POLICY DECISION

Inter-subject comparability in A level sciences and modern foreign languages

Examining the claim that these subjects are more severely graded than other A levels

ofqual

Contents

Introduction

This document presents our decision on the findings of research we have undertaken on A level grading standards in science and modern foreign language qualifications, in response to claims they are more severely graded than other A level subjects. This is the conclusion of additional work we committed to undertake when we announced our policy position on inter-subject comparability in April 2017.

Inter-subject comparability of standards in GCSEs and A levels has been a matter of debate for a long time. Our investigation into A level grading standards has drawn upon research Ofqual previously carried out on this issue in 2015, and the policy position we are arrived at as a result. The concepts and debates underpinning inter-subject comparability are complex, and readers may wish to familiarise themselves with the challenges it presents by reviewing the inter-subject comparability working papers (in particular the Ofqual Board paper *A policy position for Ofqual on inter-subject comparability*) before reading this document and the accompanying reports.

We reached our decision following detailed consideration of the evidence in the technical reports that accompany this document. We have also conducted additional research that examined perceptions of current A level grading standards amongst representatives from higher education. Whilst we provide a summary of our findings under the relevant criteria in the 'Details' section of this document, our decision should be considered alongside the full range of evidence presented in the 3 reports.

Background

There is already a significant body of research into inter-subject comparability, to which Ofqual has contributed here. In 2015, we sought to start a debate about the concept of inter-subject comparability through the publication of 6 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 Ofqual might pursue.

As a result of this work, and informed by the views of the public, we decided not to take co-ordinated action to align grading standards across all GCSE and A level subjects according to statistical measures of subject difficulty. The paper outlining this policy position can be found here. However, we also decided that we would act to adjust grading standards in individual subjects where there was an exceptional and compelling case – and that we would begin by looking at A levels in physics, chemistry and biology, and French, German and Spanish.

Our decision also recognised the need for a basket of evidence on which to make a judgement about whether to adjust grade standards in a specific subject. From our previous research and our discussions with stakeholders, we determined that this evidence would need to include:

- Statistical data (such as Comparative Progression Analysis, subject-pairs and Rasch analysis)
- Stakeholder concerns (including those from subject-associations and higher education selectors)

- Contextual data (e.g. teacher recruitment figures and A level entry data)

We received guidance from our Standards Advisory Group on how the evidence should be assembled, and to help us arrive at the criteria to be used to determine a compelling case. We were mindful from our previous work on inter-subject comparability (and in particular, that carried out by He and Stockford on Rasch models) that the comparisons between subjects on which statistical measures of subject difficulty are based rely on conceptions of attainment-linking constructs (such as ‘general intelligence’ or ‘general academic aptitude’) which some educationalists reject. The plausibility and relevance of the evidence of ‘difficulty’ produced by these linking constructs diminishes the less similar the subjects that it is used to compare. Comparisons between physics and maths, for example, would seem to be more valid than between physics and music. These constructs also don’t take into account potential differences in student motivation or teaching, which are particularly significant at A level because of the wide range of factors which may have an influence on students’ subject choices. We discuss the strengths and weaknesses of these statistical measures in greater detail in the technical reports, but have concluded we cannot be certain what those statistics are measures of or exactly what they are telling us. Therefore we determined that while statistical evidence would be a key component of any basket of evidence, it must be treated with caution.

Nonetheless, statistical measures of subject difficulty are a source of evidence which, when considered alongside evidence of possible negative impacts upon the 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 they are used in reference to the apparent difficulty of the A level subjects in question under the statistical measures of ‘Rasch’ and ‘Comparative Progression Analysis’ only.

We also considered other possible indicators of relative difficulty. The views of users of these qualifications – including, in the case of A levels, the views of universities on the standards appropriate to pursue further study – are important considerations. We also took into account the views of examiners responsible for making awarding judgements in these subjects, and of others with an interest in the qualifications. Public perceptions of severity in these subjects are longstanding, and may be having an impact on the behaviours of students and teachers due to the belief that a particular subject is difficult irrespective of the actual evidence. But where that perception also aligns with what is suggested by the statistical evidence (or may be being reinforced to some extent by it) and/or other evidence from the views of the users and awarders of these qualifications, then we recognise that there may be an actual issue with standards that we should address. This work also took place in the context of a small adjustment we made to grading standards in French, German and Spanish A levels in 2017 to reflect research into the impact of native speakers taking these qualifications.

We needed to balance different aspects of our statutory objectives and duties as we considered this issue. We are required by legislation[1] to ensure regulated

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

qualifications represent a "consistent level of attainment (including over time) between comparable regulated qualifications". To adjust grading standards in these subjects, we would need to be convinced it was appropriate to prioritise comparability of standards between A levels over the maintenance of consistent standards in a given subject year-on-year. We must also reflect on the impact of any potential action on our objective to secure public confidence in the qualifications we regulate, and our duty to have regard to the views and needs of stakeholders who are 'users' of our qualifications, such as employers and universities.

As well as assembling the evidence needed to reach a decision, we determined the criteria which we would apply to that evidence to judge whether it amounted to a compelling case to adjust the established grading standard. These criteria are presented in full in the 'Summary' section of this document.

