RESEARCH AND ANALYSIS

Inter-subject Comparability

Higher Education representatives' perception of grade standard adjustment in some MFL and science A levels

Milja Curcin, Beth Black, Qingping He

Contents

Executive summary 4

Introduction 5

Study aims and questions 7

Methodology 8

Participants 8

Scripts 9

Determining indicative qualification grade boundary adjustments 10

Selecting script samples 15

Script presentation 16

Preparation activity 17

Panel activity 17

Methodological limitations 19

Results 20

Number of responses, response coding and acceptability criteria 20

Perceptions of grade standard adjustment in the languages 26

Overall results 26

Results by question 27

Group discussion points 29

Perceptions of grade standard adjustment in the sciences 31

Overall results 31

Results by question 32

Group discussion points 34

Results summary 35

Discussion 37

References 39

Appendix A: List of participating institutions 42

Appendix B: Order of script presentation 43

Appendix C: Examples of response forms 44

Appendix D: Summary of over 50% majority views 46

Executive summary

This study was commissioned to investigate how a potential grade standard adjustment in A level modern foreign languages (French, German, Spanish), and sciences (physics, chemistry and biology) might be perceived by higher education (HE) representatives, as the key users of A level qualifications. Wider ramifications of any grade standard adjustments are important to consider, not least because they could cause unacceptable changes in interpretation of performance standards in individual subjects.

This study is part of a wider project on inter-subject comparability, being run by Ofqual in order to collect evidence on perceived disparity of grade standards between different A level subjects, flagged by a range of stakeholders as detrimental for subject take-up both at A level and at university. Statistical evidence of differences in grading standards is often cited to support stakeholder concerns although there are definitional and conceptual difficulties in interpreting statistical results. Additionally, any potential change that might be implied by a statistical analysis may or may not help the usability of results in any particular subject. This study therefore is not about inter-subject comparability *per se* but rather to help understand the nature of the impact that a closer alignment (on the basis of statistical analyses) of some subjects to others might have on the utility of their grades.

Using a Rasch-based analysis, grade boundaries were identified which would statistically align sciences with mathematics, and French, German and Spanish with geography. We then identified a sample of student work ('scripts') in each subject at and below A*, A, B and C boundaries. 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. The scripts reviewed represented the actual qualification grade boundaries, the qualification boundary implied by statistical alignment, and mark points in-between.

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 the samples of scripts. We wanted to see whether a potential grade boundary adjustment, and what magnitude of adjustment in each subject:

- would be discernible to our participants in student work that they reviewed,
- would be acceptable to them, and
- might impact on the utility of those grades for admissions purposes.

The expert panel recorded their individual judgements about each script on a recording form. After considering the scripts at and below each qualification grade boundary, the panel had the opportunity to discuss their views on the set of scripts in relation to the points listed above.

While the Rasch analysis suggested much more of a 'gap' in statistical 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, even on the 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 grade boundaries in Spanish.

In summary, this work provides some evidence from the script review of the acceptability of some grade boundary adjustments in some subjects. This is more the case for modern foreign languages than the sciences. However, where this was also the case in the sciences, the discussion by the panels did not always support making an adjustment.

Introduction

Over the last few years Ofqual has been engaged in a wide-ranging debate on inter-subject comparability of standards (ISC)[1], a long-standing issue that England and a number of other jurisdictions have been grappling with, in order to formulate its policy position on ISC. The notion of ISC refers to a debatable requirement for subjects as disparate as science and drama to have equivalent/comparable (examination) standards. Even though ISC has been discussed and researched for many years, reviews such as Ofqual (2015b) or Newton (2012) demonstrate that we have not come much closer to even defining the problem in unambiguous terms, let alone found a way to operationalise a solution to it.

Nevertheless, different approaches to addressing the problem of ISC have been proposed in the literature, most prominently a range of statistical approaches to detecting and aligning disparate standards between subjects (for example subject pairs analysis, Kelly's method, Rasch analysis, scaling methods, comparative progression analysis, etc.). These have influenced the debate on ISC to a great extent, and, arguably, even the perceptions of ISC amongst various stakeholders, who have been demanding action with respect to aligning statistical standards in specific A level and GCSE subjects in England (for example, Myers 2006; Dearing and King 2007; Coe et al. 2008; Royal Society 2008). Indeed, many who responded to Ofqual survey of policy options regarding ISC, which was run in 2015, expressed a preference for action based on Rasch-based statistical measures of subject difficulty.

Having reviewed a range of literature, and having investigated technical, practical and policy issues in relation to ISC, in November 2016 the Ofqual Board agreed that Ofqual's policy position on ISC in GCSE, AS and A level should be:

[1] See Ofqual working papers on ISC available at https://www.gov.uk/government/publications/inter-subject-comparability-2015-to-2016

> a) where there is an exceptional case that Ofqual considers to be compelling, to take action to adjust grade standards in that subject;
>
> b) having first considered with key stakeholders the implications of the evidence for, in particular, the curriculum and take-up, but
>
> c) to take no coordinated action to align standards across the full range of subjects through grading; and
>
> d) to improve the quality of assessments where it may be creating detrimental impacts in particular subjects (such as A level French) (Ofqual, 2016a).

The Board agreed that the first qualifications that should be examined to see whether an exceptional and compelling case existed to adjust grade standards in are A levels in physics, chemistry and biology, and in French, German and Spanish (Ofqual, 2016a). The decision to focus on these subjects in the first instance relates to long-standing concerns from stakeholder communities about declining entries in modern foreign languages at A level, and low take-up of these subjects at university level, as well as the concerns as to whether enough students in England, particularly girls and those of lower socio-economic status, are choosing to study A level physics in particular, but also other sciences (Ofqual, 2015a, 2016a) [2].

There is evidence that these subjects are perceived to have more severe grade standards than some other subjects (Cuff, 2017), which may be one of the reasons for the above-mentioned take-up issues (see Dearing and King, 2007; Ofqual, 2015a). Furthermore, there are indicative patterns emerging from several statistical analyses (Coe, 2010; Newton, He and Black, 2017; He, Stockford and Meadows, 2018) suggesting that some of these subjects might be consistently more severely graded than others.

