TECHNICAL REPORT

Inter-subject comparability technical report science

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

Beth Black, Jonathan Clewes, Richard Garrett, Qingping He and Milja Curcin

Contents

Executive Summary

The purpose of this report is to gather evidence to support Ofqual in making an informed, evidence-based decision about whether there is a need to make an adjustment to grading standards in A level biology, chemistry and physics. Evidence to investigate the claim that these subjects are severely graded compared with other A level subjects was collected from a range of sources, including:

- Statistical evidence
- Judgemental evidence
- Impact of grade severity on society
- The learned societies'[1] perceptions of grading severity and impact
- Awarding organisations' views
- Higher education representatives' perception of grade standard adjustment

In addition, the impact of statistical alignment of standards between A level physics, chemistry and biology and other subjects in terms of their grade distributions and performance standards is also discussed in this report.

Results from statistical analyses using a range of methods suggests that A level physics, chemistry and biology are generally more "statistically difficult" than most of the other A level subjects both in terms of a given grade, and at the overall subject level, and have been for at least the past decade. 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 to compare standards between subjects, including the strong assumptions made about the data being analysed (e.g. the unidimensionality assumption that a single underlying latent trait is assessed by examinations in different subjects required by the Rasch model and other similar models) which are seldom met by real data. Many of the factors that can potentially influence examination performance are also discussed, including motivation of students and efficiency and effectiveness of teaching and learning. It is suggested that these factors can vary substantially between subjects and that most statistical methods fail to take them into account.

The very limited research using judgemental methods to compare the standards between A level physics, chemistry and biology and other A level subjects suggests that the three subjects were aligned in performance standards to one another, and that biology and psychology were also aligned. The limitations of the subject experts-based approaches are also discussed.

Stakeholders identify several negative effects which they believe are impacting on students and society as a whole as a result of severe grading in these subjects. These include depressed uptake at A level (although in a context where entries for STEM subjects are increasing overall), but also that specific groups of students (female students, those with low prior attainment, and students from less advantaged backgrounds) are being discouraged from studying science. These are all groups whom stakeholders point out are already relatively underrepresented in STEM fields. These claims are examined through reference to wider contextual data such as A level

[1] Details of the learned societies involved are given on page 47.

entry and university acceptances, prior attainment and gender profiles, and figures on teacher supply and recruitment. Overall entries for A level sciences are increasing, and it is difficult to find evidence of the 'missing' students who might be deterred by severe grading. Where there are clear trends in entry (for instance, the male bias in A level physics entries) it is challenging to prove causation – particularly when comparisons are drawn between subjects which appear to be of similar difficulty under statistical measures but have very different entry profiles. Research into subject choices also suggests that other factors take precedence over difficulty when students select their A levels options.

The learned societies consider grading severity to be a systemic issue at A level, and reject the assertion statistical measures of subject difficulty may be measuring something significant other than grading severity. They are satisfied that the agreement between different statistical analyses confirms that there is a lack of comparability between subjects, and contend that students of the same ability should have a similar likelihood of achieving a given grade in any A level subject. They are also concerned that severe grading undermines the assumptions made by universities of equivalency of grades in different subjects for admissions purposes. Moreover, the learned societies argue that severe grading is creating barriers to access by dissuading certain groups of students from studying A level sciences, either because they perceive they will do worse in these subjects, or because schools concerned about their performance under attainment measures are setting higher entry requirements for them.

The views of awarding panels responsible for setting and maintaining standards in these subjects have also been taken into account. Overall awarders were satisfied with current standards and did not feel that these qualifications were severely graded. Some argued that at specific grades the current standards are in fact too lenient. Awarders were aware that these subjects are perceived as being 'difficult', but felt that the inherent demands of the subject meant that this will always be the case, and that it would not be acceptable to address perceived difficulty by lowering grading standards. Some also argued that the reputation of science A levels as challenging subjects means that they are better regarded by universities and employers, and were concerned that any changes might lead to them being devalued.

The awarding organisations themselves have been given the opportunity to formally state their views. The exam boards note the limitations of some of the statistical models used to measure subject 'difficulty' as a basis for achieving inter-subject comparability, particularly the extent to which they are biased by the non-random nature of students' subject choices. They also point out the current analyses were conducted on legacy qualifications, and that the relative 'difficulty' of these A levels may have changed as a result of the reforms (although the policy intention was that the difficulty would remain the same) and the decoupling of AS and A levels. Changes to content standards at GCSE were also felt to have the potential to impact on students' perceptions of A level. The boards recommend further research be carried out on the impact of these changes before any adjustment to grading standards is considered, particularly in view of the potential risks to performance standards, progression and public confidence from such a change. They also identify a tension between the stated purpose of A levels to identify attainment in a particular subject, and their use as a 'currency' for progression to higher education in general. Overall the boards are of the view that performance standards within a subject are more important than comparative standards between subjects, and that these should be prioritised.

The report also considers research conducted by Curcin, Black and He (2018) into how changes to grading standards in these subjects might be perceived by higher education, and the potential impact on the way A level grades are used in admissions. Representatives from higher education, the learned societies, and Ofqual subject experts reviewed samples of A level student work at and below several grade boundaries in each subject, indicating how those below compared to those on current grade boundaries. The participants were subsequently engaged in panel discussion on grade standards issues. The research showed that, overall, support for an adjustment to standards in a subject was inversely related to apparent difficulty under statistical measures, being stronger in those subjects which were closer to 'average' A level difficulty. Support for an adjustment to standards expressed in discussion also frequently contradicted the judgements made by participants in the script review. The research indicated that universities were likely to find adjustments to grading standards in A level languages more acceptable than in the sciences. In both subject areas, there was agreement that the range of acceptable adjustment would be relatively small.

Finally, the report demonstrated that aligning standards statistically between subjects can have substantial impact on performance standards as represented by subject grade boundaries and grade distributions. The quality of performance expected of students, for each grade in A level physics, chemistry and biology would need to be lowered considerably if they were to be aligned with other A level subjects statistically.

1. Introduction

From time to time, questions are asked as to whether some GCSE or A level subjects are harder than others and, if so, whether a better alignment should be achieved.

Ofqual has been examining this issue for some time, and there is already a significant body of research into inter-subject comparability, to which Ofqual has contributed here with a programme of work begun in 2015. Ofqual sought to start a debate about the concept of inter-subject comparability through the publication of six working papers; a number of historical research papers on the topic, including some by Ofqual's predecessor organisation the Qualifications and Curriculum Authority (QCA); and a survey of views on potential policy options.

In response to the consultation on potential policy options, Ofqual resolved not to revise grading standards in all GCSE and A level subjects to achieve inter-subject comparability on a statistical basis. However, three A level sciences and three A level languages were identified as possible exceptional cases where we might take action to address concerns about the relative difficulty of these subjects – felt by some stakeholders to be the result of more severe grading.

To determine if this was the case, we decided that it would be necessary to devise a basket of evidence upon which to make a decision about adjusting grading standards in a specific subject. This evidence would need to include not only the statistical data which initially prompted our consultation, but also the views of stakeholders (the subject associations and higher education selectors representing those subjects, and the exam boards and their awarders) as well as contextual data which might contribute to our understanding of other factors that could have an impact on perceived difficulty.

These latter two sources of evidence were felt to be particularly important because it was noted that the comparisons between subjects on which the statistical data is based rest on conceptions of attainment-related linking constructs (such as ‘general intelligence’ or ‘general academic aptitude’) which some educationalists reject. Certainly, the plausibility and relevance of the evidence of ‘difficulty’ produced by these linking constructs would seem to diminish the less similar the subjects that it is used to draw comparisons between (for instance, comparisons between physics and maths would seem to be more valid than those between physics and music). In light of this, it was recognised that whilst statistical evidence would be a key component of any basket of evidence, it must also be treated with caution. This report also explores some of the limitations of this statistical evidence in detail.

Nonetheless, statistical measures of subject difficulty are a source of evidence which, when considered alongside evidence of possible negative impacts within a subject, may contribute to a compelling case to adjust grading standards. For convenience, when discussing the evidence produced by these statistical measures we use the terms ‘severe’ and ‘lenient’, but it should be borne in mind that they are used in reference to the apparent difficulty of these subjects under Rasch and Comparative Progression Analysis only.

In considering this issue, there are different aspects of our statutory objectives and duties that we need to balance, reflecting that they are in tension with one another. We are required by legislation[2] to ensure regulated qualifications represent a “consistent level of attainment (including over time) between comparable regulated qualifications”. If we were to adjust grade standards, we would need to consider that the issues in these subjects were sufficient for us to prioritise the need to achieve comparable standards in different but comparable A level subjects over the need to maintain standards in a subject over time. We must also reflect on the impact of any potential action on our objective to secure public confidence, and our duty to have regard to the views and needs of stakeholders who are ‘users’ of our qualifications, such as employers and universities.

We have since carried out a substantial programme of work to engage with stakeholders and seek their input in determining the basket of evidence for

[2] The 2009 Apprenticeships, Skills, Children and Learning Act.

considering a possible adjustment to grading standards in A level physics, chemistry and biology, as well as revisiting the various statistical measures of subject difficulty.

The evidence we have assembled within each subject area includes statistical measures of subject difficulty such as Rasch and Comparative Progression Analysis; contextual data on issues such as teacher supply and quality; and consideration of the views of stakeholders, including subject associations and higher education but also the exam boards and awarders. We have also considered evidence such as data on changes in A level entries and university applications over time; analyses of potential changes in the ability range and gender profile of the cohorts taking particular subjects; and research into motivations behind students' subject choices. Following advice from Ofqual's Standards Advisory Group, we have also incorporated into our evidence base information about recent grade boundaries in these subjects.

We have also conducted a research study into views on the impact of possible grading standard adjustment amongst higher education, to determine what the impact might be on the utility of these qualifications for university admissions were we to make any change. A summary of the findings is presented in this report.

2. Evidence pertaining to the claim of grading severity

This section provides a brief account of the evidence collected from a range of sources that has been used to support/challenge the claim that A level Biology, Chemistry and Physics are severely graded compared with other A level subjects. This evidence has been broadly classified into the following categories:

- Statistical evidence
- Judgemental evidence
- Impact of grade severity on society
- The learned societies' perceptions of grading severity and impact
- Awarding organisations' views
- Higher education representatives' perception of grade standard adjustment

2.1 Statistical evidence

Statistical evidence can be broadly grouped into two categories, one associated with conventional statistical techniques (including the Rasch modelling approach), one associated with the comparative progression analysis (CPA) approach.

2.1.1 Evidence generated from Rasch modelling and other conventional statistical techniques

Statistical methods used to study inter-subject comparability 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. These include subject pairs analysis, common examinee linear models (including Kelly's method), latent trait models (e.g. 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 here.

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 fashion.

Common examinee linear models (including Kelly's method) derive the relative difficulties of different subjects from a matrix of examination by candidate results and involve finding the solution of a set of simultaneous linear equations.

In the case of using latent trait models such as the 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 candidate's prior attainment, gender, socioeconomic status, type of school attended, and many others) that can potentially influence examination performance.

For most of the statistical techniques, for the results to be interpreted appropriately, it is explicitly or implicitly assumed that the examinations in different subjects that are analysed together define a shared common construct or latent trait (such as "general academic ability/aptitude", the unidimensionality assumption) 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. Difference in difficulty between exams in different subjects is assumed to reflect differences in "standards" which are related to this latent trait. It is to be noted that the latent trait is inferred from the analysis of the examinations included in the analysis.

Analysis of A level subjects using the Rasch model over the past decade or so suggested that physics, chemistry and biology are harder than the majority of the subjects analysed (see Coe et al., 2008; He and Stockford, 2015). As an example, Figure 1 compares the overall difficulty and the difficulties of individual grades between the A level subjects from the 2017 exam series, with the subjects ordered by the overall subject difficulty. There is substantial variability in difficulty at individual grades between the subjects.

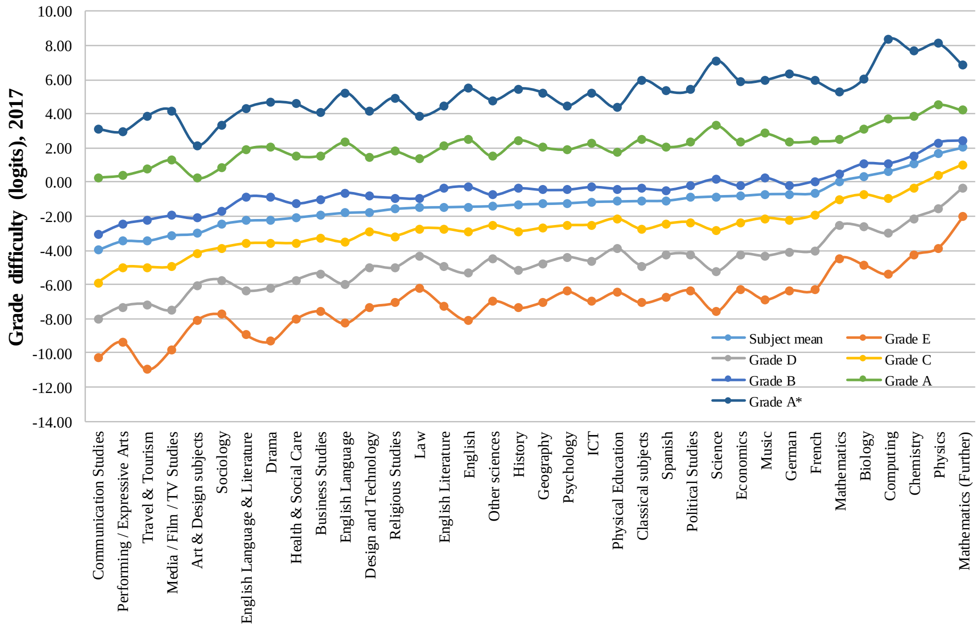


Figure 1 Comparison of the overall subject difficulty and the difficulties at individual grades for the A level examinations administered in 2017.

Table 1 below shows the overall relative difficulties and difficulties at individual grades of physics, chemistry, biology and mathematics in 2013, 2016 and 2017 respectively. Mathematics was used here as a reference subject for sciences as it is also a cognate facilitating subject required for university admissions. It is to be noted that direct comparison of the values in Table 1 *between years* is not appropriate as the data for each exam series were analysed using the Rasch model separately. At grades A and A*, both physics and chemistry were about two thirds of a grade more difficult than the average difficulty of all subjects (which was set to zero) and about half a grade harder than mathematics.

Table 1 Relative grade difficulties (relative to the mean of all subjects contained in the analysis) expressed in grade width for A level mathematics, biology, chemistry and physics from 2013, 2016 and 2017, based on Rasch analysis.

Subject	Relative difficulty (grade width)											
	2013				2016				2017			
	Overall	C	A	A*	Overall	C	A	A*	Overall	C	A	A*
Maths	0.46	0.59	0.09	0.04	0.54	0.40	0.18	0.12	0.58	0.80	0.14	0.05
Biology	0.63	0.74	0.32	0.26	0.69	0.62	0.41	0.28	0.71	0.95	0.39	0.31
Chemistry	0.84	0.86	0.55	0.64	1.00	0.79	0.71	1.00	1.02	1.16	0.67	0.84
Physics	0.98	1.03	0.71	0.70	1.19	1.05	0.94	1.14	1.28	1.50	0.94	0.98

Table 2, which was adapted from Coe et al. (2008), shows the relative grade difficulties of mathematics, biology, chemistry and physics in 2006 analysed using a range of statistical methods. Again, physics, chemistry and biology were found to be more difficult than mathematics and most of the other subjects.

Table 2 Relative grade difficulties (relative to the mean of all subjects contained in the analysis) expressed in grade width for A level mathematics, biology, chemistry and physics from 2006, estimated using different statistical methods.

Subject	Relative difficulty (grade width)						
	Rasch	SPA (unweighted)	SPA (weighted)	Kelly	Reference test	Value-added	Multilevel
Mathematics	0.52	0.25	-0.07	0.20	0.04	0.21	0.26
Biology	0.81	0.68	0.21	0.54	0.58	0.57	0.59
Chemistry	0.96	0.70	0.19	0.62	0.41	0.50	0.63
Physics	0.95	0.75	0.22	0.65	0.50	0.60	0.66

Coe et al. (2008) also reported the relative difficulties of A level subjects estimated using Kelly's method from 1994 to 2006, and Table 3 shows the values for mathematics, biology, chemistry and physics from 2000 to 2006. Again the science subjects were found to be about half a grade harder than most of the other subjects. They were also harder than mathematics. The difficulties are relatively stable over time, particularly since 2002.

Table 3 Relative grade difficulty expressed in grade width for A level mathematics, biology, chemistry and physics from 2000 to 2006 estimated using Kelly's method.

Subject	Relative difficulty (grade width)						
	2000	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**
Mathematics	0.46	0.72	0.29	0.32	0.30	0.19	0.15
Biology	0.30	0.68	0.39	0.48	0.52	0.50	0.51
Chemistry	0.48	0.95	0.54	0.61	0.61	0.61	0.61
Physics	0.48	0.81	0.49	0.58	0.60	0.59	0.64

2.1.2 Evidence from comparative progression analysis (CPA)

Newton, He and Black (2017) discussed the comparative progression analysis (CPA) approach to inter-subject comparability. CPA concerns the 'progress' made by individual students from their GCSE grade to their A level grade in the same subject area. It considers the distribution of A level grades in a subject, for students who were awarded a particular grade in the same subject at GCSE, to determine whether progression patterns are the same, or different, across subject areas. To interpret outputs from these analyses at face value – in terms of the alignment, or misalignment, of grading standards, or difference in subject difficulty – it is necessary to assume that groups of candidates ought, on average, to make the same progress across subject areas, all other things being equal; and that all other things are in fact more or less equal (see Newton et al., 2017).

Figure 2 shows the A level grade distributions for students with grades B and A at GCSE in the same subject in 2010 and 2016 for 14 A level subjects. The patterns of the distributions are similar for the two years. As is clear from the graphs, students taking the science subjects and modern foreign languages (MFLs) achieved lower A level outcomes than those taking the subjects in English and social sciences. That is, students who studied A level sciences and MFLs made less progress from GCSE to A level than those who had the same grade at GCSE in the same subject but studied English or social science subjects. However, the difference in A level grade distribution between subjects for students achieving an A grade at GCSE is not as prominent as that for students achieving a B grade at GCSE. To provide some quantitative information about the difference in A level outcomes between the subjects, Table 4 shows the average A level grade of the students in the subject with grades B and A at GCSE in 2010, 2013 and 2016 respectively. Students studying physics achieved slightly a lower average A level grade in the subject than those studying biology and chemistry. Overall, students with B at GCSE who studied the sciences achieved over 0.6 of a grade lower than those who studied English and social sciences. For students with an A at GCSE, this is about half of a grade.

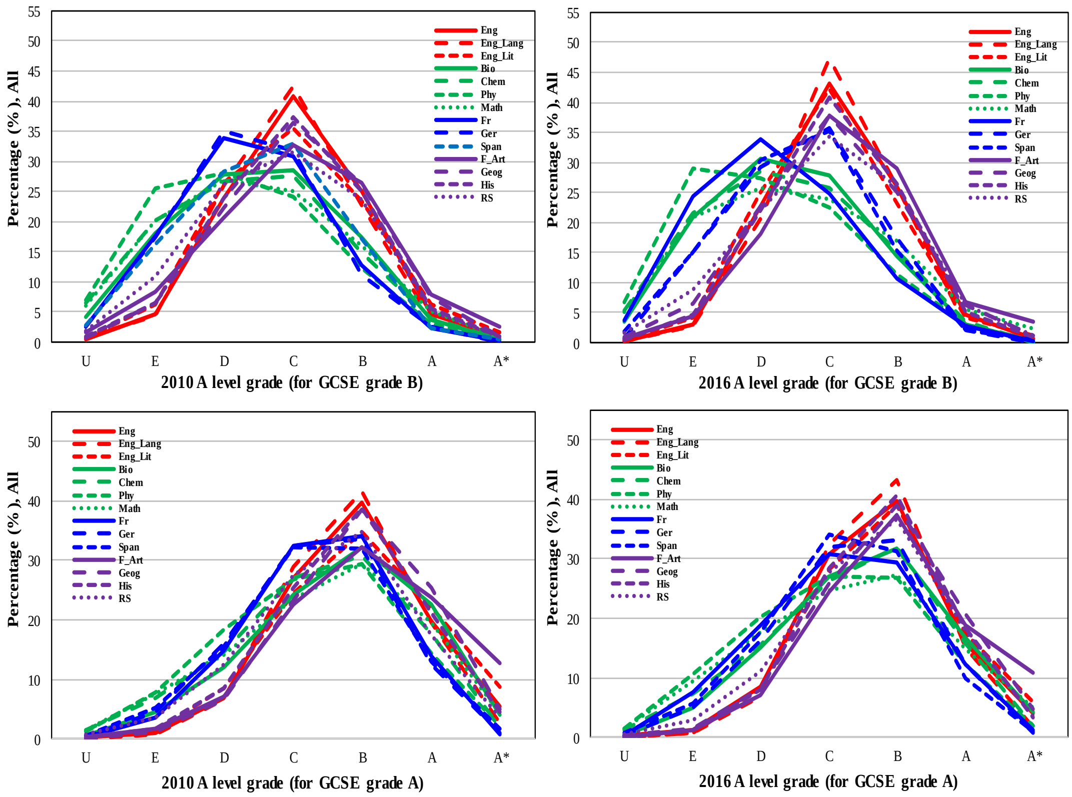


Figure 2 A level grade distributions for students who achieved grade B (top) and grade A (bottom) at GCSE in the same subject for 14 A level subjects in 2010 (left) and 2016 (right).

Table 4 Average A level grade in the subject for students receiving B and A at GCSE for 14 A level subjects (where A*=6, A=5 etc).

Subject	***Subject average A level grade***					
	Students with B at GCSE			**Students with A at GCSE**		
	2010	**2013**	**2016**	**2010**	**2013**	**2016**
English	3.02	2.96	3.07	3.86	3.77	3.72
English Lang.	2.95	2.98	3.06	3.80	3.70	3.69
English Lit	2.99	2.95	3.00	3.93	3.84	3.82
Biology	2.49	2.47	2.39	3.68	3.43	3.52
Chemistry	2.37	2.45	2.35	3.39	3.29	3.36
Physics	2.20	2.17	2.13	3.28	3.11	3.12
Mathematics	2.43	2.39	2.53	3.48	3.30	3.31
French	2.41	2.56	2.23	3.41	3.44	3.23
German	2.39	2.48	2.60	3.36	3.38	3.33
Spanish	2.54	2.49	2.55	3.35	3.36	3.23
Fine Art	3.08	3.24	3.24	4.07	4.10	3.98
Geography	3.06	3.09	3.03	3.95	3.93	3.83
History	2.99	3.07	3.06	3.82	3.80	3.77
RS	2.83	2.87	2.97	3.59	3.65	3.65

To look at how students progressed from GCSE to A level for the 14 subjects shown in Figure 2 across the range of GCSE grades, Figure 3 depicts the regression lines of A level grades against GCSE grades in the same subject. Students studying English and the social sciences generally made larger progress from GCSE to A level than those studying mathematics, the sciences and MFLs, particularly for those with low to middle grades at GCSE. As the GCSE grade moves towards higher grades, the difference in rate of progression between the subjects becomes smaller.

