RESEARCH AND ANALYSIS

Statistical evidence pertaining to the claim of grading severity in GCSE French, German and Spanish and the impact of statistical alignment of standards on outcomes

Qingping He and Beth Black

ofqual

Contents

Executive summary **3**

1 Introduction **3**

2 Statistical difficulty of GCSE French, German and Spanish **4**

2.1 Statistical methods used to study inter-subject comparability in GCSE *5*

2.2 Relative subject difficulty of GCSE French, German and Spanish *6*

2.3 Consistency of subject difficulty estimated using different methods *11*

2.4 Consistency of relative subject difficulty over time *13*

2.5 Interpretation of results from statistical analysis and implications *22*

2.6 Issues with Rasch modelling and other statistical techniques *23*

2.7 Factors not considered in statistical modelling *27*

2.8 Linking statistical difficulty measures to subject grading standards *29*

2.9 Summary *31*

3 Impact of statistical alignment of standards on outcomes **31**

3.1 Change in grade boundaries and grade distributions *31*

3.2 Summary *37*

4 Conclusion **38**

References **38**

Executive summary

This report forms part of the research undertaken by Ofqual to investigate the claim that GCSE French, German and Spanish are severely graded in comparison with other GCSE subjects. The report focuses on evidence generated from statistical analyses. Evidence from existing research as well as results from new analyses are included in the report. The report also highlights important issues with making simple inferences from this type of work.

Results from statistical analyses using a range of methods suggest that GCSE French, German and Spanish are generally "statistically more difficult" than many of the other GCSE subjects at both individual grade level and the overall subject level over the past 15 years or so. Such evidence has been frequently interpreted as an indication of grading severity in these subjects. The report discusses the major issues with the statistical methods used for comparing standards between subjects, including the strong assumptions made about the data being analysed (for example, the unidimensionality assumption that a single underlying latent trait used to define subject difficulty is assessed by examinations in different subjects) which are seldom met by real data. Many of the factors that can potentially influence examination performance are also discussed, including motivation of students, subject demand, and efficiency and effectiveness of teaching and learning. These factors can vary substantially between subjects and most statistical methods fail to take them into account. It is suggested that, although statistical analysis can provide useful information about the relationship in performance between different subjects, caution must be exercised when interpreting subject difficulty measures derived and linking difficulty measures directly to the grading standards of examinations which are generally subject specific.

The report also demonstrates that aligning standards statistically between subjects could potentially impact performance standards. Depending on the particular point of statistical alignment chosen, this would have a lesser or greater impact on the quality of performance expected of students for some grades in GCSE French, German and Spanish.

1 Introduction

The comparability of standards in A level and GCSE examinations between different subjects has been a matter of debate for a long time in England (see Coe, Searle, Barmby, Jones and Higgins, 2008; Coe, 2010; Newton, 2012, 2015; Lockyer and Newton, 2015). Concerns about inconsistency in grading standards between subjects have frequently been raised by stakeholders. In response to such concerns, Ofqual initiated a research program in 2015 to gather and evaluate relevant evidence in order to make an informed decision on whether to align standards between subjects systematically through grading. After considering carefully the evidence from existing research and the arguments for and against statistical alignment of standards between subjects, Ofqual decided not to take any coordinated action to align standards across the full range of GCSE and A level subjects through grading (see Ofqual, 2016). It has, however, committed to considering the evidence for one-off adjustments to individual subject standards. In 2018, Ofqual evaluated the

evidence collected from a wide range of sources concerning grading severity in A level sciences and modern foreign languages (MFLs, including French, German and Spanish) and decided not to adjust the grading standards of these subjects (see Black, Clewes, Garrett, He and Curcin, 2018a, b; Ofqual, 2018).

This piece of work is part of the evidence gathering to investigate the claim that GCSE French, German and Spanish are severely graded in comparison with other GCSE subjects. This report forms part of the research and focuses on evidence generated from statistical analyses. Evidence from existing research as well as results from new analyses are included in the report.

2 Statistical difficulty of GCSE French, German and Spanish

Much of the evidence underpinning the debate about the comparability of examination standards in different subjects in GCSE and A level comes from statistical analyses. While statistical analysis can provide useful information about the relationship in performance between different subjects, its important limitations must be recognised at the outset when used to compare the grading standards of examinations. For most of the statistical techniques used to study inter-subject comparability, it is explicitly or implicitly assumed that examinations in different subjects that are analysed together define a shared common construct or latent trait (such as “general academic ability/aptitude”) which is closely related to the constructs being measured by the individual examinations. The difficulty of a subject is normally defined as the amount of the common trait required to achieve a specific level of performance on the exam. Although difference in difficulty (which is statistically defined) between exams in different subjects may reflect difference in the amount of this common trait that is needed to achieve the same level of performance, it does not necessarily imply that some subjects are graded more severely (or leniently) than others. This is because examinations such as GCSEs and A levels are graded based on standards representing the expected level of knowledge and skills which are subject specific. The shared knowledge and skills (represented by the common latent trait) assessed by different subjects are therefore irrelevant in this respect. Furthermore, as will be clear, the shared knowledge and skills can vary considerably between subjects. There can also be issues with the statistical methods themselves (for example, violation of model assumptions by real data, imperfect data-model fit, missing data, and others). It is important to bear all this in mind when comparing grading standards between different subjects based on results from statistical analyses.

2.1 Statistical methods used to study inter-subject comparability in GCSEs

Statistical methods used to study inter-subject comparability in GCSEs generally involve examining the relationships of grade outcomes between different subjects taken by the same students or the relationships between subject grade outcomes and external variables that can potentially influence students' performance on the exams. Coe et al. (2008) provided a comprehensive review of the various statistical methods that have been used to investigate comparability of examination standards. These include subject pairs analysis, common examinee linear models (including Kelly's method), latent trait models (such as Rasch models), reference tests and value-added models (including multilevel modelling) (see also Lockyer and Newton, 2015). We provide a brief explanation for each of these methods below.

The subject pairs analysis (SPA) approach looks at the average of the differences between the grades achieved in two subjects taken by the same group of candidates. When a large number of subjects are involved, the difficulty of a specific subject can be calculated as the simple mean or weighted mean of the averages of differences of all possible subject pairs. This is then compared with the difficulties of the other subjects calculated in the same way.

Common examinee linear models (including Kelly's method) derive the relative difficulties of different subjects from a matrix of examinations by candidate results and involve the comparison of the average grade of candidates in one subject with the average grade in all the other subjects.

In the case of latent trait models, such as Rasch models, each examination is viewed as a polytomous item (characterised by a set of item difficulty parameters) in a test, and the grade or performance level assigned to an individual person (characterised by an ability parameter) on an exam are treated as scores on an item which represent ordered response categories. All exams contained in the analysis form a test. A mathematical function is used to describe the probability of a person (with a certain level of ability) succeeding on an item (with a certain level of difficulty). The difficulty measures of the items and ability measures of the persons can then be estimated using a range of approaches such as the conditional maximum likelihood (CML) estimation approach.

The reference test approach examines the relationship between subject grade outcomes and reference test scores (normally through regression analysis). Any difference in the relationship between subjects would suggest difference in difficulty. The value-added analysis approach (including multilevel modelling) represents an extension of the reference test approach. In this approach, the regression model can include a range of explanatory variables (such as a candidate's prior attainment,

gender, socioeconomic status, type of school attended, and many others) that can potentially influence examination performance.

2.2 Relative subject difficulty of GCSE French, German and Spanish

In this section, Rasch modelling and analysis based on relationships with prior attainment have been used to study relative difficulty of GCSE subjects.

2.2.1 Subject difficulty derived from Rasch modelling

As an example of using the Rasch modelling approach to study inter-subject comparability, Figure 1 compares the overall difficulty and the difficulties of individual grades (in logits) between 30 GCSE subjects in 2016 estimated with the subjects ordered by the overall subject difficulty. The GCSE outcome data used here were extracted from the National Pupil Database (NPD). Details of the Rasch analysis approach are provided in Coe et al. (2008) and He, Stockford and Meadows (2018). There is variability in statistical difficulty at individual grades between the subjects, with French, German and Spanish shown to be statistically harder than the majority of the subjects included in the analysis.

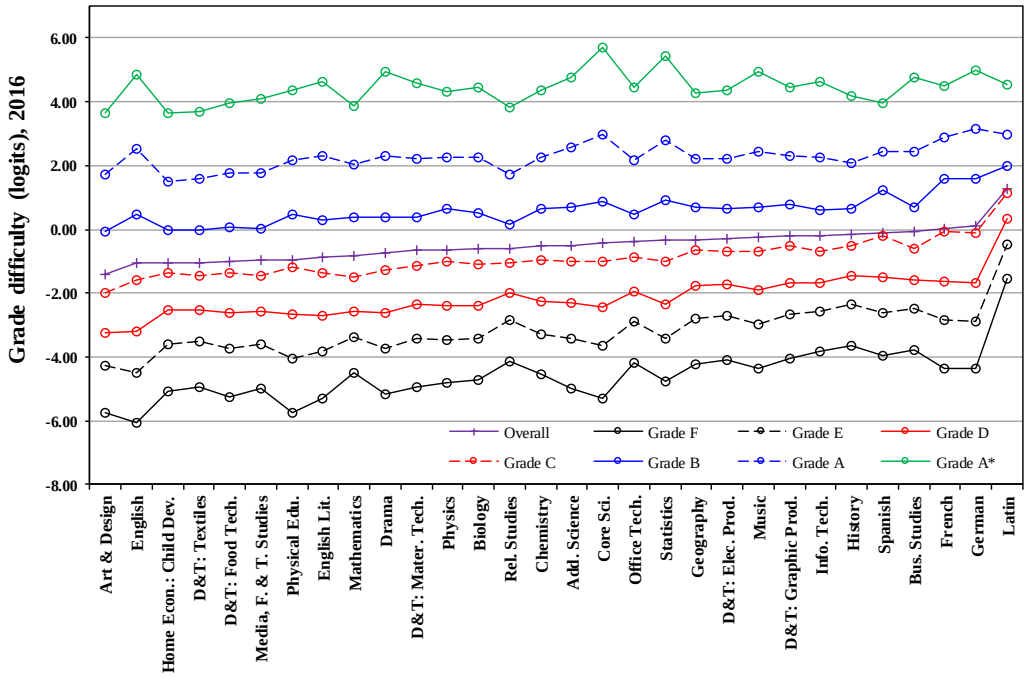

Figure 1 Distribution of overall subject difficulty and difficulties at individual grades for 30 GCSE subjects in 2016 estimated using the Rasch model.

Using the procedure detailed in He et al. (2018), which involves calculation of the mean difficulty (in logits) of all subjects at a specific grade and the difference between the grade difficulty and the mean difficulty at the same grade for each individual subject, the relative difficulties at individual grade level and the overall

subject level, in grade width (GW), for each of the 30 subjects, were calculated and are listed in Table 1. Positive values indicate that the subject is statistically harder than the average of all subjects, while negative values indicate that the subject is easier than the average. Unsurprisingly, there is not a uniform picture for all three subjects across all grades. At the overall subject level, French and Spanish are about a third of a grade more difficult than the mean of all subjects while German is about two fifths of a grade harder. Latin is the most difficult subject, while art is the easiest subject. At A*, French is close to the mean difficulty of all subjects, Spanish is about a fifth of a grade easier, but German is about a quarter of a grade harder. At grade B, Spanish is about a third of a grade more difficult than the mean of all subjects, while French and German are about three fifths of a grade more difficult. At C, they are about half of a grade more difficult than the average.

Table 1 Relative grade difficulties (relative to the mean of all subjects analysed) expressed in grade width for 30 GCSE subjects from 2016, estimated using the Rasch model.

Subject	Relative difficulty (grade width)							
	Overall	**A***	**A**	**B**	**C**	**D**	**E**	**F**
Art & Design	-0.68	-0.37	-0.25	-0.42	-0.72	-0.91	-1.06	-0.83
English	-0.42	0.20	0.11	-0.08	-0.44	-0.88	-1.26	-1.07
Home Econ.: Child Dev.	-0.42	-0.37	-0.35	-0.39	-0.30	-0.33	-0.42	-0.36
D&T: Textiles	-0.39	-0.35	-0.32	-0.39	-0.34	-0.33	-0.34	-0.27
D&T: Food Tech.	-0.39	-0.21	-0.24	-0.32	-0.31	-0.40	-0.55	-0.48
Media, F. & T. Studies	-0.34	-0.17	-0.23	-0.36	-0.36	-0.37	-0.39	-0.28
Physical Education	-0.33	-0.04	-0.06	-0.09	-0.17	-0.43	-0.84	-0.84
English Literature	-0.27	0.08	0.00	-0.18	-0.31	-0.48	-0.64	-0.51
Mathematics	-0.23	-0.26	-0.12	-0.14	-0.38	-0.37	-0.18	0.06
Drama	-0.17	0.23	0.02	-0.13	-0.23	-0.38	-0.52	-0.41
D&T: Mater. Tech.	-0.13	0.06	-0.03	-0.13	-0.16	-0.19	-0.25	-0.24
Physics	-0.10	-0.06	0.00	0.01	-0.07	-0.22	-0.27	-0.15
Biology	-0.10	0.01	-0.01	-0.05	-0.12	-0.22	-0.22	-0.09
Rel. Studies	-0.08	-0.28	-0.25	-0.28	-0.09	0.12	0.32	0.33
Chemistry	-0.03	-0.03	0.00	0.02	-0.02	-0.11	-0.09	0.01
Add. Science	-0.02	0.15	0.14	0.06	-0.06	-0.14	-0.24	-0.29
Core Sci.	0.06	0.59	0.32	0.15	-0.06	-0.24	-0.46	-0.52
Office Tech.	0.07	0.01	-0.04	-0.09	0.01	0.16	0.29	0.27
Statistics	0.11	0.46	0.24	0.18	-0.06	-0.19	-0.22	-0.14
Geography	0.13	-0.07	-0.04	0.05	0.18	0.31	0.38	0.26
D&T: Elec. Prod.	0.15	-0.04	-0.03	0.01	0.14	0.32	0.45	0.35
Music	0.17	0.23	0.08	0.05	0.14	0.21	0.21	0.16
D&T: Graphic Prod.	0.22	0.00	0.02	0.10	0.25	0.39	0.49	0.37
Info. Tech.	0.23	0.08	-0.01	-0.01	0.13	0.37	0.60	0.54
History	0.26	-0.11	-0.08	0.01	0.26	0.56	0.81	0.66
Spanish	**0.29**	**-0.21**	**0.07**	**0.36**	**0.47**	**0.50**	**0.56**	**0.43**
Bus. Studies	0.31	0.16	0.07	0.05	0.21	0.44	0.66	0.57
French	**0.36**	**0.02**	**0.28**	**0.58**	**0.56**	**0.40**	**0.30**	**0.15**
German	**0.44**	**0.27**	**0.40**	**0.59**	**0.53**	**0.39**	**0.29**	**0.16**
Latin	1.30	0.04	0.31	0.83	1.35	2.01	2.61	2.16

