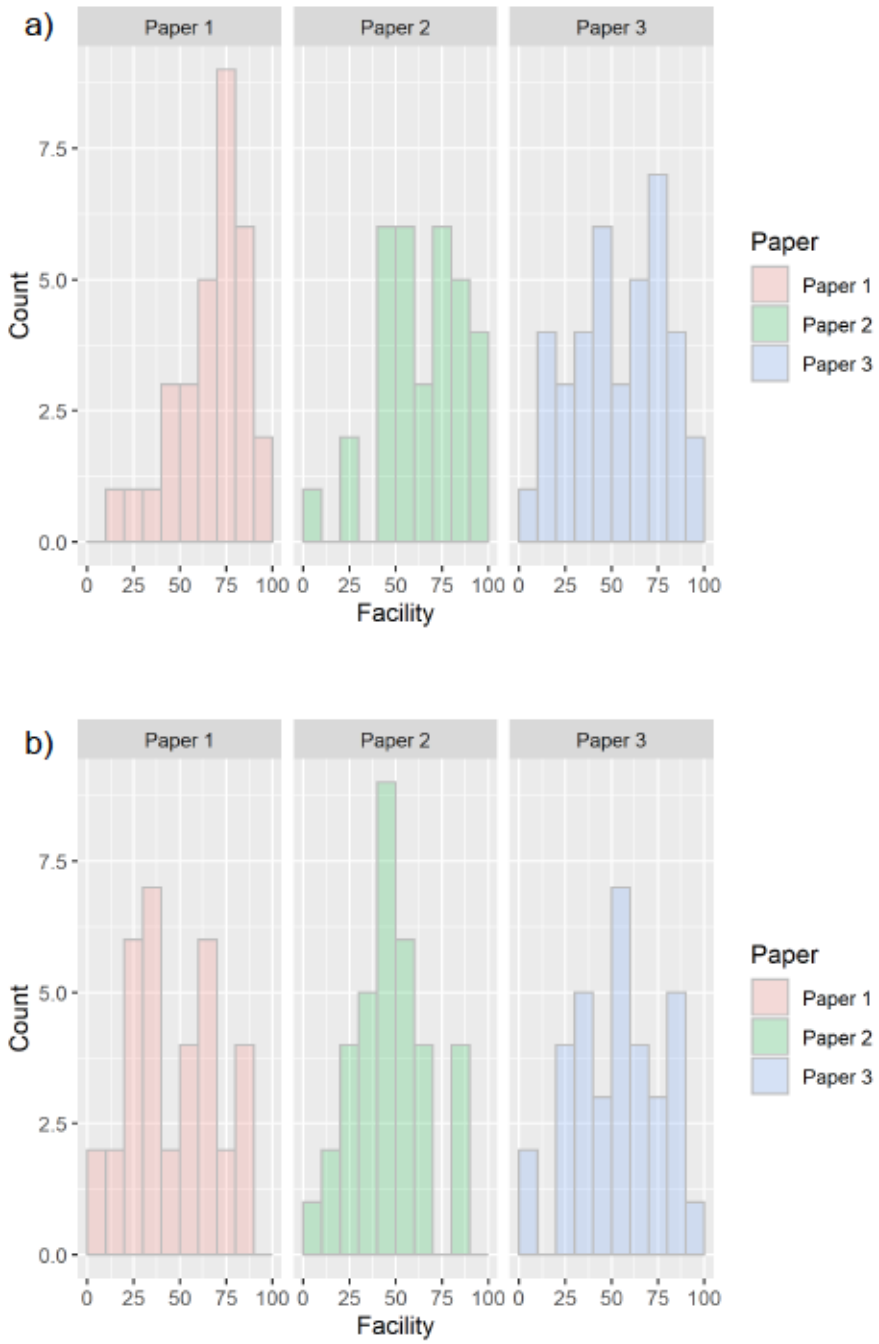


Figure 47 Item facility indices for the Pearson question papers from a) 2018 and b) 2019

ANNEX H – Template for the capture of views on the comparative judgement exercise

Reflections on paired comparisons of maths scripts

Thank you for participating in the script comparison exercise; it is an important part of the work we are doing to ensure that standards in A-level mathematics are maintained at comparable levels from series to series.

Judging performance standards from samples of candidates' scripts is recognised as a challenging task. Making paired comparisons has been proven to be a more intuitive task for judges that produces more robust data than other approaches. Nonetheless, there are still factors likely to confound judgements: some that cannot be controlled, like candidates' idiosyncrasies, and some that may be, such as those related to the structure of items and question papers.

We would like to ask you to reflect on the comparative judgement exercise you completed, so that your experience might inform our considerations of judgement uncertainty as part of this investigation, but also future question paper design and judging exercises.

Your thoughts on the first four questions may be particular to individual specifications, in which case please record them under the appropriate headings, if appropriate. The text boxes will expand as you type, should you need more space than provided.

1) Were there any aspects of the **overall question paper structure** that aided the comparison of performances between scripts? If so, what were they and how were they helpful?

AQA OCR A OCR B (MEI) Pearson General

2) Were there any aspects of the **overall question paper structure** that hindered the comparison of performance between scripts? If so, what were they and how were they unhelpful?

AQA OCR A OCR B (MEI) Pearson General

3) Were there any aspects of the **question design** that aided the comparison of performance between scripts? If so, what were they and in what way were they helpful?

AQA OCR A OCR B (MEI) Pearson General

4) Were there any aspects of the **question design** that hindered the comparison of performance between scripts? If so, what were they and in what way were they unhelpful?

AQA

OCR A

OCR B (MEI)

Pearson

General

5) What aspects of a candidate's performance **most influenced** your view of their ability? Were they the characteristics you had anticipated being important?

6) Where you found it **easy** to judge a candidate's ability from a script, what characteristics of the candidate's responses made the judgement easy?

7) Where you found it **difficult** to judge a candidate's ability from a script, what characteristics of the candidate's responses made the judgement difficult?

8) Any other comments

Published by:

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

0300 303 3344
public.enquiries@ofqual.gov.uk
www.gov.uk/ofqual

December 2019 **Ofqual/19/6567/1**