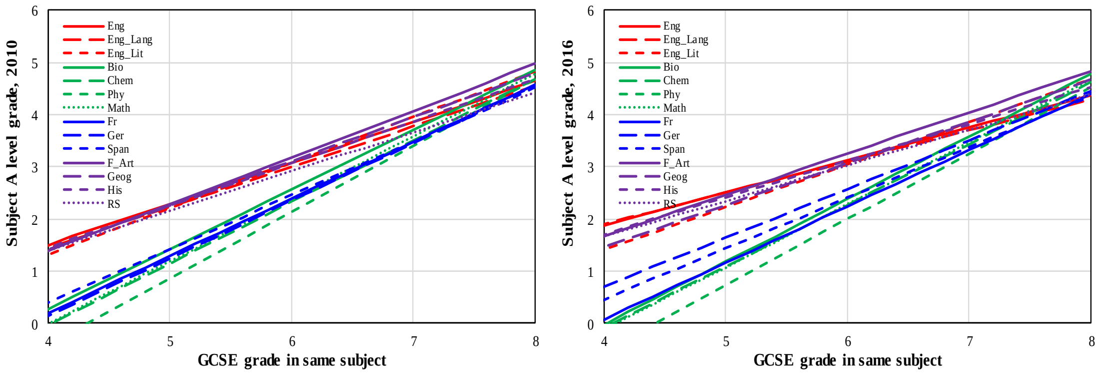


Figure 3 Regression of A level grade against GCSE grade in the same subject for the 14 A level subjects in 2010 (left) and 2016 (right).

2.1.3 Interpretation of results from statistical analysis and implications

In summary, in the case of Rasch modelling and other conventional statistical methods used to study inter-subject comparability, if we accept the assumptions involved, the differences in 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 frequently been used as evidence of inconsistency in grading standards (which are subject-specific) between A level subjects. Results from most of the analyses over the past decade or so suggested that A level biology, chemistry and physics are more difficult than most of the other subjects analysed.

In the case of CPA, again if we accept the assumptions made in the analysis, difference in the rate of relative progression from GCSE to A level would suggest inconsistency in grading standards at A level between the different subjects. Students studying the sciences showed the lowest rate of progression from GCSE to A level.

However, as is discussed below, there are limitations to the statistical techniques used to study inter-subject comparability.

2.1.4 Limitations of Rasch modelling and other statistical techniques

Coe et al. (2008) 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 also Lockyer and Newton, 2015).

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 A level 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. 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 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. 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. 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.

Because students normally take three or four A level subjects out of the 57 or so 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.

2.1.5 Limitations of CPA

For CPA, to interpret the difference in progression from GCSE to A level between subjects as difference in grading standards, it is assumed that candidates ought on average to make similar progress across subject areas, all other things being equal; and that all other things are more or less equal. However, many factors that can influence exam performance can vary substantially between subjects. Such factors can include: 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, and others, in addition to grading severity (or leniency) (also see Coe et al., 2008; Newton, 2012; Lockyer and Newton, 2015; Newton et al., 2017).

Selection of GCSE students for studying A levels

Table 5 shows the mean GCSE grades and standard deviations (SDs) of 13 GCSE subjects for all GCSE students and those who continued onto studying the 14 A level subjects discussed above in 2010 and 2016 (when calculating the mean GCSE grade, the GCSE letter grades were converted to numerical values with A*=8, A=7, B=6, … and U=0). The average GCSE grades in the same subjects for those studying the 14 A level subjects are considerably higher than those of all GCSE students, with the differences about half of a grade for English, one grade for the sciences and social sciences, and nearly two grades for the MFLs. The GCSE grade distributions for those studying the A level subjects were substantially narrower than those for all GCSE students. The average GCSE grades for those studying A level biology, chemistry, physics, mathematics and MFLs were well above 7 (GCSE grade A), substantially higher than the average GCSE grades of students studying English and the social sciences, suggesting difference in the selection from GCSE to A level between the subjects. The average GCSE scores of the students studying MFLs are also generally slightly higher than those of the students studying the sciences and mathematics. This may seem a little surprising from a learning progression perspective as the structure of knowledge and skills of the sciences and mathematics is more hierarchical than that of the languages. Differences in attainment at GCSE could also contribute to the difference in the rate of progression from GCSE to A level between the subjects.

Table 5 Average GCSE grade in the GCSE subjects for all GCSE students and those who studied the 14 A level subjects (where GCSE A*=8, A=7, B=6 etc).

A level subject	*Subject average GCSE grade*							
	2010				2016			
	All students		Studying A levels		All students		Studying A levels	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
English	5.82	1.16	6.34	0.88	5.76	1.11	6.30	0.87
English Lang.	5.82	1.16	6.37	0.87	5.76	1.11	6.33	0.85
English Lit	5.86	1.21	6.72	0.92	5.87	1.22	6.80	0.85
Biology	6.53	1.12	7.10	0.80	6.33	1.16	7.11	0.76
Chemistry	6.58	1.11	7.34	0.75	6.32	1.18	7.33	0.69
Physics	6.54	1.13	7.33	0.75	6.33	1.17	7.32	0.71
Mathematics	5.69	1.34	7.27	0.72	5.77	1.34	7.46	0.65
French	5.62	1.46	7.45	0.71	5.54	1.47	7.53	0.67
German	5.63	1.37	7.33	0.78	5.60	1.38	7.37	0.77
Spanish	5.82	1.53	7.44	0.75	5.67	1.55	7.48	0.70
Fine Art	5.99	1.28	6.83	0.96	5.93	1.32	6.94	0.96
Geography	5.86	1.45	6.74	0.99	5.78	1.48	6.78	0.92
History	5.88	1.50	6.74	0.99	5.79	1.58	6.79	0.96
RS	6.09	1.43	6.99	0.96	5.98	1.48	6.88	0.92

Figure 4 further compares in more detail the distributions of GCSE grades in the same subjects for all GCSE students and those of the students who studied English, physics, French and geography at A level in 2010 and 2016 respectively. French and physics are more selective in terms of GCSE prior attainment than English and geography. Further, French is also slightly more selective than physics.

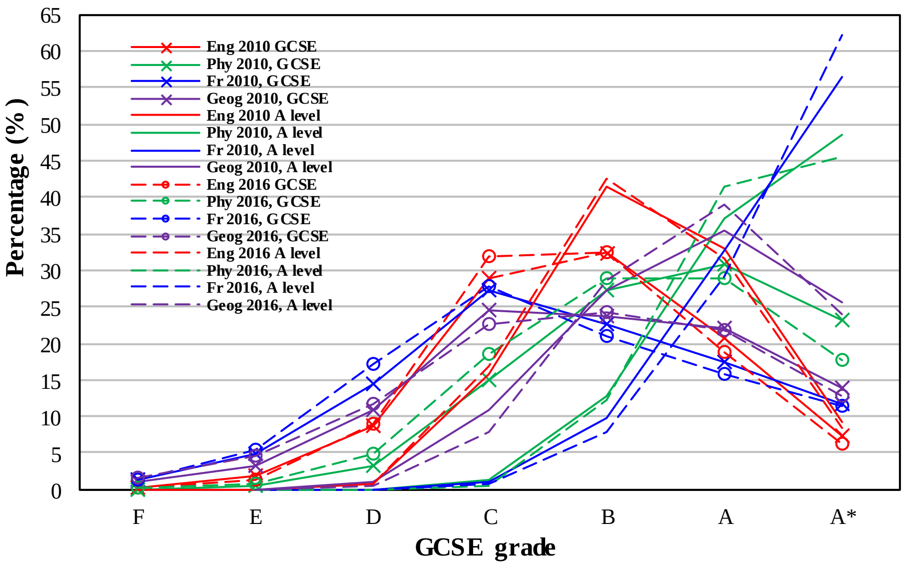


Figure 4 Distributions of GCSE grades for all GCSE students and those of the students who studied English, physics, French and geography at A level in 2010 and 2016 respectively.

Prior attainment

Figure 5 shows the distributions of mean A level scores of all students taking the 14 subjects in 2010 and 2016 and their mean GCSE scores obtained two years previously and the distributions for students with grades B and A at GCSE in the same subject. For all students, those who take the sciences, mathematics and MFLs generally achieved better outcomes at A level than those who took English and subjects in social sciences. The pattern of the distribution of mean GCSE scores is similar to that of the distribution of mean A level scores, with students taking the sciences, mathematics and MFLs having higher mean GCSE scores than those taking the other subjects. Students studying the sciences generally had considerably higher prior attainment than those studying English and the social sciences when they started their A level courses, with the average of the mean GCSE score nearly half a grade higher. Similarly, students studying the sciences achieved higher A level outcomes than those studying English and social sciences. The patterns of the distributions of mean GCSE score and mean A level grade across the subjects are similar for both years. For students who achieved grade B at GCSE, their mean GCSE scores are considerably lower than the mean GCSE scores of all students, particularly for those who studied the sciences and mathematics. The mean GCSE scores for the students who studied the sciences, mathematics and MFLs are only slightly higher than those of the students who studied English and the social sciences, but their average A level grades are slightly lower. This may suggest that students with B at GCSE who took the sciences, mathematics and MFLs generally performed similarly across their A level subjects. For students who received an A at GCSE, those who studied the sciences and MFLs had mean GCSE scores similar to those of all students, but their mean A level grades were lower. The mean GCSE scores and A level grades for students who studied English and the social sciences were slightly higher than those of all students.

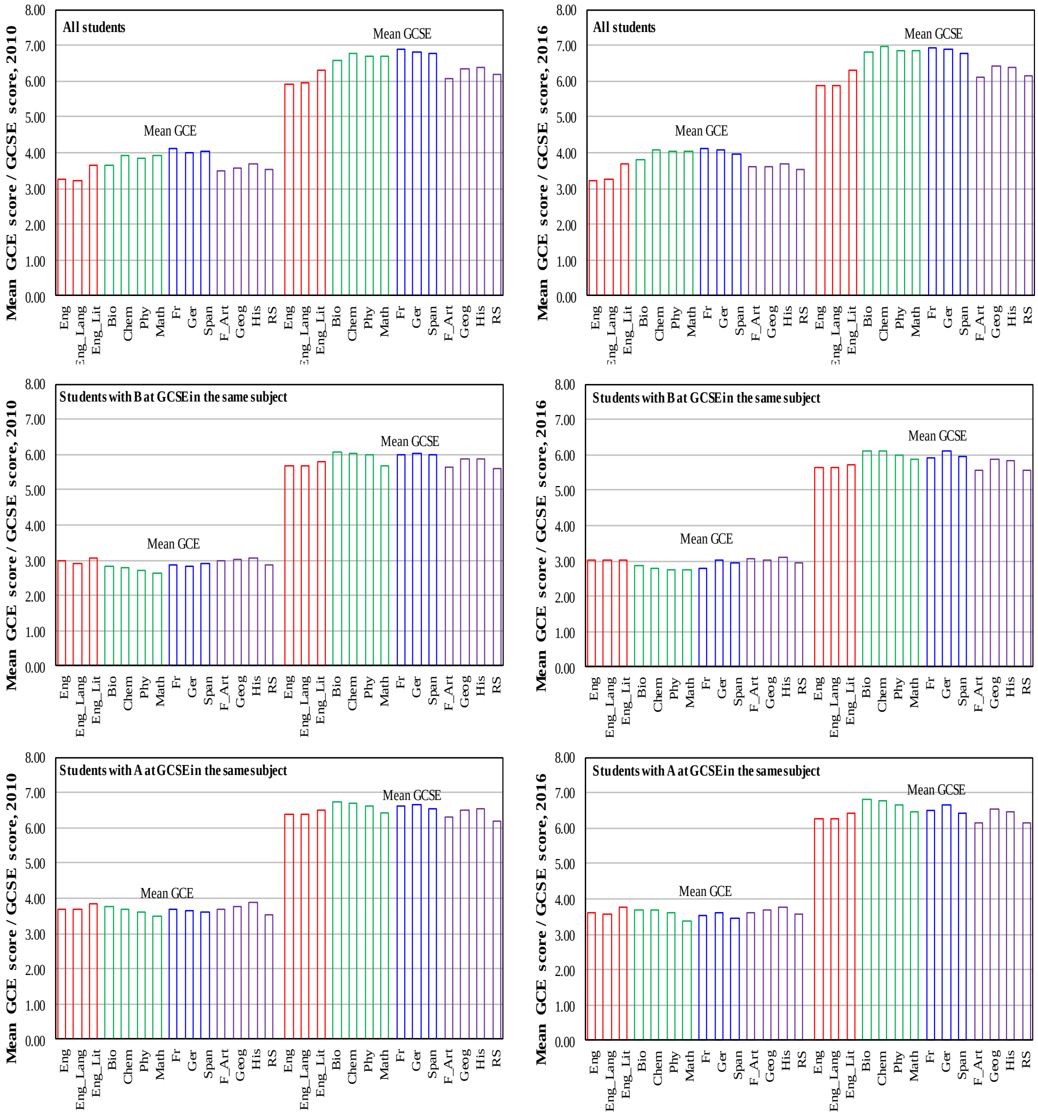


Figure 5 Distributions of mean A level scores of all students studying the 14 subjects in 2010 and 2016 and their mean GCSE scores (top), those who achieved grade B at GCSE (middle) and those who achieved grade A at GCSE (bottom).

To look at the relationship between A level outcomes and attainment at GCSE further, Figure 6 shows the mean GCSE score for all students at individual A level grades in 2010 and 2016. At individual A level grades, students taking the science subjects, mathematics and MFLs had substantially higher mean GCSE scores than students taking the other subjects, particularly at the lower grades where the difference was about half of a GCSE grade.

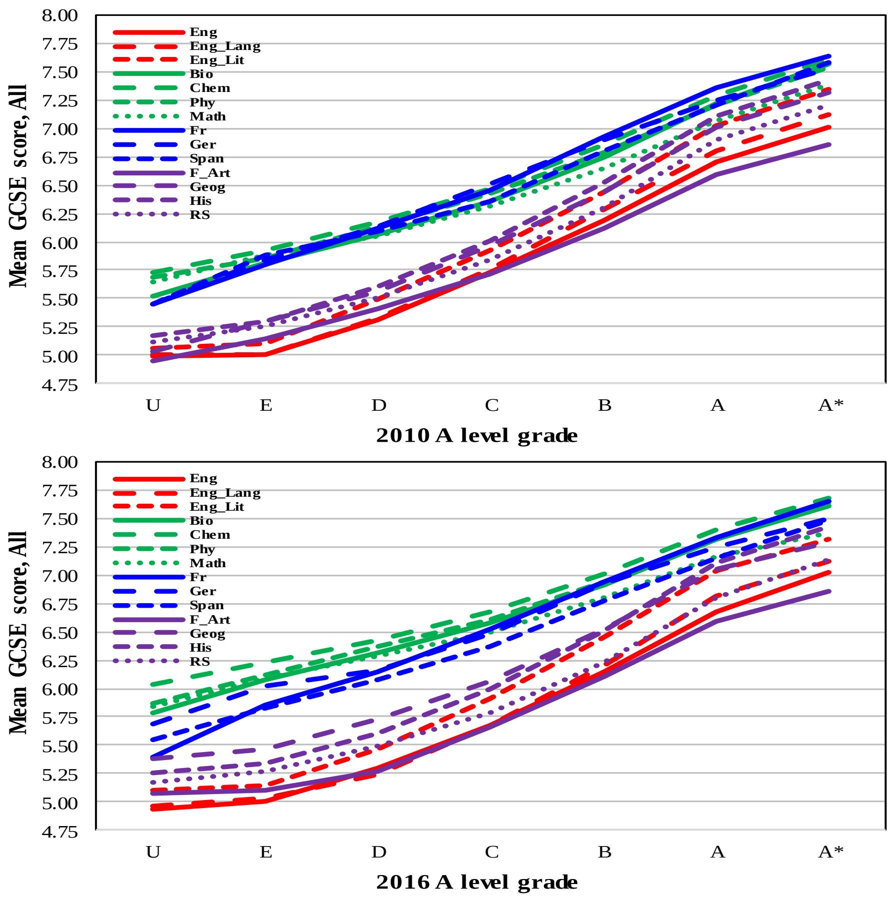


Figure 6 Average mean GCSE score at individual A level grade for all students studying the 14 subjects in 2010 and 2016.

The analysis presented above suggested that there are considerable differences in the attainment at GCSE between the 14 A level subjects. Such differences could produce different rates of progression from GCSE to A level even for students with the same grades at GCSE in the same subject.

Correlations of A level subject grades with GCSE grades in the same subject, mean GCSE score and average A level grade

Benton and Bramley (2017) demonstrated that difference in correlation between A level grade and GCSE grade in the same subject and non-random choice of subjects could produce apparent difference in the rate of progression from GCSE to A level between subjects. To look at how A level grades relate to GCSE grades and other attainment measures, Table 6 below shows the correlations between A level scores and GCSE scores, mean GCSE scores, and mean A level scores for the 14 subjects from 2013. The correlations between A level and GCSE scores in the same subject were moderate, varying from 0.54 for Religious Studies to 0.62 for Geography. The correlations between A level scores and mean GCSE scores are slightly higher than the correlations with GCSE subject scores except for German, Spanish and fine art for which the correlations are slightly lower. This suggests that for most of the subjects, mean GCSE score rather than grades at GCSE in the same subject is a better predictor of performance at A level. As expected, the correlations between A level subject scores and mean A level scores are considerably higher than those with GCSE subject scores or mean GCSE scores.

Table 6 Correlations of A level subject score with GCSE score in the same subject, mean GCSE score and mean A level grade for the 14 subjects from 2013.

Subject	*Correlation of A level subject score with*		
	GCSE score for the same subject	**Mean GCSE score**	**Mean A level score**
English (Eng)	0.57	0.66	0.83
English Language (Eng_Lang)	0.57	0.66	0.84
English Literature (Eng_Lit)	0.57	0.73	0.87
Biology (Bio)	0.60	0.69	0.89
Chemistry (Chem)	0.57	0.64	0.91
Physics (Phy)	0.59	0.66	0.92
Mathematics (Math)	0.56	0.59	0.88
French (Fr)	0.59	0.62	0.84
German (Ger)	0.64	0.53	0.82
Spanish (Span)	0.56	0.53	0.82
Fine Art (F_Art)	0.61	0.59	0.83
Geography (Geog)	0.62	0.70	0.87
History (His)	0.59	0.69	0.88
Religious Studies (RS)	0.54	0.65	0.87

2.1.6 Potential causes of difference in statistical difficulty between subjects and link to grading standards

As demonstrated above, there are considerable differences in the difficulty indices at individual grade level and the overall subject level between A level subjects derived using different statistical methods. Such variability in difficulty has remained almost the same over the past decade or so. The differences in difficulty indices between subjects have frequently been interpreted as differences in grading standards. 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. Coe (2008) and Coe et al. (2008) discussed a range of potential causes for such differences. These include differences in grading severity (or leniency), nature of the subject in terms of skills and knowledge to be learnt, level of subject demand, allocation of teaching time and other resources, motivation of students, efficiency and effectiveness of teaching and learning, and others (also see Newton, 2012; Lockyer and Newton, 2015). These factors are not considered in the conventional statistical methods and in the CPA approach.

It has also to be recognised that inter-subject comparability in A levels and other examinations has been a matter of debate for many decades. 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 such debate. Those who are against the use of statistical methods to compare different subjects argue that examinations such as A levels 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 as long as there is a theoretical basis for the analysis and the interpretation of the results is justified, statistical comparisons would still be appropriate and meaningful. They further argue that subject standards may need to be statistically aligned when grades from different subjects are used for specific purposes, particularly when they are used interchangeably or as equivalent currencies in situations such as admissions to certain university courses and within school accountability measures.

2.2 Judgemental evidence

Judgemental methods used to investigate inter-subject comparability generally involve the use of experts to compare 1) the perceived level of performance of students on the exam (or quality of students' work) and 2) the perceived level of demand of the specification and other assessment materials between the subjects (see Bramley, 2011; Lockyer and Newton, 2015). Unlike statistical methods, judgemental methods look at the consistency of grading standards between subjects directly. The difficulty with judgemental methods is that it is difficult to establish cross-curricular performance criteria that can be used for inter-subject comparison and to find experts that are qualified to make such cross-subject comparisons (Jones et al., 2011; Lockyer and Newton, 2015). As a result, there has been very limited evidence of inter-subject comparability generated using the judgemental approach.

Research carried out by the Qualifications and Curriculum Authority (QCA) in 2008 suggested that A level physics, chemistry and biology were generally well aligned to one another in performance standards:

> "The review of performance standards revealed no substantial or consistent differences in standards of performance between any subjects at any level.… At AS, the differences were again very small, with biology and chemistry almost indistinguishable and performance in physics judged to be marginally less impressive. At the lower grades in A2, the picture was again one of great consistency, with all three subjects almost perfectly aligned. At the higher grades, however, some divergence was found, with biology candidates performing slightly less well in order to achieve their grades than the chemistry candidates, with the physics candidates in between." (QCA, 2008)

The study also found that standards in biology and psychology were very well aligned across the grade range in both the AS and A2 examinations in terms of the demand of the examinations and the grading standards set:

> "Overall the analysis suggested that standards in biology and psychology were very well aligned across the grade range in both the AS and A2 examinations." (QCA, 2008)

Limitation of subject experts-based judgement

As discussed in Lockyer and Newton (2015), although judgemental methods could in theory be used to study inter-subject comparability, it is hard to construct the agreed cross-curricular performance criteria that could be used by the judges as basis to compare subjects and to make decisions. Even if such criteria can be constructed, they are likely to be too ambiguous to be usable or too constricted to reflect the core of any of the subjects to be compared (Wiliam, 1996; Coe, 2010). Although direct comparison of performance on exam tasks between subjects may not require such explicit cross-curricular criteria, judges would still need to compare students' work from different subjects on the basis of something like those criteria, even if they were unable to articulate the basis for their decisions explicitly. In summary, judgemental methods rely on the ability of experts to make inherently complex, and necessarily ill-defined, subjective decisions. Conclusions from judgemental studies can therefore be highly equivocal.

2.3. Evidence of the impact of grading severity on society

In their correspondence with Ofqual, science stakeholders have identified several negative effects which they believe are impacting on students' choices at key stage 5, and which they attribute to severe grading (see letter to Ofqual from Stevenson, Britton, McLeish, Pritchard and Price, 2016). Foremost amongst these is the assertion that perceived or actual grading severity is limiting uptake of the subject at A level, but the learned societies also point to a deterrent effect upon female students, those with lower prior attainment, and students from less advantaged backgrounds – groups that they feel are currently under-represented in A level sciences.