Even though some stakeholders suggest that the statistical evidence of severe grade standards is clear and should be acted upon (see for example annex F in Ofqual, 2016a), there are a number of issues and questions around the conceptualisation and interpretation of statistical methods and indices for measuring or identifying differences in subject difficulty. For example, there are a number of commentators arguing that the statistical differences between subjects may not actually reflect genuine differences in subject standards (for example Newton, 1997; Bramley, 2016; Newton, He and Black, 2017; Benton and Bramley, 2017) and may be statistical artefacts related to, for instance, the properties of the data being analysed (non-randomly missing data, effect of student choice, unequal correlations between outcomes in different subjects, etc.). Furthermore, as pointed out in Ofqual (2015b; see also Pollitt, 1996), commonly used statistical methods such as Rasch do not/cannot model the impact of certain factors that, probably according to most stakeholders, ought to be ‘rewarded’ by exam grades, for instance student motivation, or subject-specific attainment. All currently available statistical models assume an oversimplified definition of inter-subject comparability and/or subject difficulty (see the summary in Ofqual, 2016a; and also Bramley, 2014, 2016;

[2] See also https://www.gov.uk/government/statistics/summer-2017-exam-entries-gcses-level-1-2-certificates-as-and-a-levels-in-england

Koroboko et al., 2008), which makes it difficult to interpret their results or be confident in what they are telling us.

In addition, there are concerns that even if adjustments were to be made to grade standards, this is unlikely to be the only action that should be taken to address the aforementioned problems with take-up; the adjustments, based on the average of groups of candidates, might have limited impact at the level of the individual in HE selection process or when it comes to school accountability measures (Benton, 2016), as well as on subject take-up; and they might have other undesirable consequences (Ofqual, 2016a).

Therefore, Ofqual's intention was to gather a range of evidence for a subject or group of subjects which would include statistical evidence of subject difficulty and other pertinent research evidence, as well as contextual data (such as teacher numbers and quality, evidence of subject take-up, etc.), and grade adjustment impact data. This is presented elsewhere[3].

Study aims and questions

The purposes of A level qualifications stated in Ofqual's GCE Qualification Level Conditions and Requirements (2017) are as follows:

- 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;
- set out a robust and internationally comparable post-16 academic course of study to develop that knowledge, skills and understanding;
- permit UK universities to accurately identify the level of attainment of students;
- provide a basis for school and college accountability measures at age 18; and
- provide a benchmark of academic ability for employers.

An adjustment to grade standards, which could lead to different numbers of students achieving different grades than is currently the case in some subjects, might impact on some or all of the stated A level purposes in ways that are difficult to establish upfront. Wider ramifications of any grade standard adjustments are important to consider, not least because they could cause unacceptable changes in interpretation of performance standards in individual subjects.

In this study, we focused on the possible impact that grade boundary adjustment might have on progression and selection to higher education in French, German, Spanish, physics, chemistry and biology. By seeking the views of HE representatives, representatives of learned bodies and Ofqual subject specialists, we attempted to answer the question of how acceptable grade standard adjustment would be to them. Specifically:

[3]

- whether the standard of performance at each grade relevant for university admissions might become unacceptably low as a result of grade standard adjustment, and
- whether the grades might become less useful as the basis of university admissions criteria[4].

Given that A level grades in the subjects of interest in this study are set based on student performance and marks earned in A level examinations, we asked the participants to review samples of student A level examination work at and below different A level grade boundaries. Ultimately, we wanted to see whether a potential grade boundary adjustment, and what magnitude of adjustment in each subject:

- would be discernible to our participants in student work that they reviewed,
- would be acceptable to them, and
- might impact on the utility of those grades for admissions purposes.

Methodology

Participants

The majority of the participants in the study were representatives of higher education institutions (HEIs) from England, Wales, Scotland and Northern Ireland, with experience of teaching first year undergraduates in courses for which physics, biology, chemistry, French, German or Spanish are facilitating subjects. We also recruited learned body representatives and Ofqual subject specialists for each subject. In order to carry out the exercise, the participants were split into panels, one for each subject. The panels took place over 2 days.

The intention was to recruit 10 to 12 participants for each panel. Of these, 8 to 10 should be HEI representatives and the rest representatives of learned bodies and Ofqual subject experts. The learned body representatives were included to widen the pool of key stakeholders with expertise and interest in the subjects. Ofqual subject specialists were included to provide expertise in qualification and assessment design in each subject, as well as the knowledge of what current standards entail.

In order to ensure that in each panel we had a reasonable spread of HEIs that admit students based on a range of A level grades, we used the Guardian University Guide

[4] Many HE institutions use A level grades in facilitating subjects for initial ‘screening’ of students. However, A level grades are not the sole factor informing university admissions decisions. Other factors such as personal statements, interviews, special consideration, university inclusiveness policies, grades in other subjects, etc. can all affect admissions decisions and override universities’ published grade-related admissions criteria. Nevertheless, as long as the primary purposes of A levels are related to progression and selection to higher education (cf. also Ipsos Mori, 2012), arguably, it is necessary to have confidence in the standards of performance which different A level grades represent, even if these are not the sole indicators used for university admissions.

2017[5] to obtain the average UCAS tariff points of students on entry for each subject[6] and institution. We created lists of institutions for each subject and ranked the institutions within each list by the average UCAS tariff points of students on entry. We then divided each list roughly into 3 tiers by UCAS tariff points for the purposes of sampling (cf. Ipsos Mori, 2012). The order of institutions was randomised within each tariff point tier before sampling.

We sent invitation emails to potential participants from each tier in 2 waves. Initially we sent the invitations to the top third of each tier, followed by further invitations some time later to most of the remaining institutions. Because we were seeking participants who could represent their institutions rather than just their own personal views, the invitation emails were sent to Heads of Departments/Faculties or Program Leads, where contact details were available on university websites. They were asked to nominate a suitable representative of their institution. In a small number of cases, where we could not identify a suitable academic contact, the invitations were sent to faculty or department administration offices. In these cases, we asked for the invitation to be forwarded to the relevant person that could nominate a representative. Table 1 shows the numbers of participants for each subject. The list of participating institutions is included in Appendix A.

Table 1*. Number of participants.*

	HEI representatives	Learned body representatives	Subject specialists	Total
French	12	1	3	16
German	9	1	1	11
Spanish	10	1	1	12
Physics	7	1	2	10
Chemistry	12	1	2	15
Biology	9	1	1	11

Even though our intention was to have equal number of representatives in each panel, the final numbers depended on how many people accepted the invitation within the time frame available for recruitment. In addition, a couple of people that initially agreed to participate either needed to cancel a few days before the panels were due to take place, or were unable to attend on the day.

Scripts

For each of the 6 subjects, we requested a sample of candidate A level examination work from the summer 2017 examination series from examination boards. A level grades in these subjects are set based on the overall performance of each candidate in their examinations, i.e. based on the total mark across all of the units in a specification. For the sciences, for each candidate in the sample, we obtained the papers from all 3 units that contributed to their overall grade. For the unreformed modern foreign languages qualifications, we were only able to obtain the work for the

[5] https://www.theguardian.com/education/ng-interactive/2016/may/23/university-league-tables-2017

[6] Note that there is a single tariff per institution for modern foreign languages, rather than a separate tariff for each language.

two A2 units[7]. For each of the foreign language specifications, these included a written paper and a recording of the speaking assessment. We henceforth refer to the collection of all of a candidate's available examination papers/recordings as a 'script'. We asked for the scripts at grades A* to C only, as grades below C do not tend to be used as requirements for admission.