We also determined a set of possible actions that we might take in a subject where we deemed those criteria to be met. These potential actions took into account our previous policy position not to seek to align grading standards across all subjects based on statistical measures. Further discussion of why we did not feel this approach would be appropriate is provided in the relevant policy paper.

Summary of decisions

We have concluded that there is not a compelling case to adjust grading standards in A levels in physics, chemistry, biology, French, German or Spanish.

However, we recognise the potential for perceived grading severity to undermine public confidence in these qualifications, and we will therefore consider with exam boards how we should act to avoid the potential for these subjects to become statistically more difficult in the future. .

As we agree the regulatory requirements for awarding[2] in summer 2019, we will consult with the exam boards on using a one-sided reporting tolerance when comparing outcomes against predictions[3] at the A/B and A*/A grade boundaries in these qualifications. This would mean that exam boards could award slightly above prediction[4], but that they would need to provide additional evidence if they wished to award below prediction (or above prediction beyond the reporting tolerance threshold). This should address the perceived risk by some stakeholders within the subject communities for science and modern foreign languages that grading standards might become marginally more severe in statistical terms.

Details

The criteria for an adjustment

[2] The 2018 document is available here: https://www.gov.uk/government/publications/data-exchange-procedures-for-a-level-gcse-level-1-and-2-certificates

[3] Based on prior attainment at GCSE, used for 18-year-olds only.

[4] Within 1, 2 or 3% depending on the size of the entry.

We set criteria to identify whether there was a compelling case for an adjustment to grading standards. These criteria reflect our view that no single piece of evidence could definitively demonstrate the case for an adjustment to grading standards in a given subject (particularly in light of the limitations identified in the statistical evidence) and that we would need holistically to consider of a wide range of factors.

In summary these criteria were:

- agreement between different statistical measures of subject difficulty from several years of entry that there was evidence of persistent grading severity
- persuasive evidence of the potential detrimental impact caused by severe grading on users of the qualification and on society at large
- evidence that those who use the qualification and those responsible for maintaining the grading standard judge an adjustment to be acceptable
- that the likely benefits to users of the qualification and society as a whole from a change to grading standards outweigh any potential negative effects

We considered a range of actions we could take to address issues of misalignment of grading standards in these subjects, if we found this to be the case. We also thought about the practical considerations related to our possible actions. This included the potential impact on students from earlier and future years of any significant shift in standards in a particular year, given the impact this could have on students who might be competing with each other using their grades in that subject.

We based our decision on the defined set of criteria described above, taking into account all the relevant factors in each subject.

Before determining from the criteria whether there was a compelling case to adjust grading standards within a particular subject, we first ensured:

- There had been testing/modelling to demonstrate that the qualification would continue to support effective differentiation if an adjustment was made (i.e. through modelling the impact on qualification outcomes of the proposed change using 2018 results data)
- There had been testing/modelling to demonstrate that the impact of any adjustment in standards would be to reduce the apparent difficulty of the subject according to statistical measures, such as Rasch and Comparative Progression Analysis
- Concerns raised by stakeholders about the subject could not be better addressed through changes to the subject content or assessments.

We considered the criteria holistically: meeting one specific criterion or an absence of evidence for another was not deemed sufficient to determine whether an adjustment should or should not be made in a given subject.

Our evidence base

We applied the criteria to the broad range of evidence in our 2 technical reports on A level sciences and languages. These technical reports summarise the statistical, contextual and stakeholder evidence we gathered.

The evidence we assembled for each subject included statistical measures of subject difficulty, such as Rasch and Comparative Progression Analysis, and the concerns of stakeholders, including the views of both subject associations and other relevant stakeholders and higher education. We also included contextual evidence such as data on teacher supply and recruitment, figures showing 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 the motivations behind students' subject choices.

As A levels are widely used for university admissions, we researched the likelihood that those responsible for admitting and teaching students on undergraduate degrees would accept a grading standard adjustment in these subjects without responding in a way which would diminish the intended benefit (for example by raising entry requirements for their courses). We considered wider ramifications of any changes to grade standards because they could cause unacceptable changes in interpretation of performance standards in individual subjects, and bring into question the validity of A level qualifications in relation to their stated purposes.

Analysis

The picture presented by the evidence in each subject was mixed, with subjects which presented strong evidence under one criterion generally showing either weak or contradictory evidence under others. This made it challenging to assess whether the evidence in an individual subject presented a compelling case for an adjustment.

We took into account the strength of evidence against each criterion, but our overall judgements were holistic and based on consideration of the criteria as a whole.

We considered 4 potential responses to the evidence. These are discussed later in the paper, alongside our explanation for our decision. We also provide a brief outline of the actions we considered but rejected at an early stage.

Before reaching our decision, we undertook statistical monitoring of the likely impact of the 4 options. This enabled us to consider whether the options would have an impact on the purpose of the qualifications – for instance, by lowering grade boundaries to an extent likely to be viewed as unacceptable, or limiting the ability to effectively differentiate between candidates at certain grade thresholds. This modelling is presented in the technical reports.

The technical reports summarise the evidence we gathered in each subject area, and from which we reached our conclusions.

A brief summary of the evidence in each subject in relation to our criteria is provided below. We have summarised the evidence by subject under each criterion for ease of reference. The criteria are in italic font.