The Joint Council for Qualifications publish a detailed report on entry trends for GCEs following each summer series, which includes data on how entries in a particular subject have changed over time, and a breakdown of entries by gender. If present, we might expect to observe the effects claimed by the learned societies within the entry data – although if we accept the contentions of stakeholders about the historic nature of grading severity in A level physics, chemistry and biology, then we would find no evidence of the negative impact in entry data as the misalignment would have been carried forward for a significant period of time (possibly since the inception of the qualification).

Furthermore, if evidence were to be found it is not clear how strongly we could establish causation. There may be other factors present which could be producing similar effects, potentially either compounding those identified by stakeholders or having a direct impact on entries which is being misattributed to severe grading.

The learned societies have also argued that there are some impacts arising from severe grading which by their nature cannot be observed from entry data at all, regardless of the potential timescale over which this negative impact may have been in effect. This includes 'missing' candidates who would have studied the subject were it not perceived (rightly or wrongly) to be more difficult than others.

To some extent these missing entries might be inferred through comparison with other subjects which appear to be less 'difficult' under statistical measures. These are weak inferences however given the range of other factors which may be influencing subject choice, and require us to accept the argument (which, to some extent, conflicts with the findings of Cuff (2017)) that perceived difficulty overrides students' enjoyment and interest, and the utility of the subject for further study or preferred career, when selecting options at A level.

In looking at JCQ entry data, we have considered the significance of relative subject entries within a specific year (in comparison to other subjects which appear to be statistically less 'difficult') and of entry trends over time, to establish whether there is evidence of a persistent issue.

2.3.1 Intra-year A Level Entry Data

JCQ entry data from summer 2017 in Table 7 shows that physics was the ninth most popular A level subject out of the 39 reported subject categories, taken by 4.4% of the total entry (36578 students). This is consistent with the position of physics in summer 2016, and represents an increased share of the total entry of 0.2% since the previous year. Of the natural science A levels, physics is the least popular – but still has an entry comparable to geography, which the Institute of Physics specifically identify as an example of a subject which is likely to be more attractive to students because of 'lenient' grading. Physics entries grew by 3.5% in 2016, a faster rate of increase than in mathematics. In 2018 physics entries rose to 37806 candidates, 4.7% of the total entry.

Table 7 Top A level subjects in 2017 by entry (Joint Council for Qualifications, 2017).

Ranking		Subject	% of total	Number of candidates
1	(1)	Mathematics	11.5	95244
2	(2)	Biology	7.5	61908
3	(3)	Psychology	7.1	58663
4	(5)	Chemistry	6.3	52331
5	(4)	History	6.1	50311
6	(6)	English Literature	5.6	46411

7 (7)	Art and Design Subjects	5.3	43653
8 (8)	Geography	4.6	37814
9 (9)	Physics	4.4	36578
10 (10)	Sociology	4.2	34607

(2016 ranking in brackets)

The popularity of chemistry has increased since 2016, rising from fifth to fourth most studied A level subject. 52331 students entered A level chemistry in 2017, representing 6.3% of the total A level entry. Chemistry entries were comparable to those for history in 2017, which Rasch analysis suggests is of average difficulty. Entries have since risen again in 2018, to 54134 candidates.

Biology remained the second most popular A level subject in 2017, sat by 61908 students making up 7.5% of the total entry. Entries for biology are significantly higher than those for geography and English literature, subjects which appear to be average difficulty under Rasch analysis, and therefore substantially more lenient. In 2018 entries rose again, to 63819 candidates.

Compare these entry trends with those of A levels in performing / expressive arts, English language and literature, and communication studies shown in Table 8. These are amongst the most leniently graded subjects under Rasch analysis, but saw decreases in entry ranging from 17% to 10.6% between 2016 and 2017. Note it is not just arts subjects which are seeing an overall decline in entry – history and English literature were also less popular in 2017 than the previous year.

Table 8 A level subject decreases 2016 to 2017(Joint Council for Qualifications, 2017).

Ranking	Subject	% change	2016 candidates	2017 candidates
1 (1)	General Studies	-39.2	11759	7147
2	Communication Studies	-17.0	1851	1537
3 (10)	Other Sciences	-14.0	3304	2840
4 (12)	ICT	-12.9	8737	7607
5 (2)	English Language & Literature	-11.1	12438	11058
6 (3)	Performing / Expressive Arts	-10.6	2179	1948
7 (21)	English Language	-10.2	23575	21178
8 (4)	Music	-9.3	7089	6428
9 (3)	History	-8.1	54731	50311
10 (11)	English Literature	-4.7	48697	46411

(2016 ranking in brackets. Subjects which had a year-on-year increase in 2016 are shown as blank)

Within their cognate group, the popularity of physics, chemistry and biology correspond to their relative difficulty under Rasch analyses, with physics appearing the most difficult and biology the least difficult of the three.

2.3.2 Inter-year A level entry data

Figure 7 shows changes in percentage share of total entry in these A level subjects since 2008. Entries for physics increased steadily until 2014, before dipping slightly in 2015 and 2016. In 2017 entries returned to the level previously seen in 2014.

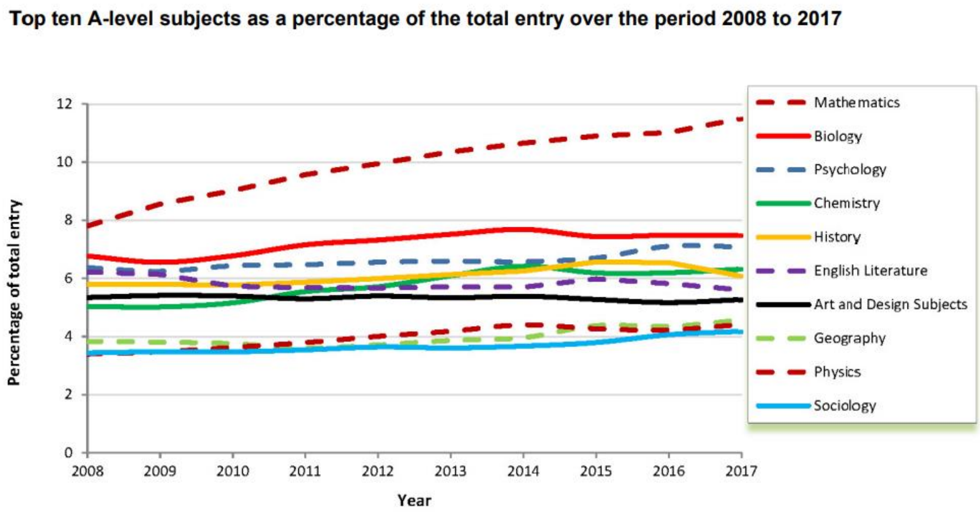


Figure 7 Top ten A level subjects as a percentage of the total entry over the period 2008 to 2017 (Joint Council for Qualifications, 2017).

Chemistry has seen a similar trend in entries to physics, although the increase in 2008 – 2014 was more pronounced and the recovery in numbers in 2017 less.

Percentage entry share for biology also rose 2008 – 2014, but saw a slight decrease in 2015 and has stayed relatively stable in 2016 and 2017.

Overall entries in all three subjects appear to be on an upward trend, and though the increase in entries has not been as pronounced as in mathematics, they do not show the relative decline 2008 – 2017 seen in potentially more leniently graded subjects such as English literature.

Figure 8 shows the total number of entries for physics, biology and chemistry over the same period. Again total entries seem in line with the apparent difficulty of these subjects under Rasch analysis, but overall trends in entry do not.

4 GCE A-level sciences, 2008-2017

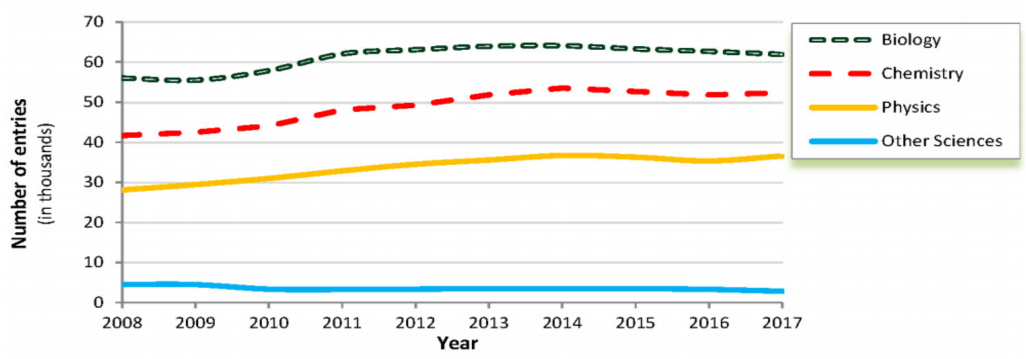


Figure 8 A level sciences 2008 to 2017 (Joint Council for Qualifications, 2017).

The most consistent increase in numbers of students has been in physics, but entries still remain the least overall. Whilst this trend was interrupted 2015 – 2016, total entries have now returned to those recorded in 2014, making physics the only subject of the three where subject entries appeared to be increasing in 2017 (although this was the case for all three subjects in 2018).

The greatest overall increase in entries since 2008 has been in chemistry, but numbers declined slightly in 2015 and after a period of relative stability have now started to increase again from 2017.

The smallest overall increase since 2008 has been in biology, although it still retains the largest total entry. Since 2014 entries declined slightly each year, but JCQ data for 2018 indicates that they have once again begun to increase

When considering the trends in A level entry outlined above, we must be mindful of some of the limitations of what this data can tell us. Whilst they may be relevant to establishing whether any potential societal impacts are persistent, they are not directly interpretable in terms of the alleged impact of grading severity because that alleged severity is claimed to have been present (and unchanged) for a much longer period – potentially since the conception of A levels as a subject. This exceeds the historic entry data published by JCQ and available for consideration in this report. Recent entry trends are relevant to consider in this context however, in that they may illustrate trends which run counter to the arguments of stakeholders about the negative impacts on A level physics, chemistry and biology of grading severity.

Furthermore, stakeholders also suggest that inaction over inter-subject comparability in recent years is serving to reinforce the perceived difficulty of these subjects through positive feedback from selection and self-selection. If this is the case, it might be something which we would expect to see evidence of in entry trends.

It is also notable that, year-on-year, entries for A level physics are lower than those for A level science subjects which are apparently less severely graded under Rasch and

Comparative Progression Analysis. If we are not convinced that recent entry trends are illustrative of negative impacts on the subject because of the contention by stakeholders that severe grading is persistent and historic, then we might assign greater weight to the fact that entries for this subject have generally been lower than others which are apparently ‘easier’ – and the fact that relative cohort size in A level sciences mirrors the relative difficulty of these subjects under statistical measures.

It is important to bear in mind that these figures do not tell us about any changes in the type of students making entries for these subjects over the period shown. For instance, they do not tell us anything about the GCSE attainment range of A level science students in a particular year, and whether this is narrowing over time as stakeholders have asserted. Nor do they tell us about changes in the relative numbers of students within the cohorts from higher or lower attainment bands. Consideration of trends over time in the relative prior attainment profiles of students in these subjects is presented in the next section of the report.

Note that the JCQ figures given above do not take into account AS entries.

2.3.3 University uptake

Stakeholder concerns about potential severe grading in A level science do not focus on access to university in the way that those expressed about modern foreign languages do. Whilst the learned societies are concerned about the number of students – especially female students or those from disadvantaged backgrounds, who they regard as making more conservative choices at A level – being put off science subjects entirely by perceived severe grading, their arguments in regard to physics in particular have been primarily about the number of students who would benefit from studying an A level in the subject but are dissuaded because of the view that the subject is one suitable *only* for those hoping to study it at university.

In this category the Institute of Physics have cited the example of those students who might go on, for example, to become nurses and technicians, courses where a C grade at A level would still reflect an understanding of physics which would be useful within their intended careers, but who are dissuaded from studying the subject due to the high number of A and A* grades awarded. In turn, stakeholders attribute the higher number of A and A* grades to the effects of selection resulting in a combination of a higher average prior attainment in the subject and a loss of discrimination amongst the top grades.

Nonetheless, as depressed university uptake is one of the criteria on which we might judge that a compelling case exists to adjust grading standards, data from UCAS on progress to university study in these subjects is presented in this report.

The Universities and Colleges Admission Service (UCAS) publish annual reports which provide data on trends in applications and acceptances within certain subject areas, reported under JAC3 – the Joint Academic Coding System. The subjects within the scope of this report sit within a range of JAC3 reporting groups. The most obviously relevant to A level science are Group C Biological Sciences and Group F Physical Sciences, but also the range of university disciplines for which at least one science A level might be expected of applicants, alongside another subject such as maths. This includes JAC3 reporting groups such as Group A Medicine and Dentistry, Group B Subjects allied to Medicine, and Group H Engineering. Figure 9 shows the overall UCAS entries by subject group in 2017.

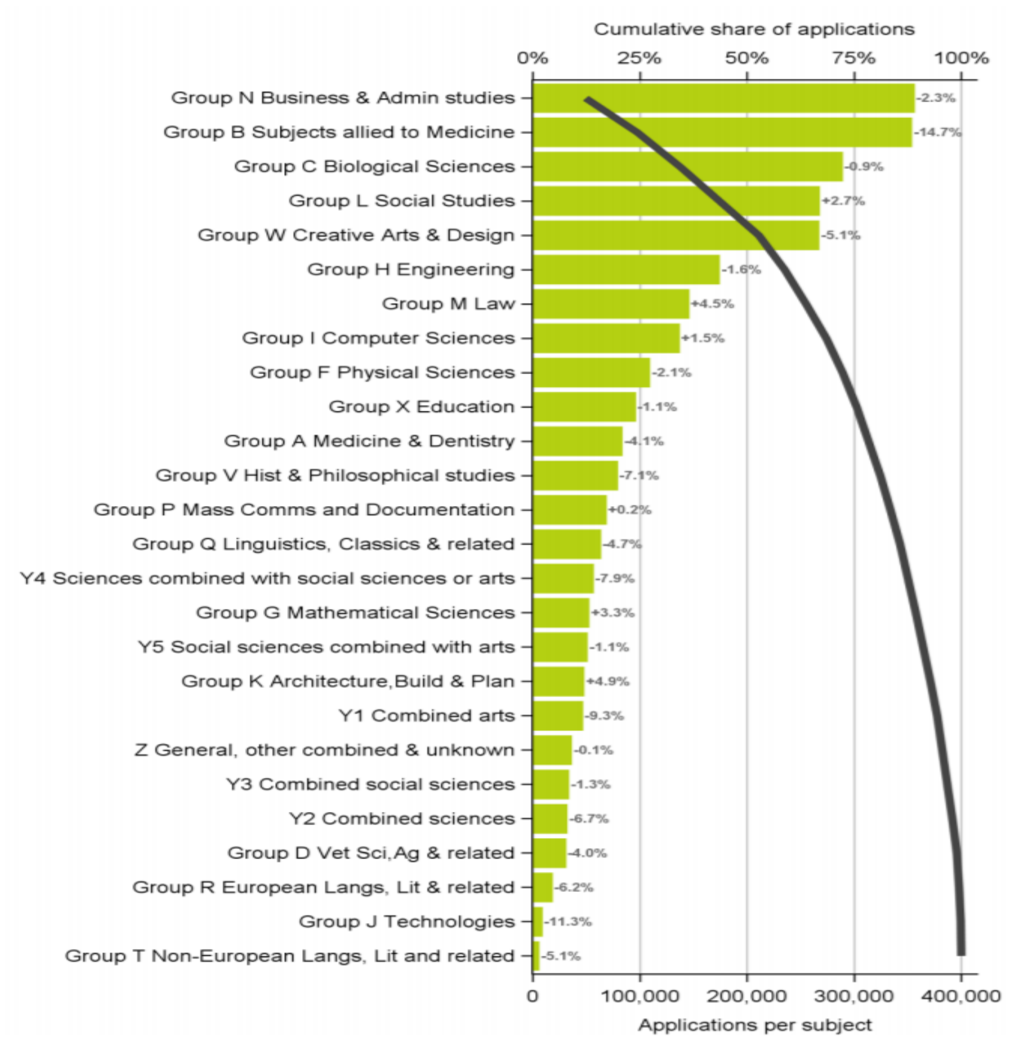


Figure 9 Applications by subject group in 2017 (with cumulative share of applications, and percentage change in applications from 2016 (UCAS, 2017)

[Note bars show number of applications per subject, line shows the cumulative share of applications, and annotations show percentage change in applications from 2016.]

All of these JAC3 groups saw a slight decrease in applications from 2016 – 2017, with the exception of Group B Subjects allied to Medicine where the fall in acceptances was more significant. It should be noted however that most groups saw a reduction in applications in 2017, with a decrease of 2.6% in acceptances overall. These JAC3 groups still all sit amongst the subjects on the table with a comparatively greater number of applications, particularly when compared to JAC3 groups such as Group R European Langs, Lit and related. Entries for Group B and Group C in particular represent a substantial proportion of the total. These figures certainly do not seem to be suggestive of any serious decline in the number of students progressing to study STEM subjects at university. Indeed, the decline in the number of students taking Group C and Group B students may be regarded as a return to a more normal level following a surge in applications in recent years. This becomes clear on looking at longer term trends in UCAS data.

UCAS also report data on the proportional change in acceptances by subject group over a period of ten years, from 2008 to 2017. This data is shown in Figure 10.

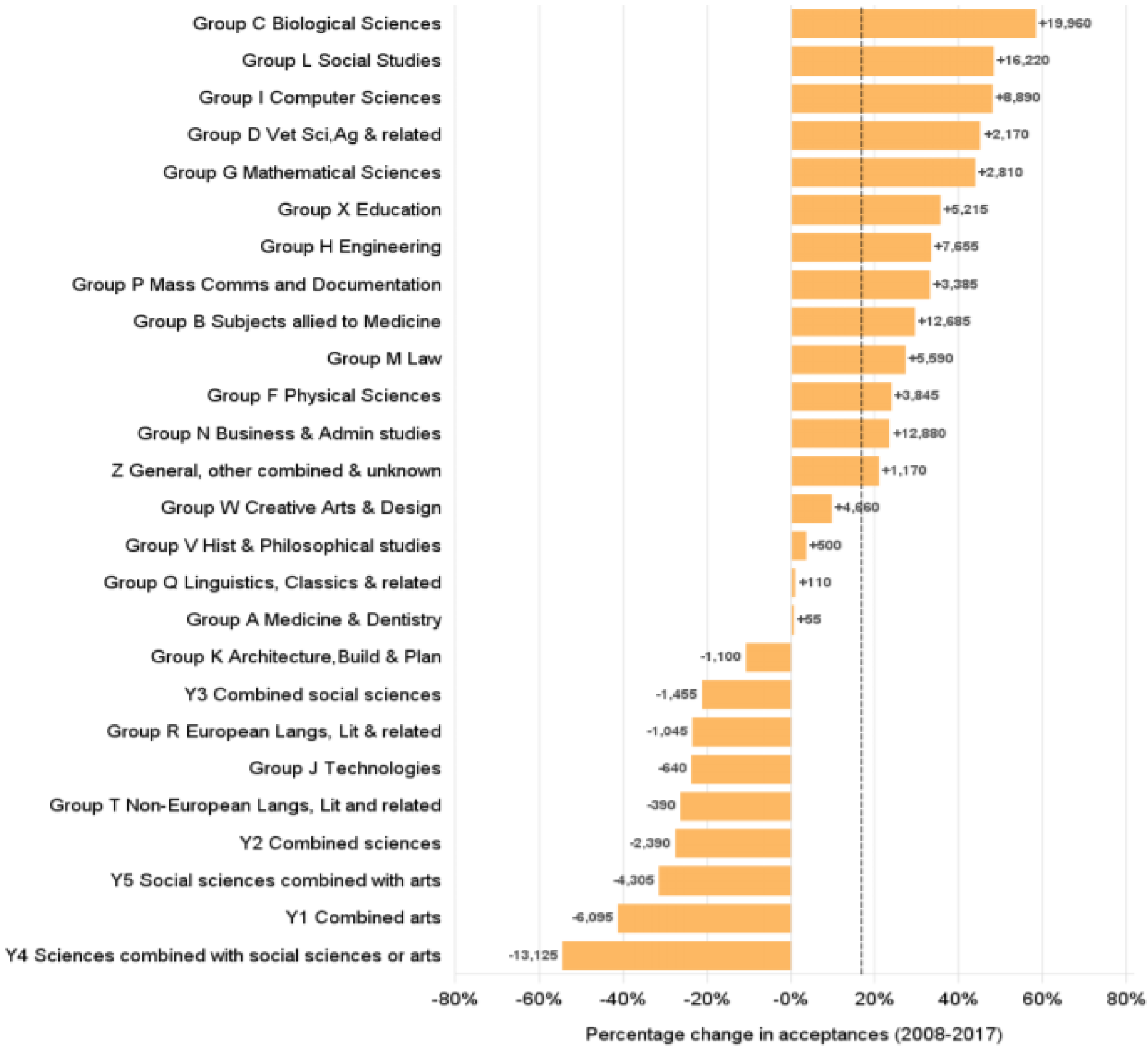


Figure 10 Proportional change in acceptances by subject group, 2008 to 2017 (UCAS, 2017).

[Note that the dotted line is the proportional increase for all subject groups (+17%). Annotations show the numerical change in acceptances between 2008 and 2017.]

With the exception of Group A Medicine and Dentistry (where the number of undergraduate medical school places is regulated by the government), the majority of JAC3 groups likely to require A level science as a condition of study have shown significant increases in the number of acceptances from 2008 – 2017. This increase has been particularly pronounced in Group C Biological Sciences, but Group H Engineering and Group F Physical Sciences have all seen an increase in acceptances which exceeds the proportional increase for all subject groups. As might be expected, this data shows that as entries for science A levels have increased since 2008 so have acceptances for related university courses.

2.3.3 Evidence of filtering (self-selection and school selection)

One of the arguments advanced by science stakeholders is that selection is occurring within A level physics, chemistry and biology which is preventing certain students from studying these subjects. Two different types of selection are claimed – informal self-selection by students who are opting not to study the subject because of perceived difficulty, and formal selection in the form of subject entry requirements introduced by schools to limit entry to students of only a certain prior attainment.

Tom Allen (2016), data analyst at the Institute of Physics, has sought to illustrate this effect using 2014 entry data from the National Pupil Database to investigate the number of school sixth forms where only students obtaining GCSE A* and A in a subject progress to study it at A level. His findings are presented in Figure 11.

