Maintaining standards in the first awards of the 9 to 1 GCSEs

Principles and evidence to inform the approach for summer 2017

December 2017

Ofqual/17/6328

Contents

1 Executive summary

This summer saw the first awards of reformed 9 to 1 GCSE qualifications in English language, English literature and mathematics. These qualifications were taken by the majority of the 16 year old cohort, since they are the only qualifications that count in school performance tables (in these subjects and at this level). In recent years, a proportion of the 16 year old cohort sat level 1/2 certificate qualifications instead of GCSEs, particularly in English language.

The awarding of GCSE qualifications is guided by prior attainment based predictions. Given the change in the cohorts this summer, there was a question of whether the predictions for the reformed 9 to 1 GCSE qualifications this summer should be based on GCSE only outcomes (as in the past), or combined GCSE and level 1/2 certificate outcomes. The latter would account for some students having previously sat level 1/2 certificate qualifications rather than GCSEs.

In considering how the predictions should be generated, there were two key questions: i) do the students who previously sat GCSEs or level 1/2 certificates perform differently, such that not including the level 1/2 certificate outcomes in the predictions would not represent the entire cohort this summer; and ii) can we be certain that the standards of the qualifications were precisely aligned in previous series. If we could not be certain that the standard of the GCSE and level 1/2 certificate qualifications were precisely aligned in previous series, then it would follow that the predictions this summer should not be based on combined GCSE and level 1/2 certificate outcomes, since this might compromise the GCSE standard.

A series of analyses were undertaken to consider these questions, focusing on English language and English literature (mathematics was not included due to the relatively small number of students sitting level 1/2 certificates in this subject). These analyses suggested that students sitting GCSEs and level 1/2 certificates did perform differently on the two qualifications (once prior or concurrent attainment was controlled for), with students generally performing better on the level 1/2 certificates. However, this was not necessarily due to the characteristics of the students taking the two qualifications, since the analyses also suggested that we cannot be certain that the standards of GCSEs and level 1/2 certificates were precisely aligned in previous series.

The findings from these analyses were used to inform discussions with the exam boards offering these qualifications, via the Standards and Technical Issues Group. Following these discussions, and on the basis of the evidence outlined above, Ofqual was of the view that the predictions this summer for the reformed 9 to 1 GCSE specifications should be based on GCSE only outcomes.

This approach was given further impetus by the following two points. First, as students left the GCSE cohort between 2013 and 2016, predictions were not

adjusted in any way to account for students moving away from the GCSE to take level 1/2 certificates. It therefore seemed difficult to justify adjusting the basis of predictions to take account of students returning to the GCSE cohort, when there was no adjustment when they left. Second, as a matter of principle, Ofqual considered that the priority this summer should be to carry forward the standard from GCSEs rather than a combined GCSE and level 1/2 certificate standard. This is due to the differences in the structure, assessment design and content of GCSEs and level 1/2 certificates, and the different methods used for maintaining standards.

The decision that the predictions for the reformed 9 to 1 GCSE specifications should be generated based on GCSE only outcomes this summer was communicated by Ofqual to the JCQ exam boards on the 21 June 2017, upon publication of the summer 2017 data exchange procedures. While the evidence reported here relates to English language and English literature, we also considered that, as a principle, the same approach should be adopted for all three subjects, hence this approach was also applied to mathematics.

2 Introduction

This summer saw the first awards of reformed GCSE 9 to 1 specifications in English language, English literature and mathematics. These qualifications were the only specifications that counted towards school performance tables in these subjects this summer, so were taken by the majority of 16-year-olds in schools and colleges in England. This signals a shift from recent summer examination series when the entries for these subjects from 16-year-olds were split between GCSE and level 1/2 certificate qualifications (commonly known as international GCSEs).

The statistical predictions used to guide the setting of grade boundaries for GCSEs typically predict the expected outcomes in a given year based on GCSE outcomes in a previous series (known as the 'reference' series)[1]. This means that, in summer 2017, the predictions for the reformed GCSE specifications would be generated based on GCSE only outcomes. However, given the number of students that sat level 1/2 certificates last summer (and that the vast majority of students sat GCSEs this summer), there was a question around whether an alternative approach to generating predictions was needed, ie whether the predictions should be based on combined GCSE and level 1/2 certificate outcomes from a previous series, rather than GCSE only outcomes[2]. The former approach would include those students that were not part of the GCSE cohort last summer in the basis of the predictions.

In preparation for the summer 2017 awards, Ofqual was in discussion with the JCQ[3] exam boards[4] via the Standards and Technical Issues Group[5] regarding the basis of the predictions for the reformed GCSE specifications. Following these discussions (which have been supported by a number of pieces of analysis), Ofqual was of the view that the predictions for all of the reformed 9 to 1 GCSE specifications should be generated based on GCSE only outcomes. This decision was communicated to the JCQ exam boards on the 21 June 2017[6], upon publication of the summer 2017 data

[1] See https://ofqual.blog.gov.uk/2017/04/21/prediction-matrices-explained/ for more information about how statistical predictions are generated.
[2] A secondary question is which year the reference series should be for generating predictions for the reformed GCSE awards in summer 2017. Given that ultimately this was a technical decision the details are not considered in this report.
[3] The Joint Council of Qualifications represents the main exam boards offering GCSEs and A levels in England, Wales and Northern Ireland: AQA, CCEA, OCR, Pearson and WJEC.
[4] Also referred to as 'boards' throughout.
[5] The Standards and Technical Issues Group comprises representatives from each of the JCQ exam boards (AQA, CCEA, OCR, Pearson and WJEC) and Ofqual. The group is established to consider technical issues.
[6] See https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/639772/Monitoring_summary_letter_-_24_August_2017.pdf

exchange procedures that set out the approach to generating predictions and reporting outcomes to Ofqual this summer[7].

This report summarises the evidence that led to this decision. This includes considering the change in entry for GCSE and level 1/2 certificate qualifications in recent years, the conditions under which it was or was not appropriate to generate predictions based on combined GCSE and level 1/2 certificate outcomes, and the evidence to support the approach taken. The rationale for the approach taken is provided in the conclusion.

The report and the analyses focus on English/English language[8] and English literature qualifications, since the entries to level 1/2 certificates in mathematics have remained small[9]. However, the evidence relating to English language and English literature was also used to inform the approach for mathematics.

2.1 GCSE and level 1/2 certificate entries in English/English language and English literature 2012-2016

From 2012 to 2016, a number of level 1/2 English qualifications have been available for learners at the end of Key Stage 4 (KS4). The GCSE suite of qualifications has comprised separate GCSEs in English, English language and English literature. The GCSE English qualification was aimed at students taking just one GCSE in this subject area and included aspects of both English language and literature, so tended to be taken by the lower ability students. The higher ability students tended to take two separate GCSE English qualifications: one in English language and one in English literature.

During the same time period, alternatives to GCSEs known as level 1/2 certificates (or international GCSEs) were also available in English language and English literature[10]. While these qualifications were also aimed at 16-year-old learners reaching the end of KS4 and were included in performance tables, they differed from the GCSE in terms of structure, assessment design and content. For example, since 2014, GCSEs in English/English language have comprised a combination of

[7] See https://www.gov.uk/government/publications/data-exchange-procedures-for-a-level-gcse-level-1-and-2-certificates

[8] Also referred to as English language throughout.

[9] See https://www.gov.uk/government/statistics/summer-2017-exam-entries-gcses-level-1-2-certificates-as-and-a-levels-in-england. The majority of entries to level 1/2 certificates in mathematics have been to the AQA further mathematics qualification. This qualification is not considered directly comparable to the GCSE or other level 1/2 certificates since it adopts a different grading scale (A^ - C).

[10] Level 1/2 certificates were available from the majority of JCQ exam boards and Cambridge International (CIE). Qualifications were available in English language and English literature (English was not available). CIE offered two level 1/2 certificate English language qualifications (0522 and 0500) but this report only focuses on 0522. Any references to the CIE specification throughout this report therefore refer to 0522 only.

controlled assessments[11] and written examination papers, while the most popular level 1/2 certificate qualification (offered by CIE) has continued to incorporate speaking and listening performance into the final grade alongside coursework and written papers[12]. Thus, the overall weighting of the assessment objectives that contribute towards the final grade are different. Similarly, the style of the question papers and the written assessment is different, reflecting differences in content. Furthermore, the CIE level 1/2 certificate qualification offers optional routes through the qualification, so that students may choose either a coursework route or a written route at each tier – no such optionality is available in the GCSE specification. Thus, while the title of the qualifications might be the same, the differences are fairly substantial. Nonetheless, both types of qualification were included in school performance tables prior to summer 2017.

Table 2.1 shows the entries for 16-year-old students to each qualification type between 2012 and 2017. As can be seen, there was a fairly significant increase in entries to level 1/2 certificates between 2012 and 2016 for both English language and English literature, with a particularly large increase between 2014 and 2015 (the entries stabilised somewhat between 2015 and 2016). This is coupled with a decrease in the entries to the GCSE versions of the qualifications over the same time period. There is a further shift in summer 2017 when the majority of students were entered to the (reformed) GCSE version of the qualifications, since these are the only qualifications that counted in school performance measures.

Table 2.1. *Entry numbers for GCSE and level 1/2 certificates summer 2012 – 2017 in England, 16-year-old students only*[13]

Year	English/English language		English literature	
	GCSE	L1/2 cert	GCSE	L1/2 cert
2012	*532,480*	4,990	*415,120*	4,200
2013	539,280	44,850	410,880	17,540
2014	375,060	110,170	395,900	51,300
2015	364,730	187,560	362,430	88,110
2016	335,110	190,720	367,880	98,690
2017	536,180	4,250	524,420	1,420

Note: 2012 GCSE data are approximate only. Entry numbers are rounded to the nearest 10.

[11] Controlled assessment replaced coursework in GCSEs in 2009. The CIE specification has retained coursework.
[12] Since summer 2014 speaking and listening has not contributed to the grade in GCSE English/English language but has been reported as a separate endorsement.
[13] Note that prior to 2014 students could make their first entry in the November examination series. The decrease in entry to the GCSE therefore might not correspond to the increase in entry to level 1/2 certificates between two years when considering just the summer examination series.

2.2 Generating predictions in summer 2017

The return of students to the GCSE cohorts in summer 2017 raised a question around how the statistical predictions used in awarding should be generated this summer. Predictions for GCSE have always been based upon GCSE only outcomes in the corresponding subject in the reference series, yet this summer there was a question of whether the predictions should be generated based on combined GCSE and level 1/2 certificate outcomes from the corresponding reference series instead. The latter approach would aim to ensure that any differences in the type of students taking the level 1/2 certificate qualifications in the reference series was accounted for this summer.

When considering which qualifications the predictions should be based on, it is worth considering under what conditions one would, or would not, want to generate predictions based on combined GCSE and level 1/2 certificate outcomes. The first issue to consider relates to the characteristics of the students. The main reason for generating predictions based on combined GCSE and level 1/2 outcomes would be because the type of students who were not in the GCSE cohort are different (and therefore perform differently) to those remaining in the GCSE cohort, ie they legitimately achieve different grades given their prior attainment. Without generating predictions based on combined outcomes, these differences would not be taken into account. As such, if the level 1/2 certificate students had (and indeed should have had) higher 'value-added' than the GCSE cohort, then their outcomes (results) would have been under-predicted based on GCSE only outcomes. However, if they had (and indeed should have had) lower value-added than the GCSE cohort, then their outcomes would have been over-predicted. Both situations are undesirable.

The second consideration relates to the standard of the qualifications themselves. It is possible that students might perform differently in the two qualifications (eg given their prior attainment), but this might be due to differences in the standards of the qualifications, rather than any differences in the students themselves. The methods via which standards are maintained are not the same in GCSE and level 1/2 certificate qualifications. For GCSE, the setting of grade boundaries is guided by statistical predictions that model the relationship between prior attainment (Key Stage 2 [KS2] results) and outcomes in a reference series, then apply this relationship to the current cohort of students. Thus, if the prior attainment of the students in the current year and the reference year are similar, then the outcomes would be expected to be similar (this is known as the comparable outcomes approach)[14]. In contrast, the setting of grade boundaries in level 1/2 certificates is not routinely driven by KS2-based predictions. This means that the relationship between

[14] The statistical predictions are generated using the same method for all exam boards that offer GCSE qualifications in the same subject, thus facilitating comparability across exam boards.

KS2 results and GCSE grades is not necessarily the same as the relationship between KS2 results and level 1/2 certificate grades.

Finally, the nature of the qualifications should be considered. While both GCSEs and level 1/2 certificates are qualifications that are aimed at learners at the end of KS4, the qualifications differ in terms of the structure, assessment design and content. Thus, even if students performed differently on the two qualifications, and the standards were considered to be precisely aligned, it still might not be considered appropriate to base the predictions on combined GCSE and level 1/2 certificate outcomes, since the nature of the qualifications are different.

In summary, the key issue here is whether or not the predictions should have been based on GCSE only outcomes or combined GCSE and level 1/2 certificate outcomes. To include the level 1/2 certificate outcomes in the predictions we would need to have been certain that these qualifications were of precisely the same standard as the GCSEs and aligned across grades. Thus, any differences in the performance of the students on the different qualifications would have been due to the nature of the students and not the standard of the qualifications (ie they did, and should, perform differently given their prior attainment). As such, the differences should be taken into account this summer. The analyses in this report aim to consider these issues (further details of the analyses are given in section 3).

Before considering the analyses themselves, it is worth considering the approach to awarding GCSE English/English language and English literature in recent years as students have left the GCSE cohort to take level 1/2 certificate qualifications. In this report we are concerned with whether predictions in summer 2017 should be based on GCSE only or combined GCSE and level 1/2 certificate outcomes. However, if there is an effect of students returning to the GCSE cohort in 2017, then equally there should have been an effect (in the opposite direction) of students leaving the GCSE cohort. It is therefore worth noting that during the period that students left the GCSE cohort no adjustments were made to the methods of generating predictions for GCSEs. This issue was discussed with the Standards and Technical Issues Group prior to the summer 2016 awards and it was decided that no action should be taken.

3 Overview of analyses

The following section provides details of the four pieces of analysis that were conducted to inform the approach to generating predictions in summer 2017 for the reformed GCSE English language, English literature and mathematics qualifications. The first two pieces of analysis (sections 4 and 5) compared the historical performance of students in the two types of qualifications – first using a syllabus pairs approach where the grades of students sitting both qualifications in the same examination series were compared, and second by considering the relationship between prior attainment and grades for students sitting either a GCSE or a level 1/2 certificate. The latter pieces of analysis (sections 6 and 7) consider whether any differences in performance between the GCSE and level 1/2 certificates identified in the first and second analyses were likely to be due to the characteristics of the students (that mean they should have performed differently), or the standards of the qualifications. While section 6 focuses on the relationship between the ability of the cohort and attainment in GCSE or level 1/2 certificate qualifications, section 7 considers a wider range of student characteristics that might potentially affect performance in English/English language and English literature qualifications. Table 3.1 summarises the key questions we considered and the corresponding analyses.

Table 3.1. *Overview of analyses*

Key question	Analyses
Did students taking GCSEs and level 1/2 certificates perform differently?	1. Syllabus pairs analysis to consider how students that entered both a GCSE and level 1/2 certificate in the same examination series performed on the two qualifications (section 4) 2. A comparison of the relationship between prior attainment and grades in GCSEs and level 1/2 certificates (section 5)
Can we be certain that the standards of the GCSEs and level 1/2 certificates were precisely aligned?	3. Inter-board comparability analysis to consider the alignment in the standards of GCSEs and level 1/2 certificates (section 6) 4. Propensity-score matching analysis to consider the performance of students with similar characteristics in GCSEs and level 1/2 certificates (section 7)

In each section the analyses relate to summer 2015 and summer 2016 data. Both years are considered since at the time the majority of the analyses were conducted the reference series for generating predictions for the reformed GCSE specifications

in summer 2017 had not been confirmed. The reference series was likely to be 2015 and/or 2016, hence the analysis of both years[15].

The analyses in this report are based on two sets of data: student-level data that is provided to Ofqual by exam boards each August, and the National Pupil Database (NPD – available from the Department of Education). The most appropriate dataset was used for each piece of analysis given the variables that were required and the availability of the data at the time that the analyses were conducted[16]. This approach aimed to ensure that, within each section, the maximum number of students were included. However, this means that the student population in each analysis differs slightly, and the figures are therefore not directly comparable.

A final point to note is that although the analyses all compare GCSEs and level 1/2 certificates, the comparisons are not necessarily all based on the same specifications. For example, for some analyses it is more appropriate to compare GCSEs with just one of the level 1/2 certificates (ie the one with the largest entry). Although this means that direct comparisons cannot be made between the findings in each section, it is possible to draw overall conclusions nonetheless.

[15] Prior to 2015 the entries in the summer examination series from 16–year-olds was less stable due to students certificating (for the first time) in the November series as well.
[16] The final NPD data for a given summer examination series is not available until the following January or later.

4 Using syllabus pairs analysis to compare the performance of students who took a GCSE and a level 1/2 certificate in English/English language

4.1 Research aim

This strand of analysis focused on comparing the grades achieved by students taking both a level 1/2 certificate qualification and a GCSE in English/English language using a syllabus pairs approach. By looking at students who have completed both awards in the same examination series we can gain an insight into how achievement differed between the two qualifications, given the assumption that the student has the same level of ability, motivation and preparedness when completing both subjects.

This strand of work initially aimed to look at both English/English language and literature specifications, however the number of students taking both a GCSE and a level 1/2 certificate specification in English literature was very low (34 students in total in 2016 and fewer in 2015), so the analysis focussed on English/English language specifications. Furthermore, given that the most popular specification among level 1/2 certificates in English language is the CIE specification, the majority of the analysis focused on this qualification.

4.2 Data

Data was taken from Ofqual's summer data request to exam boards in 2015 and 2016. This data included: details of specification codes, individual student information and grades awarded. An issue with this data is that, for the CIE specifications, standard UCI Numbers (Unique Candidate Identifier) are not available, therefore students had to be linked using other information. A unique identifier was created for each student combining: centre number, centre student number, date of birth and gender, which was used for matching across qualifications.

Analysis focused on 'English language' and 'English' specifications combined (referred to as 'English' throughout this section). Analysis therefore identified students taking all combinations of English and an English language specification. Level 1/2 certificates from Pearson, CIE, AQA and WJEC were linked with GCSE specifications from Pearson, CCEA, AQA, OCR and WJEC. Only level 1/2 certificates included in performance measures were included in the analysis.

4.3 Methodology

Using a syllabus pairs approach students who took both a level 1/2 certificate and GCSE in English in the same examination series were identified and their results across the two subjects compared. Traditionally, syllabus pairs analysis is used to compare the difficulty of two subjects taken by the same student (eg maths and

English). This type of analysis allows some comparison of the difficulty and potential difference in standards between two subjects (Nuttall, *et al.,* 1974). However, it has also been used for other purposes such as inter-board comparison. Here, we are applying this methodology to compare two different qualifications but in the same subject. This may have some bearing on the interpretation of the results, particularly as this type of analysis is inherently based on a number of assumptions about the students and their performance.