Summary

Criterion a). Statistical measures of subject difficulty show evidence of persistent grading severity over several years

To judge if this is the case, we would expect to see evidence of the following:

i. *Different forms of statistical evidence align to indicate potential grading severity in the subject, including Rasch and Comparative Progression Analysis from several years of entry*
ii. *The average level of difficulty of the qualification, as indicated by statistical measures, is above average for key and/or most grades*

All 6 subjects were of above average difficulty under both Rasch analysis and Comparative Progression Analysis.

Physics

Physics was the second most severe subject under Rasch analysis in both years considered (2013 and 2017), second only to further maths. Comparative Progression Analysis suggests that physics was consistently more severely graded for pupils obtaining either a B or A grade in the subject at GCSE than any other A level subject analysed.

Chemistry

According to Rasch analysis chemistry was the third most severely graded A level subject in both 2013 and 2017. Under Comparative Progression Analysis, for students with a B at GCSE, chemistry was more severely graded than most other A levels in 2010, 2013 and 2016 – with only physics being consistently more so. For students with prior attainment of an A at GCSE, the severity of chemistry fluctuated. It was second only to physics in 2013, but more lenient than French, German and Spanish in 2016.

Biology

A level biology appeared to become slightly more lenient under Rasch analysis between 2013 and 2017, moving from fourth most severe subject to fifth, but remaining consistently more severe than the languages. According to Comparative Progression Analysis, biology was consistently more lenient than chemistry and physics for students obtaining a B grade in those subjects at GCSE, and generally also more lenient than French. For students with an A in the subject at GCSE, biology was consistently more lenient than physics, chemistry, German and Spanish.

French

French, like the other languages, appears on Rasch analysis to be more lenient overall than physics, chemistry and biology, but is still more severe than average. It is also more severe than Spanish and just more severe than German – despite the fact that these A level languages share common subject content and a common assessment structure. It was the seventh most severe subject at A level in 2013 and 2017. In 2017 there was a +1% adjustment to predictions at grade A to take into account the impact of native speakers in the cohort. Although this contributed to an increase of 1.8% in the number of candidates achieving A* and A in that year, it did not appear to change the relative difficulty of French when compared to other adjacent A level subjects. The difficulty of French under Comparative Progression Analysis has fluctuated, although it generally appears to be less severe for pupils obtaining either a grade B or A at GCSE than physics and chemistry, as severe as, or more severe than, German, and more severe than Spanish.

German

German was the eighth most severe subject at A level in 2013 according to Rasch analysis, and the eighth most severe in 2017. German appears to be only slightly more lenient than French, but more lenient than physics, chemistry and biology and more severe than Spanish. As with French and Spanish, the native speaker adjustment of +1% to prediction at grade A contributed to more candidates obtaining A* and A grades in 2017 (0.5% and 1.4% respectively), but did not appear to change the relative difficulty of German under Rasch analysis. The difficulty of German under Comparative Progression Analysis has fluctuated for students obtaining GCSE grades A and B, with German generally appearing to be more lenient than the sciences and French (and at A, often more lenient than Spanish).

Spanish

Spanish is the most lenient of the six subjects considered according to Rasch analysis – significantly more so than the sciences. It appears to have become relatively more lenient since 2013, moving from the ninth most severe to the thirteenth most severe subject in 2017. It remains of above average difficulty however, although only just in 2017. Spanish is the only subject which saw a change in relative difficulty compared to adjacent subjects coinciding with the 1% native speaker adjustment at A, following which 2.0% more students obtained A* grades and 2.6% more achieved A grades. Under Comparative Progression Analysis Spanish appears to be relatively lenient for pupils obtaining a grade B at GCSE, with greater than average attainment than in physics, chemistry, biology, mathematics, and German in most years. However, for students who achieved a GCSE grade A, Spanish appears to have been relatively hard in recent years. Average attainment in Spanish in 2010 was lower than any subject bar physics in 2010, physics and chemistry in 2013, and physics in 2016 (with average attainment equal in Spanish and French in that year).

Criterion b) There is persuasive evidence of the potential detrimental impact caused by severe grading on those who use the qualification and on society at large over several years

To judge if this is the case we would expect to see persuasive evidence of negative impacts, which may include the following:

i. *Depressed uptake within the courses to which students taking the subject would be expected to progress*
ii. *Depressed entries within the subject*
iii. *Indications of issues in securing a sufficient supply of teachers*[5]
iv. *Indications of skills shortages related to a lack of take up of the qualification.*

Physics

Entries for physics have steadily increased over the past 10 years (it was the ninth most popular A level in 2017 and eighth in 2018). Acceptances to university courses likely to require A level physics have also increased, although stakeholders have argued that the potential rate of increase is being depressed by perceptions of difficulty. Boys significantly outnumber girls; physics was the subject with the second highest ratio of male to female students in 2017, second only to computing. This low percentage of female students has remained relatively static in recent years despite the increase in entries overall – and significant efforts within the STEM community to attract girls into the subject. This is despite the fact that girls' outcomes are comparable to or slightly better than those for boys. There is evidence that the subject may have become more selective over time (either due to schools increasing entry requirements to study the subject, or less able students being put off studying it), although not on the basis of mean GCSE physics grade. There is also evidence to suggest that physics is struggling to recruit sufficient trainee teachers to meet future requirements. Those who do begin post-graduate teacher training are less likely to hold at least an upper second degree in physics than in other subjects.