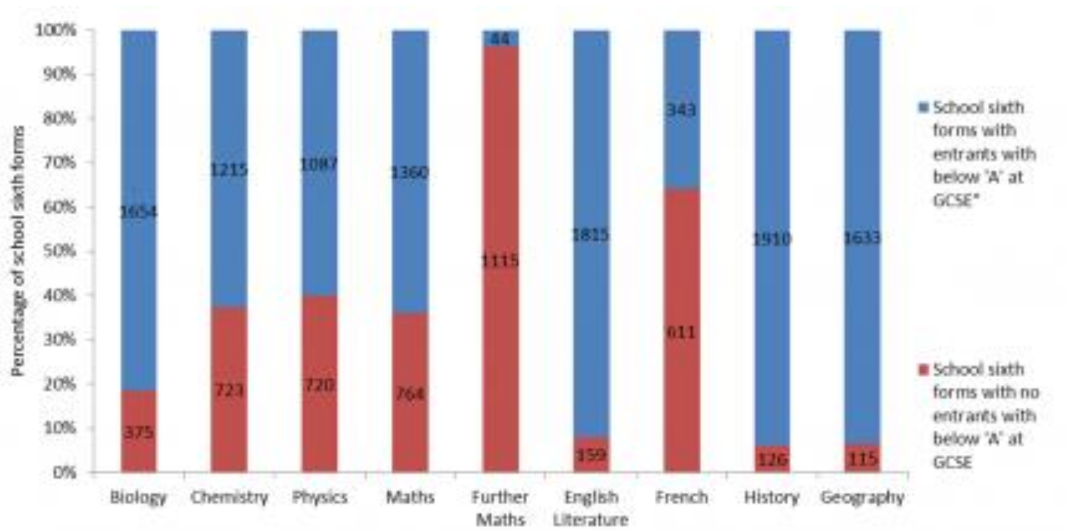


Figure 11 Percentage of school sixth forms with no A level entrants obtaining less than a GCSE grade A (Institute of Physics, 2016).

Whilst the lack of entries from students who achieved below A in a subject at GCSE in some schools may not be down to perceptions of difficulty, the significant number of schools where this occurs for STEM subjects compared to so-called ‘facilitating’ humanities subjects such as history and geography has been suggested to us as evidence that there is an external factor at work.

The data shows that for A level physics, in 2014 approximately 40% of school sixth forms had no entrants who had attained less than a grade A at GCSE.

The figure for chemistry is similar, with 37% of schools having no students progressing to A level study on less than a GCSE A grade.

For biology the number is approximately half that, with only 18% of sixth forms having no entrants with a grade below A at GCSE. This is still greater than those for English literature, history and geography however, at 8%, 6% and 7% respectively.

In A level mathematics, which appears easier than physics and chemistry under Rasch analysis, the percentages were comparable (36% for maths). For biology, which appears slightly harder than maths, the evidence of potential selection is not as strong.

Note however that the selection effect, if present, appears to be significantly more pronounced in A level French than in either physics, chemistry or biology, with 64% of schools having no entrants who had achieved less than an A.

Further insight into the possible effects of selection upon the cohorts of students sitting A level physics, chemistry and biology can be gained by looking at trends over time in the relative prior attainment profiles of students in these subjects. Were self-selection or school selection taking place, and if as stakeholders contend, this filtering was exacerbating the grading severity of science subjects and contributing to their perceived difficulty, we might expect the range of prior attainment to narrow over time due to relatively fewer A level entries from low prior attainment bands, as less able students opt out of these subjects (or are opted out by school selection policies). Table 9 shows year-on-year trends in GCSE distribution for science A level cohorts and their mean grade in the same subject at GCSE.

Table 9 Mean and SD of GCSE distributions for A level physics, chemistry and biology from 2008 to 2016 and their mean grade in the same subject at GCSE.

Subject	Year	Mean GCSE		Mean grade in subject	
		Mean	SD	Mean	SD
Physics	2008	6.648	0.818	7.26	0.82
	2010	6.677	0.791	7.33	0.75
	2013	6.786	0.741	7.38	0.73
	2015	6.811	0.744	7.35	0.73
	2016	6.861	0.717	7.32	0.71
Chemistry	2008	6.750	0.802	7.22	0.83
	2010	6.772	0.776	7.34	0.75
	2013	6.901	0.717	7.40	0.72
	2015	6.921	0.717	7.36	0.72
	2016	6.984	0.684	7.33	0.69
Biology	2008	6.582	0.792	7.02	0.87
	2010	6.595	0.767	7.10	0.80
	2013	6.746	0.718	7.16	0.78
	2015	6.773	0.707	7.13	0.77
	2016	6.827	0.687	7.11	0.76

The data shows that the mean GCSE distribution in all three subjects has increased since 2008, and that the distribution has narrowed slightly over this period. The mean grade in the same subject at GCSE also narrowed slightly.

2.3.4 A level entry gender breakdown

Another argument advanced by science stakeholders is that some groups are more likely to be deterred by the perception that a subject is ‘difficult’ than others. In particular, it has been claimed that the severe grading of physics, chemistry and biology is harming efforts to encourage more girls to study those subjects at A level.

JCQ have published data on gender participation by A level subject since 2007. Figure 12 shows the gender profiles of the cohort sitting A levels in 2017.

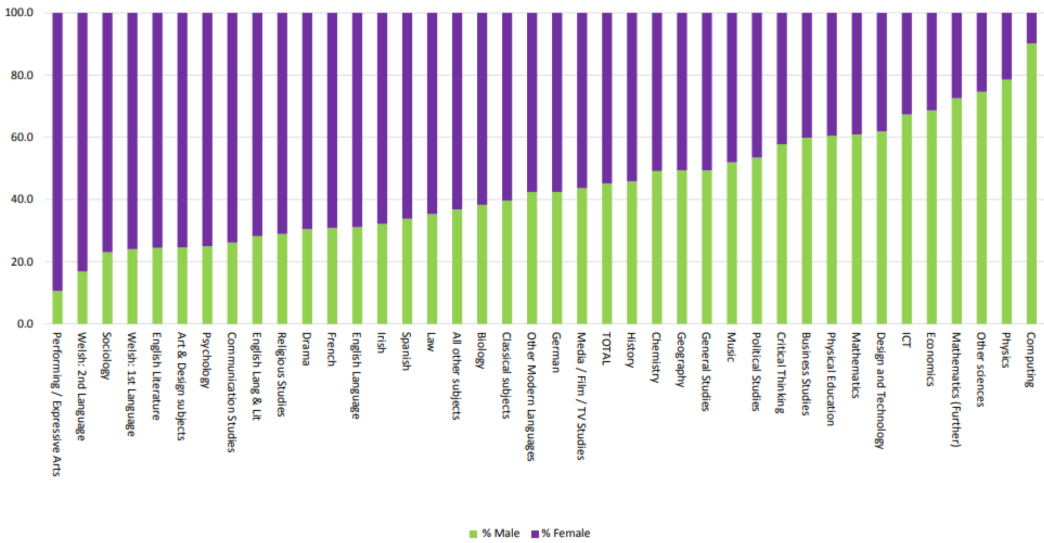


Figure 12 Differences between male and female A level subject choices 2017 (Joint Council for Qualifications, 2017).

Physics was the subject with the second highest ratio of male-to-female students in 2017. This is consistent with the male-to-female ratio in 2016, 2015, and 2014, although female participation has increased slightly (1%) since 2013. In fact, the JCQ data shows that the percentage has remained more or less consistent since 2007 – despite the increase in entry from 2008 onwards. Note that in 2017 the male-to-female gender ratio in physics was higher than in further maths, which is the A level subject which appears to be hardest according to Rasch analysis.

Chemistry entries are much more balanced, with girls making up just over half of the students studying the subject at A level. This percentage has also remained consistent in recent years according to historical JCQ data, with female participation having increased only slightly since 2007. Nonetheless, the male-to-female ratio in chemistry is still less than that in the A level cohort as a whole.

In stark contrast to physics, over 60% of students studying biology in 2017 were female, and historical JCQ data shows the proportion of female students has remained largely unchanged in recent years. Whilst participation has increased since 2007 this is only by approximately 6%, and the skewed entry profile towards female students is present as far back as JCQ publishes data. The male-to-female ratio in biology is less than that of the overall A level cohort.

Physics is a clear outlier in terms of female uptake. Focusing specifically on the gender imbalance in physics, in May 2018 the Institute of Physics published the report which looked at progress made since 2012 in attempting to encourage more girls to study the subject at A level (see Institute of Physics, 2018).

The study found that, despite performing just as well as boys in physics at GCSE, in 2016 only 1.9% of girls chose A level physics compared to 6.5% of boys. The differences in uptake between boys and girls were narrower in chemistry and maths, and in biology more girls studied the subject than boys – 9.5% of girls and 5.6% of boys – although again the gender imbalance is less than in physics. Furthermore, the report found that in 44% of schools no girls chose to study physics at A level at all, as opposed to 28% of schools where no boys chose to study the subject.

The report also looked at the extent to which GCSE performance, and in particular attainment in physics relative to other GCSE subjects, might influence girls' subject choices. It found that the fact girls perform better in all of their GCSEs on average than boys means that even though they get a good grade in physics, they also obtain high grades in other subjects. This means physics is less likely to be in their top four subjects by attainment (65% for girls compared to 81% for boys). What is more, even where a student does have physics in their *top four results, boys are three times more likely to progress to A level physics than girls. Figure 13 shows the percentage of GCSE triple science students progressing to A level study in each subject.*

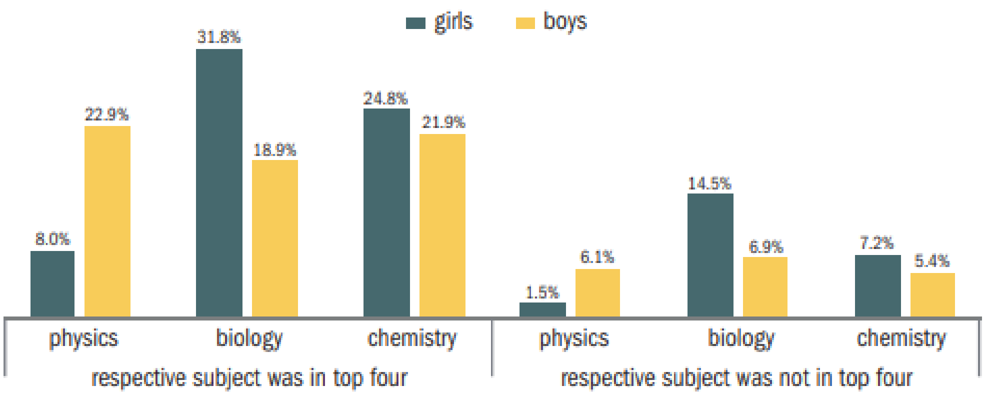


Figure 13 *Percentage of boys and girls progressing to A level physics, biology and chemistry from separate sciences (Institute of Physics, 2018).*

Whilst overall both girls and boys were more likely to study an A level science if it was amongst their top four grades at GCSE, girls were far less likely to opt to study physics than they were either biology or chemistry. For boys, there was a much more even distribution of subject choices between the three A levels. Notably, girls were more than twice as likely to choose A level biology when it wasn't in their top four grades, as they were to choose physics when it was in their top four grades. Both where a respective subject was in their top four and where it was not, girls' A level science choices aligned with the apparent relative grading severity of these subjects under statistical measures of subject difficulty. This was not the case for boys.

An interesting perspective on issues of gender within STEM subjects is proposed by Stoet and Geary (2018), which found that countries with greater gender equality see fewer female students pursuing careers in certain STEM subjects – despite possessing similar or greater aptitude for these subjects than males. Hence gender differences do not necessarily reflect grading issues. There are a range of other factors which could, individually or through their interaction, be impacting on the uptake of A level sciences by girls. Some of these are outlined in the report by the Institute of Physics referenced above.

2.3.5 Teacher supply data

One of the recommendations made in the advice submitted to the Ofqual Board by the Standards Advisory Group[3] was that, in addition to evidence from statistical measures of difficulty and evidence of the concerns of stakeholders, the basket of evidence considered to determine whether an adjustment is warranted in a particular subject should also include contextual data such as figures on teacher supply.

Stakeholder concerns about teacher supply within the subjects included in the scope of our work on inter-subject comparability are very different. In A level French, German and Spanish, uptake is declining significantly and has been for some time, with stakeholders expressing fears that this is contributing to a shortage of teachers qualified to deliver the subject (particularly dual linguists), which in turn will lead to fewer schools offering these subjects at A level. They are also concerned that a teacher shortage will contribute to the perceived 'difficulty' of A level French, German and Spanish through a decline in overall teaching quality (though it should be noted that the interaction between teacher shortages and department closures will only lead to a decline in teaching quality if less qualified teachers are appointed to roles that more qualified teachers have vacated: if fewer schools offer the subject but the remaining schools have similarly qualified teachers, then this will have no overall impact on teacher quality). These concerns are more immediate than those seen in the sciences, due to the rate at which entries are declining. In this case, we might consider declining numbers of teachers to be possible evidence of a negative impact on A level French, German and Spanish resulting from severe grading.

In A level physics, chemistry and biology however, numbers are gradually increasing, or at least remaining relatively stable. Rather, evidence of issues in teacher supply

[3] The Standards Advisory Group is a committee of the Ofqual Board that reviews research and makes recommendations about how we can maintain standards of qualifications. It is made up of members of our board and independent assessment specialists.

might be indicative of other factors influencing perceived subject difficulty which are not the result of potential severe grading. If these subjects are struggling to recruit sufficiently qualified teachers, and non-subject specialists are being required to deliver them as a result (for instance, chemistry graduates delivering A level biology courses) less effective teaching may be contributing to their apparent difficulty under statistical measures. This is supported by stakeholder concerns in the sciences, which are more focused on a decline in the pool of potential employers for non-graduate jobs roles requiring only A level, such as technicians. Thus a shortage of suitably qualified teachers in A level physics, chemistry or biology – as opposed to shortage of teachers, per se – could possibly be a potential alternative explanation of apparent grading severity, rather than evidence of it.

Definitive figures on the overall number of teachers in the education system within a specific subject area are unavailable. Furthermore, analysis of teacher supply would also require us to consider factors such as retention rates within a particular subject to determine whether there is an adequate number of qualified teachers. Nor is it clear what other factors would need to be taken into account if seeking to establish whether there had been an increase or a decline in teacher quality – would we also need to take into account education level, main degree subject, teaching experience and pedagogic ability? Is a PhD holder likely to be a more effective teacher than someone with only an MA? Is someone who has been teaching for twenty years likely to be a more effective teacher than a recent graduate?

However, the Department for Education does publish data relating to Initial Teacher Training recruitment which might give us an indication of the number of new teachers, if not their quality. The DfE uses the Teacher Supply Model (TSM) to estimate the number of postgraduate trainees required in England in each subject in a given academic year. This estimates how many newly qualified teachers are needed to maintain the stock of qualified teachers, taking into account projections of the pupil population, the effect of new policies, and estimates of teacher retention. The number of NQTs needed is then scaled up to account for trainees who do not complete their training and trainees who do not go on to secure teaching employment. This figure is used to calculate the total number of postgraduate ITT places needed. As such, data about recruitment against TSM targets should give us some evidence of the supply of teachers to deliver a specific subject.

Figure 14 shows new entrants to ITT courses as a proportion of TSM targets in 2015/2016 by subject. The data suggests that recruitment of physics teachers in that academic year was insufficient to meet the demand predicted by the TSM, with only 81% of the required number of places filled. Chemistry recruitment on the other hand was 99%, and recruitment for biology exceeded the TSM target at 115%.

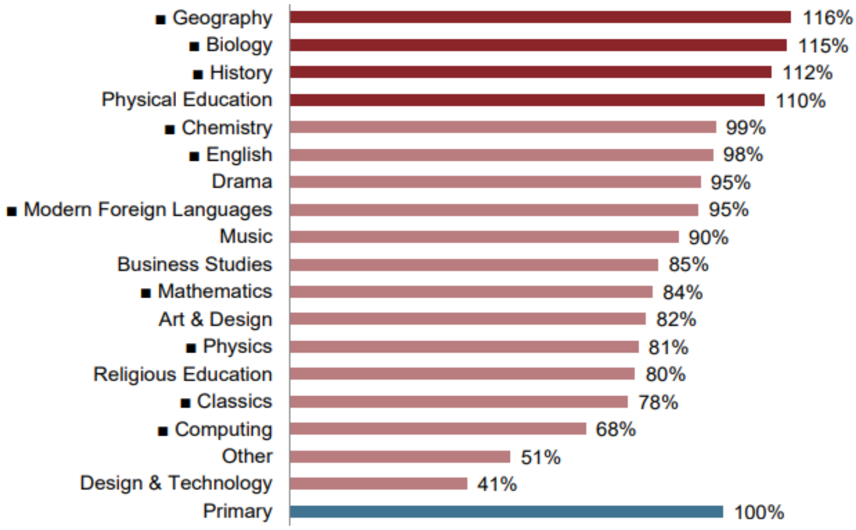


Figure 14 New entrants to postgraduate initial teacher training by subject, as a proportion of TSM targets, 2015/2016 (Department for Education, 2016).

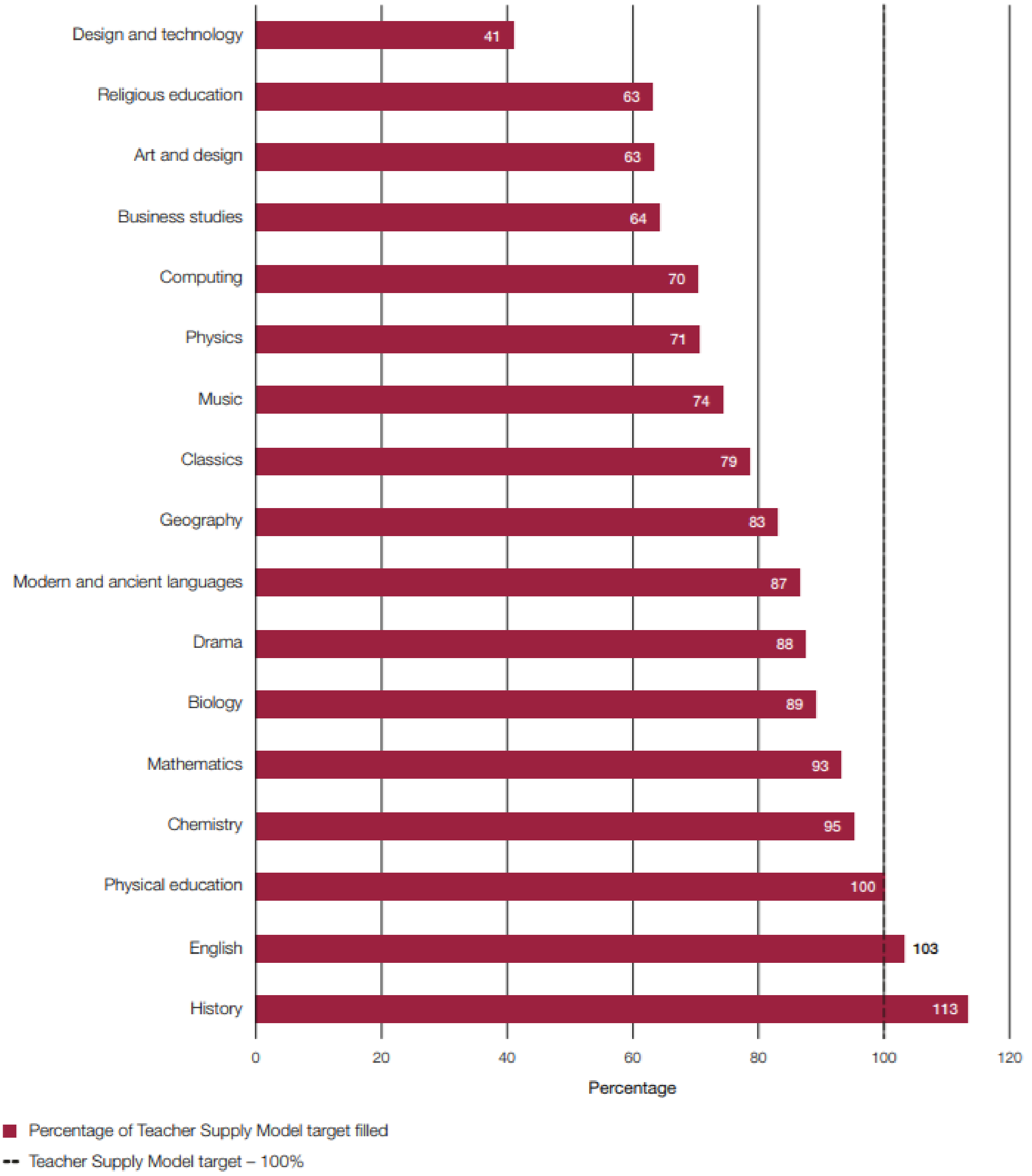
Stakeholders have also pointed to alternative sources of data on teacher recruitment, and in particular to the consideration given to Teacher Supply Model targets by the National Audit Office (NAO). Published in 2016, the NAO report *Training new teachers* sought to establish whether the Department for Education is achieving value for money through its arrangements to train new teachers.

The report identifies some shortcomings in the Department's teacher supply model, stating that whilst it "has strengths… [it] may still inaccurately predict schools' need for trainee teachers", and that "the Department is yet to demonstrate how accurate the model and its own judgements are" (Morse, 2016). The NAO report concludes that indicators suggest teacher shortages are growing, and that the number of classes in secondary schools who are taught by teachers without a relevant post-A-level qualification in that subject is growing. The report specifically physics as a subject likely to be delivered by a non-specialist, with 28% of lessons being delivered by teachers without relevant post-A-level qualifications, and notes that some subjects are harder to recruit for because they are less popular undergraduate courses for students applying to university. The NAO note that "to meet its 2014/2015 target for history trainees, the Department needed to attract 1 in 25 history graduates, for… physics, it need to attract 1 in every 5".

The NAO report provides data on the total number of training places filled against targets by subject which differs from that reported from the Department, using as its source figures from the National College for Teaching and Leadership and a different estimation of the rate at which the school age population will increase and thus the demand for teachers. Figure 15 shows the number of training places filled against targets by subject in 2015/2016 according to the NAO.

Figure 15 Training places filled against targets by subject, 2015/2016 (National Audit Office, 2016).

There is wide variation in the proportion of places filled by subject



Source: National College for Teaching and Leadership

The NAO data suggest that the situation for recruitment of science teachers is actually more severe than suggested by the Department. In particular, physics appears to have recruited only 71% of those teachers required to meet targets under the Teacher Supply Model in 2015/2016, rather than 81%, and even in subjects such as chemistry and biology which the DfE data suggests were either very close to or exceeding the recruitment required there is actually a shortage of teaching staff (95% as opposed to 99% in chemistry, and 89% as opposed to 115% in biology). This is significant as a shortage of teachers is, as outlined above, a plausible alternative explanation for the apparent difficulty seen under statistical measures of subject difficulty, and which some stakeholders attribute to severe grading.