In selecting the specifications for each subject, we endeavoured to secure those with highest entry numbers. However, the entry numbers for some of the modern foreign language specifications were quite small. In order to minimise administrative burden on examination boards, each board was asked to provide samples for one or two specifications only.

Table *2* shows which boards supplied samples for which subject and specification, and entry numbers for each specification.

Table 2. *Specifications sampled.*

Exam board	Specification title	Specification code	Entry size
AQA	Biology Advanced	7402	26470
	German Advanced	2661	1957
OCR	Physics A	H556	8947
	Chemistry A	H432	18915
Pearson	Spanish	9SP01	1554
WJEC	French	319101	1623

Determining indicative qualification grade boundary adjustments

In the A level examination system, marking and grading are two separate activities. Students' scripts are in most cases marked by independent examiners according to mark schemes. Most marking is done online and, where appropriate, scripts are split into items and distributed amongst examiners, so that a single student's script is marked by multiple examiners. When the majority of marks are in the system, the grade boundaries are determined through the 'awarding' or 'grading' process. For any particular qualification, examiners and students do not know what the grade boundaries will be in advance of an examination session.The relationship between marks and grades is illustrated in

Figure *1*.

[7] French, German and Spanish specifications still comprised AS and A2 in summer 2017. This means that students sat the first part of their examination after the first year of sixth form (the AS level), and the final part of the examination after the second year (the A2 level). The AS scripts are destroyed by examination boards, and were thus no longer available at the time when this study was taking place.

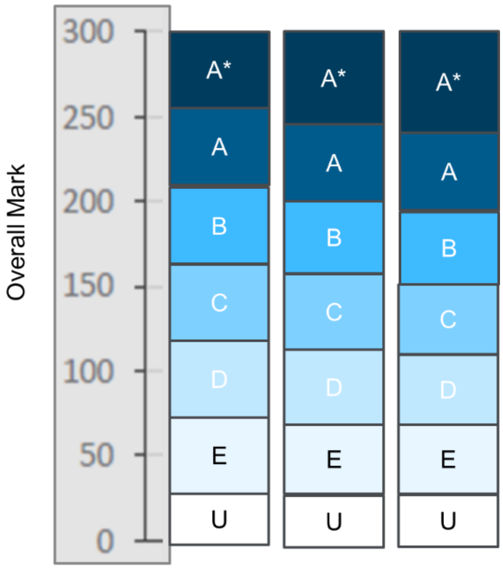


Figure 1. *The relationship between marks and grades in A level examinations for 3 different (notional) qualifications.*

Given the research questions in our study, for each subject we needed a selection of student scripts with mark points on and as far below current grade boundaries as an indicative grade boundary adjustment (see below for more details) would suggest. In this way, when reviewing the scripts, the participants would see student work that represents the current grade standard, that which represents a potential new grade standard, as well as the work at mark points or increments in between.

To enable this, we first established indicative grade boundary adjustments for each of the subjects of interest in this study. This was done using the method described in He, Stockford and Meadows (2018) and He and Meadows (2018). This method uses the partial credit Rasch model (PCM; Masters, 1982) to first establish relative difficulty of the set of subjects included in the analysis. The PCM states that, for a polytomous item with a maximum available score of m (the number of score categories minus 1), the probability $P(\theta, x)$ of an examinee with ability (latent trait) θ scoring *x* on the item can be expressed as:

$$P(\theta, x) = \begin{cases} \dfrac{exp\sum_{k=1}^{x}(\theta-\delta_k)}{1+\sum_{l=1}^{m} exp[\sum_{k=1}^{l}(\theta-\delta_k)]} & \text{for } x = 1,2,\dots m \\ \dfrac{1}{1+\sum_{l=1}^{m} exp[\sum_{k=1}^{l}(\theta-\delta_k)]} & \text{for } x = 0 \end{cases}$$

Where δ_k is the location of the *k^th^* category score on the latent trait continuum (or category threshold or difficulty, see Andrich, 2015).

In this method, each examination in a subject is viewed as a polytomous item in a test, and the grades assigned to individual examinees for an exam are treated as scores on an item which represent ordered response categories. All examinations contained in the analysis form a test. It is assumed that the examinations to be analysed together define a shared construct, usually conceptualised as ‘general academic ability’ which is related to the constructs being measured by the individual examinations. The difference in difficulty overall and at individual grades is assumed to reflect difference in grade standards between the examinations.

In this study, the A level examinations included in the analysis were from the 2017 exam series. Subjects with entries less than 1,000 were removed from the analysis. To facilitate the analysis, the letter grades were converted into numerical values representing ordered category scores: U→0, E→1, D→2, C→3, B→4, A→5, and A*→6. The data were analysed using the software Winsteps (Linacre, 2015), which implements the PCM. Inspection of model fit statistics[8] and other indicators suggested that the data fit the PCM model reasonably well overall. General studies was removed from the final analysis due to its model misfit.

In this approach, the difficulty of category or grade k, d_k, of an item (exam) is defined as the ability θ at which the expected score $E(\theta)$ on the item is $k - 0.5$:

$$d_k = \theta|_{E(\theta)=k-0.5} \tag{1}$$

This definition is similar to the definition of the item difficulty for dichotomous items. The average of the category parameters can be used to characterise the overall difficulty D of the item:

$$D = \frac{1}{m}\sum_{k=1}^{m} \delta_k \tag{2}$$

At a specific grade *k* for a specific subject, the difference between the grade difficulty d_k of this subject and that of a reference subject, d_{ref}, at this grade is defined as the relative difficulty $d_{k,R}$ of this grade:

$$d_{k,R} = d_k - d_{ref} \tag{3}$$

Although the difference in difficulty between two adjacent grades is not a constant in a subject, an average grade gap Δ (logits) can be defined across all grades and subjects as:

$$\Delta = \frac{1}{N_G N_S}\sum_{i=1}^{N_S}(d_{i,A} - d_{i,E}) \tag{4}$$

Where N_G is number of grade gaps (4, between A and E.), $d_{i,E}$ is the difficulty of grade E and $d_{i,A}$ is the difficulty of grade A. Dividing the relative grade difficulty $d_{k,R}$ by the average grade gap in logits gives the relative grade difficulty $d_{k,RG}$ in the unit of grade:

$$d_{k,RG} = \frac{d_{k,R}}{\Delta} \tag{5}$$

The relative grade difficulty $d_{k,RG}$ represents the proportion of average grade width in scaled score unit (e.g. UMS marks) to be adjusted if standards were to be aligned.

Figure 2 shows the subjects which were included in the analysis in order of difficulty (ordered by ascending mean subject difficulty). It can be seen that all of the subjects included in this study are more difficult than average according to this definition of difficulty.