The first assumption of a syllabus pairs analysis is that students taking both subjects will perform to a similar ability in both (Nuttall, *et al.,* 1974). Some factors could potentially lead to a difference in the performance of students between subjects, for example due to changes in motivation, or teaching quality. In this analysis the impact of teaching quality should be minimal as students took both qualifications in the same subject and therefore teaching should be more similar. However, there may still be differences if materials to prepare students for one specification are better than the other, or if teachers focus on content or materials for one of the specifications over the other. This may be particularly evident as the nature of the assessment in GCSEs and level 1/2 certificates in English differ and students may have been better prepared for one type of assessment over the other. As both qualifications contain controlled assessment it is also possible that students' motivation may not be consistent when completing the two assessments.

A further limitation of the syllabus pairs methodology is that the students entered for both types of qualifications in English may not be representative of the whole cohort of students taking only one qualification (Goldstein and Cresswell, 1996), ie either a GCSE or level 1/2 certificate. Inevitably, we are only looking at a small subset of the entire cohort that is non-randomly sampled. Students taking more than one qualification may come from a subset of school types or from a certain ability range. In this analysis, students who took a GCSE and the CIE level 1/2 certificate were more likely to be of average attainment. Analyses reported in Appendix A show that 22.12% of students that took a GCSE (only) achieved a grade D, while 37.15% of students that took a GCSE and a level 1/2 certificate achieved a grade D (in their GCSE). Similarly, 32.43% of students that (only) took the CIE level 1/2 certificate achieved a grade C, compared to 43.86% of students that took the CIE specification and a GCSE. Considering the importance of grade C for both schools and students, this may be a consequence of teachers entering students on the C/D borderline for more than one qualification, in order to increase their likelihood of achieving at least a C in one.

The distribution of students across school types also differed between students entered for a GCSE and the CIE level 1/2 certificate specification compared to the wider cohort (see Appendix A for details). Compared to those just entered for a GCSE, students entering both qualifications were more likely to come from comprehensive schools and academies, but were less likely to come from selective schools or FE colleges. Compared to those just taking the CIE level 1/2 certificate,

students entering both qualifications were more likely to come from comprehensive schools, but less likely to come from FE establishments. Without knowing the reason why certain students were entered for both qualifications it is difficult to definitively explain the reason for any differences between students' grades. However, despite these limitations, the data available gives some insight into the relative achievement of students in the two qualifications. Any extrapolation of the results to the cohort as a whole should, though, be treated with caution.

Initially a cross-tabulation identified the number and percentage of students achieving each combination of grades from the GCSE and level 1/2 certificate. This was first carried out for the CIE level 1/2 certificate compared to all GCSEs combined, which gives an idea of the relationship between the CIE specification and the 'average GCSE'. Considering that there are a number of GCSE specifications provided by different exam boards, in order to provide an additional level of consistency to the findings, analysis was then restricted to just look at AQA GCSE as this is the most popular GCSE English specification and the majority of students entered to more than one qualification took their GCSE with AQA.

Following this, grades were converted to point scores (A*=8, A=7, B=6, C=5, D=4, E=3, F=2, G=1 U=0) so further statistical analyses could be performed. Spearman's rank correlation coefficient was calculated to assess the level of consistency between the two awards; generally, a high correlation would provide confidence that both qualifications are measuring the same aspect of student ability. A Wilcoxon Signed Rank Test was also used to identify if there is a consistent difference between scores that students obtained in the two qualifications.

For the final analysis the difference between the two scores was calculated by subtracting the GCSE grade score from the CIE level 1/2 certificate grade score. In this case, a negative result indicates that students obtained a lower grade on their CIE specification than their GCSE, but a positive score meant that they obtained a higher score in their CIE specification than GCSE. This analysis was further broken down by score obtained on the GCSE so the mean difference in scores could be computed across the grade range. This indicates the average difference in score between GCSE and the CIE level 1/2 certificate results for students obtaining each GCSE grade.

Analysis was carried out for both 2015 and 2016 to identify any consistent patterns over time. However, between these two years there was a significant change in entry numbers for some of the awards, which may have resulted in a change in the composition of the cohort. Therefore, changes in the relationship between GCSE and CIE level 1/2 certificate grades could be due to this change in cohort, for example, if there is a change in the ability range of the students entered for both qualifications and included in the analysis.

4.4 Results

4.4.1 Summer 2016 series

Initially 715,351 individual grades were included in the dataset. Tables 4.1 and 4.2 show a breakdown by exam board of level 1/2 certificate and GCSE entries, respectively. Table 4.3 shows that there were 2,674 students who had entered more than one qualification. This represents a small proportion of total students (0.38%) who are entered for either qualification.

Table 4.1. *Level 1/2 specifications included and entry numbers, by exam board – 2016*

Exam board	*Specification code*	*Entries*
Pearson	KEA0	9,766
CIE	0522	197,634
AQA	8705	28,250
WJEC	970001	1,497
All		237,147

Table 4.2. *GCSE specifications included and entry numbers, by exam board – 2016*

Exam board	*Specification code*	*Entries*
Pearson	2EN01/2EH01	41,519
CCEA	G9290/G9310	19,468
AQA	4707/4702	283,404
OCR	J355/J350/J345	27,007
WJEC	4170LA/4190LA	106,806
All		478,204

Table 4.3. *Number of students taking a level 1/2 certificate in English and a GCSE in English – 2016*

		GCSE				
		CCEA	AQA	OCR	WJEC	Pearson
Level 1/2	CIE	29	1613	183	706	109
	Pearson	0	5	1	3	1
	WJEC	0	10	0	13	1

From Table 4.3 it is apparent that CIE 0522 is the specification with the vast majority of students also taking a GCSE in the same subject. For this reason, syllabus pairs analysis was focused on the comparison of the CIE specification and GCSEs.

The main syllabus pairs comparison analysis for 2016 is presented in Tables 4.4 and 4.5. As an aid to interpret the Tables, cells have been shaded proportionally to the numbers presented in the cell. Table 4.4 indicates the number of students obtaining each combination of CIE level 1/2 certificate and all other GCSE grades and Table 4.5 calculates row percentages, ie what percentage of students who obtained each GCSE grade obtained each level 1/2 certificate grade. This analysis is then repeated for just AQA GCSE and reported in Tables 4.6 and 4.7.

Table 4.4. *Number of students with each grade in any GCSE and in CIE level 1/2 certificate – 2016*

		CIE level 1/2 certificate									
	grade	**A***	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**U**	***Total***
All GCSE	**A***	7	7	3	0	0	0	0	0	0	*17*
	A	9	47	23	10	1	0	0	0	0	*90*
	B	7	57	102	94	8	0	0	0	0	*268*
	C	1	23	122	438	111	6	0	0	0	*701*
	D	0	2	47	470	367	83	3	0	5	*977*
	E	0	0	8	101	167	79	9	0	8	*372*
	F	0	0	2	29	38	36	18	5	4	*132*
	G	0	0	1	8	12	10	6	5	5	*47*
	U	0	0	0	1	9	5	3	4	4	*26*
	x	0	0	0	7	2	1	0	0	0	*10*
	Total	*24*	*136*	*308*	*1158*	*715*	*220*	*39*	*14*	*26*	*2640*

Table 4.5. *Percentage of students with each grade in GCSE obtaining each grade in CIE level 1/2 certificate – 2016*

		CIE level 1/2 certificate								
	grade	**A***	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**U**
All GCSE	**A***	41.2	41.2	17.6	0	0	0	0	0	0
	A	10	52.2	25.6	11.1	1.1	0	0	0	0
	B	2.6	21.3	38.1	35.1	3	0	0	0	0
	C	0.1	3.3	17.4	62.5	15.8	0.9	0	0	0
	D	0	0.2	4.8	48.1	37.6	8.5	0.3	0	0.5
	E	0	0	2.2	27.2	44.9	21.2	2.4	0	2.2
	F	0	0	1.5	22.0	28.8	27.3	13.6	3.8	3
	G	0	0	2.1	17.0	25.5	21.3	12.8	10.6	10.6
	U	0	0	0	3.8	34.6	19.2	11.5	15.4	15.4
	X	0	0	0	70.0	20.0	10.0	0	0	0

Table 4.6. *Number of students with each grade in AQA GCSE and in CIE level 1/2 certificate – 2016*

		CIE level 1/2 certificate									
	grade	**A***	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**U**	***Total***
AQA GCSE	**A***	7	6	3	0	0	0	0	0	0	*16*
	A	9	43	17	4	1	0	0	0	0	*74*
	B	7	48	70	53	4	0	0	0	0	*182*
	C	1	15	81	243	62	1	0	0	0	*403*
	D	0	2	37	260	206	53	3	0	4	*565*
	E	0	0	6	74	102	44	4	0	7	*237*
	F	0	0	2	24	26	25	7	0	4	*88*
	G	0	0	1	6	9	6	3	2	5	*32*
	U	0	0	0	0	8	2	0	1	2	*13*
	x	0	0	0	1	1	1	0	0	0	*3*
	Total	*24*	*114*	*217*	*665*	*419*	*132*	*17*	*3*	*22*	*1613*

Table 4.7. *Percentage of students with each grade in AQA GCSE obtaining each grade in CIE level 1/2 certificate – 2016*

		CIE level 1/2 certificate								
	grade	**A***	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**U**
AQA GCSE	**A***	43.8	37.5	18.8	0	0	0	0	0	0
	A	12.2	58.1	23.0	5.4	1.4	0	0	0	0
	B	3.8	26.4	38.5	29.1	2.2	0	0	0	0
	C	0.2	3.7	20.1	60.3	15.4	0.2	0	0	0
	D	0	0.4	6.5	46.0	36.5	9.4	0.5	0	0.7
	E	0	0	2.5	31.2	43.0	18.6	1.7	0	3
	F	0	0	2.3	27.3	29.5	28.4	8.0	0	4.5
	G	0	0	3.1	18.8	28.1	18.8	9.4	6.3	15.6
	U	0	0	0	0	61.5	15.4	0	7.7	15.4
	X	0	0	0	33.3	33.3	33.3	0	0	0

From Table 4.4 it can be seen that, overall, 385 (14.6%) students achieved a lower grade in their CIE level 1/2 certificate than their GCSE and 1,178 (44.8%) students achieved a higher grade in their level 1/2 certificate. 40.6% students got the same grade in both, 86.5% were within one grade and 96.5% were within two grades. Table 4.5 shows that, among students who attained a grade D at GCSE, 37.6% achieved the same grade in the CIE specification and more than 50% achieved a better grade in the CIE specification.

This pattern is more pronounced when looking at only the AQA GCSE (Tables 4.6 and 4.7), where 14.3% of students got a lower grade on the level 1/2 certificate, 38.8% obtained the same grade and 46.9% obtained a higher grade on their level

1/2 certificate. Table 4.7 shows that students who attained a grade D in the AQA GCSE were more likely to achieve a better grade on their level 1/2 certificate (52.9%) than to perform worse (around 10% achieved lower than a grade D).

When converted to point scores, grades in both GCSEs and the CIE level 1/2 certificate were highly positively correlated (r_s=0.632, N=2640, p<0.001). However, there was a significant difference in scores between the CIE specification and other GCSEs taken (Wilcoxon Signed Rank test, Z=951700, p<0.001). CIE grades were on average 0.43 (SD=1.03) grade points higher than GCSE grades, when restricted to just comparison with AQA GCSE this increased to 0.49 (SD=1.09).

This is broken down by GCSE grade in Table 4.8, indicating on average how much higher or lower students' CIE grades were at each GCSE grade. Students who got a B or above in their GCSE tended to get a lower grade in their level 1/2 certificate, however students who got a C to U in their GCSE tended to get a higher grade in their level 1/2 certificate.

Table 4.8. *Mean difference between GCSE grade and CIE level 1/2 certificate grade after conversion to grade scores – 2016*

GCSE grade	Mean difference from CIE grade	95% confidence intervals	N
*	-0.76	[-1.12, -0.41]	17
A	-0.41	[-0.59, -0.23]	90
B	-0.15	[-0.25, -0.04]	268
C	0.07	[0.02, 0.12]	701
D	0.47	[0.42, 0.52]	977
E	0.97	[0.87, 1.07]	372
F	1.47	[1.25, 1.69]	132
G	2.00	[1.53, 2.47]	47
U	2.54	[1.93, 3.15]	26

Note: For the mean difference, a negative score indicates the CIE grade was on average lower than the GCSE grade, a positive score indicates the level 1/2 certificate grade was higher.

4.4.2 Summer 2015 series

In 2015, 706,714 individual grades were included in the dataset (see Table 4.9 and 4.10 for a breakdown by specification). After linking students taking both a level 1/2 certificate and a GCSE, this left 7,934 double entries (see Table 4.11 for a breakdown by qualification).

Table 4.9. *Level 1/2 specifications included and entry numbers, by exam board – 2015*

Exam board	Specification code	Entries
Pearson	KEA0	11,552
CIE	0522	184,597
AQA	8705	23,390
WJEC	970001	1,775
All		221,314

Table 4.10*. GCSE Specifications included and entry numbers, by exam board – 2015*

Exam board	Specification code	Entries
Pearson	2EN01/2EH01/2HN01/2NN01	43,471
CCEA	G9290	19,092
AQA	4707/4702/5702/5707	284,583
OCR	J355/J350/J345	29,554
WJEC	4170LA/4190LA	108,700
All		485,400

Table 4.11. *Number of students taking a level 1/2 English and another English language GCSE with another board – 2015*

		GCSE			
		AQA	OCR	Pearson	WJEC
Level 1/2	CIE	4353	918	501	1740
	Pearson	3	0	92	145
	WJEC	18	0	0	164

As can be seen in Table 4.12, of the students entered for a GCSE and the CIE specification, 1,236 (16.5%) students attained a lower grade in their CIE specification than their GCSE, and 3347 (44.7%) students attained a higher grade on their level 1/2 certificate. At GCSE grade D, 34.7% of students received the same grade in both qualifications, but 55.8% obtained a higher grade on the CIE specification.

Table 4.12. *Number of students with each grade in any GCSE and in CIE level 1/2 certificate – 2015*

		CIE level 1/2 certificate									
	grade	**A***	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**U**	***Total***
All GCSE	**A***	49	53	9	2	0	0	0	0	0	*113*
	A	89	229	113	29	1	0	0	0	0	*461*
	B	64	285	457	284	24	4	0	0	0	*1118*
	C	10	129	553	1102	332	34	0	0	3	*2163*
	D	3	32	214	1031	795	187	15	0	15	*2292*
	E	0	4	41	230	316	202	31	3	37	*864*
	F	0	2	13	59	76	77	45	18	25	*315*
	G	0	1	6	25	28	21	16	12	17	*126*
	U	0	1	0	3	4	8	2	4	8	*30*
	x	0	0	3	12	9	4	1	0	1	*30*
	Total	*215*	*736*	*1409*	*2777*	*1585*	*537*	*110*	*37*	*106*	*7512*

Table 4.13. *Percentage of students with each grade in GCSE obtaining each grade in CIE level 1/2 certificate – 2015*

		CIE level 1/2 certificate								
	grade	**A***	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**U**
All GCSE	**A***	43.4	46.9	8	1.8	0	0	0	0	0
	A	19.3	49.7	24.5	6.3	0.2	0	0	0	0
	B	5.7	25.5	40.9	25.4	2.1	0.4	0	0	0
	C	0.5	6	25.6	50.9	15.3	1.6	0	0	0.1
	D	0.1	1.4	9.3	45	34.7	8.2	0.7	0	0.7
	E	0	0.5	4.7	26.6	36.6	23.4	3.6	0.3	4.3
	F	0	0.6	4.1	18.7	24.1	24.4	14.3	5.7	7.9
	G	0	0.8	4.8	19.8	22.2	16.7	12.7	9.5	13.5
	U	0	3.3	0	10	13.3	26.7	6.7	13.3	26.7
	X	0	0	10	40	30	13.3	3.3	0	3.3

Again, this pattern is more noticeable when restricted to only AQA GCSE (Table 4.14 and 4.15), where 15.6% of students got a lower grade on their level 1/2 certificate and 47.5% received a higher grade on their CIE level 1/2 certificate. For students who received a grade D on their GCSE, 58.7% received a grade C or higher on the CIE specification and only 8.8% received a lower grade.

Table 4.14. *Number of students with each grade in AQA GCSE and in CIE level 1/2 certificate – 2015*

		CIE level 1/2 certificate									
	grade	**A***	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**U**	***Total***
AQA GCSE	**A***	37	35	4	2	0	0	0	0	0	*78*
	A	39	143	66	11	1	0	0	0	0	*260*
	B	33	165	287	172	15	3	0	0	0	*675*
	C	8	79	329	577	171	15	0	0	2	*1181*
	D	3	27	142	598	426	102	6	0	7	*1311*
	E	0	4	37	176	198	107	12	0	24	*558*
	F	0	2	11	47	50	43	16	5	12	*186*
	G	0	0	5	14	22	11	5	3	12	*72*
	U	0	1	0	2	4	8	0	1	6	*22*
	x	0	0	0	6	1	2	0	0	1	*10*
	Total	*120*	*456*	*881*	*1605*	*888*	*291*	*39*	*9*	*64*	*4353*

Table 4.15. *Percentage of students with each grade in AQA GCSE obtaining each grade in CIE level 1/2 certificate – 2015*

		CIE level 1/2 certificate								
	grade	**A***	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**U**
AQA GCSE	**A***	47.4	44.9	5.1	2.6	0	0	0	0	0
	A	15	55	25.4	4.2	0.4	0	0	0	0
	B	4.9	24.4	42.5	25.5	2.2	0.4	0	0	0
	C	0.7	6.7	27.9	48.9	14.5	1.3	0	0	0.2
	D	0.2	2.1	10.8	45.6	32.5	7.8	0.5	0	0.5
	E	0	0.7	6.6	31.5	35.5	19.2	2.2	0	4.3
	F	0	1.1	5.9	25.3	26.9	23.1	8.6	2.7	6.5
	G	0	0	6.9	19.4	30.6	15.3	6.9	4.2	16.7
	U	0	4.5	0	9.1	18.2	36.4	0	4.5	27.3
	X	0	0	0	60	10	20	0	0	10

When converted to rank scores, grades in both GCSEs and the CIE specification were again highly positively correlated (r_s=0.673, N=7512, p<0.001). There was also a significant difference in scores between the CIE specification and other GCSEs taken (Wilcoxon Signed Rank test, Z=7931800, p<0.001). CIE grades were on average 0.42 (SD=1.1) grade points higher than GCSE grades; when restricted to just AQA GCSE this increased to 0.5 (SD=1.15). This is broken down by GCSE grade below indicating, on average, how much higher or lower CIE grades were. This suggests, for example, that students who achieved A* or A in their GCSE

tended to get a lower grade in their level 1/2 certificate, however students who got a B or below in their GCSE tended to attain a higher grade in their level 1/2 certificate.

Table 4.16. *Mean difference between GCSE grade and CIE level 1/2 certificate grade after conversion to grade scores – 2015*

GCSE grade	Mean difference from CIE grade	95% confidence intervals	N
*	-0.68	[-0.81, -0.55]	113
A	-0.18	[-0.26, -0.11]	461
B	0.06	[0.01, 0.12]	1118
C	0.20	[0.16, 0.23]	2163
D	0.56	[0.52, 0.6]	2292
E	0.89	[0.8, 0.97]	864
F	1.27	[1.1, 1.44]	315
G	2.07	[1.75, 2.39]	126
U	2.33	[1.64, 3.03]	30

Note: For the mean difference, a negative score indicates the CIE grade was on average lower than the GCSE grade, a positive score indicates the level 1/2 certificate grade was higher.