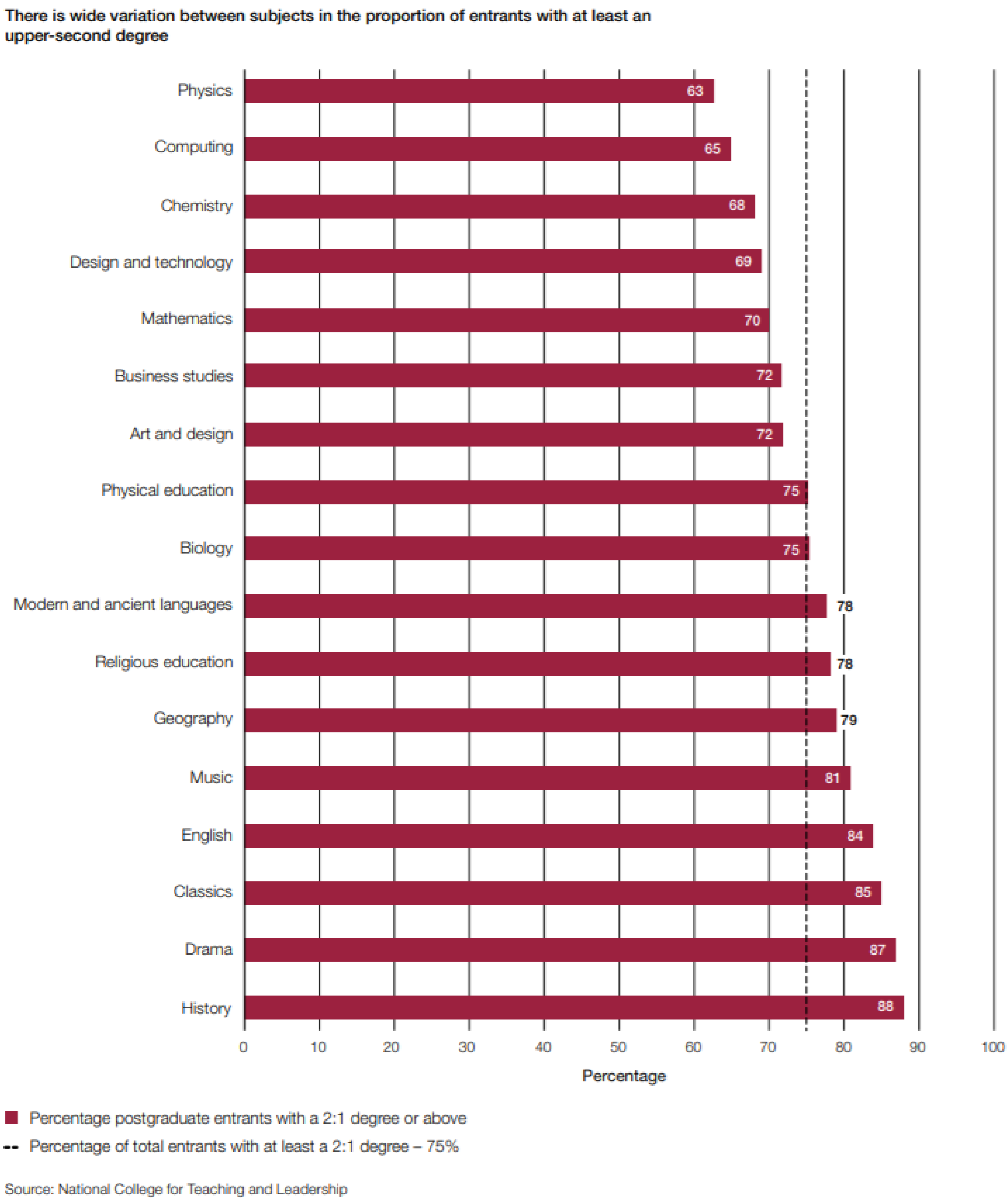
Furthermore, some stakeholders have claimed that the decline in uptake for languages at A level, which they attribute to severe grading, is contributing to difficulties with teacher recruitment in these subjects and will exacerbate the issue if prompt action is not taken to adjust grading standards.

The NAO have also attempted to measure the quality of teachers being recruited under the Teacher Supply Model, using the percentage of post-graduate entrants with a 2:1 degree or above as an indicator of subject knowledge. Although this should not be assumed to be a measure of likely teaching ability - recognised in the report, which explains that "schools and providers told us that degree class was a reasonable indicator of subject knowledge but a weak predictor of other aspects of teacher quality" (Morse, 2016) and points to the Sutton Trust publication *What makes great teaching?* which concluded that research on the link between degree class and student learning was "generally inconsistent and hard to interpret" (Sutton Trust, 2014) - it does facilitate comparison between different subjects in terms of the supply of suitably qualified prospective teachers at the beginning of their training, and allow us to consider whether any relation exists between claims of severe grading and another potential aspect of teacher supply. Figure 16 shows the proportion of postgraduate teacher training entrants with a degree class of 2:1 or above in each subject in 2015/2016.

Of the three A level sciences considered in this report, only biology achieved the 75% of postgraduate teacher training entrants with a degree class of 2:1 or above in 2015/2016 which represented the average proportion within all subjects. Both physics and chemistry returned a lesser figure, with 68% of chemistry entrants possessing a degree class of 2:1 or above and only 63% of physics entrants. It is possible that a shortage of suitably qualified teachers in these subjects could contribute to their apparent 'difficulty' under Rasch and Comparative Progression Analysis. However, whilst the relative difficulty of these subjects under statistical measures does correspond to the respective proportion of higher class degree holders entering teacher training programmes (which might be considered as a proxy for subject knowledge), we cannot draw any significant conclusions about the effectiveness of teaching from this.

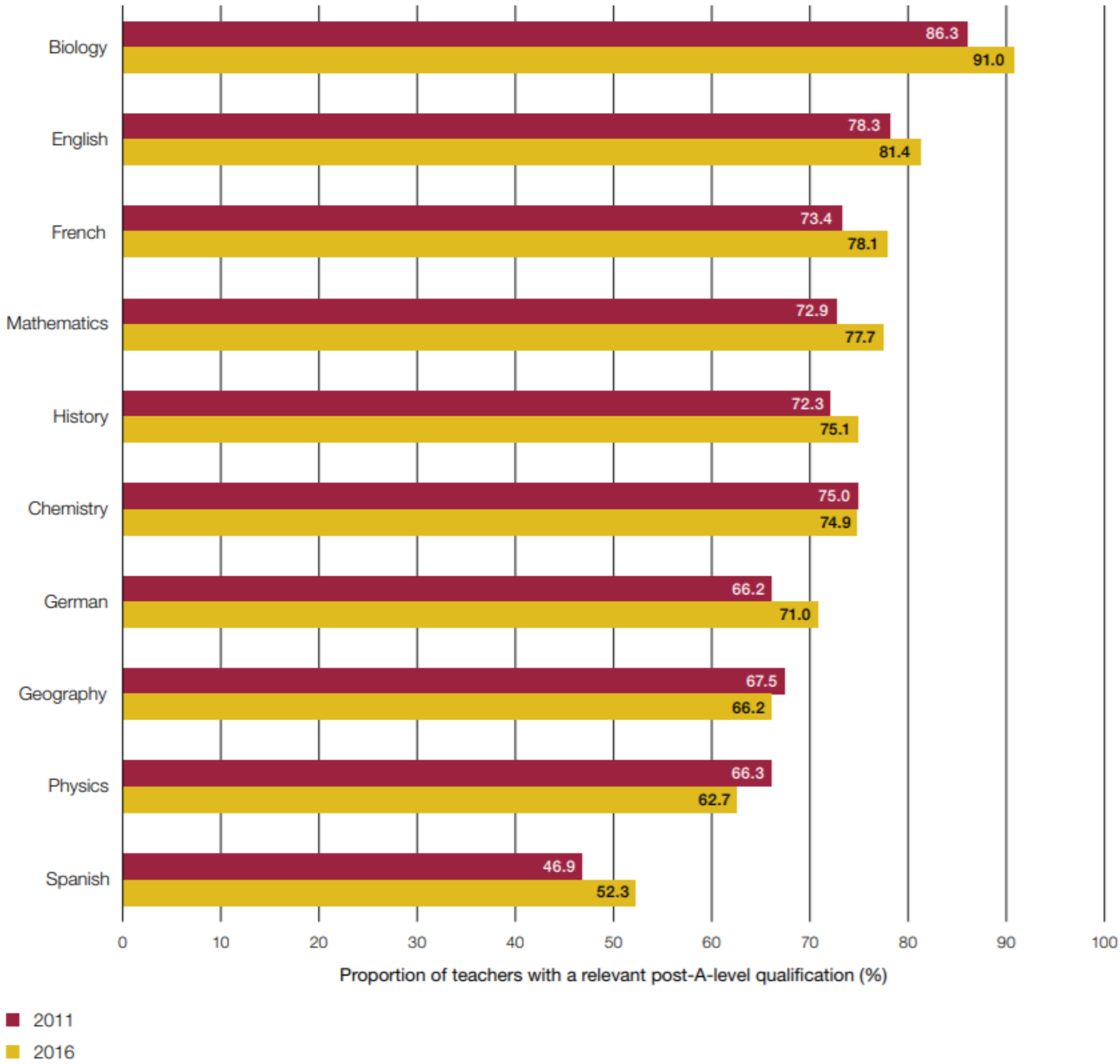
If there was weaker teaching in these subjects it would contribute to apparent subject difficulty directly, by having an adverse impact on student attainment. In addition, if the perception that some subjects are better or worse taught (or that some subjects are 'harder' because they are worse taught) could also lead to students opting away from these subjects at A level in a manner which would we be extremely difficult to distinguish from the strategic choices due to perceived subject difficulty which stakeholders claim is having an adverse impact on uptake in the sciences.

Figure 16 Entrants with at least an upper-second degree by subject, 2015/2016 (National Audit Office, 2016).



In 2017 the National Audit Office revisited the issue of teacher supply shortages in the report *Retaining and developing the teacher workforce*. Figure 17 from the report *Retaining and developing the teacher workforce* provides comparative figures which demonstrate how the proportion of secondary school teachers possessing a post-A-level qualification changed between 2011 and 2016.

Figure 17 Proportion of secondary school teachers with a post-A-level qualification and teaching in the relevant subject, 2011 and 2016 (National Audit Office, 2017).



The data shows that, whilst the proportion of teachers holding a relevant post-A-level qualification increased in biology in the period 2011-2016, in chemistry it stayed approximately the same and in physics it decreased by 3.6%. This could be considered to be evidence of an adverse social impact in physics if we were to accept that severe grading is leading students to opt away from the subject, which in turn is leading to fewer entrants in physics teaching programmes with the subject knowledge required to deliver it effectively – but it should be remembered both that overall physics A level entries are still increasing, and that the apparent ‘difficulty’ of the subject under Rasch analysis has not changed in recent years. Therefore data on teacher supply in A level physics, chemistry and biology is unlikely to present persuasive evidence of a negative social impact of potential severe grading.

2.3.6 Research into subject choice motivations

As outlined above, one of the concerns of stakeholders in the sciences is that the perception of difficulty is leading to the filtering out by selection, or implicit or explicit pressure from teachers, of students who may wish to or would benefit from studying physics, chemistry or biology but are deterred by the belief that they will secure a better outcome in another subject which is graded more leniently. The plausibility of this argument depends on the extent to which we accept that perceived difficulty (and hence outcome) is a greater motivator for students making A level choices than others such as personal interest or career aspiration.

There are several recent relevant sources of evidence which we might consider when attempting to understand the factors which influence subject choice. The first is the report by Panayiotou, Boulden, Newton and Andersson (2017), which asked 1,595 students about the motivation behind their choice of GCSE and A level subjects. The results of the survey are shown in Figure 18.

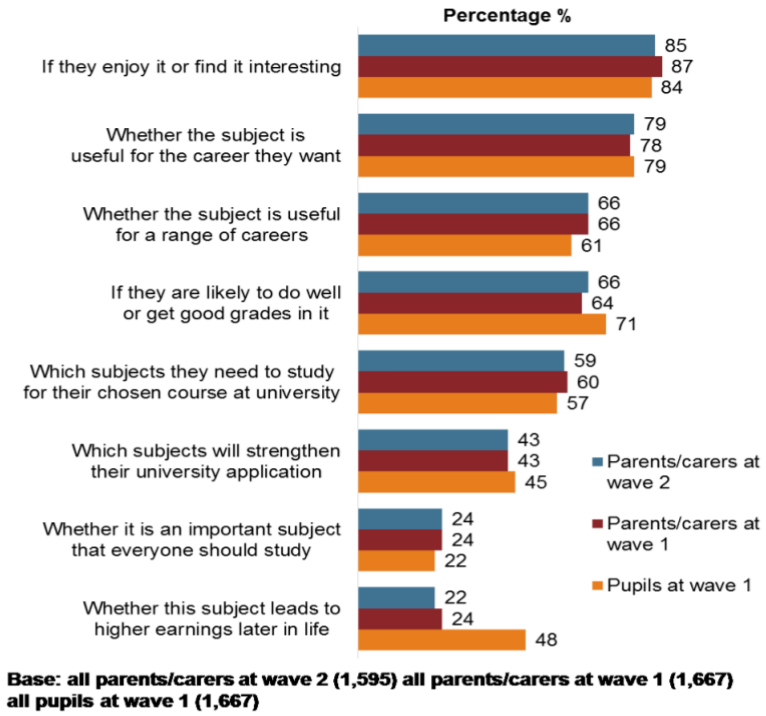


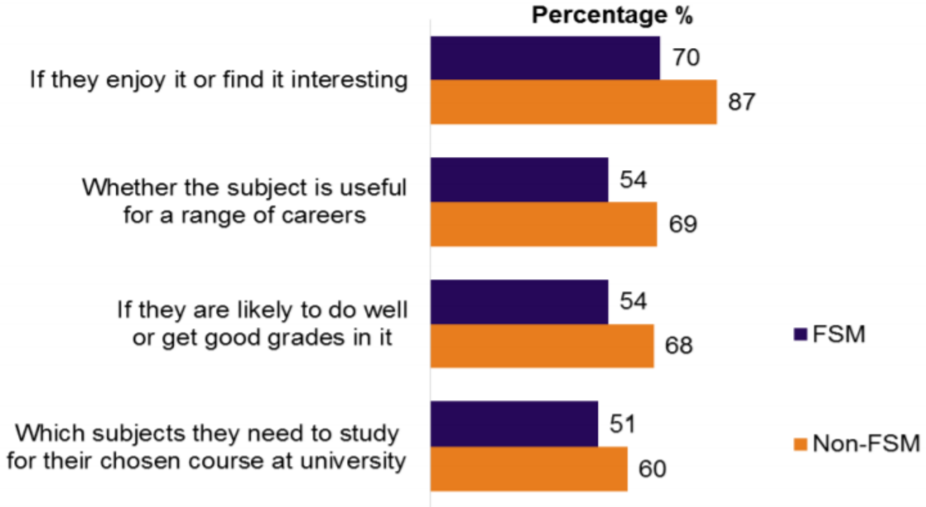
Figure 18 Reasons for choosing GCSE and A level subjects, by survey wave (Department for Education, 2017).

The survey indicates that likely outcome (and by inference, perceived difficulty) is only a tertiary factor influencing subject choice at GCSE and A level. Interest in the subject and future career prospects were both felt to be more significant considerations by students overall. This supports the conclusion of our own research into perceptions of difficulty and the impact on subject choice (see below).

What the survey does not tell us, however, is the number of those students who did not rate personal enjoyment/interest or career utility as subject choice motivations (15% and 21% respectively) who *did* indicate that they might base their decision on how likely they would be to do well in the subject. Potentially then, even though perceived difficulty may not be the main motivating factor behind the choices made by the sample as a whole, there could be a minority for whom it would have a deterrent effect when considering whether to study A level science.

Furthermore, these results might look different if GCSE and A level were reported separately. The survey also assumes the decision lies with students, not teachers or schools, who may be deterring students from studying certain subjects by enacting barriers to entry in the form of minimum grade requirements.

This survey also looked specifically at the motivations behind the choices made by 115 free school meal (FSM) pupils. Science stakeholders have asserted that the perception of severe grading is particularly likely to deter those from disadvantaged backgrounds from choosing to study A level physics, chemistry and biology. The responses of this group are shown in Figure 19, compared to those of all students.



Base: parents/carers of FSM eligible pupils (115) non-FSM eligible pupils (1450)

Figure 19 Reasons for choosing GCSE and A level subjects, by FSM eligibility (Department for Education, 2017).

For FSM pupils, respondents indicated that likely outcome was the joint second strongest motivating factor along with career utility (54% of candidates citing each as a reason behind their decision), but still ranked some way behind personal enjoyment and interest (70%), mirroring the results of the survey of all pupils.

Another source of evidence to consider is the research conducted by Cuff (2017). The purpose of this research was to explore whether teachers' and students' perceptions of subject difficulty might be having an effect on which subjects students choose to study in upper secondary education, and whether other concerns (for example subject enjoyment or usefulness) might interact with this relationship.

This qualitative survey of 112 students found that perceptions of difficulty were not the main basis of their decisions, and instead their choices were focussed more upon enjoyment and usefulness. Importantly, the majority of students stated that they were often willing to overlook subject difficulty when they enjoyed a subject and/or needed it to satisfy their university or career ambitions – though this leaves a minority who would still be deterred from studying a subject they perceived to be difficult. Students also agreed that although some subjects 'stood out' as seeming to be generally more difficult than others, whether or not they found a particular subject difficult was dependent upon their own individual strengths rather than on any commonly held notions of subject difficulty. Nonetheless students acknowledged that they did base their subject choices on perceptions of difficulty to some extent, but their conception of what made a subject difficult was based on a range of factors including style of assessment, different types of content (including how easy students found them to relate to and understand), workload, and teaching. They also recognised that they were sometimes discouraged by their teachers, parents, and friends from choosing subjects that were thought to be too difficult for them. However, although many felt that they were "discouraged" by their teachers for certain choices, students did maintain that ultimately it *was* still their choice, and rarely did they feel that they were forced into studying certain subjects.

The survey also found that some students were being prohibited from taking certain subjects at A level by school entry policies which sometimes prevented lower ability (according to their GCSE grades) students from accessing certain 'difficult' courses. These policies were sometimes driven by accountability scores, to ensure that students were able to achieve the highest possible outcomes in the subjects they chose. Teachers' decisions over which subjects to offer to their students were also sometimes affected by their perceptions of subject difficulty. For example, some schools had elected not to offer certain 'difficult' subjects, again meaning that students were prevented from choosing to study those fields of potential interest. This lends credence to the arguments offered by some stakeholders that filtering is occurring in the entry for A level science subjects as a response to perceptions of grading severity.

Figure 20 shows the range of influences on students' subject choices, and their interaction, identified in the course of the study.

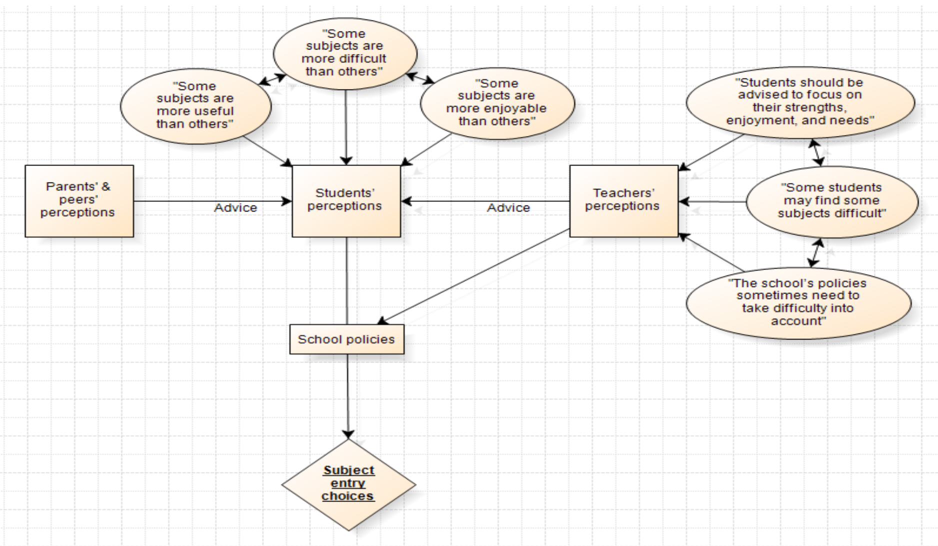


Figure 20 Thematic map of the influences of perceptions of subject difficulty on students' subject choices (Cuff, 2017).

The main conclusion drawn from this research was that subject choices appear to be primarily driven by a triad of perceptions: enjoyment, usefulness, and difficulty (with perceptions being mostly person-specific). Although perceptions of difficulty did have an influence on subject choices, they are perhaps the lesser of these three concerns – with both teachers and students agreeing that they are more driven by enjoyment and usefulness than perceptions of difficulty (cf Springate et al., 2008, who deemed enjoyment and future ambitions to be 'high influence factors', and difficulty to be a 'low influence factor'). Nonetheless, although this determinant of subject choices is perhaps not as great as previous concerns have suggested, the findings do support the argument that there are some students whose decisions are affected by perceptions of subject difficulty – although these are highly subjective perceptions, which are unlikely to be addressed by an adjustment to grading standards.

Finally, Taylor (2015) conducted research which sought to explore the drivers behind students' subject choice at A level through the psychological theory of planned behaviour, and which focused specifically on A level physics and media studies. These subjects were chosen in light of concerns about fluctuating entry in physics at a time when A level entries overall, including in those subjects which were perceived by some to be 'less academic' such as media studies, were continuing to grow – leading to fears that students were opting for subjects that were perceived as easier.

The study also acknowledged concerns expressed by the Confederation of British Industry about "fears of a skills crisis in future generations which is likely to have a detrimental effect on the UK economy" and, in addition to the economic impact, concerns expressed by the Institute of Physics about "the future provision of the subject… given an existing shortage of specialist physics teachers, which may impact upon the availability and teaching of the subject in schools". At the time the study was written, government target setting and incentives appeared to be "yielding positive results, with the Institute of Physics… reporting growth in in the number of A level entries for physics since 2006." However, there was also recent research to suggest "a change in students' attitudes towards science during secondary schooling, with physics becoming less popular than biology… and perceived as boring, irrelevant and difficult."

The aim of the study was to explore the usefulness of planned behaviour in predicting students' intentions to study AS physics and media studies, and to identify which variables (perceived behavioural control, subjective norms and attitude) were the most important predictors of intentions – as well as the importance of underlying beliefs (for instance, in likely outcome of a behaviour, or of the expectations of others) to these variables. The study was conducted on a random (although not intended to be representative) sample of secondary schools from across the UK who had made entries for AQA GCSE Combined Science, with questionnaires assessing intention to study either AS physics or AS media studies distributed equally amongst final year higher tier GCSE students. Higher tier students were targeted specifically to ensure that the sample included intenders and non-intenders for both subjects. This research was carried out in April to coincide with the time during which students selected their A level options. The questionnaires were completed by 555 students, with a reasonable split by both sex and subject intention.