[8] Residual-based model-data fit indices such as unweighted mean square (outfit mean square) and weighted mean square (infit mean square) are used in Rasch analysis to show the degree to which observed scores match the expected scores that are generated by the model. The infit mean square statistic is sensitive to an accumulation of unexpected ratings and the outfit mean square is sensitive to individual unexpected ratings. These statistics have an expected value of 1 when the data fit the Rasch model and can range from 0 to infinity (Linacre, 2002; Linacre, 2015; Myford and Wolfe, 2003). Linacre (2002) suggested that when model fit statistics are above 2.0, the measurement system would be distorted.

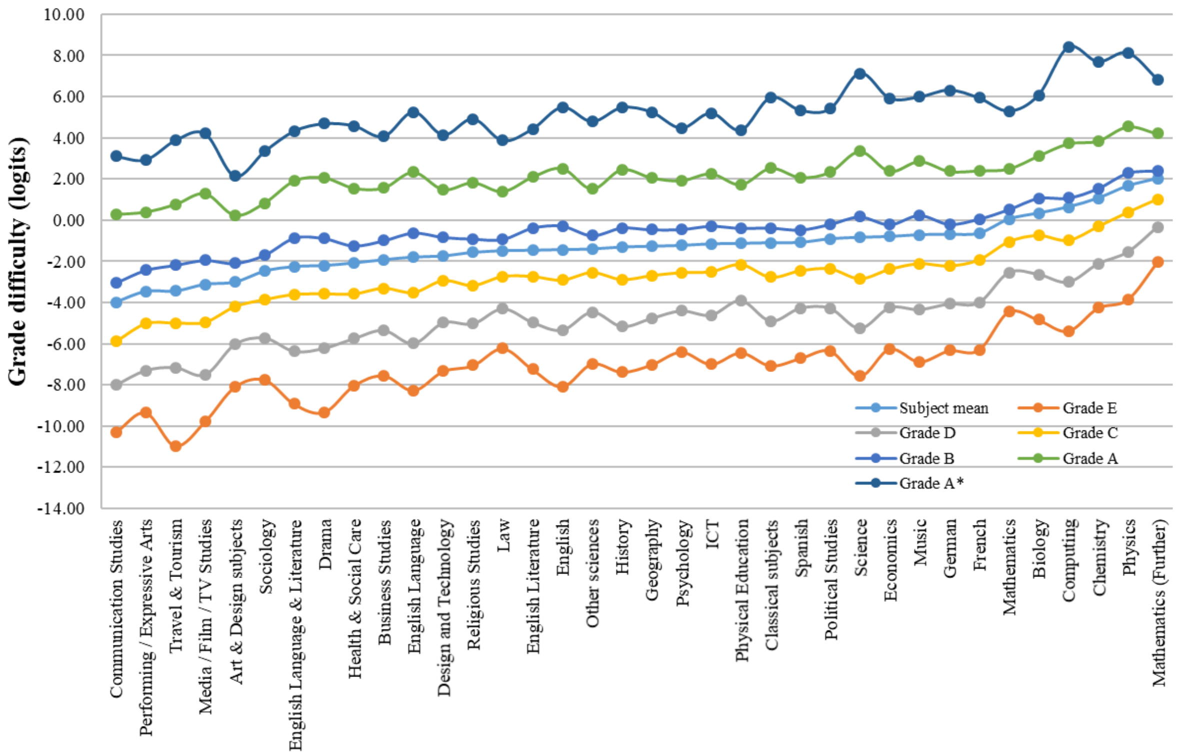


Figure 2. *Subject and grade difficulty according to Rasch analysis.*

For A levels, grade boundaries could be viewed as the operationalisation of performance standards, and aligning statistical standards between subjects would necessarily involve changing the boundary marks for certain subjects and therefore performance standards. Assuming that the original subject level grade boundary score and grade interval (grade width) at grade k are b_k and w respectively for a subject, the new grade boundary b'_k after the alignment of statistical standards with the reference subject based on results from the Rasch analysis would be:

$$b'_k = b_k - wd_{k,RG} \qquad (6)$$

This adjustment was implemented for grade boundaries A* to C in each subject in the current study. It was decided that the reference subject for the 3sciences would be mathematics, and for the 3 languages, geography. In other words, this would indicate the boundary for which the grade standards for the 3 sciences would be brought into line (i.e. become equally difficult) with mathematics, and with geography for the 3 languages.

Mathematics was chosen as the reference subject for sciences as it is also a cognate facilitating subject required for university admissions; we were of the view that if standards for the 3 sciences were to be adjusted, it would be difficult to justify lowering them further than where mathematics currently is. Geography was chosen as a reference subject for the 3 languages because it is also a facilitating subject and of medium difficulty according to the current Rasch analysis. An alternative could have been history, but we decided in favour of geography because its A level exam is more similar to the written part of the languages exams in that it contains a mixture of different item formats (short-response items as well as extended-response ones), as opposed to history, which only contains extended-response items. Ultimately, the choice of reference subjects was to some extent arbitrary, particularly for languages. However, it was necessary for the study and was judged to be reasonable in light of the abovementioned considerations and also some feedback which indicates stakeholders within mathematics and geography are satisfied with A level standards in these subjects.

Table *3* shows the original grade boundary marks and the indicative adjusted grade boundary marks[9] for each subject, as well as what proportion of the grade widths these adjustments represent. It can be seen that the size of adjustment varies by subject and grade boundary, and indeed, that there is virtually no difference between the original and indicative grade boundary in the case of Spanish grades A and B.

Table 3. *Original and adjusted indicative grade boundaries.*

	A*	A	B	C
Physics				
Original	207	178	151	124
Indicative	180	156	130	105
Proportion of grade width	0.93	0.80	0.79	0.70
Chemistry				
Original	232	198	164	130
Indicative	205	180	149	118
Proportion of grade width	0.79	0.53	0.46	0.36
Biology				
Original	174	147	124	101
Indicative	167	141	118	98
Proportion of grade width	0.26	0.25	0.25	0.15
French				
Original	180	160	140	120
Indicative	175	157	136	113
Proportion of grade width	0.24	0.13	0.22	0.34
German				
Original	180	160	140	120
Indicative	172	158	138	116
Proportion of grade width	0.39	0.12	0.10	0.21
Spanish				
Original	180	160	140	120
Indicative	179	160	140	118
Proportion of grade width	0.05	0.01	-0.02	0.09

[9] The adjustment for the three languages was implemented on A2 units, as we were only going to consider the A2 scripts in this study.