2.2.2 Subject difficulty derived based on relationship with prior attainment

Figure 2 shows the relationships between average subject GCSE grade and average attainment at Key Stage 3 (KS3), represented by the mean of National Curriculum (NC) levels attained for English, maths and science at KS3, and Key Stage 2 (KS2), represented by the mean of NC levels attained for English, maths and science at KS2, for 40 GCSE subjects taken by candidates in 2010. To produce the curves, the mean KS3 level for each candidate is calculated, and the range of mean KS3 level was then divided into 20 equal intervals. For mean KS2 level, the range was narrower and was divided into 11 equal intervals. A candidate was assigned to one of the attainment intervals based on his/her mean KS3/KS2 level. For candidates in each mean KS3/KS2 attainment interval, their average grade in the GCSE subject was calculated. These curves may be called subject grade characteristic curves (GCCs). If we assume that students with similar levels of prior attainment at KS2/KS3 should perform similarly at GCSE, then the prior attainment of candidates at KS3/KS2 may be used as a basis for comparing their performances in different GCSE subjects. The difference in the relationship between average subject GCSE grade and mean KS3/KS2 level between the subjects can be assumed to reflect differences in difficulty. This is because candidates with the same level of attainment at KS3/KS2 achieved different average grades at GCSE in different subjects, or adifferent level of KC3/KS2 attainment was required to achieve the same average subject grade in different subjects. This represents a reference test approach (or value added approach) to inter-subject comparability in which attainment at KS3/KS2 is used as a proxy of performance on a reference test. The top curves in the graphs are for subjects which are easier while the bottom curves are for subjects which are harder, since for similar level of KS3/KS2 attainment, candidates taking the subjects in the upper curves achieved higher GCSE grades than those taking the bottom subjects.

For all of the 40 subjects, an all-subjects average curve can be derived. If the GCC of a subject is above the all-subjects average curve, students performed better on this subject than the average performance of all candidates over all subjects (i.e. with positive value added) and the subject can be said to be easier than the mean of all subjects. If, on the other hand, the GCC of a subject is below the all-subject average curve, students performed worse on this subject than the average performance of all candidates over all subjects (i.e. with negative value added) and the subject can be said to be more difficult. As is clear from Figure 2, for French, German and Spanish, their GCCs in both graphs are above the all-subject average curve in the lower ability range but below in the middle and higher ability range, suggesting differentiated subject difficulty.

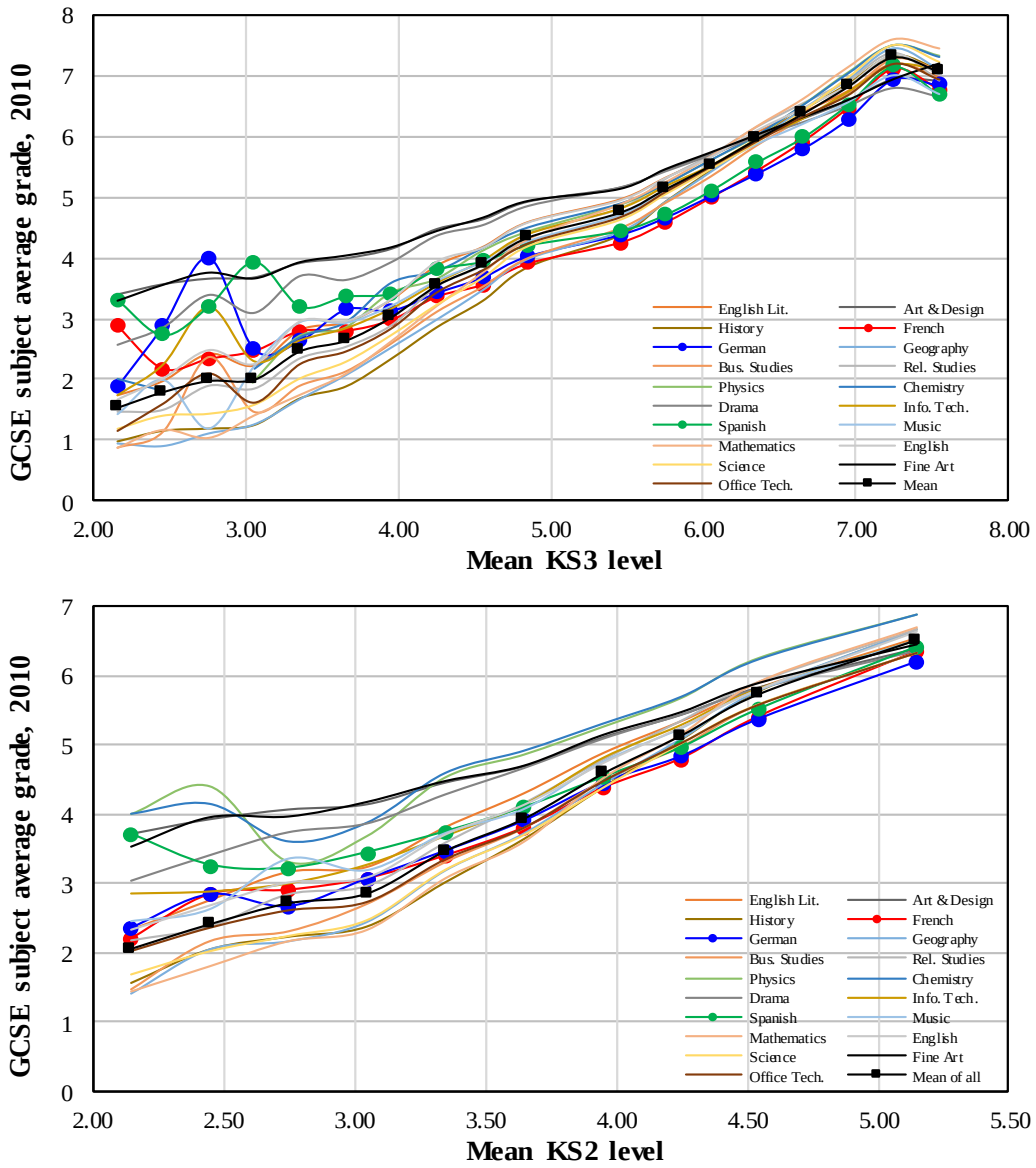

Figure 2 Relationship between average subject GCSE grade and mean KS3 level (top) and mean KS2 level (bottom) for 40 GCSE subjects from 2010.

To look at how students with different levels of prior attainment at KS2 or KS3 performed in different GCSE subjects in more detail, Figure 3 depicts the grade distributions for students with a mean KS2 NC level of 4 and 5 and a mean KS3 NC level of 4 and 7 in GCSE French, German, Spanish, geography and history from 2010. As is clear from the figure, students with a lower level of prior attainment at KS2 or KS3 performed better in French, German and Spanish than in geography and history. In contrast, students with higher level of prior attainment at KS2 or KS3 performed better in geography and history than in French, German and Spanish.

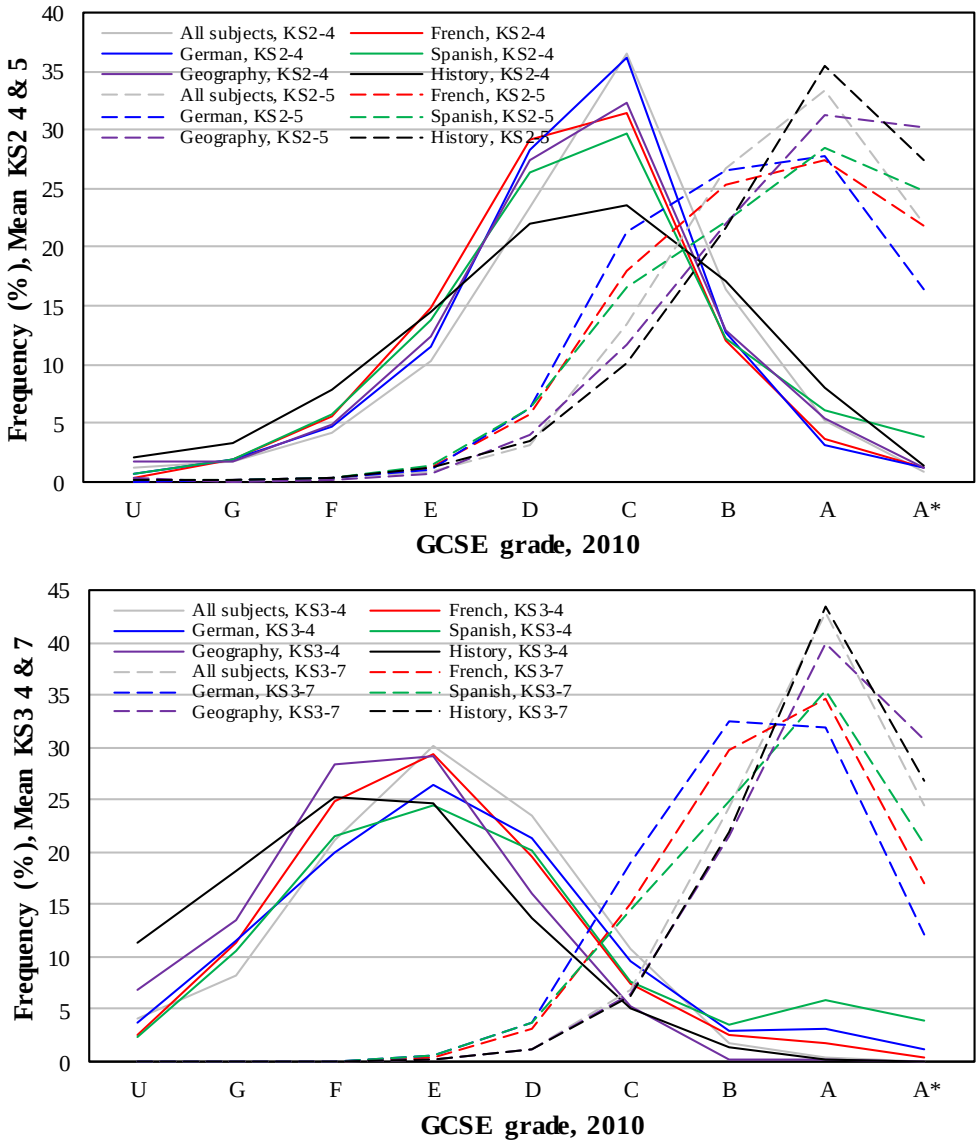

Figure 3 GCSE grade distributions for students with mean KS2 level of 4 and 5 and mean KS3 level of 4 and 7 in GCSE French, German, Spanish, geography and history from 2010.

It is possible to derive a measure of subject difficulty using the data contained in Figure 2. For a specific subject, the average subject GCSE grade in each KS3/KS2 attainment interval on the all-subjects average GCC is multiplied by the proportion of candidates taking this subject that fall into the interval and summed up over the full range of attainment. This average subject grade would be the achievable average grade by the candidates if they were to take the all-subjects average subject. The difference between the achievable average grade and the actual average subject grade is the relative difficulty of the subject in GCSE grade width (relative to the mean of all subjects). The relative difficulties thus calculated for 10 of the GCSE subjects are listed in Table 2. Although the pattern of relative difficulty between the subjects derived using KS3 and KS2 attainment is similar, the magnitude is different. Based on attainment at KS3, French and German were over two fifths of a grade

harder than the average of all subjects, and Spanish was about a quarter of a grade more difficult. Based on attainment at KS2, French and German were about a quarter of a grade harder than the average of all subjects, and Spanish was only about a tenth of a grade harder.

Table 2 Relative subject difficulties (relative to the mean of all subjects) expressed in grade width for a selection of GCSE subjects from 2010, estimated using the simple value added approach with prior attainment at KS3 and KS2.

Subject	KS3 based Difficulty (GW)	KS2 based Difficulty (GW)
D&T: Textiles	-0.52	-0.47
Drama	-0.32	-0.31
English	-0.20	-0.13
Biology	-0.17	-0.47
Mathematics	-0.03	0.05
Geography	0.16	0.03
History	0.18	0.06
Spanish	0.26	0.10
German	0.44	0.23
French	0.42	0.27

As will be discussed later, the difference in performance between GCSE subjects for students with similar levels of prior attainment at KS2/KS3 or difference in statistical subject difficulty does not necessarily suggest severity or leniency in grading standards for any particular GCSE subject since grade standards reflect the required level of attainment achieved in specific GCSE subjects. Difference in subject difficulty can be caused by a range of factors such as differences in subject demand, allocation of resources, motivation of students, effectiveness of teaching and learning, and others. It would therefore be unrealistic to require that students with similar levels of prior attainment perform similarly in different GCSE subjects or make similar rates of progression from KS2 or KS3 to GCSE.