Overall, the study found that students felt in control of their choice to study physics or media studies, and that they perceived low levels of social pressure from parents and teachers in making their choice. Of the behavioural beliefs tested in the questionnaire, the items that were most strongly related to intention to study were those concerned with the potential positive outcome from taking the subject. Those with higher intention believed that studying physics or media studies would offer more and better career opportunities, and secure them a place on their chosen degree course. Whereas the strength of the correlation between intention and utility for further study or range of job opportunities was the same for both physics and media studies, there was a significant difference between the two in terms of quality of career prospects. Potential physics students regarded studying the subject as more likely to help them secure a good job in the future than those intending to take media studies.

Of the behavioural beliefs concerning less positive associations of study, there was a positive relationship between intention to study physics and the belief that doing so would entail high workload or the study of difficult content. This correlation was negligible for media studies. Similarly, students with higher intentions were also those who perceived greater social pressure and were more motivated to comply with the expectations of others – and again the relationship between normative belief and intention tended to be stronger for physics than media studies, with the difference in the strength of the correlations being statistically significant for those items relating to expectations of parents and teachers. This indicates that students intending to study physics were more affected by external pressures than those intending to study media studies – but also the correlation suggests that this pressure is steering them towards the subject rather than away from it.

Of the control beliefs tested in the study, in both media studies and physics students with higher intentions to study the subject felt more personal control over their subject choice. Again, this is suggestive that those who do wish to study physics are not a particular risk of being influenced away from it by external factors. Comparison of the strength of the relationship between control beliefs and intentions by subject showed no statistically significant difference in the correlation for belief in either how difficult the subject would be or the likely quality of the teaching they would receive – only for how interesting the subject content would be, with the sample again showing a stronger correlation for physics than media studies.

The study concluded that “students’ subject choices at A level are most likely to be influenced by their attitude towards studying the subject and the perceptions of significant others, particularly parents and teachers.” Whilst the report noted that the opinions of parents and teachers were both important in helping form students’ decisions, it was the expectations of parents which were found to be most strongly related to intention to study both physics and media studies – not those of teachers.

The findings of the report that the beliefs most strongly related to intention were those which associated positive outcomes with selecting a particular subject (especially in terms of future career prospects), more so than beliefs about difficulty, support the conclusions of the Ofqual research into the potential impact of relative subject difficulty and the DfE survey of the motivations behind subject choice.

2.4 The learned societies' perceptions of grading severity and impact

2.4.1 Sources of stakeholder evidence

The majority of the stakeholder representation in relation to inter-subject comparability in A level physics, chemistry and biology takes the form of two open letters submitted to Ofqual by the Association for Science Education, the Royal Societies of Biology and Chemistry and the Institute of Physics, and the Royal Society in 2016 and 2017 (Stevenson, Britton, McLeish, Pritchard and Price, 2016; Cutler, McLeod, Morgan, Tracy and Mist, 2017). The first was sent to the then Chair of Ofqual Amanda Spielman in 2016, and included in an annex data taken from the report by Coe et al. (2008) which illustrates the relative difficulty of A level subjects under various statistical analyses including subject-pairs and Rasch analysis, in addition to a graph purporting to offer evidence of school 'filtering' (selection based on prior attainment) of entrants for A level courses, and an example of the analysis of variation in A level grade distribution for given GCSE prior attainment which has since been termed Comparative Progression Analysis. The second letter was sent to Ofqual's GQ Strategic Relationships team in 2017 in response to the publication of research into the validity of using Comparative Progression Analysis as a new approach to investigating inter-subject comparability, and highlighting some of its limitations.

The letters lay out the position of the learned societies on what they refer to as "the continuing problem of variable grading severity in subjects at A level", and make it clear that they regard grading severity and perceived difficulty to be an issue which is present in all three subjects.

Additional contributions to the public discourse around inter-subject comparability have been made by the Institute of Physics, who have written a series of blog posts on the topic for their website; corresponded with us in response to our call for evidence; and who delivered a presentation on what they regard as the impacts of severe grading on grading choice and uptake at the Ofqual inter-subject comparability conference in February 2016. They have also since written a further letter to us reiterating their views specifically in relation to A level physics, in response to the publication of our 2018/2019 corporate plan.

Whilst the Institute of Physics may have originated much of the stakeholder evidence, the other learned societies have made it clear at meetings of the Education Policy Alliance that they regard the issues in A level chemistry and biology to be the same and that the Institute speaks for them also.

2.4.2 Views on statistical evidence

The learned societies accept much of the summary of the discourse on inter-subject comparability presented in the 2015 Ofqual working papers, and welcomed the opportunity for public engagement on potential policy options which we offered at the time. The learned societies consider grading severity to be a longstanding issue in science subjects, suggesting that for the more than fifty years in which inter-subject comparability has been debated, the subjects which appear to be more 'difficult' than others have remained consistent.

However, they reject the possibility acknowledged in the working papers that differences in GCSE and A level outcomes between subjects may be the result of a combination of factors other than grading severity, and are satisfied that the consistency of the results of different statistical measures of subject difficulty is such that the grading severity is the most likely cause.

Indeed, they regard the statistical analysis carried out by Ofqual as confirming the argument that there are observable differences in grading severity across A level subjects, and that these variations are particularly pronounced in the sciences. They point to the fact that the same subjects appear to be 'difficult' under different methods of analysis as evidence that this is not a statistical artefact, but that a genuine lack of comparability exists between similar grades in different subjects – and that this evidence is compelling enough that it is now time to move on from debating the validity of different measures to considering how best to address the problem.

For the learned societies, the issue is fundamentally one of fairness. Whilst they accept that it is meaningless to consider absolute comparability at a subject level, or through direct comparisons of the outcomes of individuals, the learned societies argue that a feasible conception of inter-subject comparability can be achieved through considering the relative chances of a cohort obtaining a particular grade in different subjects. They contend that there should be a similar likelihood of students achieving a given grade in each subject, once aptitude and prior attainment have been controlled for.

That this is not the case according to statistical measures of difficulty they regard as a "basic inequity for students who take the subjects that are being graded more severely: they are not being awarded a just grade" (Letter to Ofqual from Cutler et al, 2017). On this basis alone, the learned societies contend that action needs to be taken. But they also argue that more severe grading in A level science subjects invalidates the assumption that the same grades in different subjects are equivalent when used for the purposes of selection by employers and Higher Education. Students in 'easier' (i.e. less severely graded) subjects receive an unfair advantage solely on the basis of subject choice if the currency of A levels is regarded as the same.

The Institute of Physics has even raised the question of whether, given what they regard as being the decreased likelihood of obtaining a 'fair' grade in their subject, "whether it would be responsible for a teacher to encourage or even allow a B grade achiever at GCSE physics to continue onto the A-level course?"

Instead, the learned societies argue, we should aspire to a grading system that does not penalise the study of certain subjects by some individuals. Choices should be made exclusively on the basis of interest and future ambition – something which they do not feel is not currently the case.

2.4.3 Potential impact on subject choice

More than just being unfair to students who take the subject, the learned societies contend that severe grading is creating barriers to entry in science subjects by creating the impression that they are difficult and "the preserve of an academic elite", and that this disproportionately impacts on certain groups. This in turn is undermining efforts to broaden participation in A level sciences where some of those groups are currently unrepresented.

In particular, the learned societies contend that the following are being negatively affected by severe grading in science subjects (Tracy, 2016):

- Perfectionist students, who shun challenging subjects for those they feel they can excel in
- Borderline students, who are choosing "safe" subject combinations to secure offers for admission to Higher Education
- "Those who lack confidence (including girls)" who avoid subjects which are perceived to be difficult (Tracy, 2016; Letter to Ofqual from Stevenson et al., 2016)
- Lower attaining students, who are directed into subjects which might secure them higher grades but perhaps limit their potential career options in the future
- Students in schools seeking to secure their position in league tables or against performance measures, which have no incentive to offer science subjects at A level.

As a matter of principle, the learned societies argue that exam results in different subjects should be comparable to allow observers to reliably determine aptitude, inform decisions (and selection decisions) about further study, and to provide a basis for accurate measures of school performance (Tracy, 2016). Comparability would also facilitate students choosing subjects based on grade outcomes, future prospects, intellectual development, and their own interest and engagement. The assertion is that the current situation does not allow this, because students do not choose – or are not allowed to choose – subjects which appear to be graded more severely.

One way in which schools and students are aware of the supposed differences in grading severity between subjects is through the widespread use of performance predictors such as ALIS (the adaptive baseline assessment delivered by the Centre for Evaluation and Monitoring), which predict the likely outcomes in different subjects for a given set of results based on historic performance data. The learned societies claim that these predictors will nearly always predict lower grades in A level sciences than other subjects, with the only exception being candidates with straight A* at GCSE.

They identify a number of adverse effects arising from the interaction between severe grading and these predictions, which they feel are restricting students' choices at A level; reinforcing myths about subject difficulty and ease; distorting the intake to the A level subjects which are perceived to be difficult; and undermining the ways in which those grades are used.

The learned societies contend that students' subject choices are influenced primarily by their likely success (though this does not take into account the necessity of science A levels for popular vocational careers such as veterinary medicine, medicine, engineering and nursing, which is also likely to be a strong driver of subject choice) and the fact that students with uniform prior attainment will be predicted lower grades in A level physics, chemistry and biology than other subjects will deter them from taking sciences in favour of those in which they are predicted to obtain a higher grade. This in spite of the fact that their grade in science may be as useful or more useful than a higher grade in a different subject. This, they claim, is a direct outcome of historic severe grading, but is also serving to reinforce its effects.

Similarly, schools may take steps to influence subject choice as a result of their desire to perform well against attainment measures. To secure a significant number of high grades at A level, schools may require higher grades for entry onto their science courses; dissuade students from taking these subjects in favour of something which is perceived to be easier; or even cease to offer the most severely graded sciences at all at A level. Whether or not a student is good at a science subject and might be suited to a career in a STEM field, "the lower grade is unequivocally of less value to the school in performance measures."

The learned societies argue that the reputation of science subjects for being 'difficult' is undeserved, but is potentially reinforced by different entry requirements (a response to severe grading), which may deter even those students who achieve them from studying the subject by giving the impression that it is exceptional in terms of the degree of challenge and thus only suitable for the brightest of the bright.

They also point to the inevitable counterpoint to this – that some subjects may be being regarded as less meaningful or worthwhile because they are being graded more leniently, and are thus regarded as 'easy'.

Even where differential entry requirements for sciences are not formal school policy, the learned societies assert that the perception of more severe grading is likely influencing the advice given by teachers about A level choices (which comes not just from those teaching the subject, but also from other staff with pastoral responsibilities), and thus in turn the decisions that students make.

2.4.4 Potential impact on grade distribution

Additionally, the learned societies argue that this selection (or self-selection) of students entering A level physics, chemistry and biology is also resulting in grade distributions being shifted towards the higher end. They point specifically to the example of A level physics, where they contend that a skewed intake as a result of selection arising from the perception of severe grading is the cause of a grade distribution for the subject where the modal grade is A/A*. The Institute for Physics argue that the resultant loss of discrimination amongst candidates and high number of top grades being awarded, combined with the status of physics as a facilitating subject regarded as being particularly ‘difficult’, means that the subject is no longer viewed as a feasible option for those students predicted C/D grades or who do not specifically require the subject to access a particular university course (or who do not intend to progress to higher education at all). As a result, A level physics has become solely a “gateway” subject to further study in a very narrow range of university disciplines; no longer a route to apprenticeships, technical careers or vocation-specific degrees with lower entry requirements (such as nursing) where candidates may benefit from knowledge of the subject, or a choice made solely on the basis of individual interest.

Allen (2016) makes the point that because the average difference in A level physics grade between GCSE A grade physics students and GCSE grade B physics students is less than one A level grade, any selection is deterring some students who would otherwise go on to achieve perfectly satisfactory A level outcomes (i.e. GCSE B grade students). Indeed, in 2014 more than 92% of students who *did* study A level physics having previously achieved a B grade at GCSE went on to pass.

These effects are said to be cumulative and self-perpetuating, with the learned societies arguing that they should be regarded as both evidence of severe grading and evidence of the negative impacts arising from it. This they claim makes a case for immediate action – not only since inaction will result in unfairness for subsequent A level cohorts, but because as the consequences and symptoms become more pronounced and entrenched over time, the effects upon cohort entry profiles arising from self- and school selection make it increasingly difficult to address issues of grading severity: as doing so would result in even more students obtaining A and A* grades and would potentially lead to accusations of grade inflation. Thus maintaining the status quo is not only undesirable, but is actively damaging to these subjects.

The learned societies also argue that the distorted entry profile in science A level subjects is also leading to the lower parts of the grade distribution being filled by “downgraded” mid-attaining students, treating them more harshly than students in other subjects and reinforcing the effect of grading severity.

2.4.5 Lack of transparency

Regardless of whether Ofqual is satisfied that severe grading is present within A level science subjects, the learned societies contend that the perception that these subjects are more difficult than others is affecting the behaviour of schools and students. Strategic decisions are being made about subject choice, and a lack of clarity over the issue of inter-subject comparability means that these decisions are not always beneficial. This is being exacerbated by the fact that no system-wide approach has been adopted to ensure that inter-subject comparability is achieved. Similarly, the lack of any public statement acknowledging differences in grading severity or attempts to tackle the implications of it means that there is a fundamental lack of transparency within the education system which is undermining the positive effects of efforts to increase participation in science A levels – both restricting the opportunities open to students and depressing entries.

Whilst stakeholders accept that university admissions tutors do probably discriminate between subjects based on their perceived value (for instance, by deeming some subjects to be ‘facilitating’), potentially conveying some advantages from having taken a subject which is regarded as being ‘difficult’, the implicit nature of these value judgements means that students making strategic decisions on the basis of perceived difficulty may in fact be damaging their own educational prospects in the long term. “Equitable grading” would remove any ambiguity about the worth of grades in different subjects.

The learned societies argue that there is an additional inequality in the fact that notions of ‘difficulty’ in A level science is contributing to a lack of diversity in these subjects. Those from more privileged backgrounds with better access to information are less likely to be influenced in their subject decisions by perceived difficulty. Rather, the deterrent effect is greatest on students who are more cautious and on those with lower ‘science capital’ – which they contend are features of two groups that are currently underrepresented in some scientific fields, and particularly in A level physics: girls and students from more deprived areas.

In this context the learned societies identify as a major issue the “unhelpful” language of difficulty, where the perception of what is “difficult” is the result of a combination of threshold aptitude, threshold achievement and objective correctness, but also the emotional response of those choosing or delivering the subject and the misattribution of particular subjects within a ‘hierarchy of knowledge’. The only measure of difficulty which is significant in the mass-participation context of examinations is how many individuals attain a given level. The Institute of Physics offer the example of a comparison between the number of individuals who achieve careers as professional footballers and those who get onto physics courses at Oxford University - “[the difficulty of both is] about the same. However no one would say that football is ‘difficult’. Similarly – physics is not difficult… in and of itself.”

Dispelling these misplaced notions of difficulty, they claim, would lead to more open and beneficial choices for post-16 education, and ensure transparency and equity when selections are made on the basis of A level grades. However, the learned societies contend that doing so is impossible while the data (whether from statistical analyses such as Rasch and Comparative Progression Analysis, or predictive indicators such as ALIS) shows that A level science subjects are graded more severely than other subjects. Were we to devise the grading system anew, they argue, this variation in likely outcome between A level subjects for given GCSE prior attainment would not be acceptable.

2.4.6 Proposed solutions to 'severe grading'

The learned societies argue that the worst response to what they feel to be the issue of severe grading in A level physics, chemistry and biology would be to do nothing. Taking no action would perpetuate inequality either in subject outcome or subject choice and would, they claim, continue to discriminate against certain pupils by limiting access to the subjects based on gender and background. In turn, this will limit the number of students entering STEM subjects at A level – thereby threatening a government priority.

To address this perceived severe grading, the learned societies propose either (Tracy, 2016):

a) calculating and rebalancing grades every year using statistical means to ensure that grade likelihoods are equal between subjects
b) shifting outliers under statistical measures of difficulty slightly so that grade outcomes become more similar
c) weighting subjects differently in performance measures, to stop schools selecting for entry to these subjects on the basis of likely grade outcomes
d) realigning all subjects on a "one-off" basis using the statistical measures of difficulty we currently have access to
e) implementing different grading systems for "harder" and "easier" subjects which claim no equivalence between them
f) weighting certain subjects so that they carry more points for UCAS
g) introducing a diploma or baccalaureate-style qualification which restricts acceptable subject combinations for certain pathways, to prevent strategic choices whereby those taking "easier" subjects achieve an advantage competing for university places against those studying "harder" ones.

The decision of the Ofqual Board not to take action to statistically align grading outcomes across all subjects at A level rules out options a) and d), and options c), f) and g) concern accountability measures, qualification tariffs and the curriculum, so are outside of the of the remit of Ofqual. They may be formal recommendations we make as a result of our investigation into inter-subject comparability however.

Option b) might seem ostensibly to be the most feasible response to the apparent grading severity suggested by statistical measures of subject difficulty, but given the limitations of Rasch and Comparative Progression models identified earlier in this paper such an adjustment may create 'new' outliers without accomplishing the desired effect – potentially also compromising performance standards in the process. It is also not clear what we would do if the relative 'difficulty' of A level subjects changed in future years for reasons other than grading severity. Would grading standards need to be adjusted every year to ensure that science subjects remained of average 'difficulty'?

Option e) poses issues because it would threaten one of the stated purposes of A levels – to "define and assess achievement of the knowledge, skills and understanding which will be needed by students planning to progress to undergraduate study at a UK higher education establishment, particularly (although not only) in the same subject area". Implement differential grading systems for sciences and other subjects would likely constrain the A level choices of students intending to study a science at university, and make it very difficult for students with one science A level amongst others to be judged fairly alongside students with all science A levels for university admission. It would also make it harder to draw comparisons between qualification standards internationally.

2.5 Awarding organisations' views

2.5.1 Awarders' views

As those ultimately responsible for maintaining grading standards in these qualifications, judgemental evidence was gathered from the A level awarding panels for these subjects via questionnaires distributed on behalf of Ofqual by the exam boards. These were completed at or shortly after the awarding meetings. Awarders were asked to reflect on the standards set in their subject in the summer 2018 series, and whether in their professional opinion it was correct. They were also asked to consider whether they felt standards were appropriate relative to other A level subjects, how acceptable an adjustment to grading standards might be in terms of the impact on candidate performance standards at the judgementally set grade thresholds of A and E, and whether the current grading standard might be having any negative impacts.

The responses received have been summarised below by subject.

Physics

Awarders from one exam board were of the view that the current grading standard in physics was correct, another that the standard was possibly slightly severe, but those from two boards that the standard was in fact too lenient. These concerns about leniency mainly focused around the E boundary (where they felt that the quality of work was very poor) with awarders generally satisfied that the standard required for an A grade was appropriate.

Awarders did not feel able to draw comparisons with grading standards in other subjects, and whilst at all four exam boards they recognised that there was a perception amongst teachers and students that this was a difficult subject, only one panel felt that it would be appropriate to attempt to address this (or that it could be addressed) through changes in grading. Awarders at the remaining three boards felt that an adjustment to grading standards was not only unacceptable (sharing the view that the perception of difficulty was unjustified), but would be actively damaging to the subject – citing concerns about perceptions of dumbing down, decreased university retention rates, and a loss of comparability with international physics qualifications. At two of the four boards awarding panels felt that the widespread reputation of physics as a challenging subject meant that it was regarded as being of greater merit for university admissions and employability, and were anxious about any change in standards which would lead to A level physics being devalued.

Chemistry

Awarders' views were received from three of the four exam boards. Of those three, opinions differed. At one board awarders were divided in their view of the standard, regarded it as either slightly severe or slightly lenient. Those at the two remaining boards felt that the standard at grade A was correct, but that the E threshold was slightly lenient and that work of this grade was not of an acceptable standard. Support for an adjustment to this standard was not strong however, with the awarding panel at one board arguing that this would be unpopular so soon after significant reforms.

Only one awarding panel offered a clear view of the comparability of standards in chemistry with other A level subjects, which they felt was acceptable (awarders at the board where views of the appropriateness were divided also differed over whether they felt chemistry was comparable to other subjects, or harder than others in a manner which is unfair because it is not reflected in the UCAS tariff). At two of the three boards awarding panels recognised that this was a subject which was perceived by candidates to be difficult, and that this might be dissuading students from STEM routes, but this was not a unanimous view – with one awarder arguing that teachers were generally satisfied with standards in the subject. At neither board did the awarding panel feel strongly that any adjustment in grading standards would be likely to address these perceptions of difficulty, with awarders arguing that the inherent demand of the subject means that less able students would regard it as challenging however it was graded, and anxious about the impact that such an adjustment would have on the value of the qualification.

Biology

Views were received from awarders at three of the four exam boards, but were complicated somewhat by the fact that one board offers two different specifications for this subject. Of the two boards that offer only a single specification, the awarding panel at one regarded the current grading standard as being slightly lenient, and the other felt that the standard was broadly correct. At the board which offers two specifications, awarders were satisfied with the standard for one but unable to reach a consensus over whether the other was graded too leniently or too severely. The awarding panels for two of the three boards suggested that there was perhaps scope to lower the standard slightly at grade A, but at neither was this a unanimous view and concerns were expressed about a potential loss of differentiation at grade A and possible damage to public confidence as a result.

None of the panels drew comparisons between grading standards in biology and those in other subjects, but awarders at one board were concerned about intra-subject comparability with the A level qualifications sat in other countries of the UK. The panels for two of the three boards felt that, overall, the potential risks of an adjustment to grading standards outweighed the beneficial impact this might have in addressing perceptions of difficulty in this subject – and at only one board did awarders express the view that this perception was having a negative impact on the subject in terms of declining A level uptake and university progression. Even this panel felt that the majority of stakeholders were interested in results not as an indicator of ability but as a measure of relative position within the cohort, and that universities would likely respond to any adjustment to grading standards in biology by increasing their offer to access undergraduate courses.

2.5.2 Viewpoints of the Awarding Organisations

In addition to gathering judgemental evidence on potential grading severity and the appropriateness of an adjustment to grading standards from their awarding panels, three of the four exam boards also offered their own analysis of the evidence presented in these reports. This is summarised below.

Statistical Evidence

Those responding recognised that the statistical evidence presented in the report seemed to suggest that A levels in physics, chemistry and biology are more difficult than the majority of A levels. However the boards also noted that the report identified some significant limitations of the various forms of statistical analysis which have been employed, which they feel present significant conceptual challenges to drawing conclusions about 'difficulty', and the advisability of adjusting grading standards.

Noting that the statistical measures of subject difficulty share the assumption that all subjects are measures of a single latent trait (i.e. a 'general academic aptitude'), the boards expressed doubts both about the unidimensionality of what is purported to be being tested in different A level subjects (whether we really can expect that a given student will perform just as well in history as they do in mathematics), and what the shared latent trait being measured actually represents. Thus whilst the statistical evidence may indicate that there are differences between A level subjects, the evidence does not provide an explanation of what these differences are – they could be the result of more severe grading, but they may also reflect differences in the resources which schools and government dedicate to particular subjects, more or less demanding subject content, or variations in overall teaching quality and effectiveness. It may also reflect differences in the 'step up' from GCSE to A level between different subjects.