4.5 Summary of the results

The results show that, relatively consistently between the two years of study, students obtained on average higher grades on their CIE level 1/2 certificate than on their GCSE. This difference was larger when only looking at AQA GCSE than when looking at all GCSEs combined. The strong correlation between results from GCSEs and the CIE level 1/2 certificate gives confidence that both qualifications are measuring student ability in the subject (Newton, 1997).

However, this pattern was not consistent across the grade range. The results suggest that the biggest differences between the two subjects occur around grades C and D. It appears that students who got a D in their GCSE on average tended to gain a slightly higher grade on their CIE specification with many students obtaining a grade C. Above a grade C, on average, students obtained a slightly lower grade in the level 1/2 certificate than the GCSE, and below a grade C students tended to obtain a higher grade on their level 1/2 certificate than the GCSE.

One explanation for these results could be a difference in standards between the two qualifications. However, there are substantial differences in the content and assessment of the two qualifications, meaning direct comparison is difficult. Differences in attainment could therefore indicate that students performed better on one type of assessment than the other. This could be due, for example, to better materials being available for the different assessments, or more practice on different

assessment types. Differences could also be caused by the method of assessment, particularly as the CIE specification contains a speaking and listening element in the final grade, which students may be more naturally familiar with. The syllabus pairs methodology does not allow for the differentiation between possible causes of differences in attainment between the two qualifications.

5 Outcome matrices and predictions

5.1 Research aim

This strand of the analyses considered two questions: i) whether there were any differences in the grades that students achieved (given their prior attainment) in GCSE or level 1/2 certificate qualifications in English/English language or English literature in summer 2015 and 2016, and ii) the implications of generating statistical predictions for awarding based on GCSE-only outcomes or combined GCSE and level 1/2 certificate outcomes in summer 2015 and 2016. While these analyses do not determine whether any differences in the grades achieved for the two qualification types are legitimate or not (ie they could be due to legitimate differences in value added due to the nature of the students, or differences in the standards of the qualifications), they provide some insight into how performance compared between the two qualifications given a student's prior attainment. Furthermore, they provide insight into the implications of any differences in student performance for generating statistical predictions to guide awarding.

The analyses reported here focus on 16-year-old students certificating in the summer examination series, since this is when the majority of 16-year-olds would be expected to complete their GCSE or level 1/2 certificate qualification. The analyses were conducted for both summer 2015 and 2016 since at the time the analyses were conducted the reference series (for generating statistical predictions to guide the 2017 English language and English literature awards) had not yet been confirmed.

5.2 Data

Two sets of data were used in each of these analyses. Student-level data that is supplied to Ofqual in advance of results day from each exam board provided the grade that students achieved at GCSE or in a level 1/2 certificate in the relevant subject (English/English language or English literature) in the summer examination series[17], and the National Pupil Database (NPD) provided students' prior attainment (KS2 score) five years previously. Prior attainment was measured as the average of a pupil's KS2 English and maths scores (APS).

Initially, the student-level data was combined across exam boards to generate a dataset showing the grade outcomes for each individual student who sat a GCSE and/or a level 1/2 certificate in the relevant subject (English/English language or English literature) with any exam board[18] in the relevant summer examination series

[17] This data is submitted to Ofqual around a week before results day. While it is nearly complete any students whose result is still outstanding will be missing from the dataset.
[18] Exam boards were AQA, CIE, OCR, Pearson, CCEA and WJEC. The analyses included all GCSE and level 1/2 certificate specifications that were available with each exam board (excluding 0500 for CIE – see introduction).

(2015 or 2016). Students were excluded from this dataset if they were not age 16 by the end of the relevant academic year, had not certificated in the summer examination series, and/or if they were from an independent or selective school. The latter students are routinely excluded from the statistical predictions used to guide awarding since they are known to have a different value added relationship based on their prior attainment (see Eason, 2010). Thus, if the proportion of students from independent and selective centres differed between GCSEs and level 1/2 certificates, this would contribute to any observed differences in the matrices.

Where students had certificated twice in either a GCSE or a level 1/2 certificate qualification in the same examination series (ie they had certificated twice in the same qualification type), their best grade was retained. Where students had certificated in both a GCSE and level 1/2 certificate, they were excluded from the analyses[19].

The student-level outcome data combined across exam boards was then matched to the NPD using a student's name and date of birth, to generate a record of the grade outcome and the prior attainment for each student who had both sets of data available. During this process, a number of students could not be matched to their prior attainment or were missing KS2 data. This included students that did not sit KS2 assessments (ie they were absent or they were not based in England at KS2[20]), and students where there were differences in the variables used for matching (eg differences in the way that their name had been recorded in the two datasets). Of the 16-year-old GCSE and level 1/2 certificate cohort from non-independent and selective centres, the match rate was reasonably high though (52%-59% in 2015, and 80-82% in 2016[21]).

For each subject (English/English language or English literature), the final datasets contained the GCSE or level 1/2 certificate grade and prior attainment for 16-year-old students from non-independent and selective centres who sat a GCSE or level 1/2 certificate in the relevant subject in summer 2015 or 2016. Thus, there were 4 datasets in total.

5.3 Methodology

Two pieces of analysis were conducted. First, matrices were generated that show the relationship between students' prior attainment at KS2 and their grade in the relevant subject (English/English language or English literature) for GCSEs and level 1/2 certificates separately – these are known as 'outcome matrices'. Second, these

[19] The proportion of students certificating in both qualifications was very small in both years – see section 4 of this report.
[20] A small number of students that were based in England at KS2 but took their GCSE at a school in another jurisdiction are included in the analyses.
[21] The lower match rate in 2015 is likely due to some schools not administering KS2 tests in 2010.

outcome matrices were used to consider the implications of generating predictions based on GCSE-only outcomes or combined GCSE and level 1/2 certificate outcomes. The analyses were repeated for 2015 and 2016.

5.3.1 Outcome matrices

To create the outcome matrices, students in the matched dataset for each subject (see section 5.2) were divided into octiles (or groups) based on their KS2 score, as shown in Table 5.1. The highest octile (octile 8) comprised the students with the highest prior attainment, and the lowest octile (octile 1) comprised the students with the lowest prior attainment. The octiles were selected for each year by splitting the students with valid English and maths KS2 scores into eight groups containing (as close as possible[22]) 12.5% of students in each, prior to the KS2 data being merged with the GCSE data[23].

A matrix was then generated for each subject cross tabulating students in each prior attainment octile with the grade achieved in their GCSE or level 1/2 certificate qualification. These matrices were generated separately for students taking a GCSE or a level 1/2 certificate. The resultant matrices showed the cumulative percentage of students achieving each grade for each octile of prior attainment, and the overall cumulative outcomes for the entire GCSE or level 1/2 certificate cohort (ie the total row at the bottom of the matrix).

The outcome matrices for GCSE and level 1/2 certificates were then compared with one another for each year to identify any differences in the grades achieved in each octile of prior attainment. When comparing the differences, each cell of the level 1/2 certificate matrix was subtracted from the corresponding cell of the GCSE matrix. As such, a positive difference suggests that the GCSE students with a particular level of prior attainment achieved higher outcomes at that grade than the level 1/2 certificate students, and a negative difference suggests that the GCSE students achieved lower outcomes at that grade than the level 1/2 certificate students[24].

[22] It was not always possible to have exactly 12.5% of students in each category, due to the number of students on certain marks.
[23] The cut-offs for each category differed slightly between years.
[24] Of exception to this is the total row at the bottom of the matrix, where the differences between qualification types shows the difference in overall outcomes (ie not controlling for prior attainment).

Table 5.1. *Octiles and corresponding prior attainment scores at KS2*

Octile	KS2 score (2010/2015)	KS2 score (2011/2016)
8	APS 83-100	APS 83.5-100
7	APS 77.5-82.5	APS 77-83
6	APS 72-77	APS 71.-76.5
5	APS 66.5-71.5	APS 65.5-70.5
4	APS 60.5-66	APS 59-65
3	APS 53-60	APS 52-58.5
2	APS 43-52.5	APS 42-51.5
1	APS 0-42.5	APS 0-41.5

5.3.2 Predictions

The outcome matrices were then used to generate predictions for the GCSE cohort each year based on GCSE-only outcomes[25], or combined GCSE and level 1/2 certificate outcomes. The resultant predictions were then compared to one another, providing an insight into the implications of generating predictions based on GCSE-only outcomes or combined GCSE and level 1/2 certificate outcomes. Any differences in the predictions are likely to reflect differences in the outcome matrices described above.

5.4 Results

The matrices and predictions are considered separately for English/English language and English literature, since the findings differ between the two subjects.

5.4.1 English/English language matrices

The outcome matrices for GCSE English/English language and level 1/2 certificates in English language in 2015 and 2016, respectively, are provided in Appendix B. The differences between each cell of the matrices and the overall outcomes (ie the bottom row of each table) for 2015 and 2016, respectively, are provided in Tables 5.2 and 5.3[26].

Comparing first the total figures at the bottom of the matrices shows that, overall, in both years, the GCSE students achieved higher grades than the level 1/2 certificate students. This is likely to reflect differences in the profile of the students taking the

[25] Note that the GCSE predictions are the same as the GCSE outcomes in a given year, since the predictions are generated for the students that they are based on (ie the prior attainment profile is identical).
[26] All figures in the tables are rounded to 1 decimal place.

two types of qualification (ie differences in their prior attainment). Therefore, it is the differences for each octile of prior attainment at each grade – ie the difference between each cell in the matrices – that is of most interest here, and not the overall differences in the outcomes.

Table 5.2 shows that for 2015 there were some small differences in the grades achieved between the students taking GCSE English/English language and level 1/2 certificates in English language, when considering the individual cells in the matrices. In particular, there are differences for the students with lower prior attainment that achieved a grade C or above, and the students with higher prior attainment that achieved a grade A or A*. For example, for students in the lowest octile (ie those with the lowest prior attainment), 1.6% fewer students achieved a grade C (or above) in the GCSE than the level 1/2 certificate. Conversely, for students in the highest octile (ie those with the highest prior attainment), 2.1% more students achieved a grade A (or above) in the GCSE than the level 1/2 certificate. Generally though, the differences are relatively minor.

Table 5.3 shows that for 2016 some of these differences are greater. In particular, for those students with lower prior attainment that were around the grade C/D borderline. For example, for students in the lowest octile (ie those with the lowest prior attainment), 6.4% fewer students achieved a grade C (or above) in the GCSE than the level 1/2 certificate. A similar difference is observed for students in the second lowest octile, where 4.7% fewer students achieved a grade C or above in the GCSE than the level 1/2 certificate.

There are two possible explanations for these differences. They might suggest that the type of students taking the level 1/2 certificate qualifications are different to those taking the GCSE – ie the level 1/2 certificate students made (and indeed should have made) more progress given their prior attainment. This would mean that there were differences in value added between the students taking the GCSE and level 1/2 certificates, that are entirely legitimate. Alternatively, it might suggest that the standard of the qualifications are not aligned, such that students with the same prior attainment receive greater reward in the level 1/2 certificate qualifications than the GCSE. The analyses reported in this section cannot disentangle these two reasons for the observed differences, but this is considered elsewhere in this report.

A final point worth considering is the differences between the two years. In 2015 the differences between the GCSE and level 1/2 certificate matrices were relatively small, yet were greater in 2016, particularly for the lower ability students. There are a number of possible explanations for this. It could be a result of some schools not administering KS2 tests in 2010, meaning that students matched to their prior attainment in 2015 are a non-representative subset of the overall cohort. Alternatively, it could be that the type of students sitting each of the qualification types changed between 2015 and 2016. These explanations both seem unlikely though: analyses conducted prior to the 2015 GCSE awards suggested a negligible

effect of some schools not administering KS2 tests in 2010 on the statistical predictions, and the increase in entry to level 1/2 certificate qualifications between 2015 and 2016 was relatively small (see Table 2.1). It could therefore be the case that the standard of one or both of the qualification types changed between 2015 and 2016, thus resulting in the differences in the matrices. Again, this cannot be determined from these analyses alone, but will be considered in the other analyses in this report.

Table 5.2. *Differences between the GCSE and level 1/2 certificate matrices (English/English language) – June 2015*

Octile	*	A	B	C	D	E	F	G	U
8	2.1	2.1	1.3	0.0	0.0	0.1	0.1	0.1	0.0
7	0.2	0.6	1.6	0.4	-0.2	-0.1	0.0	0.0	0.0
6	0.0	0.9	1.8	0.6	0.1	0.0	0.1	0.2	0.0
5	0.0	0.6	2.8	0.1	-0.3	-0.2	-0.1	0.1	0.0
4	-0.1	0.5	2.9	1.9	0.0	-0.1	0.1	0.4	0.0
3	-0.1	0.1	2.1	0.3	-0.2	-0.4	0.2	0.6	0.0
2	0.0	0.4	1.9	-0.2	0.1	-0.6	0.3	1.1	0.0
1	0.0	0.0	-0.2	-1.6	-0.9	0.4	2.1	3.2	0.0
Total	1.6	6.4	12.3	9.0	3.7	1.3	1.0	1.1	0.0

Table 5.3. *Differences between the GCSE and level 1/2 certificate matrices (English/English language) – June 2016*

Octile	*	A	B	C	D	E	F	G	U
8	2.6	-1.7	-0.6	-0.3	-0.2	-0.1	0.0	0.0	0.0
7	0.6	-0.9	-1.5	-0.6	-0.2	-0.2	0.0	0.0	0.0
6	0.3	-0.6	-0.9	-1.1	-0.4	-0.3	0.0	0.1	0.0
5	0.2	0.7	0.8	-1.1	-0.4	-0.1	0.1	0.3	0.0
4	0.1	0.1	0.5	-2.0	-0.9	-0.6	-0.1	0.3	0.0
3	0.0	0.2	0.1	-2.9	-0.8	-0.9	0.0	0.6	0.0
2	0.0	0.2	-0.2	-4.7	-3.7	-1.6	-0.1	0.8	0.0
1	0.0	-0.1	-0.4	-6.4	-8.2	-3.8	-0.3	1.6	0.0
Total	1.9	5.2	9.2	5.3	1.4	0.2	0.4	0.7	0.0

5.4.2 English literature matrices

The outcome matrices for GCSE and level 1/2 certificates in English literature for 2015 and 2016, respectively, are provided in Appendix B. Tables 5.4 and 5.5 show the differences between the matrices in each year, again subtracting each cell of the level 1/2 certificate matrix from the GCSE matrix.

For English literature, the overall outcomes each year are quite different when comparing the GCSE and the level 1/2 certificates. In both years, with the exception of A*, the raw outcomes for the GCSE students are considerably higher than the raw outcomes for the level 1/2 certificate students. This is particularly the case in the middle of the grade distribution. For example, at grade C (and above) the GCSE students outperformed the level 1/2 certificate students by around 20% in 2015 and by around 15% in 2016. Since this is likely to reflect differences in the prior attainment of the students taking each type of qualification, the focus again is on the differences between the individual cells in the matrices.

Tables 5.4 and 5.5 show that in both years there are considerable differences in outcomes between the two qualification types, particularly for the lower ability students around the middle of the grade distribution (ie for those students around the grade C/D borderline). Here, there tended to be lower outcomes for students taking a 1/2 certificate compared to students taking a GCSE. As in the previous analyses, it is not possible to determine from these analyses whether these differences reflect genuine differences in the performance of students taking each qualification, or whether they reflect something about the standards of the qualifications. Other analyses presented later in this report aim to disentangle this though.

Table 5.4. *Differences between the GCSE and level 1/2 certificate matrices (English literature) – June 2015*

Octile	*	A	B	C	D	E	F	G	U
8	-8.5	3.6	6.3	2.8	1.6	0.7	0.3	0.2	0.0
7	-4.5	3.4	10.9	6.7	3.2	1.4	0.6	0.4	0.0
6	-2.0	3.6	14.7	11.6	6.1	3.2	1.6	0.7	0.0
5	-1.4	2.9	15.0	14.6	10.1	5.9	3.7	1.6	0.0
4	-0.8	1.6	13.5	17.0	13.7	7.9	4.5	1.9	0.0
3	-0.4	0.7	10.6	19.4	17.5	11.6	5.7	2.5	0.0
2	0.0	0.5	5.8	14.3	20.8	15.0	8.0	3.3	0.0
1	0.0	0.1	1.4	7.9	19.3	23.5	15.2	6.9	0.0
Total	-0.1	7.9	19.9	21.7	18.2	12.4	6.8	3.0	0.0

Table 5.5. *Differences between the GCSE and level 1/2 certificate matrices (English literature) – June 2016*

Octile	*	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*U*
8	-5.9	3.0	4.5	1.9	0.8	0.4	0.3	0.1	0.0
7	-3.4	2.6	9.1	5.2	2.3	1.1	0.7	0.3	0.0
6	-2.7	-0.3	8.1	7.4	5.3	2.7	1.4	0.7	0.0
5	-1.6	-0.3	7.7	9.6	7.3	3.7	1.9	0.8	0.0
4	-0.6	0.3	7.5	12.0	11.1	6.4	3.0	1.2	0.0
3	-0.4	-0.6	4.9	11.3	13.2	8.2	4.0	1.5	0.0
2	-0.2	-0.3	1.7	7.8	15.1	11.4	5.9	2.6	0.0
1	-0.1	-0.3	-0.7	3.2	14.5	18.1	11.0	4.5	0.0
Total	-0.1	5.4	13.6	15.0	13.5	9.1	4.7	2.0	0.0

5.4.3 English/English language predictions

The second set of analyses consider the implications of any differences in the outcome matrices for generating statistical predictions to guide awarding. Tables 5.6 and 5.7 show the predictions for English/English language based on GCSE only outcomes or combined GCSE and level 1/2 certificate outcomes in 2015 and 2016, respectively, at each of the key grades used in GCSE awarding. As shown in Table 5.6, the 2015 predictions are very similar at all key grades regardless of which qualifications the predictions were based on. This is not surprising given that there were only minor differences in the outcome matrices for GCSE and level 1/2 certificate qualifications in 2015. Thus, however the predictions are generated, the outcomes would be very similar.

Table 5.7 shows that there are greater differences in the 2016 predictions. While the predictions based on GCSE only outcomes compared to combined GCSE and level 1/2 certificate outcomes are similar at grades A*, A and F, at grade C there is almost a 1% difference in the predictions, with the combined outcomes yielding a higher prediction. Again, this is not unexpected given the differences in the outcome matrices in 2016, where the lower ability students taking the level 1/2 certificate qualifications had higher outcomes than those taking the GCSE, particularly around the C/D borderline. Thus, when predictions are generated based on the combined outcomes, they are higher.