Chemistry

Over 50% of students studying chemistry at A level in 2017 were female, and entries at A level are high overall (though stable, rather than increasing as in physics). University acceptances to courses allied to chemistry, of which there is substantial overlap with physics in the UCAS JAC3 reporting categories, are also increasing. DfE and NAO figures indicate that postgraduate entries to teacher training are just under the Teacher Supply Model target – although a relatively low proportion of trainee chemistry teachers in 2016 had at least an upper-second degree in the subject.

Biology

Biology was the second most popular A level subject in 2016. Entries in 2017 were higher than those for both physics and chemistry, although they have declined very slightly since 2013. Acceptances for university courses allied to biology (e.g. JAC3 Group C Biological Sciences and Group B Medicine) are buoyant, and indeed Group C acceptances have shown the greatest increase of any subject group over the past decade. Over 60% of biology A level students in 2017 were female, and this gender ratio has remained stable over the past decade. DfE and NAO statistics differ on whether the number of postgraduate entrants to teacher training courses for biology exceeded or fell just short of teacher recruitment targets in 2016. The number of

[5] Whilst appreciating that a shortage of appropriately qualified teachers could also be a possible cause of the apparent severe difficulty we see under statistical measures, in addition to being considered as potential evidence of the long-term impact upon entries.

teachers entering initial teacher training with at least an upper second degree in the subject was in line with the average figure for all subjects.

French

Entries for A level French are in long-term decline, falling from approximately 15,000 candidates to under 8,000 over the past decade. Acceptances to associated university undergraduate courses (JAC3 Group R European Langs, Lit & related) have also declined, and there has been a significant decrease in the number of universities offering single and joint honours degrees in French (which almost halved between 1998 and 2015). Data on prior mean GCSE attainment of French A level students suggests that the subject has not become more selective over the past decade, but there has been a decline in the number of students studying more than one language at A level over the same period. Figures from both the DfE and the NAO indicate that the decline in uptake may now be contributing to issues with teacher supply, with recruitment to postgraduate teacher training falling short of the required number indicated by the Teacher Supply Model. However more than three quarters of postgraduate entrants to teacher training in modern foreign languages in 2016 had at least an upper second 2:1 degree or above, which was greater than the overall proportion of postgraduate trainee entrants.[6] There is evidence to suggest that the decline in the number of schools offering A level languages is more pronounced in the state sector than in independent schools, and also that state schools are finding it harder to recruit and retain language teachers.

German

Entries in German have not fallen to the same extent as in French, but the proportional decrease is equivalent. In 2018, entries for A level German were fewer than 3,000, and the subject has now been overtaken in popularity by Chinese. Like French, German is experiencing a decline in university uptake with a decreasing number of institutions offering single and joint honours degrees. German is also falling short of Teacher Supply Model targets. More than three quarters of postgraduate entrants to teacher training in modern foreign languages in 2016 possessed at least an upper second 2:1 degree or above, which was greater than the overall proportion of postgraduate trainee entrants. There is also evidence to suggest that the decline in pupils studying (and schools offering) the subject has been greater in the state sector than in independent schools. Whilst data indicates that students studying A level German have relatively high prior attainment and the overall mean GCSE achievement of students taking this subject has increased slightly, their mean German GCSE grade has remained stable. This suggests that perceived difficulty has not led to the subject becoming more selective over time.

Spanish

A level Spanish entries are bucking the trend seen in French and German, having increased gradually since 2008, although there was a slight downturn in entries in 2018. Spanish is now almost as popular as French and more than twice as popular as German at A level, and might overtake both in 2019. Specific data on university acceptances for this subject is not available, as the figures for Spanish are combined

[6] The DfE and NAO group language subjects for reporting purposes, so these figures apply to French, German and Spanish.

into the same UCAS reporting group as French and German (Group R European Languages, Lit & related). However, acceptances for this reporting group have decreased overall. As with French and German, a number of universities have stopped offering single and joint honours degrees in the language over the past decade. The fact this has happened in Spanish despite increasing A level entries may call into question the assertion of stakeholders that the negative trends in this subject are attributable in any significant way to the effects of severe grading. More than three quarters of postgraduate entrants to teacher training in modern foreign languages in 2016 had at least an upper second 2:1 degree or above, which was greater than the overall proportion of postgraduate trainee entrants.

Criterion c) There is evidence which shows that those who use the qualification and those responsible for maintaining the grading standard judge an adjustment to be acceptable

Under this criterion we would consider:

i. *The views of those who use the qualification (for A level subjects this includes those within higher education who utilise it to inform admissions decisions)*
ii. *The views of the exam boards, and specifically the judgement of those examiners responsible for making awarding decisions.*

Evidence under this criterion comes from the Ofqual higher education perceptions study, and a survey of the senior examiners responsible for awarding.

Physics

While there was weak evidence from the research study that higher education would accept a small lowering of grade boundaries at the A*/A and A/B thresholds only, support for lowering grade boundaries was not strong. A number of participants suggested that the current A* threshold was a good discriminator for university selection and they would not want to see more students awarded this grade. Some though suggested there was some scope for grade thresholds at grades B, C and D to be lowered.