Another limitation to the statistical evidence recognised by the exam boards is the extent to which the statistical models may be compromised by either non-representative samples or non-random missing data – evidenced by the simulations conducted by Bramley (2016) – which may be contributing to biased estimates of subject difficulty. This is particularly an issue given the relatively unrestricted nature of students' subject choices at A level. Rasch models assume that subjects are chosen at random, and that candidates in the years considered selected only those subjects in which they would perform well. However, the boards are not convinced that this is the case. One board suggested that the possibility that students may choose certain A level subjects because they are deemed to be 'facilitating' (that is, regarded as providing greater opportunities for further study or employability) even where it is not one of their strongest subjects makes facilitating subjects particularly problematic under this kind of analysis. All of the subjects considered in this report are so-called 'facilitating subjects'.

Boards also felt that there were differences in the expectations of prior learning between A level sciences and some of the other subjects considered within the statistical analysis of subject difficulty. In particular the fact that students must previously have studied a GCSE science (often to a minimum standard of competency) to progress to study it at A level, whereas some subjects students will be studying for the first time, was viewed as potentially being a reason that the 'difficulty' of different A levels may justifiably differ.

One board also made the point that whilst the report considered the evidence from the perspective of addressing potential grading severity by considering whether to lower grade boundaries in sciences, it did not give any consideration to whether standards in other, apparently lenient subjects should be increased. However, such a response to inter-subject comparability was specifically excluded from the scope of this research as a result of the Ofqual consultation in 2016.

In light of the potential limitations outlined above, the boards shared the view expressed by Benton and Bramley (2017) that these limitations means that is not possible to align subjects solely on the basis of statistical analysis.

Impact of reform

The boards also stressed that the statistical evidence presented in the technical reports for A level science was based upon analysis of results data from the legacy A levels sat in 2013 and 2017, rather than the reformed qualifications. Whilst they acknowledged that the comparable outcomes approach that Ofqual adopt to awarding means that grading standards from the legacy qualification have been carried forward and will continue to be maintained in the new A levels (along with any historical severity in grading which may have existed in these subjects previously - particularly since in 2017 and 2018 outcomes were closely aligned with predictions), there have been substantial changes to the design of these assessments which may arguably have resulted in a change in content standards.

In particular, one board pointed to changes such as the removal of assessment of practical skills in science and the increased emphasis on quantitative skills as having potentially impacted upon the 'difficulty' of these subjects as experienced by candidates, despite the efforts made to ensure the comparability of performance standards. They also noted that, as the reformed science A levels were awarded for the first time in summer 2017, it would take time and further research to fully understand what the impact of those changes might be. This is complicated further by factors such as the saw-tooth effect, which may compound the apparent difficulty of these subjects in the first few years of awarding as teachers adjust to delivering the new qualifications, with candidate performance progressively improving as they do. For this reason the board were of the view that any adjustment should not take place immediately, but after a period in which the new qualifications were allowed to bed in.

Similarly, it was also noted that it was still unclear what the impact of the decoupling of AS and A level might be upon the apparent ‘difficulty’ of these subjects, something which may not be evident until all of the reformed A levels have been introduced and assessed. Whereas previously students routinely began studying four AS levels (likely choosing to discontinue one subject after a year) there has now been a widespread movement in schools to offering three A levels only. Whilst this has been identified elsewhere in the report as providing a potential alternative explanation for the decline in uptake in certain subjects (A level entries being a ‘zero sum game’), one board have hypothesised that this will also reduce the number of students taking subjects across disciplines (e.g. a combination of sciences and humanities) and lead to an increased propensity to focus A level choices within the same subject group. They suggest that this may have an impact on the relative difficulty of different A level subjects under statistical measures which are predicated on the assumption of a generic linking construct such as Rasch and Comparative Progression Analysis, and recommend that further research is carried out to establish the impact of AS and A level decoupling – rather than attempting to adjust grading standards on the basis of data from the legacy qualifications.

Several of the boards also made the point that AS and A level specifications must build on the skills, knowledge and understanding proscribed at GCSE, and that there is likely to be an interaction between how effectively GCSEs prepare students for A level study and perceptions of subject difficulty at key stage 5. This interaction is likely to have changed following the reforms to GCSE, which explicitly set out to increase the content demands of these subjects (which was not the case at A level). The reformed GCSEs in science were awarded for the first time in summer 2018, and the first students progressing to A level from the reformed qualifications will take their assessments in summer 2020. Similarly, one board emphasised that policy changes associated with the reforms are likely to have an impact on school behaviours (and thus the experiences of students) which may result in changes to the apparent ‘difficulty’ of these qualifications (for instance, the change in the number of students taking combined science as opposed to three separate sciences at GCSE). Again the impact of these changes will not become clear until 2020, and one board argued that any action on grading standards should be postponed until this time.

Judgemental Evidence

The boards gathered individual feedback from members of their awarding panels via questionnaires, which have been summarised above. However, the boards also provided their own analysis of the judgemental evidence they submitted.

The boards noted that the views from awarders often differed, and that on the majority of panels there was some diversity of opinion as to whether, in their professional judgement, an adjustment to grading standards would be acceptable. Overall they reported that there was generally consensus of opinion amongst awarders in the sciences, and that where awarders were in agreement this was mostly that the current standard was correctly set. Indeed, it was interesting to note that the awarders were most confident in the standard in those subjects which appeared to be the most severely graded under statistical measures of subject difficulty (physics and chemistry). Furthermore where awarder opinion was divided, it was not always the case that the disagreement was over whether the standard was correct or too severe – some feeling that the standards were in fact too lenient.

On balance, the boards felt that overall the judgemental evidence did not provide a compelling argument that grade boundaries at A level should be adjusted, either because awarders were overwhelmingly of the view that standards were correct or because opinion was divided to the point that the evidence was inclusive. Boards did report however that the majority of awarders recognised that stakeholders perceived these subjects as being hard (rightly or wrongly), although one board pointed out this was perhaps unsurprising given the extent to which their panels are invested in the current standard.

The boards reported that their awarders did not feel that there was any compelling evidence that an adjustment would lead to any increase in entries for science subjects – which they noted had already seen substantial rises in recent years, likely as a result of recent initiatives to promote STEM subjects.

Boards also cited the research conducted by Ofqual and the DfE into students' subject choices, which indicate that the motivations behind study decisions are complex and that utility and enjoyment are often greater considerations than difficulty. One board did consider that enjoyment may be linked to some extent to perceived success in the subject, and that this may exacerbate the impact of beliefs about subject difficulty on uptake They also noted that the factors influencing student decisions are liable to change over time.

Potential Risks

The exam boards were mindful that any re-evaluation of the standard should weigh the impact of the current standard against the likely impact of an adjustment, and in particular that this should be considered in the context of the purpose of A levels.

Two purposes in particular were regarded as being at risk from an adjustment to standards, these being the requirements of A levels:

"To define and assess achievement of the knowledge, skills and understanding… needed by students planning to progress to undergraduate study at a UK higher education establishment, particularly (although not only) in the same subject area."

"To permit UK universities to accurately identify the level of attainment of students."

Here it was felt that there was a tension between the purpose of A levels to facilitate progression in a particular subject, and the purpose to facilitate progression to higher education in general. Boards were concerned that any change to the standards at A level, particularly at grade A, might result in a loss of discrimination within subjects at the top end and lead universities to respond by raising their entry requirements within that subject – and cited the findings of the research into *HE perceptions of grade standard adjustment* which they argued illustrated the limited support from higher education representatives to lower the standards within A level science subjects. However, it was also acknowledged that outside of a particular subject A levels are presented as having equal currency according to their UCAS point tariff (although admissions tutors may have their own views on the relative value of a qualification), and that this means that some students may be disadvantaged if they have taken a 'hard' subject when applying to university, compared to those who have taken an 'easy' one. Boards recognised that any change in the value which is assigned to different A levels was beyond the remit of Ofqual, and we can only seek to assure comparability of subjects through changes to grading standards. One possible solution suggested was to adjust the standard at grade E (which would have a knock-on effect on the standard of the intermediate grades) whilst maintaining the standard at grades A* and A to support progression in the subject. However it was acknowledged that this was unlikely to address the perception of difficulty in these subjects.

Increases in university entry requirements were felt to be a particular risk if changes were to be made mid-cycle rather than at the defined end-point of a specification, as universities would have to adjust their understanding of the performance standard represented by passing and higher grades to reflect the change in the difficulty of the qualification. Concerns were expressed that either some universities may not fully grasp the new standard, unfairly penalising some students and advantaging others, or that students in adjacent cohorts would be disadvantaged – either those in the cohort immediately preceding any adjustment who would be competing for jobs and university places with students who achieved the same grade for a lower level of performance, or students in successive cohorts who were regarded as holding a devalued qualification. To mitigate this risk, boards felt that any adjustment should be carried out incrementally.

Ultimately, boards were of the view that for greater comparability to be achieved between subjects, it may be necessary to prioritise one of the stated purposes of A level qualifications over the others. In view of the use of some of these subjects to decide access to socially important and demanding courses such as medicine and engineering, exam boards argued in favour of preserving the current standard to enable universities to continue to effectively differentiate between candidates.

Conclusions

Overall exam boards were supportive of re-evaluating grading standards in A level science to ensure that they remain appropriate, but did not necessarily feel that either the statistical or judgemental evidence presented a compelling case for an adjustment – or a clear indication of what a more appropriate alignment of subject difficulty might be.

However boards did feel that this was an issue which warranted further consideration, and in particular that it was reasonable to keep grading standards under review generally because the nature of comparable outcomes (which maintains the relationship between prior attainment and results in a given subject year-on-year) means that there is a risk that any contextual factors which may have contributed to an increased or decreased level of difficulty in a specific subject when comparable outcomes was first introduced have been reflected in the grading standard from that point forward, regardless of whether those factors have subsequently changed (although we would expect that awarders would identify changes in the performance of candidates over time such as these)

It was also noted that difference in difficulty between subjects was likely also present at GCSE, and that if adjustment were to made at A level it would also be necessary to consider subject alignment lower down. In doing so, there would likely be similar tensions between prioritising parity and ensuring appropriate progression from GCSE to A level study in a given subject.

Finally one exam board advocated that if any adjustment were to be made to grading standards this should not occur immediately, but as part of any future reform to A levels.

2.6 Higher Education Representatives' Perception of Grade Standard Adjustment

Curcin, Black and He (2018) investigated how a potential grade standard adjustment in A level French, German, Spanish, physics, chemistry and biology might be perceived by Higher Education (HE) representatives, as the key users of A level qualifications. This study investigated whether an adjustment, and what magnitude of adjustment in each subject: would be discernible to the participants; would be acceptable to them; and might impact on the utility of those grades for admissions purposes.

To identify the maximum possible adjustment to each grade boundary in each subject, the Rasch-based method described in He and Stockford (2015, also see next section) was used to identify the range of subject difficulty on single scale, and then identify the implied grade boundary adjustments to bring sciences in line with mathematics, and languages in line with geography. Determining the subjects to align each cognate group to was essentially an arbitrary decision, but one which was necessary for the study and which was judged to be reasonable in light of the similarity of the balance between skills and knowledge in these subjects to those considered in the study; their relative difficulty under Rasch analysis; and the views of stakeholders. In particular, the undergraduate courses which students might use these subjects to access (and which may be regarded as comparable for admissions purposes by HE) were considered, as well as feedback which indicates that stakeholders within mathematics and geography are satisfied with A level standards.

For each subject, representatives from a range of HE institutions as well as learned body representatives and Ofqual subject experts took part in panels which were asked to review samples of work from A level examinations (student scripts) at and below each of the A*/A, A/B, B/C and C/D grade boundaries, including the scripts on each indicative adjusted grade boundary as suggested by Rasch analysis. The scripts that represented the actual qualification grade boundaries were treated as 'benchmark' scripts, i.e. scripts representing the current grade boundary standard.

The scripts in the sample represented the whole work of candidates that were examined at the end of the qualification. For science subjects, for each script/student, this comprised the totality of each student's A level examined work. For the languages this included paper examinations as well as the speaking examination (audio file) but only the examined work taken at A2 level.

For each script in each grade boundary script set, the participants were asked to answer either:

- Whether the student who wrote it was equally deserving of admission to their institution as the ones who wrote the 'benchmark' scripts on the current grade boundary (this decision was only made by those participants whose institutions use the higher grade as an admissions criterion), or
- whether it was similar, better or worse than the 'benchmark' scripts

After this exercise, the participants discussed grade standards issues in panels.

While the Rasch analysis suggested much more of a 'gap' in alignment between the sciences and mathematics than between the languages and geography, the script review exercise suggested that there may be less appetite for grade standard adjustment in the sciences, in contrast with the languages.

Furthermore, more often than not in the science subjects the discussion by the panel indicated a lack of acceptance of any adjustment, perhaps surprisingly even on those few occasions where the collated experts' individual judgements indicated that some level of adjustment might be acceptable. The opposite general pattern was apparent in the languages, where, even in the few cases when the outcome of the script review suggested a lack of clear acceptance of grade standard adjustment, the discussions were overwhelmingly in favour of adjustment. This was also the case even where the statistical adjustment was not proposed at all, as was the case for some boundaries in Spanish.

In summary, this work provides some evidence from the script review of the acceptability of limited potential grade boundary adjustments in some subjects. This is more the case for modern foreign languages than the sciences. However, where this was the case in the sciences, the discussion by the panels did not always support making an adjustment. Detailed results are discussed in the report annexed to this document.

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 (for example, subjects which are either too lenient or too severe 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 some of the subjects under consideration.

3.1 Change in grade boundaries and grade distributions

3.1.1 Aligning to the mean of all subjects (based on Rasch analysis)

Table 10 shows the original grade distributions for A level physics from the 2013 examination series. A level physics qualifications were unitised specifications[4] and used uniform mark scale (UMS)[5] Percentage changes in candidates at individual grades and the changes in the cumulative percentages as well as the shift in grade boundary scores, represented using a percentage of the maximum qualification level 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 grade width determined by its relative difficulty (see Table 1; also see He and Stockford, 2015, for details of the boundary mark adjustment process). The adjusted boundary marks at grade A, B and C would be about 7%, 8% and 10% of the maximum available UMS marks lower than the original boundary marks respectively (note: A level physics was offered by different awarding bodies which all used UMS in 2013 but with different maximum available marks). Since the grade width is 10% of the maximum UMS marks, these changes in boundary marks suggest that the performance standards at the adjusted grade boundaries would be substantially lower

[4]Unitised or modular examinations or qualifications used in England: Assessment units in the examination or qualification can be taken throughout the course of study.
[5] UMS mark is a scaled score used to ensure the comparability of raw marks from different examination series. See http://www.aqa.org.uk/exams-administration/about-results/uniform-mark-scale for a detailed explanation.

than the original performance standards at these grades. The cumulative percentage of candidates at grade A would go up by nearly 17% (from about 31% to nearly 48%). At grade C, the cumulative percentage of candidates would increase by over 14% (from 74% to over 88%). As is clear, aligning standards of physics to those of other subjects statistically would make its already skewed grade distribution to skew even more. Judged by the relative difficulties shown in Table 1, the changes in both grade boundary marks and grade distributions for chemistry would be slightly smaller than those for physics. For biology, the changes would be considerably smaller than those for physics.

Table 10 Changes in percentages of students receiving individual grades and cumulative percentages for A level physics in the 2013 examination series after alignment of standards statistically to the mean of all subjects (based on Rasch analysis).

Number of candidates	Grade distribution (%) and grade boundary change (% of max UMS marks)						
		A*+A	**B**	**C**	**D**	**E**	**U**
35,781	Original (Ind.)	30.94	23.29	19.41	14.28	9.00	3.07
	New (Ind.)	47.80	22.94	17.57	9.91	1.64	0.14
	Change (Ind.)	**16.86**	**-0.35**	**-1.85**	**-4.37**	**-7.36**	**-2.93**
	Original (Cum.)	30.94	54.24	73.65	87.93	96.92	100.00
	New (Cum.)	47.80	70.75	88.31	98.22	99.86	100.00
	Change (Cum.)	**16.86**	**16.51**	**14.67**	**10.29**	**2.93**	**0.00**
	Boundary shift	**-7.13 (A)**	**-8.30**	**-10.30**	**-12.60**	**-14.41**	

A level qualifications generally contain a number of components or units which are graded individually, and it is at the unit level the standards are actually set. Since the qualification level UMS marks represent a linear combination of unit level UMS marks, the subject level percentage changes in UMS grade boundaries which were required for aligning subject standards were also assumed to represent the level of impact on unit level standards and applied equally to the unit UMS grade boundaries for all units in a qualification.

The A level physics Specification A from one of the awarding bodies has six assessment units, with Units 3 and 6 having two options and Unit 5 four options (see Table 11). Table 11 shows the changes in grade distributions for individual units after statistical alignment of subject standards. The pattern of changes in grade distributions for the units is broadly similar to that of changes at subject level. The changes in percentages of candidates at individual grades and the cumulative percentages also vary between the units, with some of the grades having changes larger than the change at subject level while others smaller.

Table 11 Changes in grade distributions at unit level for the 2013 A level physics (Specification A) from a particular exam board after statistical alignment of standards to the mean of all subjects (based on Rasch analysis).

Unit	Number of can.	Grade distribution (%)						
			A	B	C	D	E	U
PA1	14,308	Original (Ind.)	48.69	26.13	15.10	6.69	2.54	0.85
		New (Ind.)	66.03	21.18	9.40	2.61	0.50	0.28
		Change (Ind.)	**17.33**	**-4.95**	**-5.70**	**-4.07**	**-2.04**	**-0.57**
		Original (Cum.)	48.69	74.83	89.92	96.61	99.15	100.00
		New (Cum.)	66.03	87.21	96.61	99.22	99.72	100.00
		Change (Cum.)	**17.33**	**12.38**	**6.69**	**2.61**	**0.57**	**0.00**
PA2	14,306	Original (Ind.)	54.65	23.75	12.15	6.26	2.26	0.94
		New (Ind.)	69.91	18.38	8.51	2.41	0.50	0.29
		Change (Ind.)	**15.26**	**-5.36**	**-3.64**	**-3.84**	**-1.76**	**-0.65**
		Original (Cum.)	54.65	78.39	90.54	96.80	99.06	100.00
		New (Cum.)	69.91	88.29	96.80	99.21	99.71	100.00
		Change (Cum.)	**15.26**	**9.90**	**6.26**	**2.41**	**0.65**	**0.00**
PA3-1	9,599	Original (Ind.)	36.87	29.92	18.57	8.94	3.82	1.88
		New (Ind.)	57.61	22.65	14.04	4.27	1.22	0.21
		Change (Ind.)	**20.74**	**-7.27**	**-4.53**	**-4.67**	**-2.60**	**-1.67**
		Original (Cum.)	36.87	66.79	85.36	94.30	98.12	100.00
		New (Cum.)	57.61	80.26	94.30	98.57	99.79	100.00
		Change (Cum.)	**20.74**	**13.47**	**8.94**	**4.27**	**1.67**	**0.00**
PA3-2	5,234	Original (Ind.)	42.34	28.62	14.54	8.16	4.18	2.16
		New (Ind.)	63.24	17.86	12.55	4.72	1.26	0.36
		Change (Ind.)	**20.90**	**-10.76**	**-1.99**	**-3.44**	**-2.92**	**-1.80**
		Original (Cum.)	42.34	70.96	85.50	93.66	97.84	100.00
		New (Cum.)	63.24	81.10	93.66	98.38	99.64	100.00
		Change (Cum.)	**20.90**	**10.15**	**8.16**	**4.72**	**1.80**	**0.00**
PA4	14349	Original (Ind.)	34.47	17.99	16.03	11.21	8.81	11.50
		New (Ind.)	45.81	19.79	14.10	9.70	6.31	4.30
		Change (Ind.)	**11.34**	**1.80**	**-1.93**	**-1.51**	**-2.50**	**-7.20**
		Original (Cum.)	34.47	52.46	68.49	79.69	88.50	100.00
		New (Cum.)	45.81	65.59	79.69	89.39	95.70	100.00
		Change (Cum.)	**11.34**	**13.14**	**11.21**	**9.70**	**7.20**	**0.00**
PA5-1	6,518	Original (Ind.)	29.40	14.50	14.25	12.49	8.61	20.76
		New (Ind.)	37.76	17.54	15.34	10.29	9.99	9.08

		Change (Ind.)	**8.36**	**3.04**	**1.09**	**-2.19**	**1.38**	**-11.68**
		Original (Cum.)	29.40	43.89	58.15	70.64	79.24	100.00
		New (Cum.)	37.76	55.29	70.64	80.93	90.92	100.00
		Change (Cum.)	**8.36**	**11.40**	**12.49**	**10.29**	**11.68**	**0.00**
PA5-2	**1,027**	Original (Ind.)	23.97	12.50	14.27	11.94	9.05	28.26
		New (Ind.)	31.44	16.23	15.02	11.10	12.13	14.09
		Change (Ind.)	**7.46**	**3.73**	**0.75**	**-0.84**	**3.08**	**-14.18**
		Original (Cum.)	23.97	36.47	50.75	62.69	71.74	100.00
		New (Cum.)	31.44	47.67	62.69	73.79	85.91	100.00
		Change (Cum.)	**7.46**	**11.19**	**11.94**	**11.10**	**14.18**	**0.00**
PA5-3	**2,245**	Original (Ind.)	27.48	15.99	15.59	12.83	9.58	18.53
		New (Ind.)	37.02	19.11	15.77	10.78	9.53	7.80
		Change (Ind.)	**9.53**	**3.12**	**0.18**	**-2.05**	**-0.04**	**-10.73**
		Original (Cum.)	27.48	43.47	59.06	71.89	81.47	100.00
		New (Cum.)	37.02	56.12	71.89	82.67	92.20	100.00
		Change (Cum.)	**9.53**	**12.65**	**12.83**	**10.78**	**10.73**	**0.00**
PA5-4	**4,606**	Original (Ind.)	28.44	13.70	14.89	11.59	8.45	22.93
		New (Ind.)	36.58	17.93	14.11	10.51	9.18	11.68
		Change (Ind.)	**8.14**	**4.23**	**-0.78**	**-1.09**	**0.74**	**-11.25**
		Original (Cum.)	28.44	42.14	57.03	68.63	77.07	100.00
		New (Cum.)	36.58	54.52	68.63	79.14	88.32	100.00
		Change (Cum.)	**8.14**	**12.38**	**11.59**	**10.51**	**11.25**	**0.00**
PA6-1	**9,138**	Original (Ind.)	30.32	15.21	15.87	11.40	10.01	17.18
		New (Ind.)	38.26	15.64	18.91	10.01	15.20	1.98
		Change (Ind.)	**7.93**	**0.43**	**3.04**	**-1.39**	**5.19**	**-15.20**
		Original (Cum.)	30.32	45.54	61.40	72.81	82.82	100.00
		New (Cum.)	38.26	53.90	72.81	82.82	98.02	100.00
		Change (Cum.)	**7.93**	**8.36**	**11.40**	**10.01**	**15.20**	**0.00**
PA6-2	**5,314**	Original (Ind.)	35.72	15.66	13.79	12.97	7.06	14.81
		New (Ind.)	46.37	14.11	17.65	7.06	11.22	3.59
		Change (Ind.)	**10.65**	**-1.54**	**3.86**	**-5.91**	**4.16**	**-11.22**
		Original (Cum.)	35.72	51.37	65.17	78.13	85.19	100.00
		New (Cum.)	46.37	60.48	78.13	85.19	96.41	100.00
		Change (Cum.)	**10.65**	**9.11**	**12.97**	**7.06**	**11.22**	**0.00**

3.1.2 Aligning to A level mathematics (based on Rasch analysis)

It might be more appropriate to align the standards between cognate or semi-cognate subjects and Table 12 shows the original grade distributions for A level biology, chemistry and physics from the 2017 examination series, together with changes in grade boundaries and percentages of candidates at individual grades after alignment of their standards statistically to those of mathematics (based on Rasch analysis, also see Table 1. It is also to be noted that the reformed A level sciences offered from 2017 are linear specifications[6] offered by different awarding bodies with different raw or scaled mark scales). For all three subjects, percentage of students receiving A* or A would go up significantly, particularly for chemistry and physics, with an increase over 12% and 15% respectively. For biology, the change in grade distribution is smaller.