Selecting script samples

In order to allow the participants to see a range of student performance starting at each original grade boundary and including each of the indicative grade boundaries, we needed to select scripts on a number of total mark points within that range for each grade boundary. We initially selected scripts on 13 mark points (2 scripts on each)[10], including the original grade boundary mark, within this range.[11] Depending on how wide each range was, these 13 mark points included either every mark point in the range, a selection of mark points in that range, or also the mark points beyond the indicative grade boundary where it was very close or indistinguishable from the current grade boundary.

As described earlier, each script consists of multiple papers produced by one candidate, one for each unit in the examination, with separate marks for each. Given the compensatory nature of A level awarding, it is theoretically possible for a candidate to obtain a maximum mark on one paper and zero on another, and still obtain a reasonably high grade even though his/her performance was very inconsistent. However, it would be difficult to form holistic judgements of script quality based on high levels of inconsistency in performance across papers within a script (see, e.g. Scharaschkin and Baird, 2000). Therefore we first needed to determine which scripts could be described as more ‘typical’ (usually also more balanced) in terms of their mark profile on different units in order to then select a small sample of these for the study. By typical, we mean that the pattern of performance across the papers is reasonably consistent and reflects the patterns apparent in the majority of candidate performances for that specification.

For each science subject, we first selected the students within the range between the current and indicative grade boundary total mark inclusive (or beyond, up to twelve mark points from the original boundary mark where appropriate). We then used the fit statistics from a partial credit Rasch model analysis (see previous section for model details) to select typical scripts within the available scripts in each pre-selected mark range.

In the Rasch analysis, we treated the units as polytomous items of a test, and the total marks the students obtained in each unit as item scores[12]. This allowed us to estimate student ability measures and how well these fitted the Rasch model. In this conceptualisation, the scripts that fit the Rasch model were considered to be typical. We removed the scripts that had fit statistics outside of the range of $0.7 <$ Infit MnSq / Outfit MnSq < 1.3[13] (cf. Raikes, Scorey and Shiell, 2008) as atypical, and sampled the required scripts from the rest.

[10] We requested 2 scripts on each mark point to maximise the chance of at least one of those being sufficiently legible and thus usable by the participants in the study.

[11] The number of mark points requested was somewhat arbitrary, and was thought to be the maximum that he participants would be likely to go through during the time available for each grade boundary.

[12] It was not possible to run this analysis on actual item level data as these were not available.

[13] Rule-of-thumb upper and lower limits for acceptable mean square fit values have been established for identifying misfit, commonly outside the range of $0.5 <$ Infit MnSq / Outfit MnSq <1.5 (Linacre,

For languages, the only available scripts were those for A2 units. Because our panellists would not have the opportunity to see the AS scripts, we needed to try and keep the marks on those units relatively constant, while selecting the scripts from A2 units within the grade boundary ranges of interest. Given that this restricted the available scripts to a great extent, as well as because of small entry sizes for these subjects, it was not feasible to then further remove candidates based on Rasch fit statistics as this did not provide sufficient number of candidates on each total mark point of interest. However, a small number of most inconsistent performances was identified by eye and removed (for example where candidates scored close to maximum marks on one paper, and very few marks on the other).

Therefore, we first selected candidates based on restricting the range of their AS total UMS marks, although this range still needed to be quite wide in some cases (usually 20-30 marks) to allow us to ultimately be able to select candidates on each A2 total mark point of interest. In most cases, we selected the range of AS total marks around the relevant AS boundaries[14]. From these candidates, we then selected those that had the A2 total mark on the relevant grade boundary and either down to the indicative grade boundary total, or twelve marks below the boundary total, whichever was greater.

Once the relevant pool of candidates was selected for each subject and grade boundary, the scripts were sampled randomly within each grade boundary range, stratifying by mark point, so that there were 2 scripts either per each mark point or for every second or third mark point where the range between the current and indicative boundary was large.

Script presentation

Each pair of scripts on each mark point requested was reviewed for legibility and the easier one to read was chosen for the study. Student names and marks were removed from the scripts (marker annotations were not removed). The scripts were organised into sets by grade boundary, i.e. A*A, AB, BC and CD script set. Ultimately, given the amount of time it would take to review each script, and the constraints of panel duration, out of the 13 requested we presented 7 to 9 scripts per grade boundary, and the panellists ultimately reviewed 4 to 6 scripts for languages and 6 to 8 scripts for sciences per grade boundary.

Script order was randomised in each pack within 2 constraints:

- the top 2 scripts in each pack were the original boundary scripts – the ‘model’ scripts, and
- the indicative grade boundary script was always the fourth in language packs and fifth in the science packs,[15] to ensure that they get reviewed by each panel in the time available.

2002), though a more stringent range of 0.7 < Infit MnSq / Outfit MnSq < 1.3 and also wider ranges are sometimes used depending on purpose (cf. Wright and Linacre, 1994; Linacre, 2002).

[14] In the case of C boundary in Spanish, we needed to select candidates around the modal AS total rather than the boundary one as there were very few candidates around the grade boundary itself.

[15] The indicative boundary script was presented earlier in MFL packs, and there were fewer scripts in MFL packs, because it was anticipated that MFL panels would take longer to review the scripts because of having to listen to the recordings of the speaking assessment as well as review the written work.

Tables showing the order in which the scripts were presented are in Appendix B.

Preparation activity

Ahead of the main meeting, the participants needed to become acquainted with the demands and style of the A level assessments which were to be considered in the study. They were, therefore, asked to spend half a day doing a preparation activity.

For each subject, they were sent one script on each of the original A*A and CD boundaries (the same scripts were sent to all the participants), as well as the materials that were provided to the students during the examinations. The participants were informed about which script was on which grade boundary and that they could expect to see a difference in performance standard between these 2scripts.

The participants were asked to initially familiarise themselves with the question papers and note down any observations about the papers, the overall coverage of the subject/examination and the difficulty of the papers. They were subsequently asked to focus on students' work and try and form impressions of:

- the overall standard of work from A level students during a timed examination,
- how students respond to these assessment tasks,
- the differences in standard between the performances of the two students,
- what abilities (such as skills and knowledge) each student demonstrates,
- how these students' abilities appear to compare to their own students at the start of the course (if they had experience of teaching first year undergraduates), and
- the extent to which these students represent those which their institution might admit onto their courses (if they were a representative of an HEI).

Panel activity

Following the preparation activity the participants came together in person for the panel activity, which took place over 2 days. As mentioned previously, in this study we wanted to see whether a potential grade boundary adjustment, and what magnitude of adjustment in each subject:

- would be discernible to our participants in student work that they reviewed,
- would be acceptable to them, and
- might impact on the utility of those grades for admissions purposes.

This was put to the panels as 2 questions:

Question 1: Would the students who wrote the following scripts be as deserving of admission to your institution as the students who wrote the 'model' scripts?