Overall, exam boards' awarders were not in favour of lowering grading standards in physics. One awarding panel was of the view that the current grading standard in physics was correct, another that the standard was possibly slightly severe, but the remaining two panels were of the view that the standard was in fact too lenient. These concerns about leniency mainly focused around the E boundary (where they felt the quality of work was poor), with awarders generally satisfied that the standard required for an A grade was appropriate. The view was expressed that the requirement to combine scientific knowledge and understanding with advanced mathematical skills and problem solving made physics an inherently more challenging subject than other A levels. Unsurprisingly, given their view of the standard, awarders felt that in their professional judgement the grade boundaries in this subject should not be lowered.

Chemistry

The research study suggested that higher education would accept a lowering of the grade boundary at the A*/A threshold, and to a lesser extent at A/B. This contradicted the subsequent discussion amongst the participants, however, who indicated they would not support changes to the A*/A and A/B boundaries. One participant suggested there was an argument to raise their admissions criteria based on the *current* grading standard, and another stated that lowering the A* or A boundaries would have a significant impact on admissions.

Most respondents to the awarder survey did not favour a change to grading standards. One exam board's awarders were divided in their view of the standard – regarding it as either slightly severe or slightly lenient. Those at the 2 remaining boards felt that the standard at grade A was correct, but that the E threshold was slightly lenient. Support for increasing thresholds at E was not strong. The view was expressed that the inherent demand of the subject meant that students would consider it challenging regardless of any change to grading standards. The view that teachers were generally satisfied with standards in this subject was expressed.

Biology

In biology there was evidence that higher education would tolerate an adjustment at the A/B threshold only. This was reflected in comments from participants who generally thought that adjusting the standard at other grades would not be advisable (although it should be noted that it would not be possible to change the A/B or E/U thresholds without having an impact on the arithmetically calculated grade boundaries in between). Indeed, some awarders stated that they thought the A* boundary was too low.

Most biology awarders felt that, on balance, grade boundaries should not change. There was more difference of opinion seen here than for physics or chemistry, and awarders' views were complicated by the fact that one board offers 2 different specifications for this subject. Of the 2 boards that offer only a single specification, the awarding panel at one considered the current grading standard slightly lenient, and the other felt the standard was broadly correct. At the board which offers 2 specifications, awarders were satisfied with the standard for one but unable to reach a consensus about the other. The awarding panels for 2 of the 3 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.

French

For French, the study indicated that higher education would be likely to tolerate a limited adjustment to grading standards at A/B, and a smaller adjustment at B/C. There was strong support from participants for an adjustment to grading standards, particularly at the A/B threshold – although with some indication when the reasons for this were discussed that this was a response to declining university entries, rather than because the demands of the current grading standard were too high.

Awarders' views on the grading of French were mixed, with as many supportive of maintaining the current standard as felt that it might be slightly severe. At one board the panel agreed that there was scope to relax the standard marginally at both the A/B and E/U thresholds. At the other board, awarders suggested that any change should be at grade A only – although one dissented, arguing that the standard at this threshold should in fact be raised as they expected candidates obtaining an A to demonstrate a higher level of proficiency. All awarders recognised the beneficial impact that a downward adjustment to grading standards would have on students' and teachers' perceptions that the subject is difficult, and were generally of the view that a minor adjustment to grading standards could be tolerated if it addressed the declining uptake of A level French.

German

The evidence suggested that higher education would also be likely to tolerate an adjustment at the A/B and B/C threshold in German, although in both cases on a lesser scale than for A level French. Again, support for adjusting grading standards was strong, with one participant claiming the majority of scripts were awarded one grade lower than they deserved.

Whilst the individual views of exam boards' awarders encompassed severity, leniency, and that the current standard was about right, awarders from one board felt that grading standards were broadly correct whilst the other board's awarders felt that the subject was graded severely at A. Where awarders felt that the standard required of candidates was incorrect, this was attributed to changes in the assessment made as part of the reforms, rather than a longstanding misalignment of the standard. Both awarding panels felt there was a widespread perception that German is more difficult than other A levels, and that this was dissuading students from studying it in the belief that they will secure better grades elsewhere. However, only at one board did awarders feel that this viewpoint was justified. Whilst both awarding panels generally felt an adjustment to standards might help to address declining entries in the short term, this was not a unanimous view. One awarder felt that negative perceptions of the subject would likely persist regardless. Generally awarders were in favour of lowering the grading standard at the A threshold if this would help to address the decline in A level study.

Spanish

There was very strong support from participants in the higher education study for lowering grade thresholds in Spanish at A/B, and for more substantial adjustments at A*/A and C/D. There was also unanimous support for an adjustment of over five marks at the B/C threshold, though according to Rasch analysis this was already aligned with the mean A level 'difficulty'. This was the strongest support for an adjustment seen in the research study, which interestingly was in the subject where the evidence for an adjustment based on statistical measures of difficulty was the weakest.