Table 5.6. *Predictions based on GCSE-only outcomes or combined GCSE and level 1/2 certificate outcomes (English/English language) – June 2015*

	*	A	C	F
Prediction for all students based on GCSE matrix	3.04	15.37	72.02	98.75
Prediction for all students based on combined matrix	2.99	15.18	71.98	98.57
Difference	-0.05	-0.19	-0.04	-0.18

Table 5.7 *Predictions based on GCSE-only outcomes or combined GCSE and level 1/2 certificate outcomes (English/English language) – June 2016*

	*	A	C	F
Prediction for all students based on GCSE matrix	3.62	15.57	70.42	98.58
Prediction for all students based on combined matrix	3.51	15.62	71.41	98.60
Difference	-0.11	0.05	0.99	0.02

5.4.4 English literature predictions

Tables 5.8 and 5.9 show the same analyses for English literature, for 2015 and 2016, respectively. Here, the differences in the predictions are marginal at grades A* and A in each year, yet are greater at grade F and even greater at grade C. At grade

C, generating predictions based on combined GCSE and level 1/2 certificate outcomes would result in a lower prediction than generating predictions based on GCSE only outcomes in both years. This effect is greater in 2015 than in 2016, yet still exceeds 1% in 2016. This means that if the 2017 awards for English literature were guided by predictions based on combined GCSE and level 1/2 certificate outcomes, rather than GCSE only outcomes, the predictions would be lower.

Table 5.8. *Predictions based on GCSE only outcomes or combined GCSE and level 1/2 certificate outcomes (English literature) – June 2015*

	*	A	C	F
Prediction for all students based on GCSE matrix	4.29	20.46	75.45	98.50
Prediction for all students based on combined matrix	4.55	20.18	73.43	97.52
Difference	0.26	-0.28	-2.02	-0.98

Table 5.9. *Predictions based on GCSE only outcomes or combined GCSE and level 1/2 certificate outcomes (English literature) – June 2016*

	*	A	C	F
Prediction for all students based on GCSE matrix	4.32	19.23	73.84	98.48
Prediction for all students based on combined matrix	4.58	19.17	72.49	97.73
Difference	0.26	-0.05	-1.35	-0.75

5.5 Summary of the results

In summary, the analyses reported in this section suggest that there are differences in the grades achieved for students taking GCSE and level 1/2 certificate qualifications, once their prior attainment is accounted for, for both English/English language and English literature. However, the direction of the differences is different for each subject. For English/English language students taking the level 1/2 certificate achieved higher outcomes (given their prior attainment) than students taking the GCSE, while in English literature the converse is true.

The second part of these analyses showed that these differences have implications when generating statistical predictions based on GCSE-only outcomes or combined GCSE and level 1/2 certificate outcomes, mainly at grade C. For English/English language, basing predictions on GCSE-only outcomes from 2016 would result in a prediction around 1% higher at grade C than basing the predictions on the combined outcomes, while in English literature basing the predictions on GCSE-only outcomes from 2016 would result in a prediction around 1% lower than basing the predictions on the combined outcomes.

These analyses suggest that there are differences between the outcomes that students achieved in the two qualification types given their prior attainment.

However, it is not possible to ascertain from these analyses whether these differences are due to legitimate differences in value-added (ie the students taking a level 1/2 certificate English language genuinely have higher value-added than those taking the GCSE and therefore should perform better), or whether these differences are due to different standards on the two qualifications. The remaining analyses in this report aim to disentangle this further.

6 Inter-board comparability analysis

6.1 Research aim

This strand of work investigated the comparability of standards in GCSE English/English language and English literature, and level 1/2 certificates in English language and English literature, between exam boards. This strand of work aimed to provide evidence on the alignment of standards across specifications within the same subject using mean GCSE score, Rasch modelling and inter-board statistical screening.

These approaches are generally based upon the reasoning that there is a relationship between a measure of a student's ability and their score in the specification that they took. The measure of a student's ability provides a link between the scores of the students in the different specifications being compared. The specifications within the same subject are therefore considered comparable if students who demonstrate the same level of prior attainment obtain the same grade in different specifications (Elliott, 2011).

The mean GCSE score was first used as a measure of ability. Rasch analysis allows us to derive a second measure of ability (Rasch ability) to be used in addition to mean GCSE score. Furthermore, the Rasch modelling approach takes into consideration the difference in difficulty between different subjects when estimating the abilities of the students. The different approaches allow us to provide more robust evidence on the comparability of the specifications offered by exam boards.

6.2 Data

Student-level data for examinations in 16 GCSE subjects administered in 2015 and 2016 by the exam boards that provide GCSE and level 1/2 certificate qualifications were collected for this study. These subjects (listed in Table 6.1) are large entry subjects that count towards the English Baccalaureate school performance measure. In order for the results to be more accurate and reliable, students taking fewer than two subjects were excluded from the analysis, which resulted in the sample size of the data being considerably smaller than the original sample sizes. Table 6.1 reports the actual number of students included in the analysis.

The analysis focused on the comparability of standards in GCSE English/English language and English literature, and level 1/2 certificates in English language and English literature, between the exam boards. Table 6.2 lists the number of students from individual exam boards that took English/English language and English literature in the 2015 and 2016 examination series included in the analysis. From Table 6.2 it is apparent that for English/English language, the CIE level 1/2 certificate specification is by far the most popular among the level 1/2 certificates, making CIE the second most chosen exam board offering a qualification in English/English

language. For both English/English language and English literature, AQA is the most popular qualification.

Table 6.1. *Number of students taking each of the subjects studied and included in the analysis*

Subject	*2015*	*2016*
English (including English language)	607,067	607,163
English literature	499,941	497,147
French	148,705	127,913
Geography	216,177	234,902
German	51,654	48,085
History	236,335	249,529
Mathematics	610,625	601,294
Applications of mathematics	12,479	8,725
Methods in mathematics	11,569	7,405
Additional science	316,326	360,315
Biology	143,338	148,947
Chemistry	141,800	148,728
Further additional science	22,934	17,341
Physics	143,102	149,127
Science	245,830	269,042
Spanish	84,381	86,688

Table 6.2. *Number of students from individual exam boards taking English and English literature in the 2015 and 2016 included in the analysis*

		GCSE				*Level 1/2 certificates*		
		AQA	*OCR*	*Pearson*	*WJEC*	*AQA*	*Pearson*	*CIE*
English	*2015*	232,667	24,923	34,478	129,394	15,808	8,832	160,965
	2016	227,014	22,307	33,087	123,078	17,932	8,138	175,607
English literature	*2015*	232,485	26,739	36,981	109,589	36,141	40,519	17,487
	2016	230,742	24,801	35,602	110,641	46,129	30,023	19,209

6.3 Methodology

Details of the use of the Rasch model to study inter-board comparability of examination standards can be found in the report by He, Stockford and Meadows (2016). Briefly, this work extends the application of the unidimensional partial credit Rasch model (PCM) in inter-subject comparability studies to the investigation of

inter-board comparability of GCSE examination standards for individual subjects (see Masters, 1982; Wright and Masters, 1982; Coe, 2008; Coe et al., 2008; Bramley, 2011; He and Stockford, 2015; Opposs, 2015; He, Stockford and Meadows, 2016). To achieve this, exams that test the same subject (eg GCSE English and English language and level 1/2 English language) from different exam boards were treated as one item in a test which comprises all the subjects studied, and students from different exam boards were treated as different subgroups. It is assumed that these examinations together define a shared construct which is closely related to the constructs being measured by the individual examinations. Comparability of standards in examinations that test the same subject between exam boards at a specific grade can be investigated by comparing the values of the category parameters (or grade difficulty) between the different subgroups (ie differential category functioning – DCF). The existence of significant DCF at specific grades between the exam boards in a subject can be assumed to indicate inconsistency in standards at those grades.

One of the statistical approaches routinely used for monitoring and maintaining inter-board comparability of examination standards in GCSE by the exam boards is the post-award inter-board statistical screening (see Taylor, 2013). This involves, for a specific subject, establishing a relationship between the overall grade distribution of students from all exam boards and a performance measure which is assumed to represent a construct similar to the construct measured by the examination being investigated empirically first. This relationship is then examined for the grade distribution of students from individual exam boards. Significant departure from this all-boards relationship for individual boards would suggest inconsistency in standards between the exam boards which will be taken into consideration in awarding next year (see Taylor, 2013; He, Stockford and Meadows, 2016). Both the estimated Rasch ability measures and the mean GCSE score based on the 16 subjects of the students have also been used as performance measures for inter-board statistical screening to investigate inter-board comparability in English and English Literature here.

To facilitate the analysis using the partial credit model and inter-board statistical screening, the GCSE and level 1/2 certificate grades were converted into numerical values representing ordered category scores: A*=8, A=7, B=6, C=5, D=4, E=3, F=2, G=1 U=0. The maximum score on every item (subject exam) is therefore 8.

6.4 Results

Analysis of variances suggested that the total variances in the data accounted for by the Rasch model is about 79%, suggesting that the datasets could be essentially treated as unidimensional for the purpose of this investigation. An inspection of the fit statistics from an initial analysis suggested that candidates achieving a U did not fit the partial credit model well and were therefore treated as missing. Grade G was then taken to be the lowest score category, and this resulted in only a small number

of the score categories (numerical grades) with information-weighted mean square (infit MNSQ) slightly over 2.0 which was used to judge whether an item fits the Rasch model sufficiently well. To account for the effect of misfit of data to the model on the standard errors of item category measures, the model based standard errors were enlarged by a factor calculated as the square root of the infit MNSQ (when larger than 1.0) when calculating the level of significance of the DCF effects (Linacre, 2015).

6.4.1 Relationship between grade outcomes in English/English language and English literature and mean GCSE and Rasch ability

Figure 6.1 shows the relationship between mean GCSE score (or Rasch ability) and the average observed score (numerical grade) in English/English language or English literature for students taking each specification (from each exam board). To produce the graphs, the mean GCSE score for each student is calculated based on his/her achieved grades in the 16 subjects, the mean GCSE scale was divided into 20 intervals, and the students were assigned into one of the performance groups based on their mean GCSE score. For students in each mean GCSE score interval their average score in English or English literature was calculated. Therefore, the graph is based on the average score achieved for each mean GCSE score interval. For the Rasch ability scale, 30 intervals were devised. The confidence intervals of the mean observed scores are small and are not shown on the graphs, but the statistical significance of the differences across boards will be commented on in the text. The pattern of the relationship between the average observed score in English/English language or English literature and mean GCSE is similar to that between observed score and Rasch ability. The average observed score generally increases with increasing mean GCSE score or Rasch ability.

These graphs may be used to define the relative difficulty at individual grades and overall of the subject between the exam boards. If, for a specific mean GCSE or Rasch ability interval, the average observed score in the specification is higher or lower than the average observed scores for the other specifications, then the examination for that specification may be viewed as a different standard to the other examinations. For example, Pearson level 1/2 English Language for both 2015 and 2016 could be viewed as a different standard to the average observed scores of all boards, as its average observed scores are different to the average observed scores of all boards. Similarly, the 2015 CIE level 1/2 certificate in English literature can be viewed as a different standard to the average of all boards. In general, for English/English language, the level 1/2 qualifications could be viewed as a different standard at the top or middle to top grades.

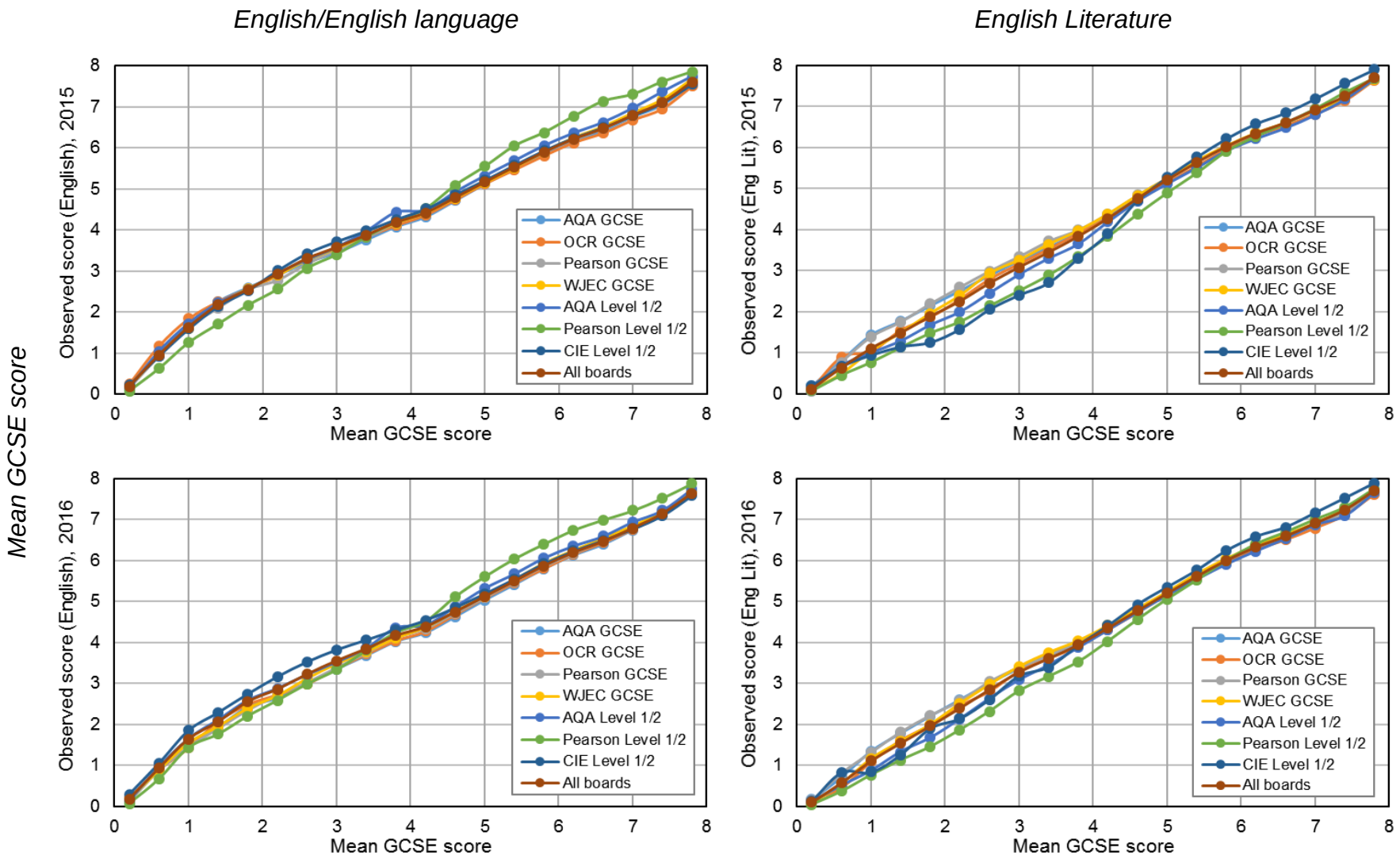

Figure 6.1. *Relationship between average observed point grade in English/English language and English literature and mean GCSE and Rasch ability, 2015 and 2016*

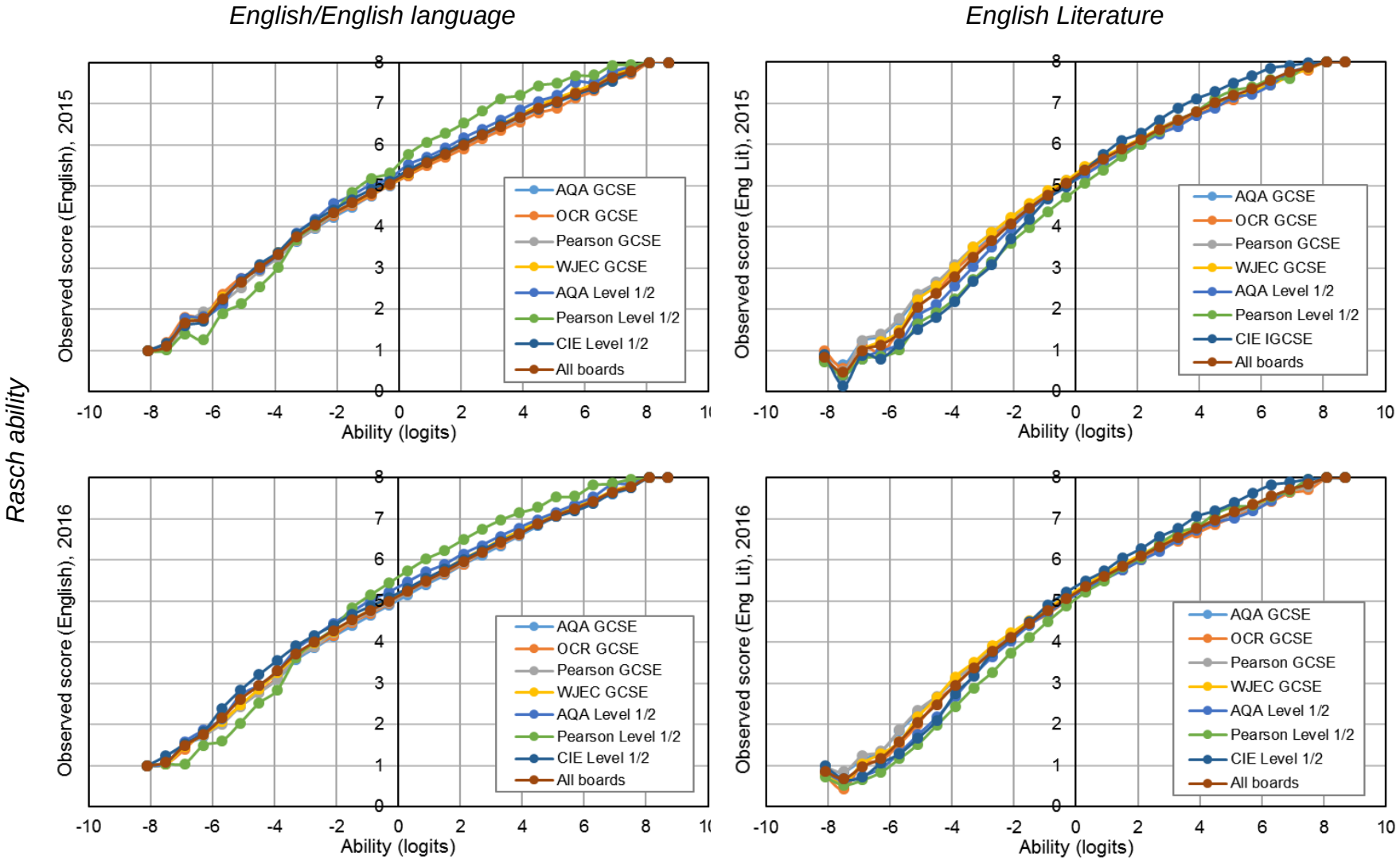

Figure 6.1. (ctd) *Relationship between average observed point grade in English/English language and English literature and mean GCSE and Rasch ability, 2015 and 2016*

6.4.2 Relative inter-board grade difficulty for English and English literature based on Rasch analysis

In the Rasch model, each numerical grade in a subject can be represented by a difficulty parameter in the unit of logits which is estimated using the model and the grades achieved by students in individual subjects. The relative difficulty at a specific grade for a subject from a specific exam board can be defined as the difference between its grade difficulty and the all-boards grade difficulty estimated based on students from all exam boards. By using the average grade gap in logits, the relative grade difficulty can be expressed in the unit of grade (see He, Stockford and Meadows, 2016).