2.3 Consistency of subject difficulty estimated using different methods

Coe et al. (2008) used different statistical methods to estimate the difficulty of GCSE subjects from 2006 and found that although the magnitude of subject difficulty is slightly different for different methods, the patterns of difficulty distribution between the subjects are similar across the methods, with French, German and Spanish shown to be consistently more difficult than many of the other subjects. Table 3 shows the relative subject difficulties (in grade width) for French, German and Spanish in 2006 estimated using six different methods. French, German and Spanish are about two fifths of a grade more difficult than the mean of all subjects based on the Rasch model and slightly less than a third of a grade harder based on the reference test approach. Geography and history were also listed in the table to provide some context of the magnitude and variability in difficulty estimates in other subjects between the different methods. French, German and Spanish were more

difficult than history which in turn was harder than geography. Geography was slightly harder than the mean of all subjects.

Table 3 Relative subject difficulties (relative to the mean of all subjects) expressed in grade width for GCSE French, German, Spanish, geography and history from 2006, estimated using different statistical methods.

Subject	Relative difficulty (grade width)					
	Rasch	SPA (unweighted)	SPA (weighted)	Kelly	Reference test	Value-added
French	0.38	0.35	0.37	0.41	0.29	0.34
German	0.43	0.37	0.38	0.42	0.26	0.32
Spanish	0.41	0.34	0.34	0.36	0.22	0.33
Geography	0.17	0.04	0.07	0.12	0.07	0.00
History	0.36	0.10	0.14	0.20	0.18	0.08

Figure 4 below shows the relationship between subject relative difficulties derived using the Rasch model and those using measures of prior attainment at KS3 and KS2 for the 40 GCSE subjects in 2010. The patterns of difficulties based on the Rasch model and those based on prior attainment are broadly consistent. The correlations between difficulties based on the Rasch model and attainment at KS3 and between the difficulties based on attainment at KS3 and attainment at KS2 are high, with R_2=0.88 and 0.72 respectively. However, the correlation between Rasch difficulties and KS2 attainment based difficulties are considerably lower, with R_2=0.53. It is to be noted that KS2 tests were taken 5 years before GCSE, while KS3 tests two years before GCSE. It would therefore be expected that attainment at KS3 would have a much higher power in predicting performance at GCSE than attainment at KS2.

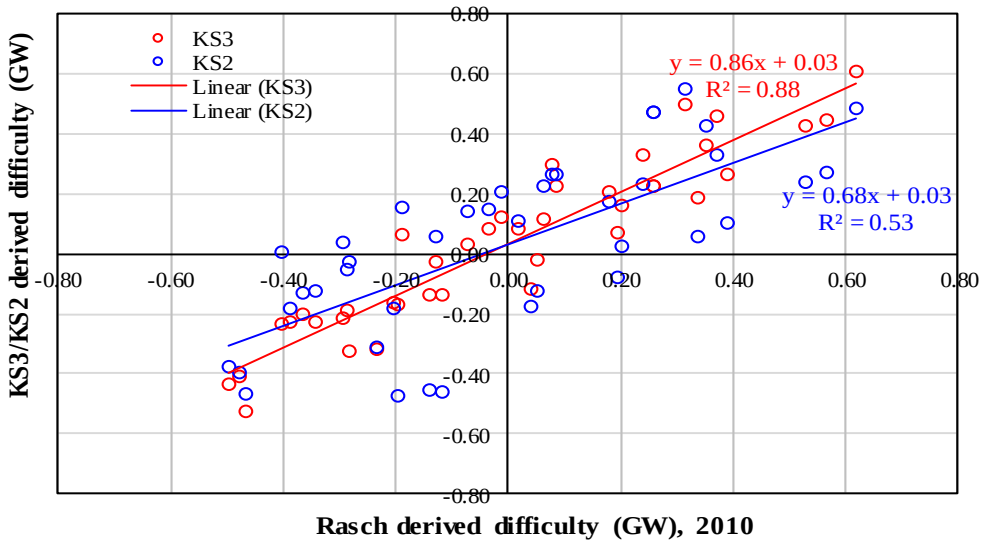

Figure 4 Relationship between subject relative difficulties derived using the Rasch model and prior attainment at KS3 and KS2 for 40 GCSE subjects in 2010.

2.4 Consistency of relative subject difficulty over time

2.4.1 Analysis based on Rasch modelling

To investigate the variability of relative subject difficulty over time, GCSE outcome data from 2006 to 2016 extracted from the NPD were analysed using the Rasch model. (As both letter grades and number grades were used in 2017, the 2017 NPD data were not analysed. Data from 2002 to 2005 were also not analysed since only a limited number of GCSE subjects were included in the data in these years). Figure 5 shows the distribution of the overall subject difficulty in grade units for a selection of years. The overall subject difficulty of a subject in grade units in a particular year was derived by dividing the difference between the subject difficulty (in logits) and the average difficulty (in logits) of all subjects by the all-subject average grade gap (in logits) between grade A and grade F. The patterns of difficulty distribution are similar over the years of analysis, with French, German and Spanish shown to be consistently among the most difficult subjects.

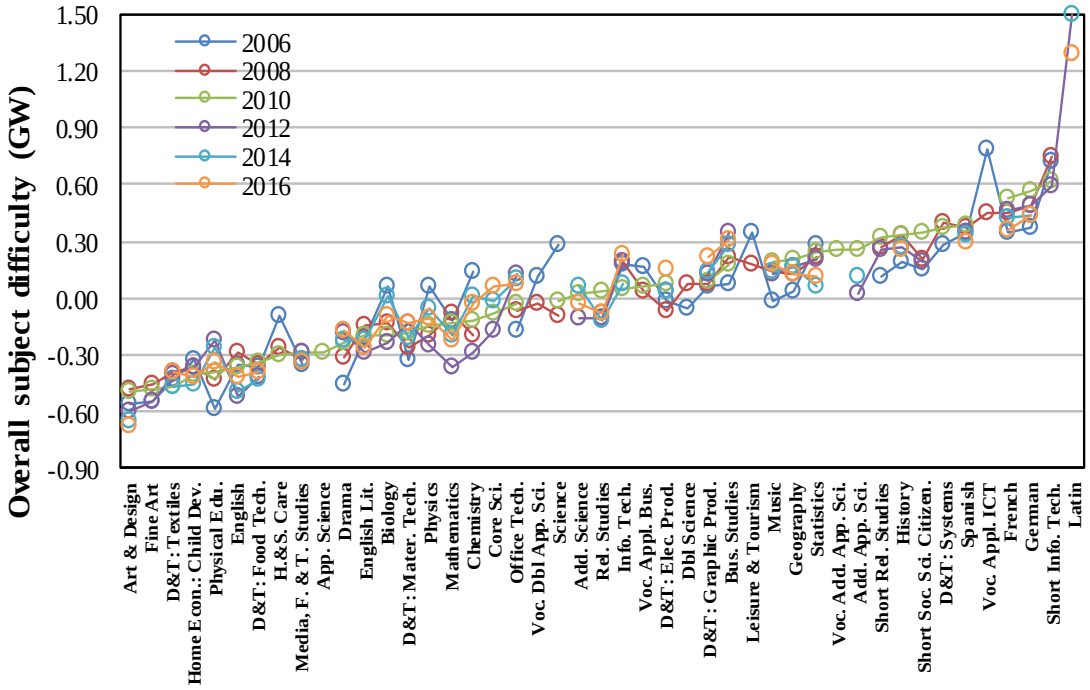

Figure 5 Distribution of the overall subject difficulty (in grade width) GCSE subjects from 2006, 2008, 2010, 2012, 2014 and 2016 based on NPD data and using the Rasch model (subjects are ordered by overall difficulty in 2010).

Figure 6 further depicts in detail the overall relative difficulties, relative to the mean difficulty of all subjects, and the difficulties at grades A*, A and C in grade width (GW) for GCSE French, German and Spanish from 2006 to 2016 estimated using the Rasch model. Again geography and history were included in the graphs to

provide some context of the magnitude and variability of relative difficulty over time in other subjects. It is stressed that variability in relative grade difficulty of the same subject over time shown in Figure 5 does not imply variation of grading standards of the subject as the data for each year were analysed separately using the Rasch model. The relative difficulties at the overall subject level and individual grade level for French, German and Spanish were generally higher than the average difficulties of all subjects and the difficulties of geography and history over the years.

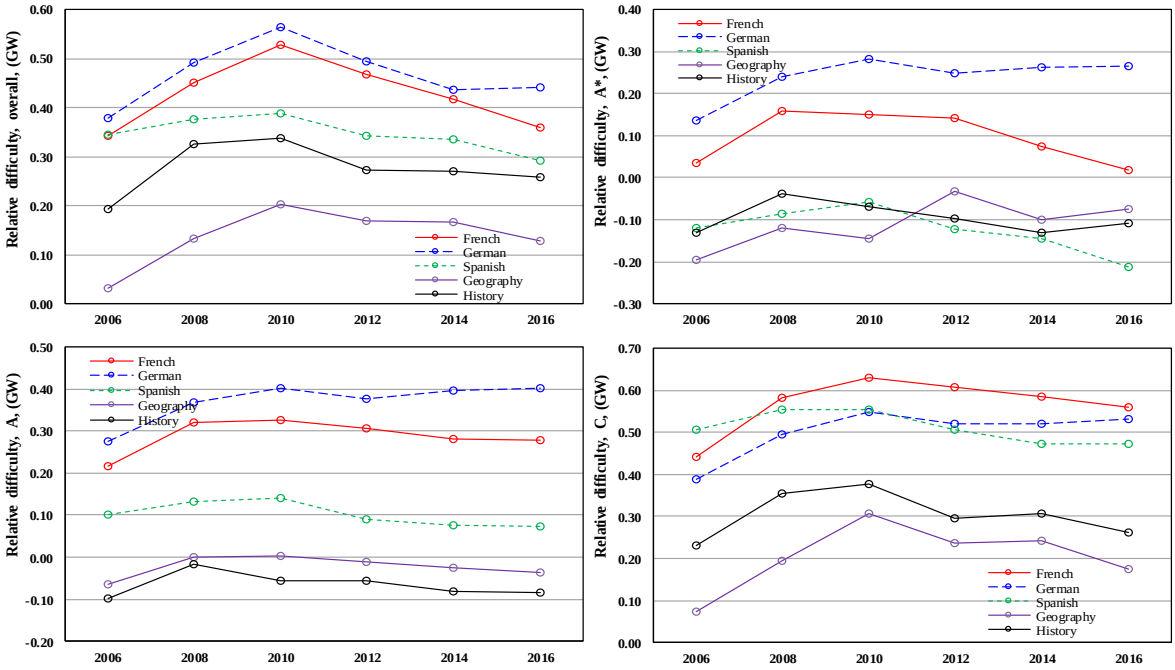

Figure 6 Distribution of overall subject difficulty and difficulties at A*, A and C for GCSE French, German, Spanish, geography and history from 2002 to 2016 based on NPD data and using the Rasch model.

At the overall subject level, German was slightly harder than French, and Spanish was easier than both French and German over the years. The patterns of variation over time of subject difficulty for French, German and Spanish are similar: the difficulty increased slightly from 2006, peaked at 2010 and decreased gradually from 2010 to 2016, with Spanish showing the least variability. By 2016, the overall relative subject difficulty is similar to that in 2006 for French, slightly higher for German but lower for Spanish.

At A*, both French and German were harder than the average of all subjects but Spanish was easier. The difficulty of German went up slightly from a tenth of a grade higher than the average difficulty of all subjects in 2006 to nearly a third of a grade higher in 2010, decreased slightly from 2010 to 2012, and remained about a quarter of a grade higher than the average from 2012 to 2016. For French, the difficulty in 2006 was close to the average of all subjects, went up slightly to about a sixth of a grade higher in 2008, and decreased gradually from 2008 to the average difficulty in

2016. The difficulty of Spanish was about a tenth of a grade lower than the average in 2006, went up very slightly in 2010, and decreased gradually from 2010 to nearly a fifth of a grade lower than the average in 2016.

A grade A, German was harder than both French and Spanish. It was also about a quarter of a grade harder than the average of all subjects in 2006. Its difficulty increased to about two fifths of a grade higher than the average difficulty in 2010 and remained at that level since. French was about a fifth of a grade harder than the average in 2006. Its difficulty increased to nearly a third of a grade higher than the average in 2010 but decreased gradually from 2010 to slightly over a quarter of a grade higher than the average in 2016. Spanish was about a tenth of a grade harder than the average in 2006. Its difficulty went up slightly to over a tenth of a grade harder than the average in 2010 and decreased to below a tenth of a grade harder than the average in 2016.

At grade C, French, German and Spanish were statistically harder than the average of all subjects. French was slightly over two fifths of a grade harder than the average of all subjects in 2006. Its difficulty went up to slightly over three fifths of a grade above the average in 2010 and decreased gradually from 2010 to over half of a grade above the average in 2016. German was nearly two fifths of a grade harder than the average in 2006. Its difficulty went up to slightly over half of a grade above the average in 2010 and remained broadly at this level since. Spanish was about half of a grade harder than the average in 2006. Its difficulty increased slightly in 2010 and decreased gradually from 2010 to below half of a grade above the average in 2016.