Awarders were split in their opinions on the current grading standard in Spanish. One board's awarding panel was of the view that the current standard in A level Spanish is correct (with students obtaining the 'right' grade in this subject). Awarders at another board felt that the subject was graded severely at grade A. The third board's awarders were divided over whether the standard was correct (particularly now that an adjustment had been made to take into account the impact of native speakers) or slightly severe. All 3 panels felt the standard at grade E was correct. All of the panels felt that Spanish was considered to be more difficult than other subjects by students, particularly in terms of achieving grades A* and A, and that this was leading them to study alternative A levels which they considered 'easier'. This was attributed to the impact of native speakers within the cohort, rather than a misalignment of standards. All were generally of the view that lowering grading standards would improve take-up of the subject.

Criterion d) The likely benefit to users of the qualification and society as a whole from a change to grading standards must outweigh any potential negative effects

To judge if this is the case, we would expect that:

i. *There is evidence of support from users of the qualification for any change*
ii. *There is no reason to believe that there would be a detrimental impact on the extent to which the subject fulfils the defined purpose of the qualification*[7]
iii. *There is no reason to believe that any change in standards would have a detrimental impact on performance standards, for example by decreasing the level of cognitive demand in the subject in comparison to other cognate subjects*
iv. *There is no reason to believe that there would be a significant detrimental impact to other parts of the education system as a result of an adjustment.*

The higher education perception study suggested only small adjustments would be acceptable. These would be unlikely to have a discernible impact on the order of relative difficulty according to Rasch.[8] It is unlikely that, if we changed grading standards in the modest way suggested by the higher education study, we would create new Rasch outliers which would prompt objections from other subjects.

The exam boards were concerned 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. The boards felt 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. They cited the findings of 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 the exam boards 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

[7] In the case of A levels, the purposes defined in the *GCE Qualification Level Conditions* include providing the knowledge, skills and understanding needed by students for progression to higher education; and permitting universities to accurately identify student attainment.
[8] Adding 1% to the prediction at A/B in French, German and Science in 2017 as a result of the native speakers research led to an increase of 1.8%, 1.4% and 2.6% at A and 1.8%, 0.5% and 2.6% at A* respectively, but did not change the position of French and German as the 7th and 8th most 'difficult' subjects under Rasch 2013 – 2017. Spanish appeared to become more lenient over this period, moving from 9th to 13th most severe subject.

have taken a ‘hard’ (i.e. potentially more severely graded) subject when applying to university, compared to those who have taken an ‘easy’ (i.e. less severely graded) one.

Increases in university entry requirements were considered a particular risk if changes were made mid-cycle, rather than at the defined end-point of a specification, as universities may adjust their expectation of the performance standard of a pass 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. Those in the cohort immediately preceding any adjustment may be competing for jobs and university places with students who achieved the same grade for a lower level of performance, and students in successive cohorts might be regarded as holding a devalued qualification. To mitigate this risk, boards felt that any adjustment should be made incrementally.

It was noted by the exam boards that differences in difficulty between subjects are also likely to be present at GCSE, and that if adjustments were to be made at A level it would be necessary to consider subject alignment lower down as well. In doing so, there would be similar tensions between prioritising parity and ensuring appropriate progression from GCSE to A level study.

Ultimately, boards were of the view that, to achieve greater comparability between subjects, it might 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 to differentiate between applicants.

Views were also gathered from the higher education research study and exam board awarding panels on the potential impact of a change to grading standards. These are summarised below by subject.

Physics

Physicists participating in the higher education perception study were concerned that lowering the threshold at A*/A might complicate admissions decisions. They were also concerned about the public response to a decision to lower grading standards, as the cumulative percentage of A* and A grades achieved in the subject is already high in comparison to other subjects.[9] Some may view any adjustment as being akin to grade inflation.

Awarders for A level physics felt that an adjustment to grading standards was not only unacceptable, but would actively damage the subject – citing concerns about perceptions of dumbing down, decreased university retention rates, and a loss of comparability with international physics qualifications.

Chemistry

Some chemists participating in the higher education research study argued against an adjustment to grading standards in the subject at A*/A because this would have

[9] [9] In 2018 9.3% of A level physics students achieved A*, and 29.2% grade A. In comparison. 5.8% and 23.5% of students sitting A level history obtained A* and A respectively, and only 4.5% and 17.7% of students studying A level psychology.

an impact on their ability to discriminate between candidates when making admissions decisions. This sentiment was generally restricted to representatives of those selective institutions recruiting on A*.

None of the exam boards' awarding panels felt strongly that any adjustment in grading standards would address perceptions of difficulty in this subject, and were anxious about the impact of an adjustment on the perceived value of the qualification.

Biology

Most awarders felt that, overall, the potential risks of an adjustment to grading standards outweighed the beneficial impact this might have in addressing perceptions of difficulty. They also believed most stakeholders in higher education 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 the A level grades on which they based their offers.

This reflected the views of the biologists participating in the higher education research study, who generally thought that adjusting the grading standard at thresholds other than A/B would not be advisable. Some voiced a concern that the A*/A boundary was already too low.

French

Awarding panels disagreed over whether an adjustment to standards would be necessary to address the decline in entries in French, with some arguing that the perception of difficulty may wane in coming years as reformed GCSEs provided better preparation for A level study. Awarders were also concerned that lowering standards at grade A would likely make it more challenging for universities to identify students with the necessary grammatical knowledge to cope with undergraduate courses.

This was not supported by the findings of the higher education perceptions survey however, where participants were generally in favour of an adjustment at the A/B threshold.