Table 6.3 shows the relative grade difficulty in the unit of grades between the exam boards for English/English language and English literature for the 2015 and 2016 exam series. Most of the relative grade difficulties (differences between the grade difficulties of individual boards and the average grade difficulties of all boards) are significant at the level of $p<0.05$. There is variability in relative grade difficulties in both GCSE qualifications and level 1/2 qualifications. Again, Pearson level 1/2 English language at grades A and A* could be seen as a different standard to the average of all boards. The relative grade difficulties shown in Table 6.3 are broadly consistent with the patterns of the relationship between the average observed grade and mean GCSE score or Rasch ability. The relative grade difficulty distribution for 2015 is also similar to that for 2016. In general, for English/English language, for both 2015 and 2016, from A* to C, the level 1/2 certificate qualifications (shaded) appear to be a different standard to the GCSE qualifications. For English literature, the CIE level 1/2 certificate qualification appears to be a different standard to the other specifications at the top grades for both 2015 and 2016.

Table 6.3. *Relative grade difficulty (in unit of grade) in English and English literature for the 2015 and 2016 examination series*

Board	*Relative grade difficulty (in unit of grade)*						
	F	*E*	*D*	*C*	*B*	*A*	*A**
English, relative grade difficulty, 2015							
All boards	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AQA	0.19	0.08	0.07	0.10	0.04	0.07	0.07
OCR	0.03	0.04	0.06	0.07	0.10	0.14	0.17
PEARSON	0.21	0.12	0.06	0.03	0.02	-0.04	-0.14
WJEC	0.09	0.03	0.00	-0.01	0.02	-0.04	-0.06
AQA Level 1/2	0.15	0.06	-0.03	-0.19	-0.20	-0.18	-0.15
PEARSON Level 1/2	0.40	0.27	0.12	-0.14	-0.41	-0.64	-0.90
CIE	-0.04	-0.10	-0.10	-0.10	-0.07	-0.06	-0.03
English Literature, relative grade difficulty, 2015							
All boards	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AQA	-0.28	-0.20	-0.09	-0.02	-0.02	0.03	0.07
OCR	-0.09	-0.07	-0.03	0.05	0.12	0.10	0.19
PEARSON	-0.33	-0.29	-0.20	-0.08	-0.02	-0.05	-0.08
WJEC	-0.13	-0.15	-0.16	-0.13	-0.06	-0.01	0.11
AQA Level 1/2	0.22	0.19	0.12	0.02	0.13	0.15	0.04
PEARSON Level 1/2	0.38	0.37	0.37	0.33	0.19	-0.10	-0.43
CIE	0.38	0.47	0.38	0.16	-0.03	-0.29	-0.61
English, relative grade difficulty, 2016							
All boards	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AQA	0.12	0.11	0.12	0.15	0.10	0.09	0.04
OCR	0.12	0.09	0.09	0.11	0.06	0.05	0.00
PEARSON	0.10	0.12	0.10	0.08	0.03	-0.04	-0.13
WJEC	0.04	0.07	0.06	0.00	-0.03	-0.05	0.00
AQA Level 1/2	0.01	-0.01	-0.07	-0.20	-0.24	-0.16	-0.04
PEARSON Level 1/2	0.22	0.23	0.12	-0.19	-0.55	-0.61	-0.68
CIE	-0.28	-0.24	-0.19	-0.16	-0.08	-0.06	0.04
English Literature, relative grade difficulty, 2016							
All boards	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AQA	-0.25	-0.17	-0.07	0.00	0.01	0.05	0.07
OCR	-0.01	-0.01	0.04	0.07	0.07	0.11	0.25
PEARSON	-0.16	-0.13	-0.08	-0.01	0.01	-0.03	0.07
WJEC	-0.07	-0.10	-0.12	-0.10	-0.07	-0.04	0.07
AQA Level 1/2	0.22	0.22	0.14	0.04	0.06	0.07	0.01
PEARSON Level 1/2	0.39	0.36	0.33	0.23	0.05	-0.21	-0.44
CIE	0.22	0.32	0.22	-0.01	-0.15	-0.28	-0.47

Note: A difference suggests that the board appears to be of a different standard. Level 1/2 certificates are shaded.

6.4.3 Relative between-board grade difficulty for English/English language and English literature based on inter-board screening with mean GCSE score and Rasch ability

The inter-board statistical screening procedure is based on a relationship established between subject grade outcomes and a performance measure representing the construct which the examination in the subject is intended to measure. Both the mean GCSE score and the Rasch ability measure were used as a performance measure for the inter-board screening analysis reported here.

Figure 6.2 below shows changes in grade outcomes after aligning standards between the exam boards based on inter-board screening with mean GCSE and Rasch ability for English/English language and English literature (negative values indicate lenient grading while positive values harsh grading). To produce the graphs, the mean GCSE score for each student is calculated, the mean GCSE scores (or Rasch ability) of all students taking English/English language or English literature are divided into 10 deciles (with similar numbers of students in each decile), and the grade distribution of students in each mean GCSE score decile is generated, resulting in an outcome matrix showing the relationship between mean GCSE score (or Rasch ability) profile and outcomes (numerical grades) across all specifications or exam boards. This outcome matrix is then used to generate a prediction of grade outcomes for each specification or exam board (given the mean GCSE score or Rasch ability profile of their cohort), that is then compared against the actual grade outcomes achieved to produce the changes in outcomes when the exams are aligned in standards.

The distribution of changes in grade outcomes after aligning standards between the exam boards using the inter-board screening procedure for both English/English language and English literature is similar for both 2015 and 2016. The pattern of changes based on mean GCSE score is closely similar to that produced using Rasch ability. For both 2015 and 2016, Pearson level 1/2 English language could be seen as a different standard to the examinations from the other boards. The CIE level 1/2 certificate English language qualification also appears to be of a slightly different standard at grade C compared to the other exam boards. In English literature, the Pearson and CIE specifications appeared to be of a different standard at the higher grades.

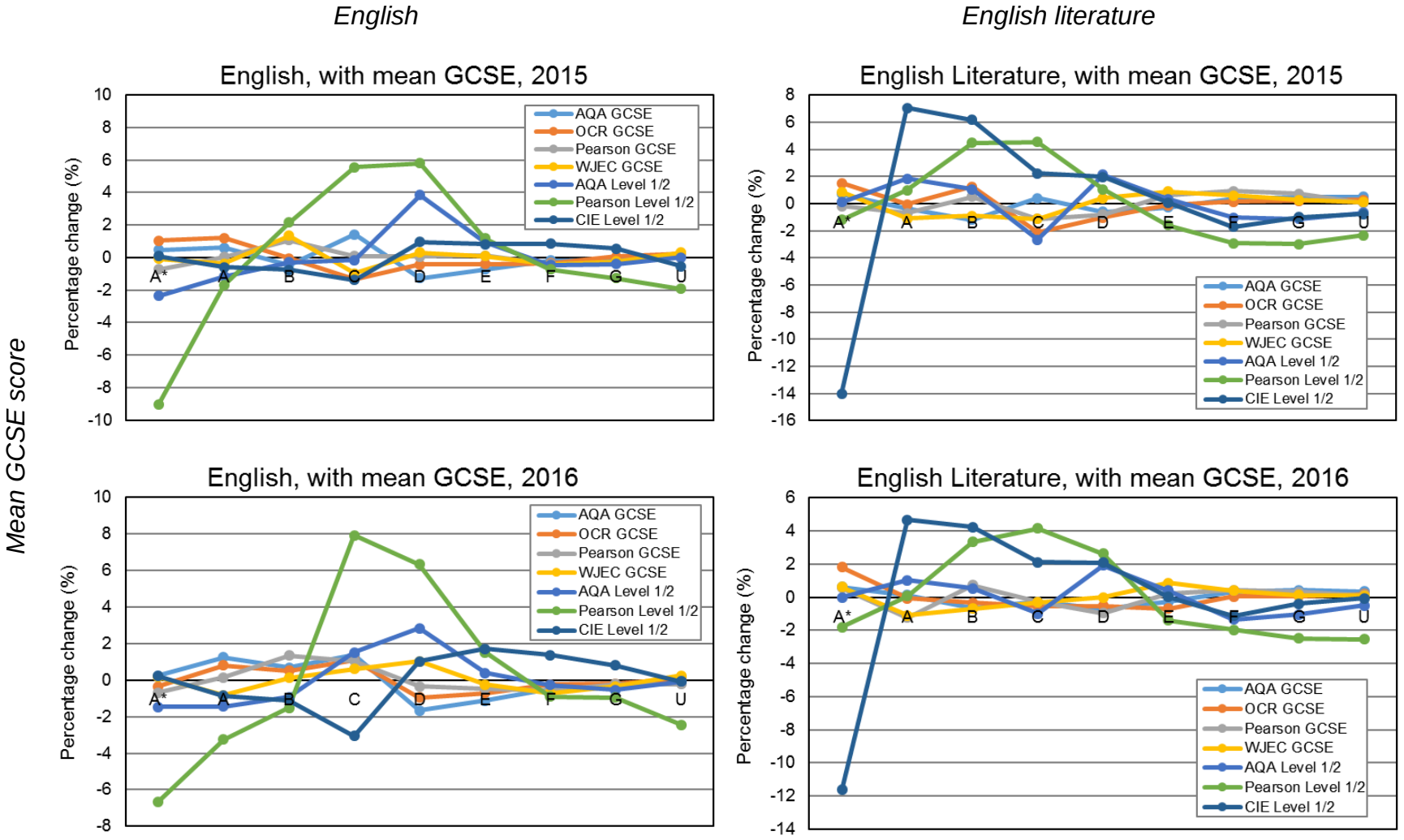

Figure 6.2. *Changes in grade outcomes in English/English language and English literature after aligning standards based on the inter-board statistical screening analysis*

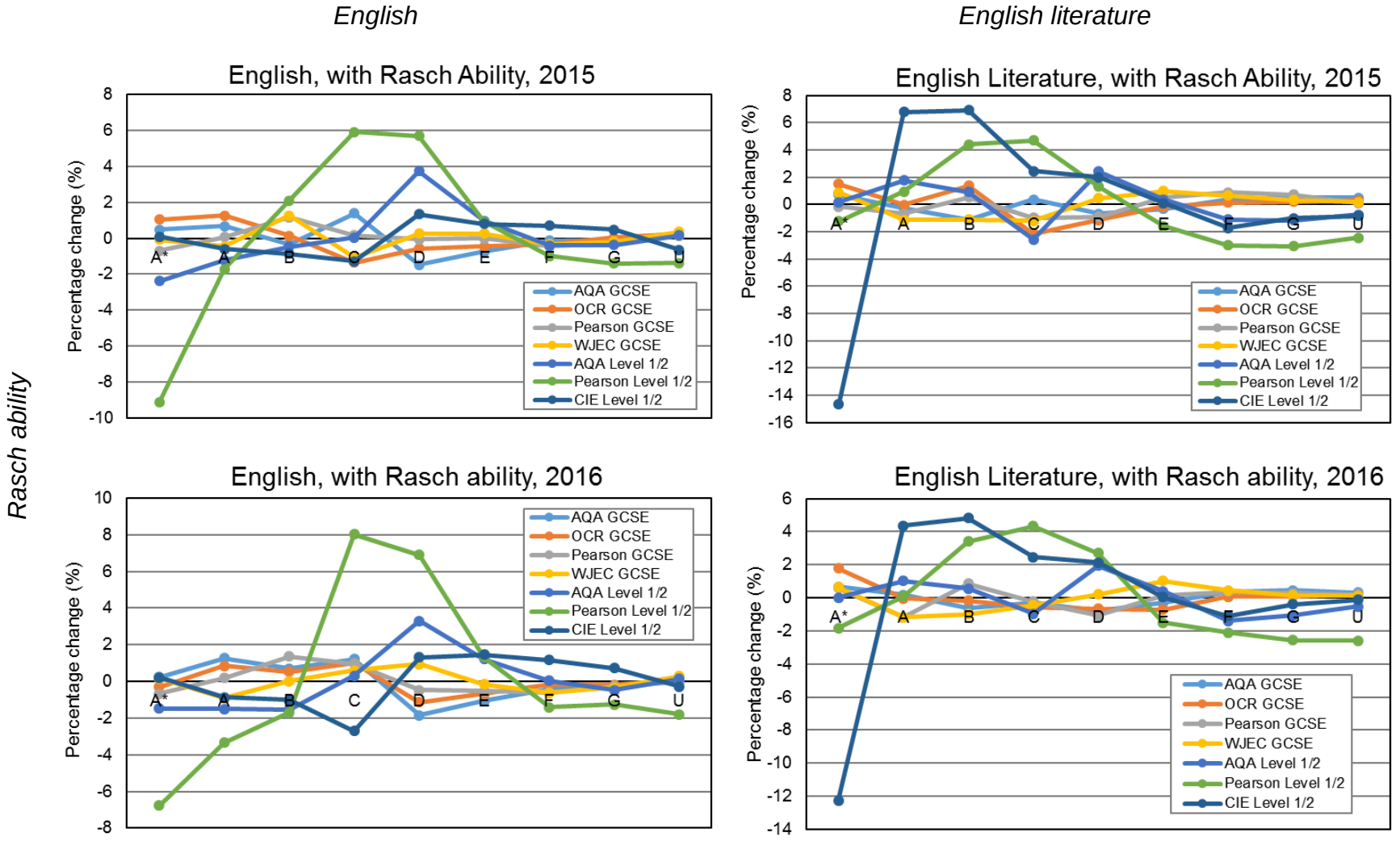

Figure 6.2. (ctd) *Changes in grade outcomes in English/English language and English literature after aligning standards based on the inter-board statistical screening analysis*

6.5 Summary of the results

This strand of work looked at the comparability of standards in GCSE and level 1/2 certificate qualifications in English/English language and English literature between exam boards.

Different approaches, all relying on the comparison of the relationship between grade outcome and an underlying ability measure established empirically or through a theoretical model, have been used. It is to be noted, however, that direct comparison between the results from the different approaches may not always be appropriate. This is because the different approaches conceptualise and operationalise examination standards or grade difficulty slightly differently. The inter-board statistical screening (Figure 6.2) is based on the observed grade distribution of the students in the 10 mean GCSE deciles. The relationship between students' ability and outcomes (Figure 6.1) is based on the mean observed grades within each mean GCSE score interval for which the actual English/English language or English literature grade distribution for different exam boards could be different even if they had the same average English/English language or English literature score. The relative grade difficulty derived through Rasch modelling (Table 6.3) was produced by applying the partial credit model to the grades achieved by the students in individual subjects.

However, from the different analyses performed, it is possible to observe some common patterns. In particular, although there is variability in difficulties in both GCSE qualifications and level 1/2 certificate qualifications, for both English/English language and English literature, the CIE level 1/2 certificate specification and other level 1/2 certificates generally appear to be of a different standard at the top grades or middle to top grades compared to the GCSE specifications. Overall, among GCSE specifications there is greater alignment in standards than between GCSEs and level 1/2 certificates.

The use of mean GCSE score, the Rasch model, as well as the Rasch measure of ability, to compare the outcomes of students taking specifications provided by different boards, suggests that standards are not precisely aligned between GCSE and level 1/2 certificate qualifications in English/English language and English literature.

7 Investigating the comparability of standards in GCSE and level 1/2 certificates using propensity-score matching

7.1 Research aim

The analyses presented thus far on English/English language and English literature suggest that standards across qualification types might not have been fully aligned. More specifically, section 6 shows that, whilst the standards seem aligned across GCSE specifications, there might be some differences between GCSE and level 1/2 certificates in both subjects in 2016 and, to a lesser extent, in 2015. The inter-board screening analysis that pointed towards this conclusion relies on the fact that two examinations are comparable if students with the same level of attainment (or ability) obtain the same grade in the two examinations. While this may not necessarily be the 'optimal' definition, it offers a tangible foundation for assessing comparability across different qualifications and/or specifications[27]. As such, inter-board screening is routinely performed after each summer series in order to inform any decision regarding the maintenance of standards.

An alternative approach to investigate the comparability of examination standards is one based on the statistical comparison of students sitting alternative examinations who are similar to each other with respect to a broader set of characteristics, including but not restricted to the same level of students' attainment (or ability). Zanini (2016) proposed to employ propensity-score matching as a tool to operationalise this broader definition of comparability.

In this section the application of propensity-score matching to investigate the comparability of GCSE and level 1/2 certificates in English language and English literature in both 2015 and 2016 is presented. The findings of this exercise aim to provide further evidence on the comparability of these specifications and, therefore, to inform the maintenance of standards in 2017.

7.2 Data

In this section, data from the National Pupil Database (NPD) provided by the Department for Education was used. The NPD is a linked administrative data archive compiled using data supplied by examination centres and exam boards. In addition to data on exam performance, the NPD provides a rich set of data on students, including demographic characteristics, socio-economic background indicators and information on the school attended. Thanks to the wide range of student characteristics available in the NPD, this data archive is particularly suitable when

[27] For a broad discussion on comparability see Newton (2007) and Elliot (2011).

performing propensity-score matching. As will be explained in the methodology subsection, this method relies on the availability of information that allows us to identify students that are very similar to each other, apart from the fact that they took different qualifications/specifications. The only downside connected to the use of the NPD is that, for a subset of students, some of the required variables are missing or incomplete. This forces a relatively small proportion of students to be excluded from the analysis.

Table 7.1 shows the number of students included in the analysis. As in the other sections of the report, the focus is on English/English language and English literature. The comparability of examination standards across English/English language specifications is explored in both the 2015 and 2016 June examination sessions. For English literature the study is limited to 2016. The number of students included in the analysis is slightly smaller than in other sections of this report. This is due to a number of reasons, mainly the restrictions imposed by using the NPD, as well the aim of focussing on 'typical' students in order to ensure like for like comparisons.

Table 7.1. *Numbers of students per each English/English language and English literature specification included in the analysis, by type of qualification and exam board*

Exam board	*English*			*English literature*	
	Spec	*2015*	*2016*	*Spec*	*2016*
All GCSEs		*300,423*	*305,563*		*215,988*
AQA	4707	180,739	182,357	9717	134,494
OCR	J355	19,947	18,179	J360	13,624
Pearson	2EN0	25,830	26,727	2ET0	20,798
WJEC	9700	73,907	78,390	4200	47,072
All level 1/2		*182,036*	*186,426*		*60,733*
AQA	8705	11,760	11,622	8710	27,688
CIE	0522	162,373	167,654	0486	14,556
Pearson	KEA0	6,955	6,661	KET0	18,489
WJEC	9700	948	489	-	-

Only students in England who took at least four qualifications, either GCSEs or level 1/2 certificates, were included in the analysis. Students up to the age of 16 were included in the analysis and year group used to match students. Only grades A* to U were considered (pending or incomplete grades were not included) and for the small number of students entering multiple qualifications in the same subject (see section 4) only the entry with the highest grade was taken into consideration.

The set of variables used to identify students that were similar to each other was selected in order to avoid the exclusion of a significant number of students. For the same purpose, some categorical variables with missing values were recoded and an additional category was created. Students were matched against the recorded version of such variables.

Overall, the restrictions on the data caused only a small proportion of candidates to be excluded. Therefore, it seems reasonable to say that the validity of the research design was not compromised and that the findings can be generalised to the whole cohort of typical students entering these examination without any loss of generality.