The reformed GCSE qualifications which are graded using number grades (9 to 1) have been introduced in phases, with English language, English literature and mathematics awarded to students in 2017. A substantial proportion of the subjects were still graded using letter grades (A* to G) in 2018. In 2019, only a few subjects with small entries were graded using letter grades, and the vast majority of the subjects were graded using the new number grading system. This makes it possible to examine the relative subject difficulty of the reformed GCSE French, German and Spanish (first awarded in 2018) at individual grades in comparison with other GCSE subjects in 2019 using the Rasch model. Figure 7 shows the distributions of difficulties derived from the Rasch model at the overall subject level and individual grade level for 29 reformed GCSE subjects (based on the summer awarding data provided to Ofqual by the awarding organisations). In the analysis, U was treated as missing and grade 1 the lowest score category. French, German and Spanish again were found to be among the most difficult subjects at individual grade level and the overall subject level. It is noticed that the variability in the gap between two adjacent grades estimated for the 2016 NPD data (using letter grades) is different from that estimated for the 2019 awarding data (using number grades). For the 2016 data, the gap between A* and A is the largest and the gap between two adjacent grades

decreases from A* to E, with the gap between D and E the smallest. The gap between E and F is larger than that between D and E. In contrast, for the 2019 awarding data, the gap between grade 2 and grade 3 is the largest. The gap between two adjacent grades decreases from 2 to 6, with that between 5 and 6 the smallest. The gap increases slightly from grade 6 to grade 8. The gap between grade 8 and grade 9 is similar to that between grade 3 and grade 4. The gap between two adjacent grades from grade 4 and grade 8 is considerably narrower than that for the other grades. This difference in grade gap between the letter grading system and the number grading system reflects the design feature of the latter. Table 4 shows the relative difficulties at individual grades and the overall subject level (in grade width – GW) for each of the 29 subjects analysed. At the overall subject level, German is about three fifths of a grade more difficult than the average, French is over two fifths of a grade harder, and Spanish is about three tenths of a grade more difficult. At grade 9, Spanish is very slightly above the mean of all subjects, French is over two fifths of a grade more difficult, and German is about three fifths of a grade harder. At grade 7, Spanish is nearly a fifth of a grade more difficult, French is about two fifths of grade harder, and German is about two thirds of a grade more difficult. At grade 4, Spanish is slightly over half of a grade more difficult than the average, and French and German are about three fifths of a grade harder. The increase in the magnitude of relative difficulty in GW for some grades in Spanish, French and German in 2019 compared with that in 2016 reflects the narrower grade gap associated with the new number grading system.

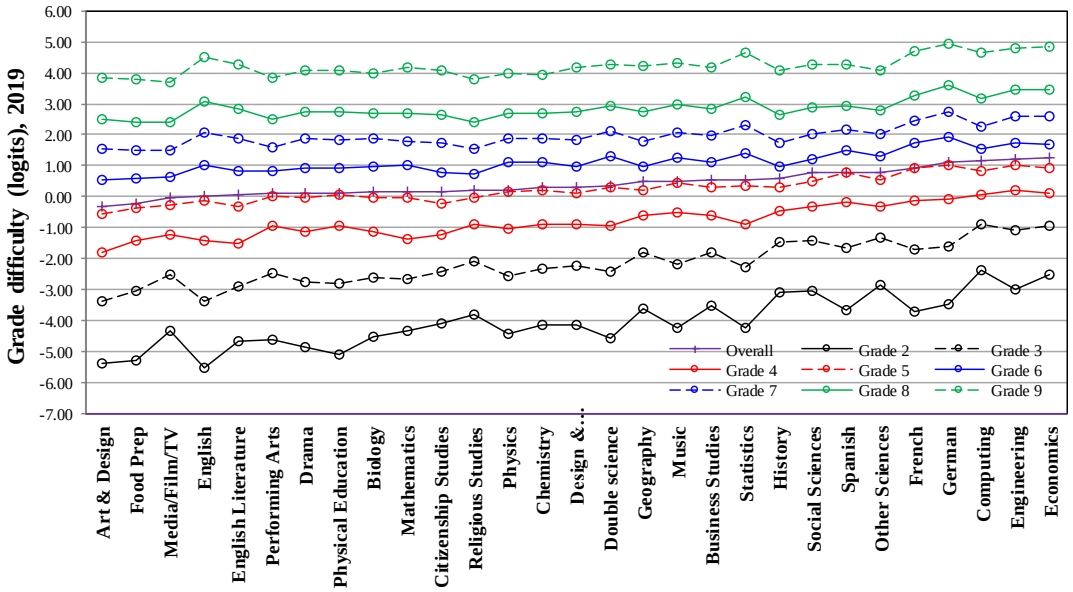

Figure 7 Distribution of overall subject difficulty and difficulties at individual grades for 29 reformed GCSE subjects from 2019 using the Rasch model.

Table 4 Relative grade difficulties (relative to the mean of all subjects analysed) expressed in grade width for 29 GCSE subjects from 2019, estimated using the Rasch model.

Subject	Relative difficulty (grade width)								
	Overall	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**
Art & Design	-0.67	-0.34	-0.32	-0.37	-0.51	-0.68	-0.88	-1.06	-1.16
Food Prep	-0.57	-0.36	-0.38	-0.42	-0.48	-0.52	-0.55	-0.76	-1.07
Media/Film/TV	-0.39	-0.46	-0.40	-0.39	-0.44	-0.46	-0.41	-0.31	-0.26
English	-0.35	0.26	0.18	0.08	-0.11	-0.32	-0.55	-1.06	-1.30
English Literature	-0.33	0.05	0.01	-0.10	-0.26	-0.47	-0.64	-0.64	-0.56
Performing Arts	-0.30	-0.32	-0.31	-0.31	-0.28	-0.20	-0.16	-0.28	-0.50
Drama	-0.28	-0.13	-0.09	-0.10	-0.18	-0.24	-0.30	-0.53	-0.69
Physical Education	-0.28	-0.11	-0.10	-0.11	-0.15	-0.14	-0.15	-0.54	-0.91
Biology	-0.25	-0.23	-0.14	-0.10	-0.14	-0.24	-0.33	-0.37	-0.44
Mathematics	-0.23	-0.05	-0.15	-0.15	-0.11	-0.24	-0.52	-0.42	-0.26
Citizenship Studies	-0.23	-0.12	-0.16	-0.22	-0.31	-0.39	-0.38	-0.21	-0.03
Religious Studies	-0.20	-0.39	-0.39	-0.39	-0.35	-0.24	-0.08	0.08	0.18
Physics	-0.17	-0.20	-0.13	-0.06	-0.02	-0.06	-0.21	-0.37	-0.34
Chemistry	-0.11	-0.24	-0.14	-0.06	-0.01	-0.03	-0.12	-0.16	-0.10
Design & Technology	-0.10	-0.06	-0.09	-0.13	-0.15	-0.13	-0.09	-0.06	-0.08
Double science	-0.05	0.03	0.08	0.14	0.15	0.04	-0.13	-0.23	-0.45
Geography	0.06	0.01	-0.09	-0.15	-0.12	-0.01	0.14	0.30	0.37
Music	0.08	0.10	0.10	0.09	0.11	0.18	0.22	0.00	-0.19
Business Studies	0.11	-0.05	0.00	0.00	-0.02	0.04	0.16	0.31	0.44
Statistics	0.11	0.36	0.30	0.28	0.25	0.09	-0.09	-0.12	-0.16
History	0.14	-0.15	-0.19	-0.20	-0.12	0.05	0.28	0.60	0.84
Social Sciences	0.29	0.03	0.01	0.03	0.09	0.20	0.39	0.67	0.87
Spanish	**0.30**	**0.03**	**0.08**	**0.18**	**0.34**	**0.48**	**0.53**	**0.45**	**0.31**
Other Sciences	0.30	-0.13	-0.05	0.04	0.15	0.27	0.39	0.71	1.04
French	**0.45**	**0.43**	**0.38**	**0.41**	**0.53**	**0.59**	**0.58**	**0.41**	**0.27**
German	**0.61**	**0.61**	**0.64**	**0.67**	**0.71**	**0.69**	**0.60**	**0.49**	**0.49**
Computing	0.63	0.37	0.26	0.26	0.36	0.50	0.72	1.09	1.43
Engineering	0.69	0.51	0.52	0.54	0.55	0.67	0.84	0.95	0.92
Economics	0.73	0.53	0.55	0.53	0.51	0.59	0.77	1.06	1.32

2.4.2 Analysis based on the profile of prior attainment at KS2

Analysis of the 2010 NPD data discussed previously indicated that GCSE French, German and Spanish were more difficult statistically than many other subjects on the assumption that students with similar level of prior attainment at KS2 or KS3 should perform similarly in different GCSE subjects. This section further examines how the relationship between performance at GCSE and attainment at KS2 varies between subjects in more recent time through analysis of the 2016 and 2018 GCSE performance data extracted from the NPD. (Note that attainment data at KS3 are no longer available after 2010). To represent the level of attainment at KS2 more accurately, students' scores in the KS2 mathematics and English reading tests were used to derive a measure of attainment (the science test became a national sampling test from 2010 and the English writing test was assessed internally from 2012). For each of the two years, the KS2 maths test scores and English reading test scores for all GCSE students were converted to a normal distribution with a

mean of 0 and a standard deviation of 1 separately and the mean of the two normalised scores for each student calculated. Students from a specific year were then grouped into 10 attainment bands (or deciles) based on their mean scores, with 1 representing the lowest level of attainment while 10 the highest level of attainment. Each attainment band contains a similar number of students. Figure 8 shows the average GCSE grade distributions for 30 subjects for students who were classified into different attainment bands at KS2 in 2016. Students classified into higher KS2 deciles achieved higher levels of average performance at GCSE. For each subject, the average grade distributions with respect to KS2 attainment levels shown in Figure 8 can be used in conjunction with its KS2 attainment profile to produce an expected grade distribution which can be compared with its observed grade distribution to derive a difficulty measure.

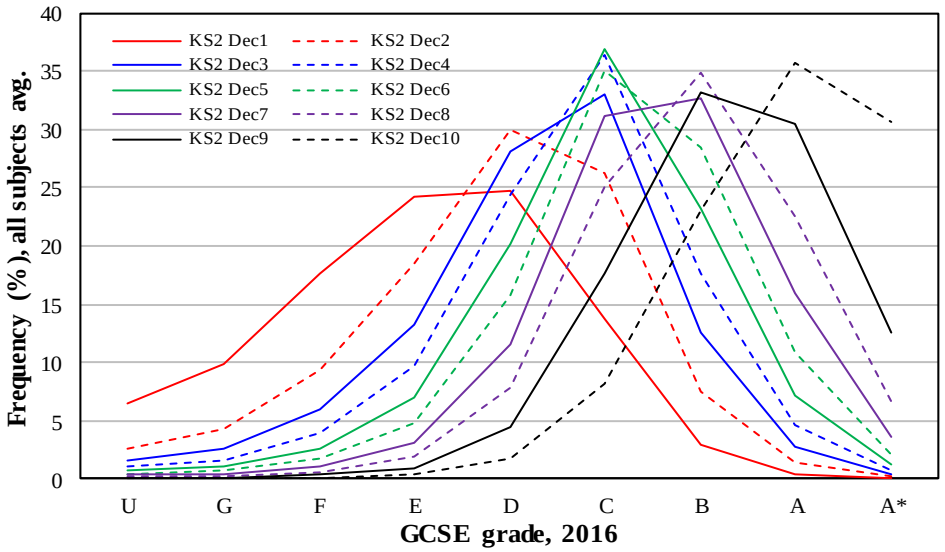

Figure 8 Average GCSE grade distributions of 30 subjects for students classified into different prior attainment bands at KS2 (KS2 deciles) in 2016.

Figure 9 further compares the grade distributions of GCSE French, German, Spanish, geography and history in 2016 for students classified into KS2 attainment deciles 3, 6 and 9 respectively. Similar to the grade distributions seen in Figure 3 for 2010, students with middle to high levels of attainment at KS2 performed noticeably better in geography and history than in French, German and Spanish.

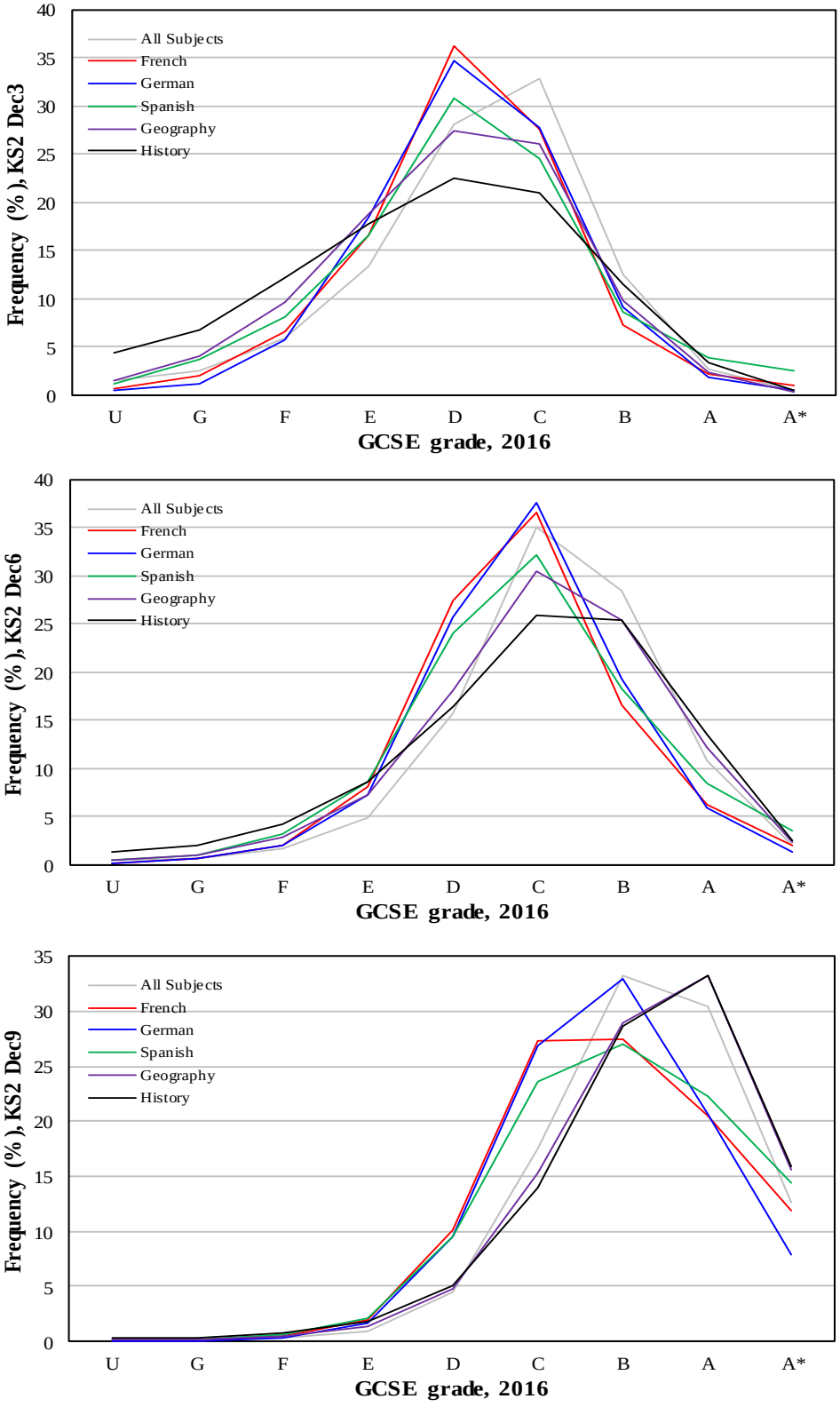

Figure 9 GCSE grade distributions in French, German, Spanish, geography and history for students classified into KS2 attainment deciles 3 (top), 6 (middle) and 9 (bottom) in 2016.