German

Some awarders felt that intervention to adjust grading standards might be unnecessary, as the reforms to GCSE will better prepare candidates for progression to A level and their experience of the subject will improve.

Higher education representatives participating in the research study were broadly supportive of minor changes at the A/B and B/C thresholds, but not at A*/A or C/D.

Spanish

Whilst there was overall support for relaxing grading standards if this would help to address declining uptake, awarders were also mindful of the need to ensure there was meaningful progression from GCSE to A level. This was felt to be particularly important so students were adequately prepared for further study and universities continued to value the A level for admissions purposes.

University representatives participating in the HE perceptions study however were strongly in support of lowering grading standards at all thresholds considered.

Options considered

On the basis of the evidence outlined above, we rejected some options at an early stage as inappropriate. For completeness, we have summarised these below.

One option considered but eliminated early on was to make significant adjustments to all 6 subjects to move them to the point which represents average ‘difficulty’ under statistical measures such as Rasch. The scale of the adjustment that would be required in subjects such as physics and chemistry (between 7% to 10% of the total mark, almost a grade width) would be similar in practice to a wholesale realignment of standards of the type ruled out by our policy decision in 2017, and would create a new set of ‘outlier’ subjects. In particular, questions might be raised about the apparent difficulty of A level mathematics, as another facilitating STEM subject which would then appear to be out of alignment with the sciences. It is also not clear what we would do if the relative ‘difficulty’ of these subjects changed in future years. Would we adjust further to ensure they remained of average ‘difficulty’?

A variation of this was the potential to align sciences and languages to similar but non-cognate subjects such as maths (for the sciences) and geography (for languages), which some argue are treated comparably for admissions purposes. This is the approach favoured by some languages stakeholders for French, German and Spanish (and we used these subjects as the basis for the maximum potential adjustment to grading standards modelled in the higher education perspectives study). However, the actual similarity of these subjects is questionable, and their selection ultimately arbitrary. Furthermore, the same issue would arise if in subsequent years the ‘difficulty’ of these subjects appeared to change – would we in essence be pegging A level science and languages to them, potentially requiring further adjustments in the future.

We considered decreasing the predictions used when setting grade boundaries at A*/A and A/B in Spanish (thus making the subject more severe), to bring it into closer alignment with French and German, if we were also adjusting standards in those subjects. This would effectively move all 3 towards a new median difficulty. We decided this was not a feasible option however, as it would mean deliberately increasing the difficulty of Spanish when the statistical evidence suggests that it is already harder than the average A level. Furthermore, the higher education perception study showed the same or greater support for lowering thresholds in Spanish as in French and German. We did not feel that we could justify increasing the apparent difficulty of Spanish contrary to the evidence of that study, whilst simultaneously citing support from higher education as the basis for lowering thresholds in French and German.

We also considered decreasing the prediction at the A*/A and A/B threshold in biology if we adjusted standards in physics and chemistry, to match the action suggested above for Spanish. We disregarded this for the same reason: it would entail increasing the difficulty of a subject which statistical evidence suggests is already harder than most A levels. We were also less persuaded than in the case of languages by the argument that as cognate subjects, science A levels should be of similar difficulty. Science A levels do share the same subject content and assessment structure as modern foreign language A levels do, and biology differs substantially in field of knowledge from physics and chemistry. As such, it is harder to argue that they should be appear to be more closely aligned under statistical measures of ‘difficulty’. Biology also already appears to be more severe than all 3 languages. It would be difficult to justify making biology more severe whilst simultaneously easing standards in French and German on the basis that the evidence suggested these language subjects were inappropriately severe.

We considered 4 other actions that seemed to offer a reasonable response to the evidence according to our criteria, before rejecting them in favour of the recommendation presented earlier in the paper. These alternative options are provided below to illustrate our thinking.

Option A

Take no action to adjust grading standards in A level French, German or Spanish on the basis that our criteria for a compelling case have not been met. There is a lack of persuasive evidence for criterion a, and the evidence under criteria d and to some extent c is mixed. The evidence under criterion b is apparently strong, but causation is questionable.

Take no action to adjust grading standards in physics, chemistry or biology on the basis of lack of persuasive evidence for criterion b, and mixed evidence under criteria c and d. The evidence under criterion a appears strong, but there limitations to the evidence which lead us to question its validity.

In taking no action, we would endorse current grading standards in these subjects.

Option B

Adjust prediction +1%[†] at the A/B threshold (and therefore also the semi-arithmetical A*/A threshold) in physics and chemistry.

Take no action to adjust grading standards in biology, French, German or Spanish on the grounds that our criteria for a compelling case have not been met.

Option C

Adjust prediction +1%[†] at the A/B threshold (and therefore also the semi-arithmetical A*/A threshold) in physics only.

Take no action to adjust grading standards in chemistry, biology, French, German or Spanish on the grounds that our criteria for a compelling case have not been met.

Option D

† Potentially increasing to 2% dependent upon the outcome of modelling, if this were to suggest that a 1% adjustment would not have a meaningful impact on apparent difficulty.

Adjust prediction +1%† at the A/B threshold (and therefore also the semi-arithmetical A*/A threshold) in physics and chemistry. Do the same at the A/B threshold in French and German to achieve greater inter-subject comparability between these languages and Spanish, which have common subject content and share the same assessment design.