7.3 Methodology

7.3.1 Research design

In order to provide evidence on the comparability of standards between GCSEs and level 1/2 certificates and to inform the maintenance of standards in 2017, a number of comparisons were performed. Considering that the question underlying this investigation is whether or not the standard in level 1/2 certificates is precisely aligned with GCSEs, the main focus of the analysis was on the comparison between these two types of qualification. However, it can be argued that differences across specifications within the same qualification type are also possible and could potentially be misleading, ie differences in the comparability of standards in different GCSEs or different level 1/2 certificates may exist but be cancelled out. Therefore, in addition to considering the comparability of examinations within the same qualification type, it is necessary to compare specific level 1/2 certificates with specific GCSE specifications. Given the unequal share of students taking alternative English specifications, this analysis considered the most popular level 1/2 certificate and the most popular GCSE specification, as highlighted by Table 7.1.

A further comparison was performed to explore the comparability of standards within the same qualification type over time. Considering that the standard of level 1/2 certificates was considered against the standard of GCSEs, it is important to assess whether the GCSE standard has remained stable over time. Given that some minor fluctuations over time are possible, this additional analysis also provides us with benchmark figures to understand the magnitude of such fluctuations in standards over time.

7.3.2 Statistical techniques

In order to perform these comparisons two different methods were employed. Although the main focus of this section is to perform propensity score matching to compare students taking different qualifications who are similar to each other with respect to a broad set of characteristics, multilevel modelling was also employed to provide preliminary evidence on the comparability of standards when only concurrent attainment is used as a link between qualifications.

The aim of using multilevel modelling in addition to propensity-score matching is twofold. First, it performs an analysis similar to the inter-board screening presented in the previous section, but based on a different statistical technique, so that it is possible to provide robustness checks for the main findings in the previous section. Second, it allows us to operationalise two alternative definitions of comparability and therefore to check whether or not the findings are dependent on these definitions.

Both statistical techniques are employed to explore the relationship between student performance and the examinations sat, once the comparison is restricted to 'similar' students. Within multilevel modelling, similar students are defined as those with the same level of concurrent attainment[28]. In using propensity-score matching, students taking different qualifications will be deemed to be similar if they have a balance of similar values for a larger set of variables, including level of concurrent attainment, demographic characteristics, socio-economic background indicators and information on the school attended (see Appendix C for the full list of variables used).

In both cases, the level of concurrent attainment is measured as the average GCSE point score (where grades are converted as follows: A*=8, A=7, …, G=1, U=0) and computed excluding the focus subject (ie English or English literature). Two measures of attainment are considered in both the modelling and matching: *i)* whether a student attained a grade C or above; and *ii)* whether a student attained a grade A or above.

The main features of the two statistical techniques employed in this section are described below. More technical details are presented in Appendix C.

Multilevel modelling

Multilevel modelling is used in this section to study the relationship between student performance and the examination taken, once the level of concurrent attainment is accounted for. Given that the dependent variable is a dichotomous variable, equal to 1 for students who attained a grade C/A or above and 0 otherwise, a logistic regression is necessary. Furthermore, it has to be considered that students are clustered within schools. As it is fair to assume that two students in the same school are more likely to be similar than those students in different schools, a multilevel approach was taken. Failing to recognise the hierarchical structure of the data could lead to potentially misleading results.

A set of multilevel logistic regressions was then performed to model the probability of attaining a grade C/A or above given the specification taken and the level of

[28] It must be noted that, theoretically, the modelling approach allows the consideration of a broad set of student characteristics. The decision to use the level of attainment as the only student characteristic was led by the intention to compare alternative definitions of comparability as explained above.

concurrent attainment. The outcome of this method is an estimate of the regression coefficients. These allow one to check whether the specification taken is a significant predictor of the performance, once the level of concurrent attainment and the hierarchical structure of the data are taken into account. The fitted regression equations were used to retrieve the probability of attaining a grade C/A or above according to the level of concurrent attainment and the examination taken. In this way, it was possible to compare whether or not attaining a grade C/A is more likely for students taking a certain type of qualification.

In order to obtain the regression coefficients it is necessary to specify a regression model first. Although data inspection has been used to inform the choice of the regression specification, it is important to stress that this implies imposing a functional form to the relationship between the dependent variable (ie performance) and the independent variables (ie level of concurrent attainment and examinations sat). This can be a limitation of the method that can be overcome, as an example, by using a more data driven approach.

Propensity-score matching

Propensity-score matching is a data driven approach. This statistical technique can be used to match each student sitting a certain examination with a 'similar' student sitting another examination. Similarity across students in the two groups is based on a broad set of characteristics that potentially affect attainment in the examination under scrutiny. Prior attainment (ie KS2 results) and concurrent attainment (ie GCSE results in other subjects) as well as gender, socio-economic background, and attributes of the school/college attended are all important determinants of attainment and, therefore, have to be considered in the matching procedure.

Students for whom a match was not found are excluded from the comparison. For those matched, it is possible to look at their performance and retrieve an estimate of the average effect of taking one specification rather than another one on the probability of attaining a grade C/A or above. The difference in the average attainment of the two groups of matched students (those taking one examination and those taking the other specification) provides an indication of whether the standards were comparable in the two examinations. In other words, this estimator tells us how much more likely a student was to attain a certain grade in one examination rather than the other one. Clearly, the closer to zero this difference is, the more comparable the standards of the two qualifications/examinations are.

The use of matching stems from the fact that the students taking different qualifications/specifications can be quite different from one another in many respects. They may not only differ in terms of ability (measured through prior and/or concurrent level of attainment), but also because of the different composition of the two groups in terms of schooling and other background characteristics. This compositional difference is known as selection bias. If the two groups of students are

different with respect to characteristics that can affect the performance in English/English language and English literature, it is impossible to attribute any difference in the attainment to the specification taken, as it may be due to compositional differences between the two groups.

Matching allows us to overcome this issue, simply limiting the comparison of students taking different qualifications to those who are deemed to be 'similar' to each other. It has to be noted that under specific assumptions, that in this case evidence suggests do hold, matching enables us to retrieve the causal effect of taking a certain examination rather than another examination, avoiding the risk of selection bias. In practice, matching allows us to evaluate the difference in the attainment of students who took a certain specification, compared to what would have happened had they taken another specification. In the context of causal inference the difference retrieved from the subset of matched students is also referred to as the average treatment effect on the treated, ATT (see Appendix C for more details).

7.4 Results

7.4.1 English language

The first comparison carried out considered the CIE level 1/2 certificate qualification, the most popular level 1/2 certificate, and all GCSE specifications. Results of the multilevel modelling are shown in Figure 7.1, where the predicted probability of attaining a grade C/A or above is shown for both the qualifications under scrutiny (left hand side), along with the differences between these probabilities (right hand side). Figure 7.1 shows how these probabilities vary according to students' concurrent attainment.

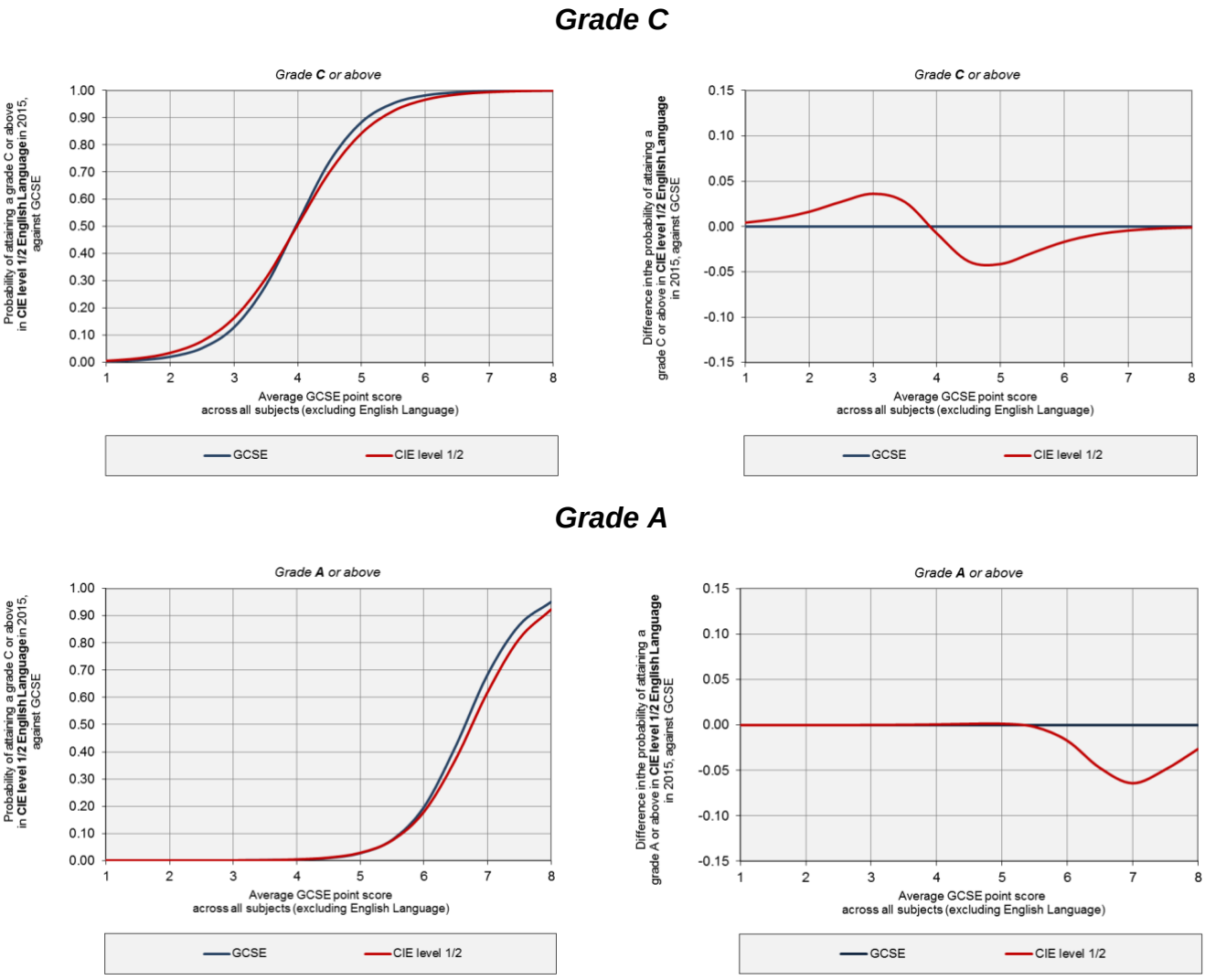

Figure 7.1. *Probability of achieving a grade C/A or above in English/English language (CIE level 1/2 vs. All GCSE English/English language specifications) according to average GCSE point score, 2015*

Table 7.2. *Propensity-score matching estimates of the average effect of attaining grade C/A or above in English language: CIE level 1/2 vs. All GCSE English/English language specifications, 2015*

Specifications compared	*Grade*		*Difference in the probability of attainment*		*Confidence interval*
			Mean	*S. E.*	
CIE level 1/2 vs. All GCSEs	C or above	Before matching	-16.95%	0.14	
		After matching - ATT	**+2.65%**	0.21	[2.25 ; 3.07]
	A or above	Before matching	-9.05%	0.11	
		After matching - ATT	**+1.31%**	0.11	[1.10 ; 1.53]

From Figure 7.1 it emerges that, in 2015, students with an average GCSE point score of 4 (an average D grade) who took the CIE English language specification were more likely to attain a grade C or above than those taking a GCSE in English/English language. The opposite is true for students with an average GCSE point score greater than 4. In both cases differences were below 5%. A difference of a similar size is also apparent for the probability of attaining a grade A or above, where CIE students with an average GCSE point score of 7 (an average grade A) were less likely than those taking a GCSE in English to achieve a grade A or above.

These differences, however, may be due to a different composition of the cohorts of students taking the two qualifications. The propensity-score matching estimates that control for such compositional differences are presented in Table 7.2. The 'before matching' estimates are not very informative and can be misleading as they show the overall raw differences in attainment between the two groups of students taking these qualifications. In this case, it is clear that CIE students performed less well (almost 17% at grade C and 9% at grade A) than GCSE students, though this may be due to a different profile of the cohorts taking the two qualifications.

Table 7.2 shows that, when compositional differences are removed, the probability of attaining a grade C/A or above reduced notably. The 'after matching - ATT' estimates show that CIE students were actually more likely to achieve at least a grade C/A than their matched GCSE counterparts. The estimate of the difference in attainment between the two groups is statistically significant, though the size, 2.65% at grade C and 1.31% at grade A, suggests minor differences in standards in 2015.

Results of the same comparison for the 2016 examination session are reported in Figure 7.2 and Table 7.3. The multilevel modelling analysis shows a similar pattern to the one described with reference to 2015, with one important exception. The size of the difference in the probability of attaining a grade C or above for students with an average GCSE point score between 3 (an average E grade) and 4 (an average D grade) was greater than 10%, whereas in 2015 this figure was less than 5%. This means that, in 2016, CIE students in this ability range were more than 10% more likely than GCSE students to achieve at least a grade C.

The findings of propensity-score matching (Table 7.3) confirm that differences in the probability of achieving a grade C/A (or above) were not simply due to compositional differences not accounted for by the level of concurrent attainment. Once only matched students were considered, a grade C or above was achieved by 6.64% more CIE students. At grade A the mean difference was smaller, close to 2%, but still statistically significant. At both grades, the estimate of the effect of taking the CIE level 1/2 certificate rather than GCSE was greater in 2016 than in 2015.

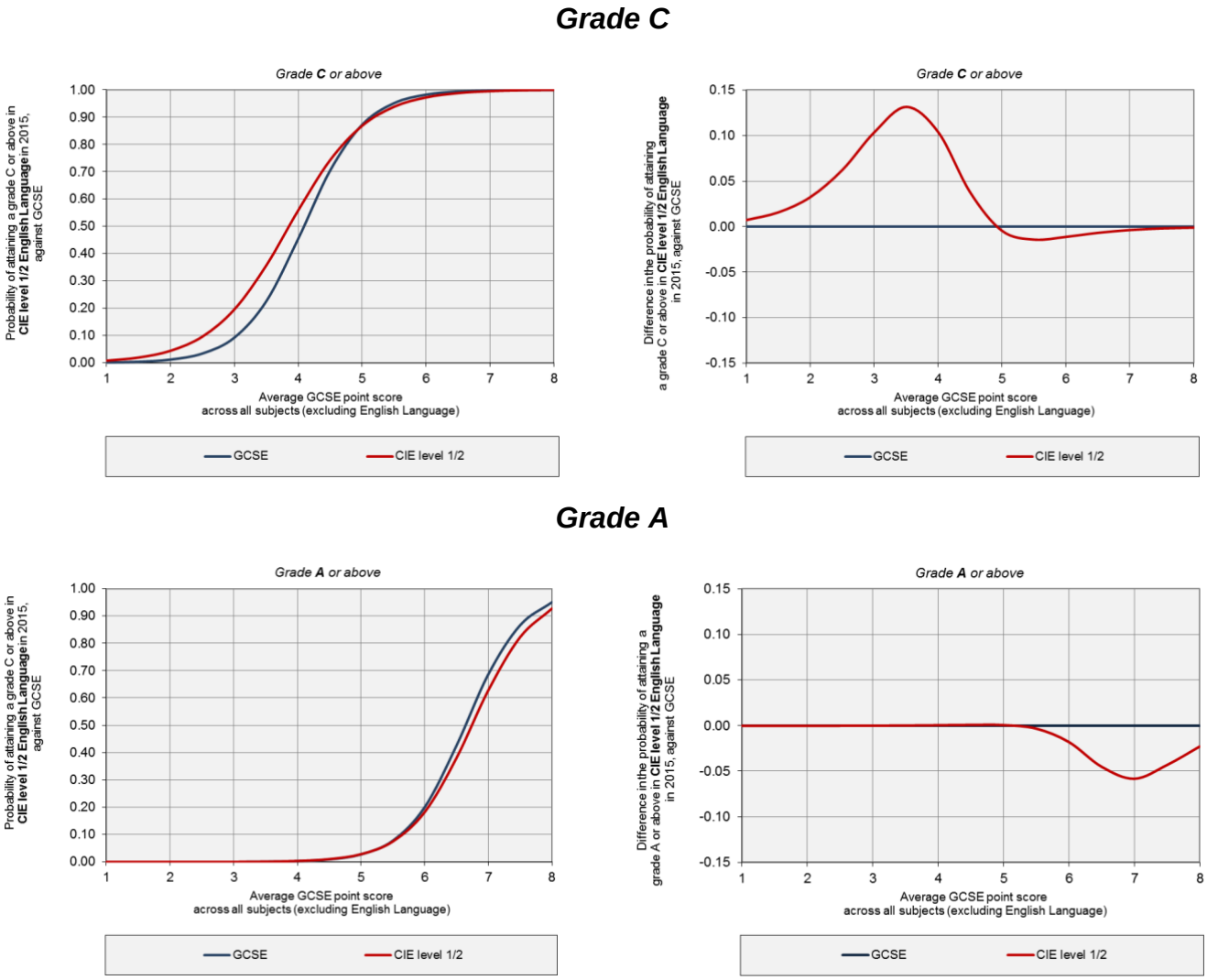

Figure 7.2. *Probability of achieving a grade C/A or above in English language (CIE level 1/2 vs. All GCSE English/English language specifications) according to average GCSE point score, 2016*

Table 7.3. *Propensity-score matching estimates of the average effect of attaining grade C/A or above in English language: CIE level 1/2 vs. All GCSE English/English language specifications, 2016*

Specifications compared	*Grade*		*Difference in the probability of attainment*		*Confidence interval*
			Mean	*S. E.*	
CIE level 1/2 vs. All GCSEs	C or above	Before matching	-10.22	0.14	
		After matching - ATT	**+6.64%**	0.21	[6.24 ; 7.05]
	A or above	Before matching	-8.37%	0.11	
		After matching - ATT	**+1.93%**	0.11	[1.72 ; 2.10]

In order to check the robustness of the above estimates, a further comparison was carried out. Tables 7.4 reports the propensity-score matching estimates of the average effect of taking the CIE level 1/2 certificate rather than the AQA GCSE specification. Focusing on the most popular GCSE English specification allows us to refer to a specific cohort, rather than to a mix of students sitting a number of examinations (though of the same type). This comparison shows even greater differences than those shown above. Once only matched students are compared, the probability of attaining a grade C or above for CIE students is 8.58% greater than for their AQA counterparts. At grade A, the estimate of the effect of taking the CIE specification rather than the AQA GCSE specification was 2.59%.

Table 7.4. *Propensity-score matching estimates of the average effect of attaining grade C/A or above in English language: CIE level 1/2 vs. AQA GCSE English/English language, 2016*

Specifications compared	*Grade*		*Difference in the probability of attainment*		*Confidence interval*
			Mean	*S. E.*	
CIE level 1/2 vs. AQA	C or above	Before matching	-11.01%	0.10	
		After matching - ATT	**+8.58%**	0.23	[8.14 ; 9.03]
	A or above	Before matching	-8.36%	0.09	
		After matching - ATT	**+2.59%**	0.12	[2.36 ; 2.83]

With the aim of checking whether standards were misaligned across GCSE specifications within the same subject and/or over time, two additional comparisons were performed. Table 7.5 reports the propensity-score matching analysis between the two most popular GCSE specifications, ie those provided by AQA and WJEC. Findings suggest that the difference in the probability of attaining a grade C or above was statistically not different from zero. At grade A, AQA students were 1.56% less likely than their WJEC counterparts to achieve this level of attainment. It should be noticed that these two figures are smaller than those found when the CIE level 1/2 certificate was the focus of the analysis.