Figure 10 below shows the distributions of mean GCSE grade against KS2 attainment for 10 GCSE subjects and the average of all subjects in 2016 and the observed grade distributions for GCSE French, German, Spanish, geography and history and their expected grade distributions on the all-subjects average subject based on their KS2 attainment profiles. The outcomes of GCSE French, German and Spanish were considerably lower than the expected outcomes based on their KS2 profiles. Similar to Figure 2, performances of students with middle to high level of attainment at KS2 in French, German and Spanish were lower than the average performance of all subjects and the performances in geography and history.

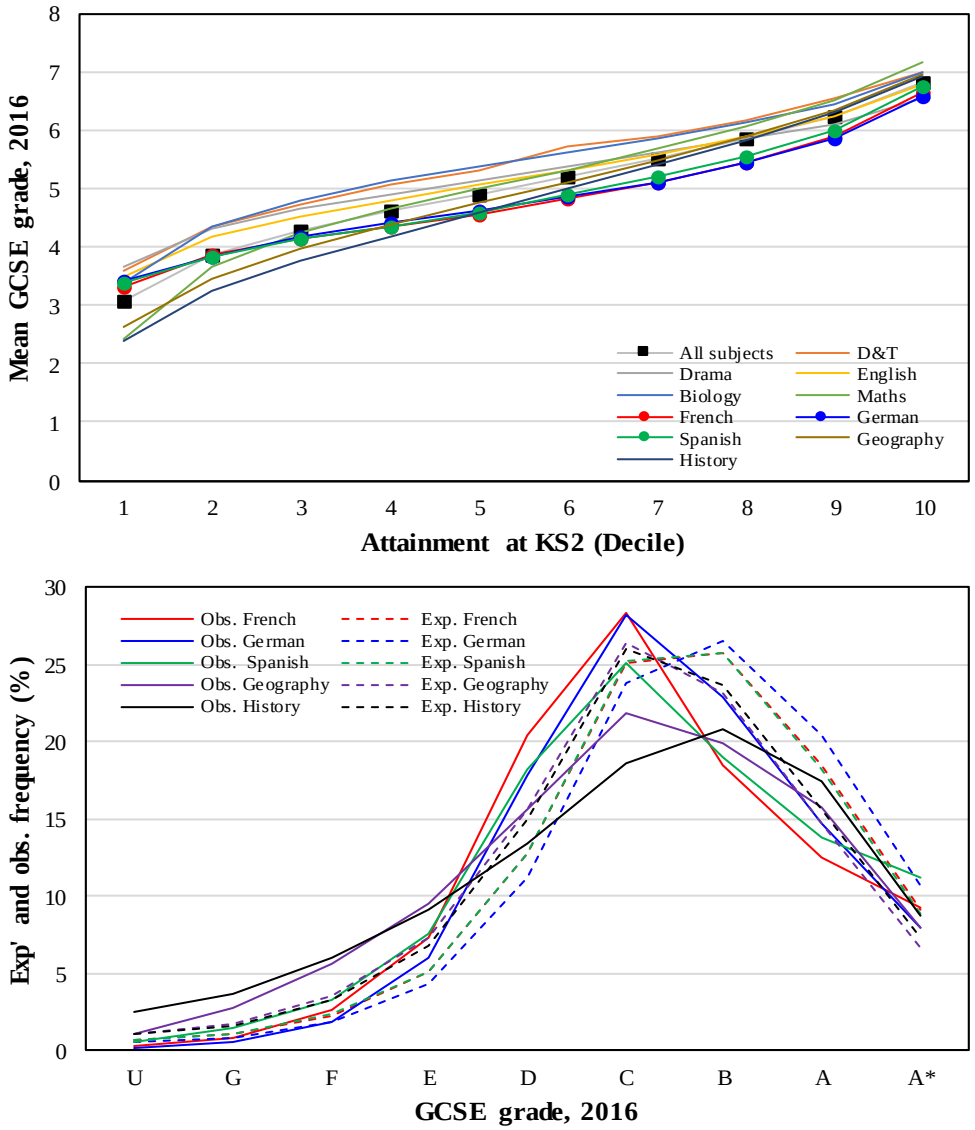

Figure 10 Distributions of mean GCSE grade against KS2 attainment for 10 GCSE subjects and the average of all subjects in 2016 (top) and the observed grade distributions for GCSE French, German, Spanish, geography and history and their expected grade distributions on the all-subjects average subject (bottom).

Using the procedure described above, an overall relative subject difficulty can be derived (relative to the mean of all subjects), and Table 5 shows the values of subject difficulty for the 10 GCSE subjects shown in Figure 10. The pattern of the distribution of relative difficulty between the subjects is similar to that shown in Table 2. However, it appears that the difference in difficulty between GCSE MFLs and geography and history in 2016 is slightly smaller than that in 2010.

Table 5 Relative subject difficulties (relative to the mean of all subjects) expressed in grade width for a selection of GCSE subjects from 2016, estimated using the simple value added approach with prior attainment at KS2.

Subject	Difficulty (GW)
D&T	-0.40
Biology	-0.28
Drama	-0.16
English	-0.14
Maths	-0.04
Geog.	0.12
History	0.22
Spanish	0.22
French	0.29
German	0.30

Both letter and number grades were used in the award of GCSE subjects in 2018, which makes it difficult to directly compare all subjects based on students' prior attainment at KS2. A selection of GCSE subjects awarded with number grades were however used to provide some indication of subject difficulty for GCSE French, German and Spanish in relation to other subjects. The top graph in Figure 11 shows the relationship between mean GCSE grade and attainment at KS2 for 10 GCSE subjects, and the bottom graph shows the grade distributions of French, German, Spanish, geography and history for students classified into KS2 attainment decile 9 (the second highest level of prior attainment). When calculating the mean GCSE grade, U was converted to 0 and the other number grades retained their numerical values. For students with middle to high level of attainment at KS2, their performances in GCSE French, German and Spanish were lower than in all the other subjects included in the analysis. This again suggests that GCSE, French, German and Spanish were more difficult than the other subjects based on attainment at KS2. For students who were classified into KS2 decile 9, the average grade in French, German, Spanish, geography and history are 5.63, 5.55, 5.63, 6.37 and 6.26 respectively. This suggests that students with prior attainment decile 9 at KS2 who studied GCSE MFLs achieved about 0.7 grade lower than those who studied geography and history.

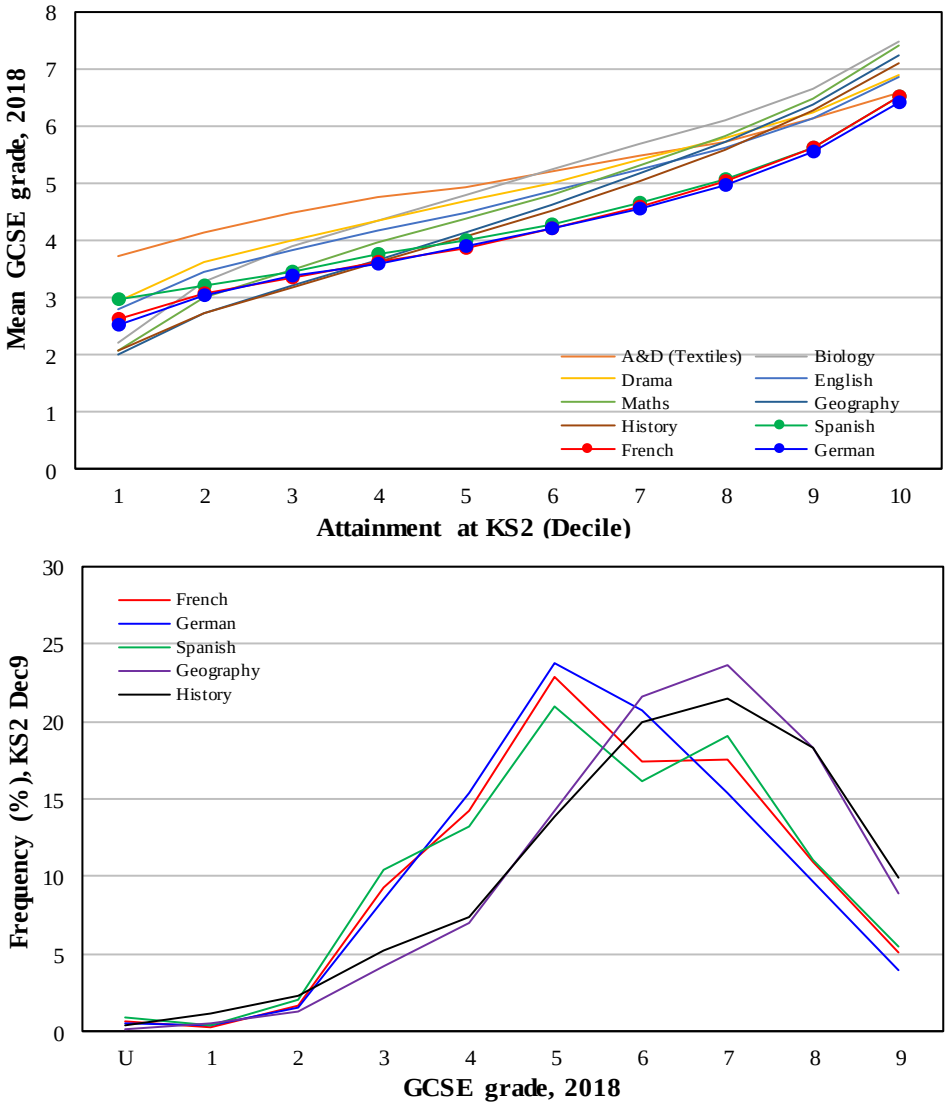

Figure 11 Distributions of mean GCSE grade against KS2 attainment for 10 GCSE subjects in 2018 (top) and the grade distributions of GCSE French, German, Spanish, geography and history for students who were classified into KS2 attainment decile 9 (bottom).

In summary, GCSE French, German and Spanish were found to be consistently more difficult over time than many of the other subjects based on prior attainment at KS2.

2.5 Interpretation of results from statistical analysis and implications

In the case of Rasch modelling and other conventional statistical methods, such as the value added approach used to study inter-subject comparability, if we accept the assumptions involved, the differences in subject difficulty derived could be interpreted as differences in standards related to some traits assumed to be shared by the examinations in the different subjects. Such differences have, however, been frequently used as evidence of inconsistency in grading standards (which represent

the level of knowledge and skills required and are generally subject-specific) between GCSE subjects. Results from most of the statistical analyses over the past decade or so suggested that GCSE MFLs, particularly French and German, are more difficult than many of the other subjects analysed. However, as is discussed below, there are limitations to the statistical techniques used to derive subject difficulty measures.

2.6 Issues with Rasch modelling and other statistical techniques

Coe et al. (2008) (see also Lockyer and Newton, 2015; He et al., 2018; Black et al., 2018b) discussed the major issues with statistical methods used for studying inter-subject comparability. These, among others, include the strong assumptions made about the data being analysed (e.g. the unidimensionality assumption of the underlying latent trait shared by the examinees and assessed by the different examinations required by the Rasch model and other similar models) which are seldom met by real data, unrepresentativeness of samples used, missing data, imperfect data-model fit, sub-group effect, and different results from different statistical models for the same dataset (see, for example, discussion in Section 2.1.2). Some of the major limitations are discussed below.