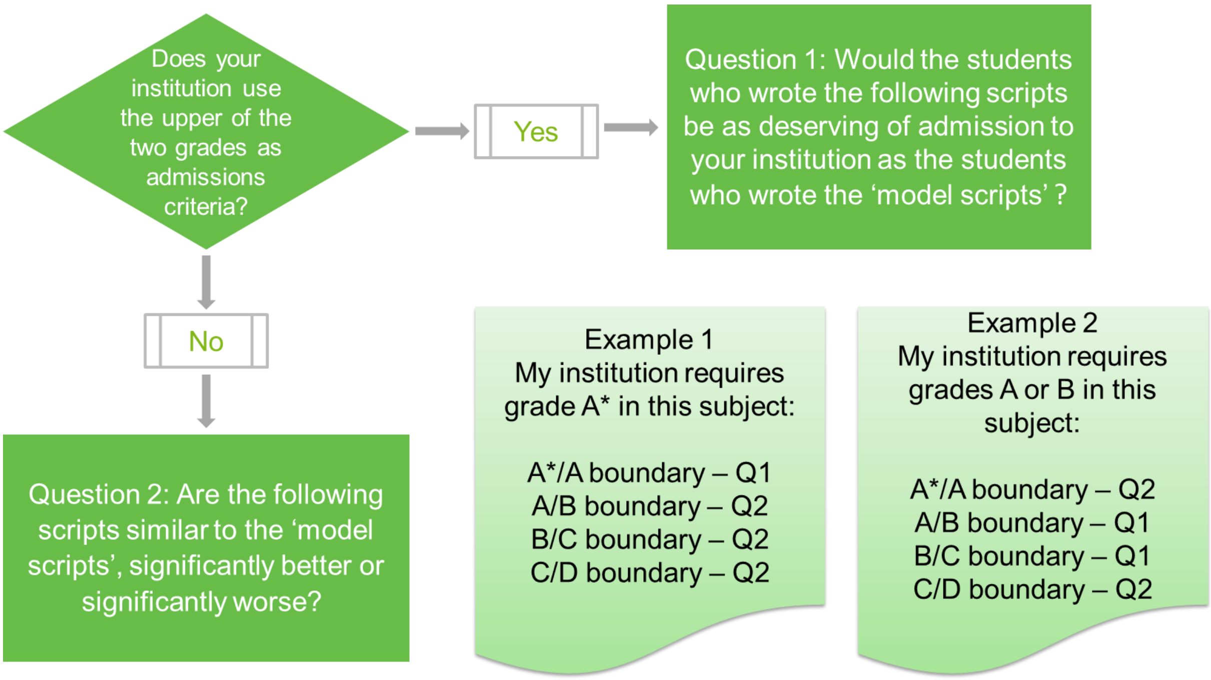
Question 2: Are the following scripts similar to the ‘model’ scripts, significantly better or significantly worse?

For question 1, the panellists were asked to choose one of the 4 options: ‘yes’, ‘maybe yes’, ‘maybe no’, and ‘no’. For question 2, they were asked to choose one of three options: ‘better’, ‘similar’, or ‘worse’.[16]

- In the panels there were representatives from a range of HEIs, with a range of admissions criteria, as well as subject specialists and representatives of learned bodies, for whom university admissions are not relevant. For each grade boundary pack, the participants were, therefore, asked to choose which question was more appropriate for them to answer.

Figure *3* summarises the guidance given to the panellists on how to choose which question to answer.

Question 1 was only relevant for a subset of the panellists, those who represented HEIs, depending on which boundary script set they were reviewing, and on their institution’s admissions criteria. Even though we could have asked both groups of panellists to answer question 2, we believed that it was important, where possible and relevant, to tap into their views of the utility of grade standard adjustment specifically with respect to university admissions, given the primary purpose of A levels.



Does your institution use the upper of the two grades as admissions criteria?

Yes

Question 1: Would the students who wrote the following scripts be as deserving of admission to your institution as the students who wrote the ‘model scripts’ ?

No

Question 2: Are the following scripts similar to the ‘model scripts’, significantly better or significantly worse?

Example 1
My institution requires grade A* in this subject:

A*/A boundary – Q1
A/B boundary – Q2
B/C boundary – Q2
C/D boundary – Q2

Example 2
My institution requires grades A or B in this subject:

A*/A boundary – Q2
A/B boundary – Q1
B/C boundary – Q1
C/D boundary – Q2

[16] The panellists were given the less determinate response options of ‘maybe yes’ and ‘maybe no’ for question 1, as well as both ‘better’ and ‘worse’ in addition to ‘similar’ as opposed to just similar/different as we felt this would make their task more straightforward, meaningful and capture their opinions more closely. However, we anticipated that, given the small sample size, the data analysis would need to collapse some of these categories in order to allow us to see more general patterns in the responses.

Figure 3. *Summary of guidance for choosing the appropriate question.*

Each panellist was given a set of student scripts per grade boundary, a set of instructions and 2 forms on which to note their responses, one for each question. These are presented in Appendix C.

The panels first reviewed the scripts on A*/A boundary, then A/B boundary, etc. They were asked to work independently. There were opportunities for discussion after each grade boundary set had been reviewed and at the end of each panel.

At the top of each boundary pack were 2 scripts on the original grade boundary – the model scripts (one of these on A*/A and C/D boundaries had already been reviewed as part of the preparation activity, but was also presented during the panel activity). The panellists were asked to first review the model scripts, form a holistic picture of the 2 performances where possible, and treat that as a model of performance that would just earn the relevant original grade. Then, going through the pack in the order in which the scripts were presented, they were asked to review each following script, compare it to the model and, for each script, answer the question that they had chosen for that boundary script set. We asked the panellists to compare the scripts on mark points below current grade boundary to those on the current grade boundary, i.e. 'model' scripts as this was likely to make their task cognitively easier. This is based on research suggesting that it is easier to make comparative rather than absolute judgements (e.g. Thurstone, 1927; Laming, 2004).

The participants were asked to try and form a holistic judgement of the student who produced each script, having reviewed all of the work available, and to consider knowledge, skills and understanding apparent in the scripts. Inevitably, partly due to the compensatory nature of A level mark schemes, the evidence available within and between scripts was inconsistent, which made the task challenging. It was acknowledged that A level grades are not the sole factor informing university admissions decisions. The panellists were, therefore, asked to put the other factors to one side and assume that those are equal across students whose scripts they reviewed.

Each panel was chaired by an Ofqual representative, and observed by another. The chairs led the script review activity as well as the group discussions, which happened after each grade boundary set had been reviewed as well as at the end of each panel. The chairs and observers made notes of the discussions and of any pertinent points and observations made during other panel activities.

Methodological limitations

An important limitation of this study is a relatively small number of participants, and related to it, the representativeness of the participant sample. This is especially the case when the responses are broken down by question (there were very few participants in a position to answer question 1 for A* boundary, for instance). Also, there was a relatively small number of scripts that could be reviewed in the short time available. This does limit the extent to which the results can be generalised and the level of confidence that can be placed on them.