Figures of a similar size to those in Table 7.5 are also reported in Table 7.6, which contains the estimates of the propensity-score matching analysis for all GCSE English specifications over time. Also in this case the ‘after matching’ estimates show differences in the probability of attaining a grade C/A of around 1%.

Table 7.5. *Propensity-score matching estimates of the average effect of attaining grade C/A or above in English/English language: AQA vs. WJEC, 2016*

Specifications compared	*Grade*		*Difference in the probability of attainment*		*Confidence interval*
			Mean	*S. E.*	
AQA vs. WJEC	C or above	Before matching	3.24%	0.24	
		After matching - ATT	**-0.45%**	0.30	[-1.04; 0.14]
	A or above	Before matching	2.79%	0.22	
		After matching - ATT	**-1.56%**	0.29	[-2.13; -0.99]

Table 7.6. *Propensity-score matching estimates of the average effect of attaining grade C/A or above in GCSE English/English language, 2016 against 2015*

Specifications compared	*Grade*		*Difference in the probability of attainment*		*Confidence interval*
			Mean	*S. E.*	
GCSEs 2016 vs. GCSEs 2015	C or above	Before matching	-3.60%	0.11	
		After matching - ATT	**-0.82%**	0.17	[-1.11; -0.48]
	A or above	Before matching	-0.78%	0.11	
		After matching - ATT	**+1.23%**	0.14	[0.94; 1.52]

The analysis presented in Tables 7.5 and 7.6 seem to suggest that some fluctuations within GCSE English specifications and over time are possible, but are usually quite small. The size of such fluctuations in standards may be quantified as around 1%. This has to be considered as a very minor change that is likely to happen, once the whole methodology to set and maintain standards is considered[29]. The comparisons involving CIE led to differences in the probability of attaining at least a grade C greater than 8%. Although there is not a clear threshold between what is considered a normal fluctuation and what is a difference in standards, evidently, a change of this size cannot be treated as a normal fluctuation.

[29] As discussed in section 5, statistical predictions led the awarding process. However, there are tolerances (up to 3%, depending on the entry size) applied to the predicted grade boundaries within which awarders may apply their professional judgment to set a grade boundary slightly different from the one indicated from the statistical methodology employed.

7.4.2 English literature

Considering that for English language standards were not precisely aligned for 2016, it is for this year that the analysis for English literature is focused. Table 7.7 shows that CIE students were 10.16% more likely to achieve a grade A or above and 5.44% to achieve a grade C than their GCSE counterparts that were deemed to be similar. Differences in attainment were confirmed also by the estimate of the propensity-score matching procedure of CIE with WJEC. As in the case of English/English language, for English literature no difference in standards was highlighted by the comparison of two GCSE specifications.

This is additional evidence that, also for English literature, difference in standards was found when level 1/2 certificates (in this case CIE) are compared to GCSEs. Similar to English/English language, GCSE English literature specifications in 2016 appeared to be comparable to each other.

Table 7.7. *Propensity-score matching estimates of the average effect of attaining grade C/A or above in English literature: various specifications compared, 2016*

Specifications compared	*Grade*		*Difference in the probability of attainment*		*Confidence interval*
			Mean	*S. E.*	
CIE level 1/2 vs. All GCSEs	C or above	Before matching	1.62	0.64	
		After matching - ATT	**5.44%**	0.94	[3.58; 7.30]
	A or above	Before matching	11.55	0.55	
		After matching - ATT	**10.16%**	0.92	[8.34; 11.98]
CIE level 1/2 vs. AQA	C or above	Before matching	-1.86%	0.62	
		After matching - ATT	**3.25%**	1.67	[-0.03; 6.52]
	A or above	Before matching	9.86%	0.57	
		After matching - ATT	**8.42%**	1.58	[5.33; 11.52]
AQA vs. WJEC	C or above	Before matching	3.24%	0.24	
		After matching - ATT	**-0.45%**	0.30	[-1.04; 0.14]
	A or above	Before matching	2.79%	0.22	
		After matching - ATT	**-1.56%**	0.29	[-2.13; -0.99]

7.5 Summary of the results

In this section the comparability of examination standards between level 1/2 certificates and GCSE was investigated through the use of propensity-score matching. The use of this methodology allowed us to make like-for-like comparisons, where only students who took alternative qualifications and were deemed to be

similar to each other, with respect to a broad set of characteristics potentially affecting their performance, were compared.

For both English and English literature the comparison of CIE students and their GCSE counterparts in 2016 highlighted differences in the probability of attaining a grade C/A or above. The size of such differences were much greater than those emerging from the comparisons of GCSE English specifications (both in 2016 and over time). In other words, whilst the analysis suggests alignment of standards between GCSE specifications, there is evidence highlighting the presence of a non-negligible effect on attainment of taking the CIE specification rather than a GCSE.

It is therefore possible to conclude that, once compositional differences of the students sitting alternative examinations are accounted for, a potential lack of comparability in examination standards exists between the CIE specification, the most popular level 1/2 certificate, and GCSE, for both English/English language and English literature in 2016.

8 Summary and conclusion

The purpose of this report was to document the evidence that has informed the approach determined by Ofqual for generating predictions for the reformed GCSE 9 to 1 specifications this summer[30] (as outlined in the summer 2017 data exchange procedures[31]). That is, to generate predictions for each subject based on GCSE only outcomes in the corresponding subject in the reference series. The findings of the analyses were therefore initially considered with reference to the conditions under which one would, or would not, want to generate predictions based on combined GCSE and level 1/2 certificate outcomes.

The purpose of generating predictions based on combined GCSE and level 1/2 certificate outcomes this summer would have been to ensure that any differences in the nature of the students that previously took a level 1/2 certificate are accounted for, now that the majority of 16-year-old students take the GCSE qualification. While the analyses in this report suggested that historically students taking a GCSE or a level 1/2 certificate did perform differently (particularly in 2016), the evidence also suggested that this is likely to be due to differences in the standards of the qualifications, rather than the characteristics of the students. Indeed, once a large number of socio-demographic characteristics are accounted for, the differences in outcomes between those taking the GCSE and level 1/2 certificates remained. This suggested that the outcomes of level 1/2 certificates should not be included in the basis of predictions this summer since we cannot be confident that the standards of the qualifications are precisely aligned. It should be noted that the approach to awarding level 1/2 certificates is different to the GCSE since KS2-based predictions are not routinely used to drive the maintenance of standards. This means that the relationship between KS2 results and grades might not be the same across the two types of qualification.

While these analyses therefore suggested that predictions should not be generated on combined GCSE and level 1/2 certificate outcomes, there are a number of other points to consider. The first relates to what the predictions for the reformed GCSE specifications this summer are attempting to achieve – ie whether they are trying to maintain the GCSE standard, or the overall cohort standard. As a matter of principle, Ofqual consider that the priority is to maintain the GCSE standard. This is due to the differences in the structure, assessment design and content of the GCSE and level 1/2 certificates, and the different methods used for maintaining standards (as

[30] Note that the position of the JCQ exam boards did not align with Ofqual's position.
[31] See https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/620964/Summer_2017_data_exchange_procedure.pdf

outlined above). This provides further impetus for not generating predictions based on combined GCSE and level 1/2 certificate outcomes.

A second point to consider is the method that has been used to generate predictions for GCSE awarding in recent years as students have moved from the GCSE cohort to the level 1/2 certificate cohort. While it might be argued that using only 2016 GCSE outcomes to generate predictions does not reflect the whole candidature sitting GCSEs in summer 2017, this seems a difficult argument to sustain. In 2013-2016 predictions were not adjusted in any way to account for students moving away from the GCSE to take level 1/2 certificates. It therefore seems difficult to justify adjusting the basis of predictions to take account of students returning to the GCSE cohort, when there was no adjustment when they left.

A final point to consider relates to the consistency of approach across subjects. The analyses reported here have focused on English language and English literature since this is where entries to the level 1/2 certificate qualifications have been greatest. However, as a principle, Ofqual considered that the same approach should be applied to all three of the reformed GCSE subjects this summer. Thus, the same approach is set out in the data exchange for mathematics and will also be taken for all reformed GCSEs as they are awarded.

9 References

Adelson, J.L. (2013). Educational research with real-world data: reducing selection bias with propensity score analysis. *Practical Research & Evaluation*, 18.

Bramley, T. (2011). Subject difficulty – the analogy with question difficulty. *Research Matters: A Cambridge Assessment Publication*, Special Issue 2: Comparability, 27-33.

Caliendo M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1): 31–72.

Coe, R. (2008). Comparability of GCSE examinations in different subjects: An application of the Rasch model. *Oxford Review of Education* 34, pp. 609 – 636.

Coe, R., Searle, J., Barmby, P., Jones, K. and Higgins, S. (2008). *Relative difficulty of examinations in different subjects, Report for SCORE (Science Community Supporting Education)*, CEM Centre, Durham University. Available online at: http://www.score-education.org/media/3194/relativedifficulty.pdf

Eason, S. (2010).Predicting GCSE outcomes based on candidates' prior achieved Key Stage 2 results. Guildford, UK: CERP.

Elliott, G. (2011). A guide to comparability terminology and methods. *Research Matters: A Cambridge Assessment Publication*, Special Issue 2: 9-19.

Goldstein, H. (2011). *Multilevel Statistical Models (4th edition)*. Chichester: John Wiley & Sons.

Goldstein, H. and Cresswell, M.J. (1996). The comparability of different subjects in public examinations: a theoretical and practical critique, *Oxford Review of Education*, 22, pp. 435-442.

He, Q. and Stockford, I. (2015). Inter-Subject Comparability of Exam Standards in GCSE and A Level. Ofqual: Coventry, UK. Available online at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/486936/3-inter-subject-comparability-of-exam-standards-in-gcse-and-a-level.pdf

He, Q., Stockford, I. and Meadows, M. (2016). Using step functioning analysis and Rasch modelling to compare GCSE exam standards between exam boards. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/572103/Inter_Board_Comparability_-_Rasch_modelling.pdf

Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression*. Chichester: Wiley.

Linacre, J. (2015). Winsteps® Rasch measurement computer program User's Guide. Beaverton, Oregon: Winsteps.com.

Masters, G. (1982) A Rasch model for partial credit scoring. *Psychometrika* 47, 149-74.

Newton, P.E. (1997). Measuring comparability of standards between subjects: why our statistical techniques do not make the grade. *British Educational Research Journal*, 23(4), 433-449.

Nuttall, D.L., Backhouse, J.K. & Willmott, A.S. (1974). Comparability of Standards Between Subjects. *Schools Council Examinations Bulletin* 29 (London, Evans/Methuen Educational).

Opposs, D. (2015). Inter-subject comparability: an international review. Ofqual: Coventry, UK. Available online at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/486939/4-inter-subject-comparability-an-international-review.pdf

Rosenbaum, P. and Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70: 41-50.

Rosenbaum, P. and Rubin, D.B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. *The American Statistician*, 39: 33-38.

Taylor, M. (2013). GCSE (and level 1/2 project and functional skills) statistical screening. Assessment and Qualifications Alliance (AQA).

Wright, B. and Masters, G. (1982). *Rating scale analysis: Rasch measurement*. Chicago, USA: MESA Press.

Zanini, N. (2016). *The comparability of comparability methods: screening, modelling and matching*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Appendix A

Table A.1. *Percentage of students achieving each grade from whole cohort taking either a GCSE or CIE level 1/2 certificate and the subset of cohort taking both subjects in 2016*

Grade	GCSE all	GCSE double entry	CIE all	CIE double entry
*	3.45	0.65	2.53	0.91
A	10.57	3.42	8.52	5.15
B	20.20	10.19	17.82	11.67
C	26.40	26.65	32.43	43.86
D	22.12	37.15	23.71	27.08
E	9.53	14.14	9.18	8.33
F	3.70	5.02	2.94	1.48
G	1.93	1.79	1.04	0.53
U	2.11	0.99	1.83	0.98

Table A.2. *Percentage of students at each school type from cohorts taking a GCSE only, CIE level 1/2 certificate only and entered for both subjects in 2016*

Centre Type	CIE only	GCSE only	Double entry
Comprehensive	30.18	41.93	47.73
City academy	46.93	24.33	47.39
Other	3.33	3.00	1.02
6th/FE/Tertiary college	12.11	17.21	0.19
Independent	3.45	3.39	0.72
Secondary Modern	2.74	3.09	1.17
Secondary selective	1.00	6.78	1.78

Table A.3. *Percentage of students achieving each grade from whole cohort taking either a GCSE or CIE level 1/2 certificate and the subset of cohort taking both subjects in 2015*

Grade	GCSE all	GCSE double entry	CIE all	CIE double entry
*	2.90	1.51	2.01	2.86
A	11.91	6.16	7.79	9.80
B	22.85	14.94	18.18	18.76
C	28.25	28.91	32.40	36.97
D	20.77	30.63	23.78	21.10
E	7.82	11.55	9.67	7.15
F	2.77	4.21	3.11	1.46
G	1.41	1.68	1.10	0.49
U	1.31	0.40	1.96	1.41

Table A.4. *Percentage of students at each school type from cohorts taking a GCSE only, CIE level 1/2 certificate only and entered for both subjects in 2015*

Centre Type	CIE only	GCSE Only	Double Entry
Comprehensive	34.33	46.56	38.45
City academy	48.28	25.66	56.12
Other	2.41	2.53	0.79
6th/FE/Tertiary college	8.09	12.10	0.12
Independent	3.03	4.08	1.20
Secondary Modern	3.25	3.58	2.57
Secondary selective	0.60	5.29	0.76

Appendix B

Table B.1. *Cumulative percentage of GCSE English/English language grades achieved by students at each octile of KS2 attainment – June 2015*

		GCSE grade								
Octile	N	*	A	B	C	D	E	F	G	U
8	27,650	17.5	58.7	90.8	98.8	99.8	99.9	100.0	100.0	100.0
7	27,401	5.1	31.9	75.4	95.9	99.2	99.7	99.9	100.0	100.0
6	29,159	2.2	18.9	60.3	91.2	98.6	99.5	99.8	99.9	100.0
5	27,785	1.0	10.9	45.8	84.4	97.1	99.0	99.5	99.8	100.0
4	26,748	0.4	6.1	33.4	76.5	95.3	98.6	99.3	99.8	100.0
3	26,784	0.2	2.8	21.4	64.7	91.4	97.3	98.9	99.6	100.0
2	22,654	0.1	1.2	11.1	47.9	83.9	95.0	97.9	99.2	100.0
1	21,448	0.0	0.2	2.8	22.1	60.4	85.2	94.9	98.3	100.0
Total	209,629	3.5	17.3	44.8	74.9	91.8	97.2	98.9	99.6	100.0

Table B.2. *Cumulative percentage of level 1/2 certificate English language grades achieved by students at each octile of KS2 attainment – June 2015*

		Level 1/2 certificate grade								
Octile	N	*	A	B	C	D	E	F	G	U
8	7,058	15.5	56.6	89.5	98.7	99.7	99.8	99.9	99.9	100.0
7	8,361	4.9	31.2	73.8	95.6	99.5	99.8	99.9	99.9	100.0
6	10,366	2.2	18.0	58.5	90.6	98.5	99.5	99.7	99.7	100.0
5	11,881	1.0	10.3	42.9	84.2	97.4	99.2	99.6	99.7	100.0
4	13,502	0.5	5.6	30.5	74.6	95.3	98.6	99.2	99.4	100.0
3	15,880	0.2	2.7	19.3	64.5	91.6	97.7	98.7	99.0	100.0
2	16,321	0.1	0.8	9.2	48.1	83.9	95.6	97.6	98.1	100.0
1	17,837	0.0	0.2	3.0	23.6	61.3	84.8	92.8	95.1	100.0
Total	101,206	1.9	10.9	32.5	65.9	88.1	95.9	97.9	98.5	100.0

Table B.3. *Cumulative percentage of GCSE English/English language grades achieved by students at each octile of KS2 attainment – June 2016*

		GCSE grade								
Octile	N	*	A	B	C	D	E	F	G	U
8	34,868	20.9	59.1	90.0	98.5	99.7	99.8	99.9	100.0	100.0
7	38,738	6.2	32.1	73.8	95.1	99.1	99.6	99.8	99.9	100.0
6	38,182	2.5	18.3	57.2	89.5	98.1	99.3	99.7	99.9	100.0
5	34,150	1.2	11.2	44.1	82.6	96.6	98.9	99.5	99.8	100.0
4	36,077	0.5	5.9	30.9	73.1	94.0	97.9	99.0	99.7	100.0
3	31,898	0.2	2.9	19.1	61.3	90.0	96.6	98.6	99.5	100.0
2	29,717	0.1	1.2	10.2	45.6	81.7	94.1	97.6	99.2	100.0
1	27,882	0.0	0.2	2.8	21.2	58.6	84.0	94.6	98.1	100.0
Total	271,512	4.2	17.4	43.4	73.3	90.9	96.7	98.7	99.6	100.0

Table B.4. *Cumulative percentage of level 1/2 certificate English language grades achieved by students at each octile of KS2 attainment – June 2016*

		Level 1/2 certificate grade								
Octile	N	*	A	B	C	D	E	F	G	U
8	10,605	18.3	60.8	90.6	98.8	99.8	99.9	99.9	100.0	100.0
7	14,009	5.6	33.0	75.3	95.7	99.3	99.8	99.9	99.9	100.0
6	16,370	2.2	18.9	58.1	90.6	98.5	99.6	99.7	99.8	100.0
5	16,671	1.0	10.5	43.2	83.7	97.0	99.0	99.4	99.6	100.0
4	20,601	0.4	5.9	30.3	75.0	94.9	98.5	99.1	99.4	100.0
3	20,948	0.1	2.6	18.9	64.1	90.8	97.4	98.6	98.9	100.0
2	22,669	0.1	1.1	10.4	50.3	85.4	95.7	97.7	98.4	100.0
1	24,933	0.0	0.3	3.2	27.5	66.8	87.8	94.8	96.5	100.0
Total	146,806	2.3	12.3	34.2	68.0	89.5	96.5	98.3	98.8	100.0

Table B.5. *Cumulative percentage of GCSE English literature grades achieved by students at each octile of KS2 attainment – June 2015*

		GCSE grade								
Octile	N	*	A	B	C	D	E	F	G	U
8	29,025	19.6	62.3	91.6	98.5	99.6	99.8	99.9	99.9	100.0
7	28,871	6.9	37.5	79.0	95.5	98.8	99.5	99.7	99.8	100.0
6	30,627	3.6	24.1	66.6	91.5	97.9	99.2	99.6	99.8	100.0
5	29,154	1.6	15.6	54.1	85.2	96.2	98.6	99.3	99.7	100.0
4	27,766	0.8	9.7	41.7	77.0	93.5	97.7	99.0	99.5	100.0
3	27,071	0.4	5.3	29.4	66.3	89.0	96.3	98.4	99.2	100.0
2	22,045	0.2	2.5	17.0	49.3	80.6	93.0	97.4	98.8	100.0
1	18,196	0.1	0.7	5.5	24.8	57.3	82.7	93.4	97.1	100.0
Total	212,755	4.5	21.5	51.6	77.0	91.0	96.6	98.6	99.3	100.0