2.6.1 Unidimensionality

In the case of using the Rasch model (and similar models) to study inter-subject comparability, the difficulty of a subject is defined as the amount of the trait common to students taking different subjects that is required to achieve a specific level of performance on the exam. Even if we assume that the GCSE data analysed meet the unidimensionality requirement of the Rasch model and fit the model, the interpretation of the latent trait specified in the Rasch model is not entirely clear (also see Bramley, 2011). This is because the construct represented by the data implied from the Rasch analysis is inferred from the analysis of the set of examinations included in the analysis. This does not involve an actual measurement process in which the construct in relation to the purpose of the test to be measured must be specified and used to guide the development of the test (in this case defined by the complete range of GCSE subjects). The Rasch analysis is based on the relative frequencies of candidates receiving different grades in different examinations and it is difficult to interpret the latent trait inferred. Such a trait is likely to be influenced primarily by the subjects that are correlated well and have large entries. For example, inter-subject correlation analysis of the 2016 GCSE data from the NPD and the 2019 GCSE summer awarding data indicated that the correlation between the subjects included in the Rasch analysis varies considerably between the subjects (see Tables 6 and 7). For the 2016 NPD data, correlations varied from 0.41 between drama and physics to 0.86 between core science and additional science. For the 2019 summer awarding data, correlations varied from 0.41 between mathematics and art and design to 0.89 between biology and chemistry and between chemistry

and physics. Therefore, to a certain degree, the extent of this shared common trait measured by the exams will likely vary between the subjects in relation to the traits which the individual examinations are designed to measure. Caution must therefore be exercised when trying to link Rasch grade difficulties to the standards of examinations in different subjects if they are based on subject-specific performance levels such as GCSEs and A levels (see later discussion). With respect to the skills and knowledge assessed by the different examinations, although it is likely that some common skills will be assessed by examinations in different subjects, aspects of knowledge and understanding are generally subject specific and can vary considerably between subjects. Although the concept of “general academic ability” has been proposed to interpret the latent trait specified in the Rasch model (and other similar unidimensional statistical models), there has been limited research regarding the usefulness of the Rasch ability measures in relation to the specific uses the examinations are to be put to.

Table 6 Correlations between 20 GCSE subjects with large numbers of candidates from 2016.

	ELIT	ART	HIS	GEO	FRE	GER	BUS	RS	PE	PHY	CHE	BIO	DRA	IT	SPAN	MAT	ENG	STAT	CORESCI	ADTSCI
ELIT	1.00	0.59	0.77	0.76	0.61	0.61	0.69	0.76	0.60	0.56	0.57	0.61	0.65	0.63	0.57	0.69	0.78	0.67	0.70	0.69
ART		1.00	0.59	0.62	0.51	0.48	0.55	0.57	0.51	0.46	0.47	0.50	0.55	0.50	0.47	0.55	0.57	0.53	0.53	0.52
HIS			1.00	0.83	0.65	0.64	0.76	0.78	0.62	0.65	0.67	0.69	0.62	0.65	0.62	0.72	0.74	0.71	0.74	0.75
GEO				1.00	0.66	0.65	0.78	0.77	0.66	0.69	0.70	0.73	0.64	0.68	0.62	0.76	0.74	0.74	0.78	0.78
FRE					1.00	0.69	0.55	0.59	0.49	0.59	0.61	0.61	0.52	0.53	0.68	0.62	0.61	0.58	0.58	0.59
GER						1.00	0.57	0.59	0.50	0.59	0.60	0.60	0.50	0.56	0.65	0.63	0.61	0.59	0.59	0.61
BUS							1.00	0.71	0.60	0.64	0.64	0.67	0.57	0.66	0.54	0.72	0.67	0.67	0.73	0.73
RS								1.00	0.58	0.57	0.60	0.62	0.62	0.61	0.55	0.68	0.72	0.68	0.71	0.69
PE									1.00	0.53	0.54	0.56	0.54	0.52	0.46	0.63	0.60	0.60	0.60	0.59
PHY										1.00	0.85	0.83	0.41	0.61	0.57	0.76	0.55	0.70		
CHE											1.00	0.84	0.43	0.61	0.59	0.74	0.57	0.69		
BIO												1.00	0.47	0.61	0.59	0.73	0.61	0.69	0.81	0.80
DRA													1.00	0.52	0.47	0.57	0.63	0.55	0.58	0.55
IT														1.00	0.50	0.65	0.61	0.64	0.63	0.62
SPAN															1.00	0.56	0.54	0.56	0.50	0.53
MAT																1.00	0.70	0.82	0.79	0.78
ENG																	1.00	0.66	0.70	0.67
STAT																		1.00	0.75	0.73
CORESCI																			1.00	0.86
ADTSCI																				1.00

Table 7 Correlations between 20 GCSE subjects with large numbers of candidates from 2019.

	MAT	BUS	HIS	ENG	ELIT	SCI(DBL)	SPAN	ART	FRE	GEO	PE	RS	COMP	BIO	CHE	PHY	GER	D&T	DRA	STAT
MAT	1.00	0.76	0.74	0.72	0.69	0.84	0.55	0.56	0.63	0.81	0.73	0.68	0.83	0.82	0.83	0.84	0.64	0.72	0.63	0.88
BUS		1.00	0.80	0.73	0.72	0.76	0.52	0.54	0.59	0.85	0.72	0.74	0.75	0.73	0.69	0.69	0.60	0.71	0.67	0.79
HIS			1.00	0.79	0.80	0.75	0.59	0.59	0.66	0.87	0.72	0.82	0.73	0.73	0.70	0.69	0.65	0.74	0.72	0.76
ENG				1.00	0.83	0.69	0.53	0.60	0.61	0.80	0.67	0.77	0.68	0.66	0.60	0.59	0.61	0.71	0.73	0.70
ELIT					1.00	0.69	0.53	0.60	0.61	0.79	0.67	0.79	0.67	0.65	0.61	0.59	0.61	0.71	0.73	0.67
SCI(DBL)						1.00	0.48	0.53	0.58	0.83	0.71	0.68	0.79				0.60	0.70	0.63	0.79
SPAN							1.00	0.41	0.65	0.59	0.49	0.54	0.57	0.58	0.58	0.57	0.57	0.53	0.51	0.57
ART								1.00	0.48	0.63	0.59	0.58	0.53	0.55	0.50	0.49	0.47	0.66	0.61	0.58
FRE									1.00	0.67	0.57	0.61	0.64	0.62	0.62	0.61	0.69	0.60	0.57	0.61
GEO										1.00	0.78	0.80	0.80	0.79	0.76	0.75	0.68	0.80	0.74	0.82
PE											1.00	0.67	0.70	0.69	0.66	0.65	0.55	0.70	0.65	0.73
RS												1.00	0.66	0.65	0.63	0.60	0.61	0.72	0.71	0.71
COMP													1.00	0.78	0.77	0.79	0.65	0.72	0.64	0.80
BIO														1.00	0.89	0.87	0.61	0.69	0.56	0.79
CHE															1.00	0.89	0.62	0.67	0.53	0.79
PHY																1.00	0.61	0.65	0.52	0.80
GER																	1.00	0.59	0.56	0.57
D&T																		1.00	0.70	0.72
DRA																			1.00	0.58
STAT																				1.00

2.6.2 Missing data and representativeness of sample

Because students normally take eight or nine subjects at GCSE out of about 40 available, most statistical analyses involve large proportions of missing data which may not be assumed to be random. Using simulations, Bramley (2016) demonstrated that the existence of non-random missing data could produce biased estimates of subject difficulty using the subject pairs analysis approach. It is likely that similar bias in difficulty estimates generated using other statistical methods would also exist. This may partly reflect the fact that differences in correlation between the subjects are different (or different subjects measure different constructs) and that missing data will influence correlations. One way to partially alleviate this problem might be to analyse semi-cognate or cognate subjects together to investigate relative subject difficulty.

2.7 Factors not considered in statistical modelling

It has been demonstrated that there are considerable differences in statistical difficulty indices at individual grade level and the overall subject level between GCSE subjects derived using different statistical methods, with French, German and Spanish shown to be more difficult than most of the other subjects. Such variability in difficulty has remained almost the same over the past decade or so. Differences in difficulty indices between subjects have been interpreted as differences in grading standards by many stakeholders. However, even if we accept the underlying assumptions of the statistical models and the level of model-data fit, these differences in difficulty between subjects can be caused by many factors and cannot be simply attributed to severity/leniency in grading. Coe (2008) and Coe et al. (2008) discussed a range of potential causes for such differences in addition to grading severity/leniency (also see Newton, 2012; Lockyer and Newton, 2015; Black et al., 2018b). These, among others, include:

- Nature of the subject in terms of skills and knowledge to be learnt
- Relationship between GCSE performance and prior attainment
- Level of subject demand
- Allocation of teaching time and other resources
- Motivation of students
- Efficiency and effectiveness of teaching and learning
- Uptake by various population subgroups
- The influence of the importance attached to the subject (for example, whether a subject contributes to the EBacc measure or the Progress 8 measure)

These factors are not considered in statistical methods but they can vary substantially between subjects. They need to be taken into consideration when interpreting statistical difficulties and trying to link statistical difficulties to the performance standards of examinations.

As an example, Figure 12 shows the mean subject GCSE grade of students taking French, German, Spanish, geography and history from 2006 to 2016 and their mean KS3 level and mean KS2 level attained two and five years before respectively. Variation of mean subject grade outcomes over time for the five subjects is similar to that of mean KS3 level or mean KS2 level, with gradual increase from 2006 to 2012 and decrease from 2012 to 2014. It is noticed that, from 2006 to 2012, mean subject grade for Spanish was higher than the mean grade for German which in turn was higher than the mean for French. Mean subject grade went up slightly from 2014 to 2016 for the three subjects. The level of prior attainment at KS3 and KS2 of the students taking the different GCSE subjects also varies between the subjects, particularly at KS3 where the average KS3 level of students taking German was about 0.1 NC higher than those taking Spanish and 0.3 NC level higher than those taking geography. At KS2, the difference is about 0.15 NC level between those taking German and those taking geography. Results from the work by Coe et al. (2008) and He and Stockford (2015) suggested that the relationship between GCSE outcomes and prior attainment may not be linear and can vary between subjects. Differences in prior attainment, coupled with differences in other factors such as subject demand, motivation of students, allocation of resources and effectiveness of teaching and learning between subjects could produce different rates of progression or value added from KS3/KS2 to GCSE between the subjects, leading to apparent differences in subject difficulties that are estimated based on measures of prior attainment at KS2 explored in this report.

Table 8 further shows the correlations between GCSE subject grade and mean KS3 level and mean KS2 level for the five subjects (for KS2, the 2016 mean KS2 level was based on the NC levels attained on English and maths only). These are reasonably high for KS3 and only moderate for KS2 and also vary slightly between the subjects. Furthermore, the variances in the GCSE subject grade outcomes that can be attributed to differences in prior attainment range from about 35% for Spanish to 59% for geography based on mean KS3 level and from only about 23% for Spanish to 41% for geography based on mean KS2 level. Therefore caution needs to be taken when interpreting the difficulties derived using these prior attainment measures.

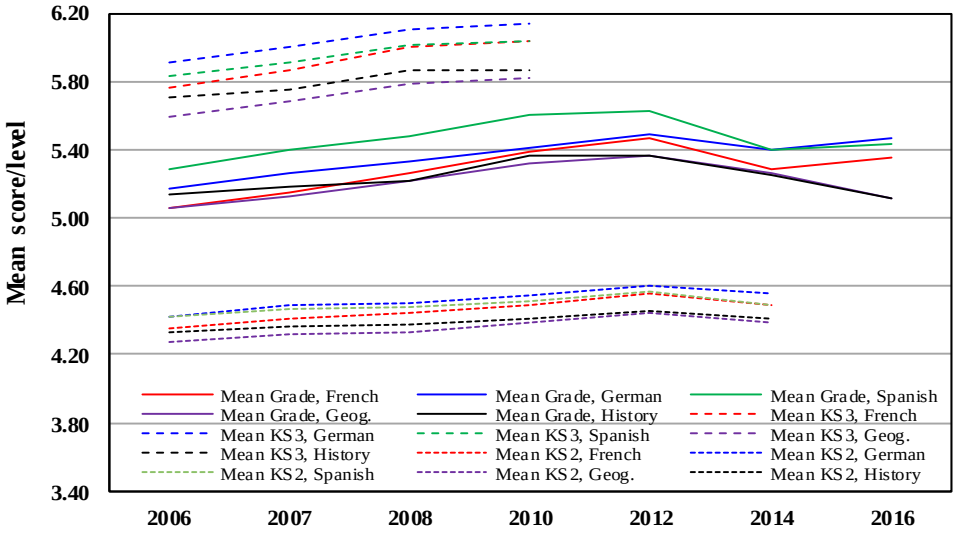

Figure 12 Distribution of mean subject grade for students taking GCSE French, German, Spanish, geography and history from 2006 to 2016 and their attainment at KS3 and KS2 represented using mean KS3 level and mean KS2 level.

Table 8 Correlations between GCSE subject grade and mean KS3 level and mean KS2 level for French, German, Spanish, geography and history over time.

	Year	FRE	GER	SPAN	GEOG	HIS
Mean KS3 level	2006	0.72	0.70	0.65	0.78	0.75
	2008	0.69	0.67	0.64	0.78	0.74
	2010	0.68	0.64	0.59	0.77	0.74
Mean KS2 level	2006	0.62	0.59	0.57	0.65	0.63
	2008	0.59	0.57	0.55	0.66	0.62
	2010	0.56	0.53	0.50	0.64	0.61
	2012	0.53	0.51	0.48	0.65	0.62
	2014	0.51	0.50	0.48	0.64	0.61
	2016	0.52	0.52	0.50	0.63	0.60

2.8 Linking statistical difficulty measures to subject grading standards

Although debate on inter-subject comparability in GCSEs and other examinations in England has been longstanding, there has been no consensus on how inter-subject comparability should be conceptualised, defined and measured (see Lockyer and Newton, 2015). Whether or not statistical difficulty indices should be linked to subject-specific grading standards of individual examinations has been the focus of the debate. Those who are against the use of statistical methods to compare different subjects argue that examinations such as GCSEs are graded based on standards that are subject-specific and that the shared knowledge and skills assessed by different examinations (the unidimensional assumption which is made implicitly or explicitly by most statistical models) is irrelevant, meaning that the

between-subject comparison is of limited meaning. Further, they argue that there are many factors that can affect performance in exams which must be considered when comparing standards in different subjects. In contrast, those who support the use of statistical approaches argue that in cases where there is a theoretical basis for the analysis and the interpretation of the results can be justified, statistical comparisons may still be appropriate and meaningful. For example, it is argued that subject standards may need to be statistically aligned when grades from different subjects are used for specific purposes, particularly when they are used as equivalent currencies in situations such as admissions to certain university courses and within school accountability measures. However, the work by Benton (2016) suggested that statistical alignment of GCSE subject grades would not substantially affect the ranking of schools.