We intended to keep the range of AS component marks constant within each boundary set, however this proved to be impossible as, given small entry numbers for languages, this would not have allowed us to select scripts on each relevant total A2 mark point. We therefore needed to widen the range of AS component marks more than desired. However, given that the boundary adjustment investigated here relates only to the A2 component and the participant only saw the A2 scripts, we believe that this should not have affected the results of the exercise significantly.

Despite being asked to form a holistic picture of the pairs of model scripts in each script set, the participants sometimes thought that, even though each pair of model scripts had the same total mark, the actual candidate performance between them differed a great deal. Some of them, therefore, found it impossible to form a holistic picture of model performance across the 2 scripts. They were asked to note where that was the case. We investigated any differing patterns in responses where enough participants used different model scripts for comparison. Often the amount of inconsistency within the responses which used the same model was similar to the amount of inconsistency across different model options. In other cases, though participant comments suggested that they chose the ‘weaker’ of the 2 model scripts, their responses were more likely to be endorsing the negative options than those of the participants who used the other, arguably ‘stronger’, model. These situations suggest that the effect of participant severity may have been stronger than that of the difference in performance levels between the 2 scripts. Ultimately, given the inevitable difficulties in making consistent judgements based on inconsistent evidence presented in student performance in general, alongside inevitably different standards that different participants no doubt had, it was decided to treat all the responses equally, irrespective of the model used.

Finally, there is the issue of the choice of reference subject. As already mentioned, even though we did have a rationale for our choice, the choice remains to some extent arbitrary and open to challenge. It demonstrates one of the problems that would likely arise if a method adopted for this study was to be used for grade adjustment operationally, i.e. finding a consensus regarding the reference subject for adjustment.

Results

In this section we present the results of the study. We first present a summary of the number of responses by question and the approach to data analysis. We then present, for the languages and then sciences, the outcomes of the script review – overall and by question, followed by a summary of the main points that emerged from panel discussions, and a results summary.

Number of responses, response coding and acceptability criteria

Table *4* shows the number of responses per question and mark point (excluding the model scripts) in each grade boundary set for each of the 3 languages.

Table 5 shows the same information for the three sciences. The shaded cells are the indicative grade boundaries. The tables are sorted in descending order of mark point within each grade boundary set.

The aim was for all the participants to review a minimum of the first 4 scripts in each set for languages and 5 for sciences as per the original randomised order. This was achieved for all subjects and boundaries, except for some missing data on a few mark points where either a participant was unable to complete a set due to illness, was unable to participate on the second day of the panel, or where we have unclassifiable decisions because a participant did not clearly choose one option on the form). In languages, the participants ultimately reviewed 4 to 7 scripts per grade boundary set and 6 to 9 in the sciences.

Language panels generally took longer to review the scripts because one of the units was a speaking unit, where panels needed to listen to audio recordings of the assessments, which in some cases took longer than to review written papers. In sciences, there are few responses on certain mark points. The scripts on these mark points were presented later in the sets, therefore only some participants managed to reach and review those before the following set was produced.

Table 4. *Number of responses per question and mark point reviewed – languages.*

French				German				Spanish			
Grade	Mark point	Q1 N	Q2 N	Grade	Mark point	Q1 N	Q2 N	Grade	Mark point	Q1 N	Q2 N
A*A	177	1	15	A*A	178	0	11	A*A	179	1	11
	176	1	15		177	0	11		178	1	11
	175	1	15		176	0	11		177	1	11
	174	1	15		175	0	11		175	1	11
AB	159	5	11		174	0	11		174	1	11
	158	5	11		173	0	11	AB	159	7	5
	157	5	11		172	0	11		158	5	5
	156	5	11	AB	159	2	9		157	4	5
	155	5	11		158	1	9		156	5	5
	154	5	11		157	2	9		155	5	5
BC	138	7	8		156	2	9	BC	139	8	3
	137	8	8		155	2	9		138	8	3
	136	8	8		154	2	9		137	8	3
	135	8	8	BC	139	7	4		136	8	3
	134	7	7		138	7	4		135	8	3
CD	118	2	14		137	7	4	CD	119	4	7
	117	2	13		136	7	4		118	4	7
	113	2	14		135	7	4		117	4	7
	112	2	14		134	7	4		116	4	7
				CD	119	1	10		115	4	7
					118	1	10				
					117	1	10				
					116	1	10				
					115	1	10				
					114	1	10				
					113	0	10				

Table 5. *Number of responses per question and mark point reviewed – sciences.*

Physics				Chemistry				Biology			
Grade	Mark point	Q1 N	Q2 N	Grade	Mark point	Q1 N	Q2 N	Grade	Mark point	Q1 N	Q2 N
A*A	202	5	5	A*A	230	2	1	A*A	173	0	4
	197	5	5		224	4	11		172	0	9
	192	3	2		218	4	11		171	0	10
	190	5	5		212	4	11		170	0	9
	185	5	5		209	2	2		169	0	10
	182	3	1		207	3	11		168	0	4
	180	5	5		205	4	11		167	0	11
AB	174	3	7	AB	196	5	9	AB	145	4	6
	170	3	7		192	5	10		144	4	6
	166	3	7		188	5	10		143	5	6
	164	3	7		186	5	10		142	4	5
	160	3	7		184	5	10		141	5	6
	158	3	7		182	5	9		140	5	6
	156	3	7		180	5	10	BC	122	8	2
BC	148	2	8	BC	162	5	10		121	9	2
	144	2	8		158	5	10		120	9	2
	142	2	8		154	4	5		119	8	2
	140	2	8		152	4	10		118	9	2
	136	2	8		151	5	10		117	8	2
	134	2	8		150	5	10	CD	100	2	9
	132	2	8		149	5	10		99	2	9
	130	2	8	CD	128	2	12		98	2	9
CD	120	3	6		124	2	12		97	2	9
	116	3	7		122	2	8		96	2	9
	112	2	3		121	2	12		95	2	9
	110	3	7		120	0	4		94	2	9
	108	3	6		119	0	3				
	106	2	4		118	2	12				
	105	3	7								

The response options on the forms given to the participants were yes (Y), maybe yes (Y?), maybe no (N?) and no (N) for question 1, and better (B), similar (S) and worse (W) for question 2. Some respondents did not give a straightforward answer by selecting only one of the response options for each script. For instance, some selected both better and similar (B/S) and some better and worse (B/W) for a particular script. Their comments show that this was typically done in the cases where they could not form a holistic picture of a script, but characterised one part as better and another as similar or worse.

Given the small sample size, especially when broken down by question, we collapsed some of the categories when analysing the data in order to be able to see clearer patterns where there were any.