Take no action to adjust grading standards in biology or Spanish, on the grounds that our criteria for a compelling case have not been met.

Our view

We are not convinced that any of the above options B – D would be appropriate. No subject presented strong evidence under all 4 criteria, that could be considered a compelling case to make an adjustment to grading standards.

Physics, and arguably chemistry, meet criterion a (although the evidence from Comparative Progression Analysis is inconsistent for chemistry), but the evidence for criterion b and criterion c is weak. On the other hand, the evidence under criterion c for French and German is much stronger when compared to physics and chemistry (and even biology), but the evidence under criterion a is weak. Apparently persuasive evidence under criterion b in the form of declining A level entries for these 2 languages becomes less convincing when considered alongside Spanish. Finally, Spanish and Biology both fail to demonstrate any compelling case at all under separate criteria respectively: criterion a in the case of Spanish, and criterion b in the case of biology. It is in these 2 subjects however, that we see some of the strongest support for an adjustment under criterion c.

We also recognised the need for consistency in the weighting we apply to a given criterion when considering it in relation to different subjects, and to be aware of the bigger picture of how these subjects compare to one another.

For instance, physics appears only a little more severe than chemistry under Rasch analysis (and not significantly more severe according to Comparative Progression Analysis for students who obtained either a grade B or grade A at GCSE). However, the gender balance in chemistry is much closer to the A level cohort as a whole – approximately 50% of students are female, compared to only 22% in physics. It seems unlikely that the relative unpopularity of physics with girls is the result of the marginal difference in difficulty. There are also a comparable number of schools with no students attaining less than an A at GCSE entering chemistry as in physics. If this is evidence of school (and/or self) selection as some have suggested, then the male-to-female ratio suggests that severe grading is not dissuading girls from studying chemistry, making the causal link dubious. Given that the statistical evidence suggests that physics and chemistry are of similar difficulty, if we were to adjust standards in physics on the basis that it is ‘difficult,’ it would be challenging to justify why we were not doing this in chemistry as well.

Similarly, French and German both appear to be more lenient under the various statistical measures of subject difficulty than physics, chemistry and biology – all of which are experiencing an increase in entry overall. The number of universities offering joint and single honours languages has also decreased for Spanish, despite increasing A level entries. This raises questions that the negative trends in this subject are attributable to severe grading. If we had decided that severe grading was having an impact upon uptake in modern foreign languages, the extent to which these subjects appear more lenient than physics, chemistry and biology means that logically we should also make an adjustment to grading standards in the sciences.

However, given that at least one of our criteria has been met in each of the subjects considered, we were not satisfied that option A (the 'no action' option) by itself would be sufficient – especially in light of the fact that the perceptions of students and teachers may be having an impact on entry in these subjects regardless of whether they are actually more severely graded than others. Whilst we did not find compelling evidence to lower grading standards on the basis of inter-subject comparability, we do think limited action to address stakeholder concerns that these subjects could become more severely graded in the future is appropriate.

Therefore, we have concluded that we should not make an adjustment to lower grading standards in subjects. Whilst we have however decided that we should act in relation to concerns of stakeholders that the apparent difficulty of these A levels might become more pronounced in the future. We will do this by consulting with the exam boards on a proposal to use 'one-way' positive reporting tolerances when awarding these subjects.

Exam boards use predictions to guide awarders when they set grade boundaries. Reporting tolerances reflect what we think is 'normal' variation around those predictions. In using a positive-only tolerance for these subjects, our rationale is that unless the awarders want to make a case for moving away from the grade boundaries suggested by predictions (and there is a separate mechanism for doing so) then we would expect exam boards to avoid negative variation, given the concerns of stakeholders.

Positive-only tolerances can prevent these subjects from becoming more severely graded at A/B and A*/A in statistical terms. Reporting tolerances for a specification are determined by entry size. This approach could potentially result in a positive variation from the prediction of +1% to +2% in A level physics, chemistry and biology and +1% to +3% in French, German and Spanish in 2019.

The reference series[11] used in future years to generate predictions based on cohort-level prior attainment will be baselined to the first awards of these new specifications. Reporting tolerances will apply each year to these baselined predictions, so any positive changes in the proportion of students achieving the key grades will not be cumulative in subsequent years, unless exam boards can provide evidence to support moving away from prediction.

[11] Reference series used for predictions are anchored back to the start of that specification, and is generally the average of outcomes from the first two years. This is to avoid the cumulative effect of small changes over time without any evidence of improved (or declining) levels of performance.

Implementation timescales

We will consult with the exam boards in spring 2019 on the awarding process for these A level subjects, with a view to putting in place the requirement for exam boards to use positive-only tolerances for the summer 2019 awards in these subjects.

We will review the impact of our decision on A level sciences after the summer 2019 awards, and on modern foreign language A levels following summer 2020.

Future work on inter-subject comparability

We announced our plans to expand our work on inter-subject comparability to include GCSE French, German and Spanish in February 2018. This will require us to gather and consider a range of new evidence and include consideration of the findings of the evaluation of reformed qualifications, and analysis of the first year of results.

We intend to announce our decision on inter-subject comparability in relation to GCSE modern foreign languages in autumn 2019.

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/18/6450/1