We further examine below the appropriateness of linking statistical difficulty measures to exam grading standards from the perspective of the purposes of GCSE qualifications as currently defined. While the aim of a course leading to a qualification should be to help the learners acquire the required knowledge and skills within a specified domain of content and skills, the purpose of the assessment itself is to provide an accurate measurement of the level of attainment or proficiency that a learner has achieved at the end of the course of study in relation to the purpose of the qualification. The purposes of the qualification therefore, to a large extent, determine how examination results should be interpreted and reported, which in turn will affect the kind of comparison in standards between different qualifications that one can make validly and effectively. The purposes of GCSEs are currently defined as follows (see Ofqual, 2019):

- To provide evidence of students' achievements against demanding and fulfilling content;
- To provide a strong foundation for further academic and vocational study and for employment; and
- To provide (if required) a basis for schools and colleges to be held accountable for the performance of all of their students.

In line with these purposes, the GCSE results reporting system is standards-based – there are subject content criteria, assessment objectives and grade descriptions (which may be used to articulate performance standards) defined for individual qualifications. Therefore, grades should represent the levels of skills and knowledge that the candidates have achieved in specific subject areas. The existing examination processes are aimed at producing evidence of validity to support the standards-based interpretation of exam results. Cross-subject comparison of examination standards using statistical analyses would seem to be of limited meaning in this regard. Any discussion about subject grade/grading difficulty should be primarily focused on the appropriateness of the established grade standards in relation to the defined purposes of the qualification and whether the performance

standards set on the assessment represent an accurate reflection of the grade standards.

2.9 Summary

Results from statistical analyses using a range of methods suggest that GCSE French, German and Spanish are generally statistically more “difficult” than many of the other GCSE subjects at both individual grade level and the overall subject level over the past 15 years. However, given the limitations associated with these methods discussed above, this cannot be interpreted simply as an indication of grading severity in these subjects in comparison with other subjects.

3 Impact of statistical alignment of standards on outcomes

If standards were to be aligned for different subjects, based on inter-subject comparability studies using statistical methods, the grade boundaries for some of the examinations would need to be changed. As grade boundaries can be viewed as the operationalisation of performance standards, any change in grade boundaries would imply a qualitative change in performance standards from those established for some of the examinations (e.g. subjects which are either too “easy” or too “hard” based on statistical comparisons). The distributions of grades will also change accordingly. Some of the distributions may become less effective in differentiating the candidates (for example, if the resultant grade distribution becomes highly skewed towards the top grades or bottom grades). This would have an impact on the interpretation of grades. This section looks at the impact of statistically aligning standards between subjects on the quality of candidate performance at adjusted grade boundaries and on grade distributions for GCSE French, German and Spanish.

3.1 Change in grade boundaries and grade distributions

3.1.1 Analysis based on Rasch modelling

Table 9 shows the original grade distributions for GCSE French, German and Spanish in 2016 (see also Figure 13). All three qualifications were unitised and use a uniform mark scale (UMS)[1]. Percentage changes in candidates at individual grades and the changes in the cumulative percentages as well as the shifts in qualification level UMS grade boundaries (GBs) expressed as the percentages of maximum available UMS marks after alignment of standards statistically to the average of all subjects based on Rasch analysis, are also shown in the table. The alignment involved shifting the original grade boundary mark at a specific grade by the proportion of the grade width determined by its relative difficulty (see He et al., 2018,

[1] UMS mark is a scaled score used to ensure the comparability of raw marks from different examination series. See https://filestore.aqa.org.uk/admin/results-days/AQA-UMS-GUIDE.PDF for a detailed explanation.

for details of the boundary mark adjustment process). For French, the number of students receiving A* would not change, but there would be about 4% more students receiving A and 11% more students receiving B. Cumulatively, the percentage of students receiving a grade C or above would increase from about 69% to 82%. For German, the percentage of students receiving A*, A and B would go up by about 3%, 5% and 8% respectively, with the cumulative percentage of students receiving C and above increasing by about 11%. For Spanish, the percentage of students receiving A* would decrease by about 3%. However, those receiving A, B and C would go up by about 4%, 7% and 2% respectively. Overall, there would be about 10% more students receiving a grade C or above.

Table 9 Changes in percentages of students receiving individual grades and cumulative percentages and shifts in grade boundaries for GCSE French, German and Spanish in 2016 after alignment of standards statistically to the mean of all subjects based on Rasch analysis.

Subject (N)	Change in grade distributions (%) and boundaries (% of max UMS marks)								
		A*	A	B	C	D	E	F	G+U
French 125,563	Original (Ind.)	10.12	12.99	18.58	27.72	19.94	6.90	2.57	1.19
	New (Ind.)	10.12	17.07	29.48	25.64	10.68	4.33	1.66	1.01
	Change (Ind.)	**0.00**	**4.08**	**10.90**	**-2.07**	**-9.26**	**-2.57**	**-0.90**	**-0.18**
	Original (Cum.)	10.12	23.11	41.69	69.41	89.35	96.25	98.81	100.00
	New (Cum.)	10.12	27.19	56.68	82.32	93.00	97.33	98.99	100.00
	Change (Cum.)	**0.00**	**4.08**	**14.98**	**12.91**	**3.65**	**1.08**	**0.18**	**0.00**
	Boundary shift	**-0.02**	**-2.80**	**-5.80**	**-5.60**	**-4.00**	**-3.00**	**-1.50**	
German 47,779	Original (Ind.)	7.72	14.55	23.03	28.30	17.89	5.98	1.83	0.72
	New (Ind.)	10.81	19.53	31.29	23.13	9.77	3.63	1.26	0.57
	Change (Ind.)	**3.10**	**4.99**	**8.26**	**-5.17**	**-8.12**	**-2.34**	**-0.57**	**-0.15**
	Original (Cum.)	7.72	22.26	45.29	73.59	91.48	97.45	99.28	100.00
	New (Cum.)	10.81	30.35	61.64	84.77	94.54	98.17	99.43	100.00
	Change (Cum.)	**3.10**	**8.08**	**16.34**	**11.18**	**3.06**	**0.72**	**0.15**	**0.00**
	Boundary shift	**-2.70**	**-4.00**	**-5.90**	**-5.30**	**-3.90**	**-2.90**	**-1.60**	
Spanish 87,199	Original (Ind.)	12.61	14.21	19.03	24.25	17.59	7.33	3.10	1.89
	New (Ind.)	9.64	18.24	25.96	26.01	12.37	4.68	1.93	1.17
	Change (Ind.)	**-2.97**	**4.03**	**6.93**	**1.75**	**-5.22**	**-2.65**	**-1.17**	**-0.72**
	Original (Cum.)	12.61	26.82	45.84	70.09	87.68	95.01	98.11	100.00
	New (Cum.)	9.64	27.88	53.84	79.85	92.22	96.90	98.83	100.00
	Change (Cum.)	**-2.97**	**1.07**	**8.00**	**9.75**	**4.54**	**1.89**	**0.72**	**0.00**
	Boundary shift	**2.10**	**-0.70**	**-3.60**	**-4.70**	**-5.00**	**-5.60**	**-4.30**	

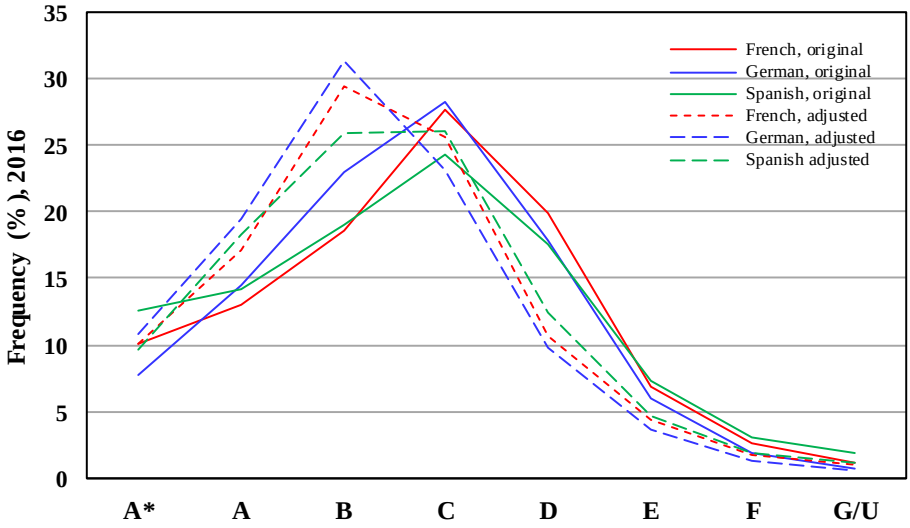

Figure 13 Original grade distributions of GCSE French, German and Spanish in 2016 and the grade distributions expected after aligning their standards statistically to the means of all subjects based on Rasch analysis.

To look at how the alignment of standards statistically affects the grade boundary locations on the mark distribution in more detail, Figure 14 shows the total UMS mark distributions, original grade boundaries and the adjusted grade boundaries for GCSE French, German and Spanish offered by one of the awarding organisations in 2016 after statistical alignment of standards with the means of all subjects. The maximum UMS mark for all the three subjects was 300, with a grade width of 30 marks. The effect of aligning standards statistically on grade boundary locations varies between subjects and grades in the same subject. At A*, no change in grade boundary would be needed for French, but the boundary mark would need to be lowered by 8 marks for German and increased by 3 marks for Spanish. Grade boundaries at A would need to be lowered by about 8, 12 and 2 marks respectively for French, German and Spanish. At grade C, the boundary mark would need to be reduced by about 15 marks for all the three subjects. All this suggests that statistical alignment of standards would result in the performance standards at some of the adjusted grade boundaries to be notably different from the standards represented by the original grade boundaries. The adjusted standards at grade C would be considerably lower than the original standards for all three subjects. This would affect the interpretation of GCSE grades in relation to the defined purposes and uses of GCSE qualifications. Furthermore, from a measurement perspective, the adjusted grade boundaries could affect the appropriateness of the apportionment of mark ranges to individual grades and the associated grade distribution, or the redesign and development of new assessments would be required.

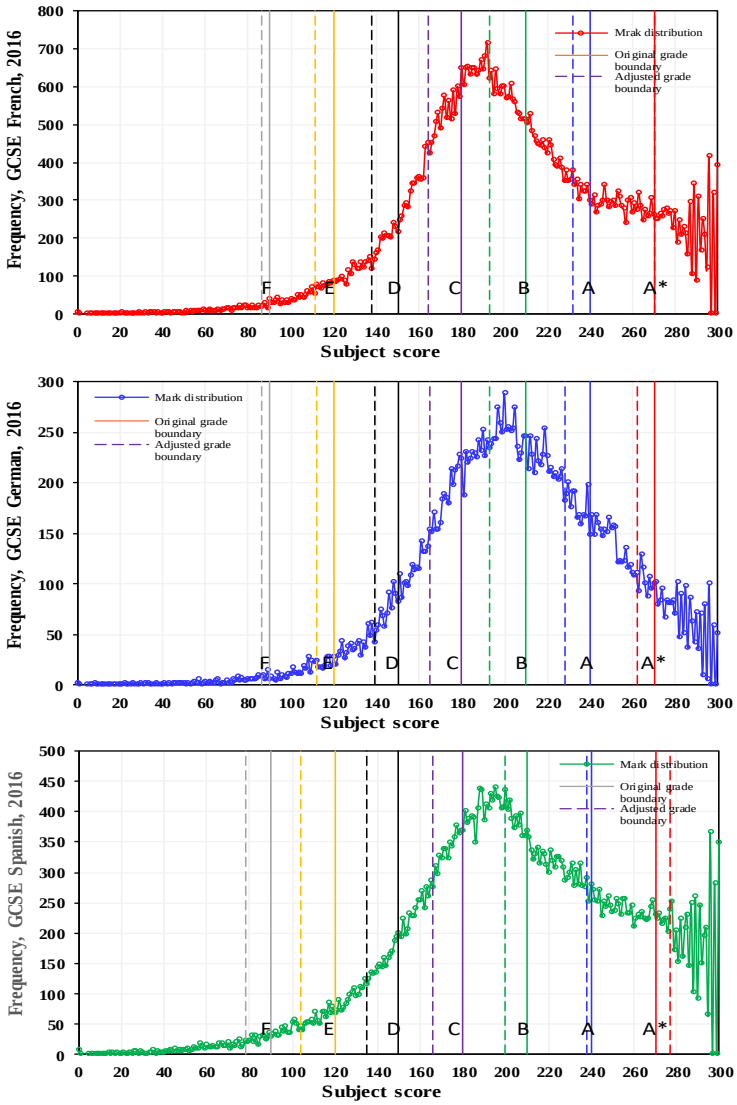

Figure 14 Qualification level UMS mark distribution for GCSE French, German and Spanish in 2016 offered by one of the awarding organisations, the original UMS grade boundaries and the adjusted grade boundaries after statistical alignment of standards with the means of all subjects.

To look at the impact of statistical alignment of standards on grade outcomes for the reformed GCSE French, German and Spanish, Table 10 shows changes in cumulative frequency at individual grades for the three subjects in 2019 after aligning their standards with the mean of all subjects based on grade difficulties derived using the Rasch model. The percentage of candidates receiving grade 9 would increase by about 2.7% for French and 3.9% for German. For Spanish, there would be no change. The percentage of candidates receiving a grade 7 or above would increase by 5.1%, 9.3% and 2.3% respectively for French, German and Spanish. Candidates receiving a grade 4 or above would increase by about 7.3%, 7.9% and 6.7% respectively for the three subjects. In Table 10, changes in cumulative frequencies at individual grades for French and German after aligning their standards to those of

Spanish (which is marginally harder than the average at grades 9 and 8 but considerably easier than French and German) are also shown. Candidates receiving grade 9 would increase by 2.4% for French and 3.6% for German after aligning their standards to those of Spanish. There would also be 2.8% more French candidates and 6.5% more German candidates achieving a grade 7 or above. At grade 4, there would be almost no change in cumulative frequency.