Table B.6. *Cumulative percentage of level 1/2 certificate English literature grades achieved by students at each octile of KS2 attainment – June 2015*

		Level 1/2 certificate grade								
Octile	N	*	A	B	C	D	E	F	G	U
8	3,494	28.1	58.7	85.2	95.7	98.0	99.1	99.5	99.7	100.0
7	3,515	11.4	34.1	68.1	88.8	95.6	98.1	99.1	99.4	100.0
6	4,172	5.5	20.4	51.9	79.9	91.7	95.9	98.0	99.1	100.0
5	4,702	3.0	12.8	39.1	70.6	86.0	92.7	95.7	98.1	100.0
4	5,305	1.6	8.1	28.1	60.0	79.8	89.8	94.5	97.5	100.0
3	6,304	0.8	4.6	18.8	46.9	71.6	84.7	92.7	96.7	100.0
2	6,507	0.3	2.0	11.2	35.1	59.8	78.0	89.3	95.5	100.0
1	7,220	0.1	0.6	4.0	16.9	38.1	59.2	78.3	90.2	100.0
Total	41,219	4.6	13.6	31.7	55.2	72.9	84.2	91.9	96.4	100.0

Table B.7. *Cumulative percentage of GCSE English literature grades achieved by students at each octile of KS2 attainment – June 2016*

			GCSE grade							
Octile	N	*	A	B	C	D	E	F	G	U
8	38,207	20.8	61.2	90.9	98.5	99.5	99.7	99.8	99.9	100.0
7	43,226	7.5	36.7	77.1	95.3	98.8	99.4	99.7	99.8	100.0
6	43,238	3.3	22.6	62.7	90.1	97.6	99.1	99.5	99.8	100.0
5	38,863	1.8	14.8	50.4	83.8	95.7	98.4	99.2	99.7	100.0
4	41,361	0.9	9.1	38.3	75.2	92.9	97.6	98.8	99.4	100.0
3	36,321	0.4	4.8	26.4	64.3	88.4	96.2	98.4	99.2	100.0
2	33,528	0.1	2.3	15.6	49.2	80.5	93.4	97.4	98.8	100.0
1	29,354	0.0	0.5	5.2	24.9	59.4	83.9	94.2	97.6	100.0
Total	304,098	4.6	20.1	48.3	75.2	90.5	96.5	98.6	99.3	100.0

Table B.8. *Cumulative percentage of level 1/2 certificate English literature grades achieved by students at each octile of KS2 attainment – June 2016*

			Level 1/2 certificate grade							
Octile	N	*	A	B	C	D	E	F	G	U
8	5,296	26.7	58.3	86.4	96.6	98.7	99.4	99.6	99.8	100.0
7	6,373	10.9	34.1	68.1	90.1	96.5	98.3	99.0	99.5	100.0
6	7,305	6.0	22.9	54.6	82.8	92.3	96.4	98.2	99.1	100.0
5	7,470	3.4	15.1	42.7	74.2	88.4	94.7	97.3	98.8	100.0
4	9,275	1.5	8.8	30.7	63.2	81.8	91.2	95.9	98.2	100.0
3	9,332	0.8	5.4	21.5	52.9	75.2	87.9	94.3	97.7	100.0
2	10,132	0.3	2.6	13.9	41.4	65.4	82.0	91.4	96.2	100.0
1	11,005	0.1	0.9	5.9	21.7	45.0	65.8	83.2	93.1	100.0
Total	66,188	4.6	14.7	34.8	60.2	76.9	87.4	93.9	97.4	100.0

Appendix C

Multilevel modelling and propensity-score matching: technical details

Multilevel modelling

Detailed discussions of the implementation and outcomes of this technique can be found in Hosmer and Lemeshow (2000) and Goldstein (2011). In very brief summary, multilevel logistic regression aims to predict which category of the dependent variable a case is most likely to fall in given the information from the independent variables and the hierarchical clustering of the data. The outcome of the analysis is a regression equation that represents the best prediction of the dependent variable using the independent variables (ie specification and average point score (*aps*)), while controlling for the hierarchical clustering.

The formal representation of the regression equation takes the following form:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 specification_{ij} + \beta_2 aps_{ij} + \beta_3 (specification * aps)_{ij} + u_j$$

where: p_{ij} is the probability of student *i* in centre *j* attaining a grade C/A or above; β_k are the regression coefficients and u_j is a random variable at the centre level which follows a normal distribution with mean zero.

The estimates of the regression coefficients from this multilevel logistic regression equation can be quite difficult to interpret directly. However, the equation can be used to calculate estimates of the probabilities of students with specific average GCSE point scores attaining a grade C/A or above in each individual specification. These calculations are conducted by re-arranging the logistic regression equation as follows:

$$\hat{p}_{ij} = \frac{EXP(\hat{\beta}_0 + \hat{\beta}_1 specification_{ij} + \hat{\beta}_2 aps_{ij} + \hat{\beta}_3 (specification * aps)_{ij})}{1 + EXP(\hat{\beta}_0 + \hat{\beta}_1 specification_{ij} + \hat{\beta}_2 aps_{ij} + \hat{\beta}_3 (specification * aps)_{ij})}$$

These are the probabilities shown in the figures presented in section 7.

Propensity-score matching

The matching procedure can be used to evaluate the comparability of a specification (or a group of specifications) under scrutiny, say specification ‘A’, and one or more different specifications, say specification ‘B’. In this context the matching procedure provides an estimate of the difference in the attainment of students who had taken specification A compared to the attainment of students who had taken specification B, once compositional differences between the two groups of students are accounted for.

More formally, let *D* be the binary variable denoting the specification taken, with *D*=1 for students who took specification ‘A’ (also referred to as ‘treated’) and *D*=0 for

students who took specification 'B' (also referred to as 'controls')[32]. Let (Y_A, Y_B) be the two individual potential outcomes that would be realised if a student takes specification 'A' and 'B', respectively. In the context of this report, Y represents the attainment in the GCSE specification measured as a dichotomous variable, with Y=1 if a student attained a grade C/A or above and Y=0 otherwise.

The causal effect of taking specification 'A' rather than specification 'B' is then defined at individual level as the difference between these outcomes, Y_A - Y_B. This quantity is not observable at individual level, since taking specification 'A' reveals Y_A but conceals Y_B. In fact, if a student took specification 'A' (D=1), then Y_A will be realised and Y_B will be a counterfactual outcome and vice versa. Matching students 'A' with students 'B' with respect to a large set of characteristics X that can potentially influence the outcome Y identifies two groups of students that are comparable except for the specification taken. Focusing on matched students, we can therefore identify the *average causal effect of taking specification 'A' rather than 'B'* as the difference in the average attainment of those taking specification 'A' and those taking 'B':

$$ATT = E[Y_A|D = 1, X] - E[Y_B|D = 0, X].$$

This is the so-called Average Treatment Effect on the Treated (ATT) and can be estimated substituting the empirical counterparts of the quantities above, ie the mean values of the attainment of students taking specification 'A' and of those taking specification 'B' that had been matched.

Under specific assumptions, matching enables us to retrieve the *causal effect* of taking specification A rather than specification B, avoiding the risk of selection bias. In practice, matching allows us to evaluate the difference in the attainment of students who took a certain specification, compared to what would have happened had they taken another specification.

The unbiasedness of the matching estimator for the average causal effect we are looking for crucially rests on the so-called *ignorability condition* (also known as *conditional independence assumption*):

$$Y_A, Y_B \perp D \mid X,$$

which amounts to assuming that compositional differences in the pool of students taking specification 'A' and 'B' are solely limited to the observable characteristics X used to match students.

As, in practice, it would be difficult to find students taking the two different specifications having exactly the same observable characteristics, the matching procedure could be unfeasible. To solve this dimensionality issue, it is possible to match students on the basis of a single summary score describing the propensity

[32] Note that, in the counterfactual model of causality, it is common practice to use the terms 'treated' and 'controls' to refer to the two groups under scrutiny.

towards taking specification ‘A’ rather than ‘B’. This is known as the *Propensity Score* (p-score). Statistically, the p-score is the conditional probability of being treated given all the characteristics that can influence the decision to take a certain specification and related to the outcomes of interest (attainment):

$$e(X) = Pr(D = 1|X).$$

Rosenbaum and Rubin (1983) proved that the conditional independence assumption based on *X* is equivalent to the one based on $e(X)$:

$$Y_A, Y_B \perp D \mid e(X).$$

This is the p-score matching identifying restriction claiming that compositional differences between students taking specifications ‘A’ and ‘B’ are solely due to observable characteristics used to estimate the p-score. In other words, once the p-score is accounted for, the distribution of the covariates between the two groups has to be balanced.

Therefore, the matching procedure can be based, without any loss of generality, on the p-score rather than on each single observable characteristic. The ATT based on the p-score is identified by:

$$ATT = E[Y_A|D = 1, e(X)] - E[Y_B|D = 0, e(X)].$$

and estimated by the empirical counterparts to the quantities above, computed for the subsamples of students matched using the p-score.

A second assumption for the p-score to be employed is that there should be enough students in the two groups with the same (or at least similar) values of the p-score. This is known as *common support assumption*, as it ensures that there are comparable units to be compared and it could be written as:

$$0 < Pr(D = 1|e(x)) < 1.$$

Operatively, in order to obtain p-score matching estimates, Rosenbaum and Rubin (1985) suggested following a two-step procedure.

The first step is to estimate the p-score. For each student *i* the p-score estimate is given by the predicted value of a logistic regression, where the specification taken is the dependent variable (*D*) and the observable covariates *X* are the independent variables:

$$\hat{e}_i = \widehat{e(X)}_i = Pr(\widehat{D_i = 1}|\, X_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}}}.$$

The criteria for the selection of the covariates to be employed in the p-score estimation have been discussed extensively in the research literature. Most of the recent studies in this field encourage the use of a rich set of covariates, including interactions and higher-order terms. Here we considered that: i) characteristics connected (even if weak predictors) to the outcome should always be included in the p-score model (Brookhart *et al.*, 2006); ii) p-score cannot be computed only from

predictors of convenience, selected in order to ensure common support (see Adelson (2013) for further references).

As a result, the list of covariates employed here included a wide range of individual and school characteristics. At individual level we considered socio-demographic characteristics (gender, year group, ethnicity, language), students' socio-economic background (eg free school meal eligibility), prior school career (attainment at KS2) and concurrent attainment (average GCSE point score excluding the focus subject, ie English language or English literature). At school level we considered: size; gender; type; average GCSE point score; number of GCSEs and other qualifications obtained in the school; KS2 attainment; percentage of pupils in the school entitled to free school meals; average number of pupils in the school with English as their mother tongue. Some interactions between covariates have also been considered in order to improve the goodness-of-fit of the regression models and then the p-score estimates. In order to check if there are enough students in the two groups with the same values of the p-score, the p-score distribution of students taking specification 'A' and 'B' is evaluated in order to verify the presence of an overlap.

The second step is to match students taking specification 'A' and specification 'B' on the basis of the estimates of p-score obtained in the first step. The p-score matching procedure can be carried out following different strategies (Caliendo and Kopeinig, 2008). The most common choice is the so-called *nearest neighbour matching* technique. Using this technique each treated unit is matched to the control unit with the closest p-score. As a control can be the nearest neighbour of several treated, in order to not lose good matches it is possible to allow for *replacement*, which permits the use of the same control unit more than once. However, in order to avoid the risk of bad matches (when the closest control is far away), it is common practice to set a *caliper*, a tolerance level on the maximum propensity score distance. Caliper matching provides a way to impose the common support assumption, as it ensures that a student in the control group is chosen as matching partner only if s/he lies within the caliper.

Results presented in the report were obtained with the *nearest neighbour within caliper* procedure. For each treated student taking specification 'A', the student taking specification 'B' with the closest p-score was selected. In order to improve the unbiasedness of the estimates, only students with an estimated p-score differing less than 0.01 from each other were considered; those not in the caliper range were not considered as possible matches. The p-score matching procedure for the estimation of the ATT was performed using the SAS macro %PSMatching. Standard errors of the ATT were obtained through a bootstrap procedure that makes use of 200 replications. Confidence intervals for the ATT were based on this estimate of the standard error.

Figure C.1 refers to the comparison between the estimated propensity score distribution for the CIE specification (treated, represented by pink bars) and all GCSE English specifications (controls, represented by blue bars) in 2016. The figure

shows that there was good overlap between the two groups of students in terms of propensity score, which means we are likely to find adequate matches for most of the 'treated' students. This is also showing us that a relatively small proportion of students at the extremes of the [0, 1] range have been discarded from the comparison. Nevertheless, we can conclude that the common support assumption is verified.

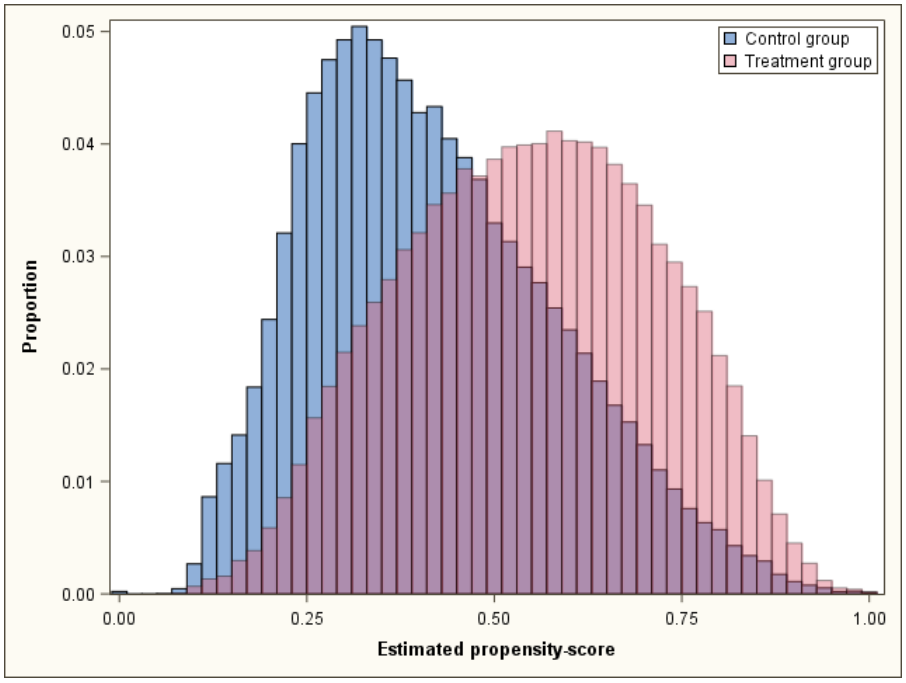

Figure C.1. *Estimated propensity score distributions, 2016*

Figure C.2 shows the distribution of the standardised differences between students taking the CIE specification and GCSE in English language, computed on the whole sample of students (before matching) and only for those matched (after matching). Results confirm that, whilst before matching treated and control students were considerably different, after matching the distribution of the standardised differences of the covariates appeared to be much smaller. This is a clear indication that, although some differences remain, the matching procedure was able to reduce the compositional differences in the cohorts taking both qualifications and therefore evidence suggesting that the conditional independence assumption holds.

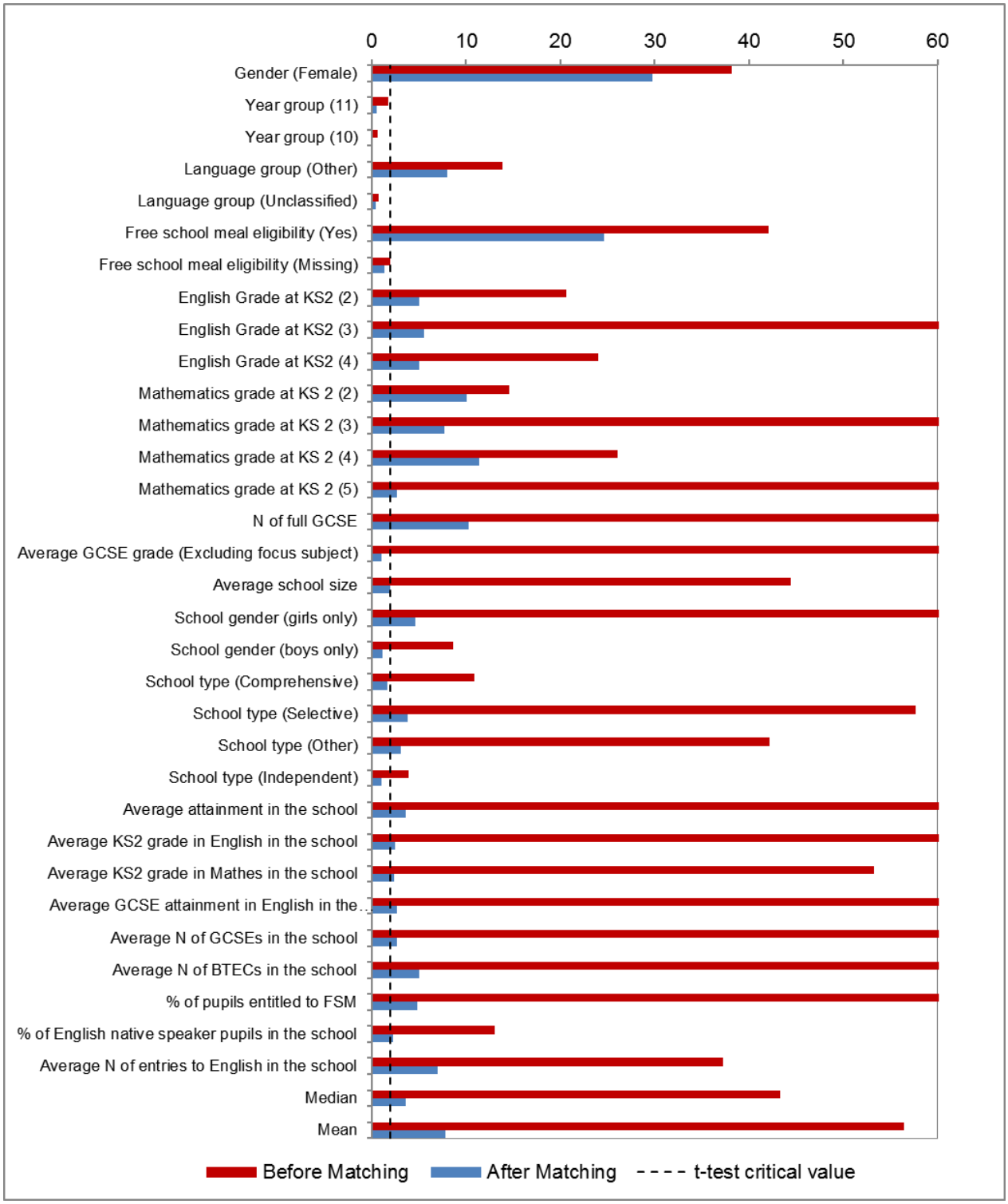

Figure C.2. *Distribution of the absolute values of the standardised differences before and after matching CIE level 1/2 certificate vs. All GCSE English language specifications, 2016*

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

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

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

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

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