Table 12 Changes in percentages of students receiving individual grades and cumulative percentages for A level biology, chemistry and physics in the 2017 examination series after alignment of standards statistically to those of mathematics (based on Rasch analysis).

Subject	Grade distribution (%) and boundary change (proportion of grade width)						
		A*+A	**B**	**C**	**D**	**E**	**U**
Biology 52840	Original (Ind.)	26.31	21.77	21.91	17.86	8.84	3.31
	New (Ind.)	30.69	22.28	19.71	14.60	8.36	4.37
	Change (Ind.)	**4.37**	**0.51**	**-2.20**	**-3.26**	**-0.48**	**1.06**
	Original (Cum.)	26.31	48.08	69.99	87.86	96.69	100.00
	New (Cum.)	30.69	52.97	72.68	87.28	95.63	100.00
	Change (Cum.)	**4.37**	**4.88**	**2.68**	**-0.58**	**-1.06**	**0.00**
	Boundary shift	**-0.25 (A)**	**-0.25**	**-0.15**	**0.04**	**0.17**	
Chemistry 45349	Original (Ind.)	31.81	23.42	19.74	14.43	7.51	3.09
	New (Ind.)	43.91	20.67	16.04	10.66	5.90	2.82
	Change (Ind.)	**12.10**	**-2.75**	**-3.70**	**-3.77**	**-1.61**	**-0.27**
	Original (Cum.)	31.81	55.23	74.98	89.40	96.91	100.00
	New (Cum.)	43.91	64.58	80.63	91.28	97.18	100.00
	Change (Cum.)	**12.10**	**9.35**	**5.65**	**1.88**	**0.27**	**0.00**
	Boundary shift	**-0.53 (A)**	**-0.46**	**-0.36**	**-0.19**	**-0.08**	
Physics 31997	Original (Ind.)	29.39	20.24	19.41	16.44	10.18	4.34
	New (Ind.)	44.66	20.01	15.99	10.09	6.20	3.05
	Change (Ind.)	**15.27**	**-0.23**	**-3.42**	**-6.35**	**-3.98**	**-1.29**
	Original (Cum.)	29.39	49.63	69.04	85.48	95.66	100.00
	New (Cum.)	44.66	64.67	80.66	90.75	96.95	100.00
	Change (Cum.)	**15.27**	**15.04**	**11.62**	**5.27**	**1.29**	**0.00**

[6] Linear examinations or qualifications: All assessment units in the examination or qualification are taken at the end of the course of study.

	Boundary shift	**-0.80 (A)**	**-0.79**	**-0.70**	**-0.44**	**-0.26**	

To look at how the alignment of standards statistically affects the grade boundary locations in more detail, Figure 21 shows the total scaled mark distributions, original grade boundaries and the adjusted grade boundaries for A level biology, chemistry and physics offered by two of the awarding organisations in 2017 after statistical alignment of standards with mathematics (also see Table 13). As is clear from the figure, the effect of aligning standards on grade boundary locations is relatively small for biology but considerable for chemistry and physics. The A and A* grade boundaries would need to be lowered by about 20 and 27 respectively marks for both chemistry and physics. At grade C, the boundary mark would need to be reduced by about 19 marks for physics and 12 marks for chemistry. All this suggests that statistical alignment of standards would result in the performance standards at the adjusted grade boundaries being considerably lower than the standards represented by the original grade boundaries. These adjusted new standards would need to be judged by the primary users of qualifications (e.g. universities and employers) to ensure that they meet the necessary requirements (see Curcin et al. 2018). 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.

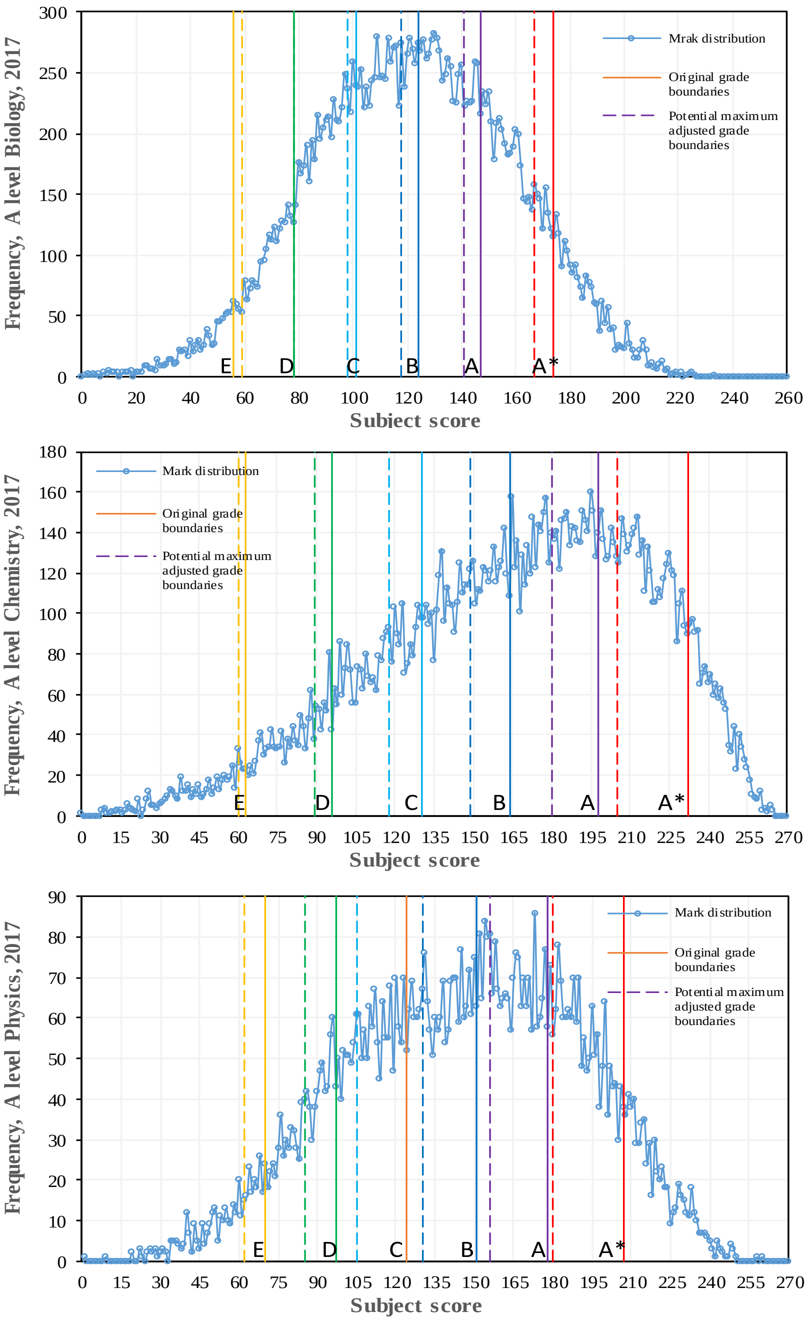


Figure 21 Mark distributions, original grade boundaries and the potential maximum adjusted grade boundaries for A level biology, chemistry and physics offered by two awarding organisations in 2017 after statistical alignment of standards with mathematics.

Table 13 Original grade boundaries and adjusted grade boundaries for A level biology, chemistry and physics offered by two awarding organisations in 2017 after statistical alignment of standards with mathematics.

Grade	Original and adjusted grade boundary marks					
	Biology		Chemistry		Physics	
	Original	Adjusted	Original	Adjusted	Original	Adjusted
A*	174	167	232	205	207	180
A	147	141	198	180	178	156
B	124	118	164	149	151	130
C	101	98	130	118	124	105
D	78	78	96	89	97	85
E	56	59	63	60	70	62

Table 14 shows changes in grade distributions for biology, chemistry and physics when the shifts in boundary marks were set as half of those required by fully aligning their standards with those of mathematics. The grade distribution for biology would be similar to that of mathematics. For chemistry and physics, the percentage of students receiving A* or A would increase by 6% and 7% respectively compared with the original grade distributions.

Table 14 Changes in percentages of students receiving individual grades and cumulative percentages for A level biology, chemistry and physics in the 2017 examination series after partial alignment of standards statistically to those of mathematics (based on Rasch analysis).

Subject	Grade distribution (%) and boundary change (proportion of grade width)						
		A*+A	B	C	D	E	U
Biology	Original (Ind.)	26.31	21.77	21.91	17.86	8.84	3.31
	Change (Ind.)	**1.83**	**0.13**	**-1.00**	**-1.54**	**0.05**	**0.53**
52840	Original (Cum.)	26.31	48.08	69.99	87.86	96.69	100.00
	Change (Cum.)	**1.83**	**1.96**	**0.96**	**-0.58**	**-0.53**	**0.00**
	Boundary shift	**-0.13 (A)**	**-0.13**	**-0.08**	**0.02**	**0.09**	
Chemistry	Original (Ind.)	31.81	23.42	19.74	14.43	7.51	3.09
	Change (Ind.)	**6.06**	**-1.47**	**-1.81**	**-1.83**	**-0.83**	**-0.12**
45349	Original (Cum.)	31.81	55.23	74.98	89.40	96.91	100.00
	Change (Cum.)	**6.06**	**4.58**	**2.78**	**0.95**	**0.12**	**0.00**
	Boundary shift	**-0.27 (A)**	**-0.23**	**-0.18**	**-0.10**	**-0.04**	
Physics	Original (Ind.)	29.39	20.24	19.41	16.44	10.18	4.34
	Change (Ind.)	**7.36**	**0.10**	**-1.56**	**-3.28**	**-2.04**	**-0.58**
31997	Original (Cum.)	29.39	49.63	69.04	85.48	95.66	100.00
	Change (Cum.)	**7.36**	**7.45**	**5.89**	**2.61**	**0.58**	**0.00**
	Boundary shift	**-0.40 (A)**	**-0.40**	**-0.35**	**-0.22**	**-0.13**	

3.1.3 Alignment of standards based on CPA analysis

To expand the work on CPA reported by Newton et al. (2017), Benton and Bramley (2017) looked at the relationship between achievement at A level and attainment at GCSE in the same subject for all A level examinations administered in 2014 using data extracted from the National Pupil Database (NPD). They then investigated the impact of aligning within subject progression from GCSE to A level to the overall progression of all subjects for A level mathematics, biology, physics, French and history. Table 15 below shows the original cumulative grade distributions and the expected (adjusted) grade distributions for mathematics, biology and physics. For biology and physics, the cumulative percentage at grade A, B, C and D would have to increase considerably, with over 10% at B and C for physics.

Table 15 Comparison of expected cumulative grade distribution after aligning progression from GCSE to A level to the overall progression of all subjects for A level mathematics, biology and physics in 2014 (Based on Benton and Bramley, 2017).

A level grade	Mathematics		Biology		Physics	
	Expected	Actual	Expected	Actual	Expected	Actual
A*	13.30%	16.40%	10.80%	12.40%	13.40%	12.10%
A	40.10%	41.30%	33.80%	33.50%	40.00%	34.90%
B	69.00%	64.40%	62.70%	58.10%	68.60%	57.90%
C	87.70%	80.70%	84.20%	77.30%	87.30%	76.10%
D	96.30%	91.80%	95.10%	91.00%	96.10%	89.30%
E	99.30%	97.70%	99.10%	98.50%	99.30%	97.60%
U	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Benton and Bramley further looked at the impact of such alignment on grade boundary marks for OCR A level physics awarded in 2015. Table 16 below shows the original and adjusted raw boundary marks at individual grades for the A2 unit (maximum of 200 marks). As is clear, at B and C, the grade boundary would have to be reduced by over half of the respective grade width. This suggests that the performance standards would be lowered substantially at those grades.

Table 16 Comparison of expected grade boundary marks and original boundary marks for OCA A level physics awarded in 2015 after aligning progression from GCSE to A level to the average progression of all subjects.

A level grade	Original grade boundary	Adjusted grade boundary
A*	163	160
A	144	141
B	127	118
C	109	94
D	91	75

E	67	53

3.1.4 Impact of small changes in outcomes

To see the impact on grade boundaries of small changes in grade outcomes, Tables 17-20 show the entry size, maximum mark and grade boundaries at A*/A, A/B and B/C for the science specifications that had entries greater than 500 in 2018 and the changes in grade boundaries at A*/A, A/B and B/C if outcomes at A* and A were to increase by 0.5%, 1.0% and 1.5% respectively. Specifications have been anonymised in these tables. An increase of 1.0% at A* and A would require boundary marks being reduced by 2 marks at A*/A and 3 marks at A/B respectively for most of the specifications.

Table 17 Entry size, maximum mark and grade boundaries at A*/A, A/B and B/C for the science specifications with entries greater than 500 in 2018.

Specification	Number of candidates	Max Mark	Original grade boundary mark		
			A*/A	A/B	B/C
BIOLOGY A	29230	260	166	138	116
BIOLOGY B	1800	300	197	167	141
BIOLOGY C	4145	300	197	170	145
BIOLOGY D	2460	300	212	180	152
BIOLOGY E	19336	270	187	159	135
CHEMISTRY A	21589	300	241	198	163
CHEMISTRY B	4530	300	250	214	177
CHEMISTRY C	20399	270	237	205	169
CHEMISTRY D	2860	270	209	172	142
PHYSICS A	9093	250	183	150	125
PHYSICS B	940	250	181	147	122
PHYSICS C	2554	250	184	152	127
PHYSICS D	6236	250	181	149	124
PHYSICS E	3421	300	207	174	146
PHYSICS F	747	320	259	221	183
PHYSICS G	9480	270	219	188	158
PHYSICS H	1936	270	206	173	145

Table 18 Changes in grade boundaries at A*/A, A/B and B/C if outcomes at A* and A were to increase by 0.5% for the science specifications in 2018.

Specification	Max Mark	Boundary shift		
		A*/A	A/B	B/C
BIOLOGY A	260	-1	-1	-1

BIOLOGY B	300	-1	-2	-1
BIOLOGY C	300	-1	-1	-1
BIOLOGY D	300	-1	-1	-1
BIOLOGY E	270	-1	-1	-1
CHEMISTRY A	300	-1	-1	-2
CHEMISTRY B	300	-1	-2	-2
CHEMISTRY C	270	-1	-1	-2
CHEMISTRY D	270	-1	-1	-1
PHYSICS A	250	-1	-1	-1
PHYSICS B	250	-2	-1	-1
PHYSICS C	250	-1	-2	-2
PHYSICS D	250	-1	-1	-1
PHYSICS E	300	-1	-2	-1
PHYSICS F	320	-1	-1	-2
PHYSICS G	270	-1	-1	-1
PHYSICS H	270	-1	-1	-1

Table 19 Changes in grade boundaries at A*/A, A/B and B/C if outcomes at A* and A were to increase by 1.0% for the science specifications in 2018.

Specification	Max Mark	Boundary shift		
		A*/A	**A/B**	**B/C**
BIOLOGY A	260	-2	-2	-2
BIOLOGY B	300	-2	-3	-3
BIOLOGY C	300	-2	-2	-2
BIOLOGY D	300	-2	-2	-2
BIOLOGY E	270	-2	-3	-2
BIOLOGY F	270	-2	-2	-2
CHEMISTRY A	300	-3	-3	-3
CHEMISTRY B	300	-2	-3	-4
CHEMISTRY C	270	-2	-3	-3
CHEMISTRY D	270	-2	-2	-2
PHYSICS A	250	-2	-3	-2
PHYSICS B	250	-3	-2	-3
PHYSICS C	250	-2	-3	-3
PHYSICS D	250	-2	-2	-3
PHYSICS E	300	-3	-3	-3
PHYSICS F	320	-2	-3	-4

PHYSICS G	270	-2	-3	-3
PHYSICS H	270	-2	-3	-2

Table 20 Changes in grade boundaries at A*/A, A/B and B/C if outcomes at A* and A were to increase by 1.5% for the science specifications in 2018.

Spec Title	Max Mark	Boundary shift		
		A*/A	A/B	B/C
BIOLOGY A	260	-3	-3	-3
BIOLOGY B	300	-3	-4	-4
BIOLOGY C	300	-3	-4	-3
BIOLOGY D	300	-3	-4	-3
BIOLOGY E	270	-3	-4	-3
CHEMISTRY A	300	-4	-5	-5
CHEMISTRY B	300	-3	-4	-6
CHEMISTRY C	270	-3	-4	-5
CHEMISTRY D	270	-2	-4	-4
PHYSICS A	250	-3	-4	-4
PHYSICS B	250	-4	-4	-4
PHYSICS C	250	-3	-4	-4
PHYSICS D	250	-3	-4	-4
PHYSICS E	300	-4	-4	-4
PHYSICS F	320	-2	-4	-7
PHYSICS G	270	-3	-4	-4
PHYSICS H	270	-3	-4	-4

3.2 Change in performance standards as judged by experts

Jones (2004) studied the impact of aligning standards statistically on the performance standards of A level physics and other subjects using subject pairs analysis (SPA) and expert judgement. The study compared 10 large entry specifications provided by the Assessment and Qualifications Alliance (AQA). A level physics Specification A was found to be about 0.64 of a grade harder than the average of the 10 specifications investigated. Aligning standards at subject level between physics and the other subjects based SPA would require lowering boundary marks at the two judgemental grades (A and E) by 50 UMS marks for physics which represents 5/6 of the subject level grade width (which is 60 UMS marks – 10% of the maximum UMS marks of 600). To examine the impact of new grade boundaries on the quality of performance at the judgemental grades, the adjusted subject level UMS marks were divided between the contributing units to estimate the adjusted unit level raw boundary marks. Scripts near the adjusted subject level boundaries of grades “A” and “E” were drawn and reviewed by the principal examiners (PEs) for their performance standards by indicating their likely position on a notional A level grade profile. It was found that none of the new “A” scripts were identified as having the performance standard required at grade A. Most of the scripts were assigned to the notional grade B. That is, statistical alignment of standards between the subjects would result in a considerable decrease in performance standards at the adjusted grade boundaries for physics.

3.3 Summary

It has been shown that aligning standards statistically between subjects can have substantial impact on the performance standards and grade distributions. For A level biology, chemistry and physics, the quality of performance expected of students, for each grade, would need to be lowered considerably if they were to be aligned with other subjects statistically.

4. Decision

The document detailing our decision on inter-subject comparability in A level physics, chemistry and biology can be found here.

5. References

Allen, T. (2016). What sort of selection stops students studying physics. Institute of Physics Blog: http://www.iopblog.org/role-of-selection/

Benton, T., and Bramley, T. (2017). Some thoughts on the 'Comparative Progression Analysis' method for investigating inter-subject comparability. Cambridge Assessment Research Report. Cambridge Assessment: Cambridge.

Bramley, T. (2011). Developing Methods to Address Comparability Issues. Paper presented at the Perspectives from Cambridge Assessment Seminar 'Comparability of examination standards', 6 April 2011. London.

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. (2010) Understanding comparability of examination standards. *Research Papers in Education* 25, 271–284.

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.

Cuff, B. (2017). Perceptions of subject difficulty and subject choices: Are the two linked, and if so, how? Ofqual: Coventry.

Curcin, M., Black, B. and He, Q. (2018). Inter-subject Comparability – Higher Education representatives' perception of grade standard adjustment in some MFL and science A levels. Ofqual: Coventry.

Cutler, M., McLeod, L., Morgan, N., Tracy, C. and Mist, R. (2017). Open letter to Ofqual from Association for Science Education, Institute of Physics, The Royal Society, Royal Society of Biology and Royal Society of Chemistry.

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

Institute of Physics (2018). Why not physics? A snapshot of girls' uptake at A-level. Available on line at: http://www.iop.org/publications/iop/2018/file_71495.pdf

Jones, B. E. (2004). Inter-subject standards: An investigation into the level of agreement between qualitative and quantitative evidence in four apparently discrepant subjects. Paper presented at the annual conference of the International Association for Educational Assessment, June 13-18. Philadelphia, USA.

Jones, B.E., Philips, D. and van Krieken, R. (2011). Inter-subject standards: An insoluble problem? *AQA:* Manchester.

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

Morse, A. (2016). Training new teachers. National Audit Office: London.

Morse, A. (2017). Retaining and developing the teacher workforce. National Audit Office: London.

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., He, Q., and Black, B. (2017). Progression from GCSE to A level: Comparative Progression Analysis as a new approach to investigating inter-subject comparability. Ofqual: Coventry.

Office of Qualifications and Examinations Regulation (Ofqual) (2015). GCE qualification level conditions and requirements. Ofqual: Coventry.

Panayiotou, S., Boulden, K., Newton, S. and Andersson, D. (2017). Omnibus Survey of Pupils and their Parents/Carers (Research report wave 2). Department for Education: London.

Qualifications and Curriculum Authority (QCA). (2008). Inter-subject Comparability Studies. Qualifications and Curriculum Authority: London.

Stevenson, C. Britton, P., McLeish, T., Pritchard, J. and Price, G. (2016). Open letter to Ofqual from Association for Science Education, Institute of Physics, The Royal Society, Royal Society of Biology and Royal Society of Chemistry.

Stoet, G. and Geary, D. (2018). The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education. *Psychological Science* 29, 581-593.

Taylor, R. (2015). Using the theory of planned behaviour to understand students' subject choices in post-compulsory education. Assessment and Qualification Alliance: Manchester.

Tracy, C. (2016). Grading severity, choice and uptake. Presentation delivered at the Ofqual inter-subject comparability conference in February 2016 held in London.

Wiliam, D. (1996). Meaning and consequences in standard setting. *Assessment in Education* 3, 287–307.

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 2018

Ofqual/18/6450/3