Table 10 Changes in cumulative frequency at individual grades for reformed GCSE French, German and Spanish in 2019 after alignment of their standards to the mean of all subjects and those of Spanish.

Alignment	Subject	Change in cumulative frequency (%)							
		9	8+	7+	6+	5+	4+	3+	2+
Aligning with mean	French	2.70	3.63	5.13	5.00	9.59	7.35	4.23	1.19
	German	3.92	6.58	9.32	9.55	12.38	7.89	3.70	1.50
	Spanish	0.00	0.49	2.27	3.09	7.18	6.74	4.69	1.48
Aligning with Spanish	French	2.40	2.67	2.82	1.75	1.14	0.11	-0.68	0.33
	German	3.58	5.48	6.55	5.19	3.86	0.78	0.25	0.49

3.1.2 Analysis based on application of cohort prior attainment profile

To maintain standards over time, Ofqual introduced the "comparable outcome" approach to awarding of GCSEs and A levels from 2012, which involves maintaining the consistency over time of the relationship between prior attainment and GCSE grade outcomes established in a particular year (or reference year) for a specific subject (see Ofqual, 2014). This approach has also been used to address the issue of inter-board comparability (see Taylor, 2013). This section explores the impact of aligning standards using the KS2 attainment profile approach on grade outcomes for GCSE French, German and Spanish. This involves establishing the relationship between KS2 profile and grade outcomes for the reference subject first. Operationally, this is achieved by calculating the proportion of candidates in each KS2 decile who achieved each of the GCSE grades. These proportions form a matrix. To align other subjects to the reference subject, this matrix is applied to the KS2 profiles of the subjects to produce the predicted grade outcomes. The predicted outcomes can then be compared with the original grade outcomes to assess the impact of alignment. This approach is different from the Rasch approach discussed in 3.1.1 where changes in grade boundaries were required to align standards.

Figure 15 below shows the original cumulative grade distributions of GCSE French, German and Spanish for KS2 matched candidates in 2016 and their predicted grade distributions after aligning their standards to those of history based on using the KS2 attainment profile approach discussed above[2]. The predicted cumulative grade distributions of French and German after aligning their standards to those of Spanish are also shown in the figure. Table 11 shows changes in cumulative frequency at

[2] It is worth noting this analysis does not use the exact same methods as those used in awarding and are thus indicative rather than definitive. For example, the prediction matrix is based on 2016 outcomes rather than previous outcomes; and the predicted outcomes were calculated for the whole cohort rather than on a board-by-board basis.

individual grades after the alignment of standards. At Grade B, students achieving a grade B and above would increase by 13.6%, 12.7% and 9.4% for French, German and Spanish respectively after aligning their standards to those of history. If French and German were aligned with Spanish, the percentage of candidates receiving A and A* would increase by 3.6% and 5.8% respectively for French and German. At grade C, there would be very little change for both subjects.

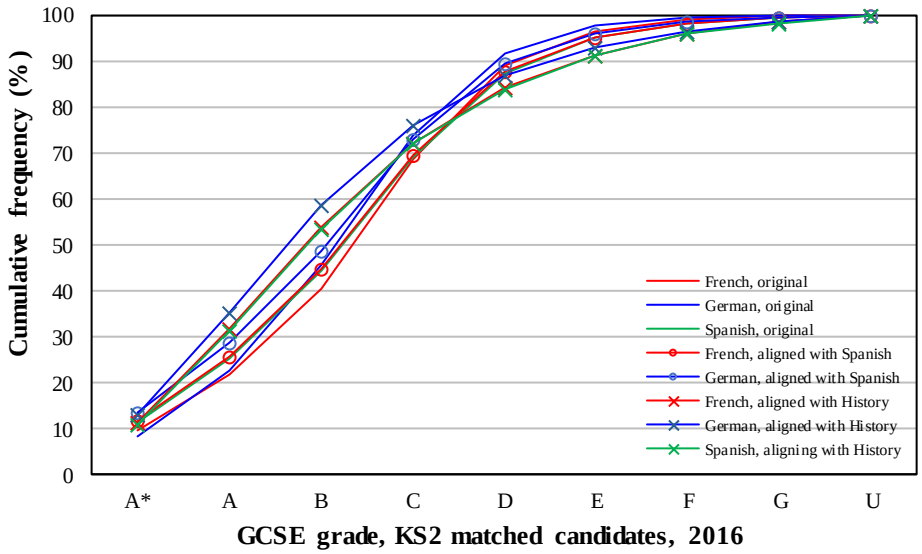

Figure 15 Original grade distributions of GCSE French, German and Spanish for KS2 matched candidates in 2016 and their predicted grade distributions after aligning their standards to those of history and Spanish.

Table 11 Changes in cumulative frequency at individual grades for French, German and Spanish in 2016 after alignment of their standards to those of history and Spanish (based on KS2 matched candidates).

Alignment	Subject	Change in cumulative frequency (%)							
		A*	A+	B+	C+	D+	E+	F+	G+
Aligning with History	French	1.70	9.64	13.60	3.50	-5.07	-5.00	-3.13	-1.40
	German	4.84	12.62	12.76	2.04	-4.97	-4.67	-2.75	-1.22
	Spanish	-0.54	5.93	9.45	2.71	-3.48	-3.63	-2.35	-1.15
Aligning with Spanish	French	2.23	3.67	4.17	0.80	-1.62	-1.40	-0.80	-0.27
	German	5.23	5.87	2.76	-1.03	-2.22	-1.74	-0.86	-0.30

Similarly, Figure 16 shows the original cumulative grade distributions of the reformed GCSE French, German and Spanish for KS2 matched candidates in 2018 and their predicted grade distributions after aligning their standards to those of history and Spanish. Table 12 shows changes of cumulative frequencies after the alignment of standards. At Grade 7, students achieving a grade 7 or above would increase by 9.2%, 13.6% and 3.1% for French, German and Spanish respectively after aligning their standards to those of history. At grade 4, there would be very little change for all the three subjects. If French and German were aligned with Spanish, percentage of candidates receiving a grade 7 or above would increase by 2.7% and 5.9% for

French and German respectively. At grade 4, there would be very little change in cumulative frequency for French but a decrease of 1.9% for German.

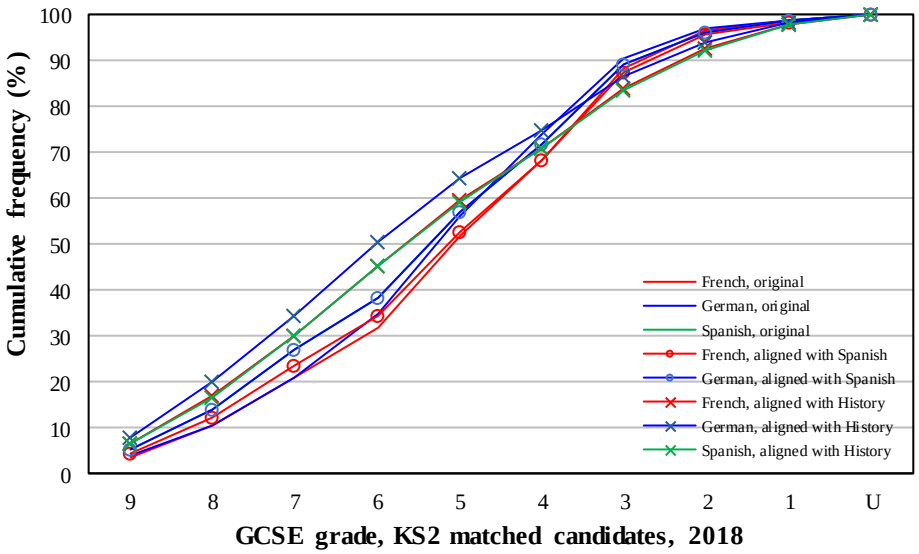

Figure 16 Original grade distributions of GCSE French, German and Spanish for KS2 matched candidates in 2018 and their predicted grade distributions after aligning their standards to those of history and Spanish.

Table 12 Changes in cumulative frequency at individual grades for French, German and Spanish in 2018 after alignment of their standards to those of history and Spanish (based on KS2 matched candidates).

Alignment	Subject	9	8+	7+	6+	5+	4+	3+	2+	1+
Aligning with History	French	2.81	6.34	9.23	13.66	7.48	2.65	-4.38	-3.93	-0.82
	German	4.08	9.41	13.63	15.40	8.31	1.16	-4.18	-3.11	-0.62
	Spanish	1.25	2.81	3.10	6.90	2.34	-1.06	-5.65	-3.90	-0.64
Aligning with Spanish	French	0.73	1.53	2.73	2.55	0.79	-0.05	-0.70	-0.70	-0.30
	German	1.32	3.37	5.92	3.31	1.03	-1.94	-1.23	-0.53	-0.33

3.2 Summary

It has been shown that aligning standards statistically between subjects might impact the performance standards and grade distributions. For GCSE French, German and Spanish, particularly French and German, the quality of performance expected of students at some grades would need to be relaxed if they were to be aligned with other subjects statistically. Care would need to be taken to ensure the statistical alignment of standards between subjects would not invalidate the interpretation of the results from GCSEs as implied by their purposes and uses as currently defined.

4 Conclusion

Results from analyses using a range of statistical methods indicate that there is variability in statistical difficulty between GCSE subjects which appears to be consistent over the past 15 years or so. GCSE MFLs were found to be harder statistically than many of the other subjects. Differences in difficulty between subjects have frequently been used as evidence of severity (or leniency) in subject-specific grading standards. However, the limitations of statistical techniques used to study inter-subject comparability must be realised. Whilst differences in statistical difficulty between subjects may reflect differences in the amount of some shared common trait that is required to achieve the same level of performance, it does not necessarily imply that some subjects are graded more severely (or leniently) than others. This is because examinations such as GCSEs and A levels are graded based on standards representing the expected level of knowledge and skills which are subject specific. The shared knowledge and skills (represented by the shared common latent trait) assessed by exams in different subjects are irrelevant in this regard. Moreover, the shared knowledge and skills can vary considerably between subjects. There can also be issues with the statistical methods themselves (e.g. violation of model assumptions by real data, imperfect data-model fit, missing data, and others). Most of the statistical techniques also fail to take account of some of the important factors that can affect the performance of candidates in exams. Those factors can vary considerably between subjects. All this makes it difficult to interpret the difficulty measures derived and to link difficulty measures directly to the grading standards of examinations which are subject specific.

It has been demonstrated that aligning standards statistically between GCSE subjects could potentially have important impacts on performance standards. Depending on the particular point of statistical alignment chosen, this would have a lesser or greater impact on the quality of performance expected of students at some grades in GCSE French, German and Spanish. This could in turn affect the interpretation of GCSE grades in relation to its defined purposes. It should be stressed that any attempt to align standards statistically between subjects should consider the limitations of statistical methods and the impact of such alignment on performance standards and the interpretation of GCSE results. Furthermore, any new standards resulting from statistical alignment would need to be judged by the primary users of GCSE qualifications in order to ensure that they are acceptable and meet their requirements.

References

Benton, T. (2016). On the impact of aligning the difficulty of GCSE subjects on aggregated measures of pupil and school performance. *Research Matters: A Cambridge Assessment publication*, *22*, 27-30.

Black, B., Clewes, J., Garrett, R., He, Q. and Curcin, M. (2018a). The evidence pertaining to the claim of grading severity in A level physics, chemistry and biology and the impact of statistical alignment of standards on outcomes. Ofqual: Coventry.

Black, B., Clewes, J., Garrett, R., He, Q. and Curcin, M. (2018b). The evidence pertaining to the claim of grading severity in A level French, German and Spanish and the impact of statistical alignment of standards on outcomes. Ofqual: Coventry.

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

Bramley, T. (2016). The effect of subject choice on the apparent relative difficulty of different subjects. Cambridge Assessment Research Report. Cambridge Assessment: Cambridge.

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

He, Q., and Stockford, I. (2015). Inter-Subject Comparability of Exam Standards in GCSE and A Level. Ofqual: Coventry.

He, Q., Stockford, I. and Meadows, M. (2018). Inter-subject comparability of examination standards in GCSE and GCE. *Oxford Review of Education* 44, 494-513.

Lockyer, C., and Newton, P. E. (2015). Inter-subject comparability: a review of the technical literature. Ofqual: Coventry.

Newton, P. E. (2012). Making sense of decades of debate on inter – subject comparability in England. *Assessment in Education* 19, 251-273.

Newton, P. E. (2015). Exploring implications of policy options concerning inter-subject comparability. Ofqual: Coventry.

Office of Qualifications and Examinations Regulation (Ofqual) (2019). GCSE (9 to 1) Qualification Level Conditions and Requirements. Ofqual: Coventry.

Ofqual. (2014). Setting GCSE, AS and A level grade standards in summer 2014 and 2015, Ofqual/15/5759. Ofqual: Coventry.

Office of Qualifications and Examinations Regulation (Ofqual) (2016) A policy position for Ofqual on inter-subject comparability. Ofqual: Coventry.

Office of Qualifications and Examinations Regulation (Ofqual) (2017). Grading new GCSEs. Ofqual: Coventry.

Office of Qualifications and Examinations Regulation (Ofqual) (2018). Inter-subject comparability in A level sciences and modern foreign languages: examining the claim that these subjects are more severely graded than other A levels. Ofqual: Coventry.

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

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

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

November 2019

Ofqual/19/6557/4