Table *6* shows how we treated the predetermined response options as well as different combinations of these when analysing the data.

Table 6. *Response coding.*

Question	Responses	Coding by question	Coding overall
Q1	Y	Y	Y
	Y?	Y	Y
	Y/Y?	Y	Y
Q2	S	S	Y
	B	S	Y
	B/S	S	Y
	B/W	S	Y
Q1	N	N	N
	N?	N	N
	N?/N	N	N
Q2	W	W	N
	S/W	W	N
Q1	Y?/N?	Unclassified	Unclassified

The overall question for this study was about acceptability of grade standard adjustments, whether this was indicated through acceptability of students to institutions or similarity of performance standard at a current grade boundary to standard at mark points below the boundary. Our criteria for considering that a set of responses indicated acceptability of grade standard adjustment, and of the magnitude of acceptable adjustment, were as follows:

- For an individual script/mark point – over 60% of participants endorsed a positive response (i.e., as acceptable to university courses as the model script; or, performance standard in the script better than or similar to the model script)
- As an indication of magnitude of adjustment – bottom of a range of consecutive scripts/mark points with over 60% positive endorsement rate.

We have chosen 60% majority as our criterion because, in our view, for important change decisions more than a simple majority is arguably required. In Appendix D we also present a summary of findings based on over 50% majority. It will be up to those involved in the decision-making process to decide which of these is ultimately more appropriate for this purpose.

Perceptions of grade standard adjustment in the languages

Overall results

In this section we present the summary of responses across the 2 questions (as per column 4 of

Table *6* and our acceptability criteria) for the languages.

Figure *4* shows the proportion of positive and negative endorsements by language and grade boundary. The first column in each chart represents the model script(s) and the indicative boundary mark point is denoted accordingly on each chart.

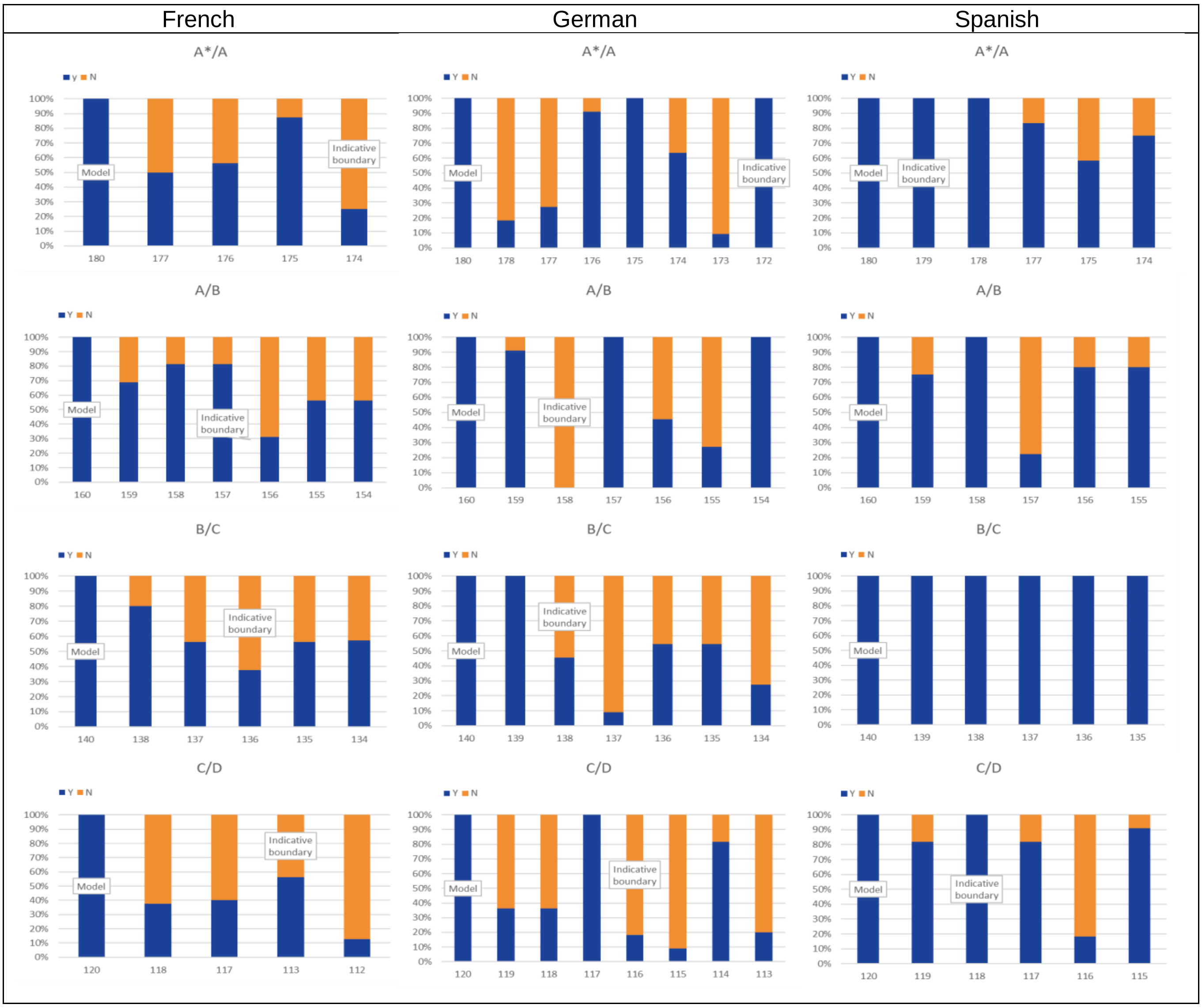


Figure 4. *Proportion of positive and negative endorsements by language and grade boundary.*

It can be seen that the pattern of responses in Spanish is different from the other 2 languages. Even though, based on Rasch analysis, only a small grade standard adjustment is proposed for A*A and CD boundary, and no adjustment for the other 2, the responses suggest that there might be scope to adjust the grade standard across all 4 grades, and to a greater extent than suggested by Rach analysis.

Evidence for acceptability of grade standard change is weak for A*A grade boundary in French and inconsistent for this boundary in German. There is evidence that an adjustment might be acceptable for AB boundary in French and part way down the range suggested by Rasch analysis for BC boundary in French and for both of these boundaries in German. An adjustment would not seem acceptable for CD boundary in either French or German.

Overall, it can be seen that there is a fair amount of inconsistency in the patterns of responses. One might expect to see acceptability decrease with each mark, however, some of the graphs are more 'noisy'. This suggests that the participants were not always happy with the rank ordering of candidates according to their overall mark and illustrates how difficult it is to make fine judgements about performance when faced with some inevitably inconsistent evidence.

Results by question

Figure *5* shows positive endorsements by question as a percentage of the total number of responses given for each question.

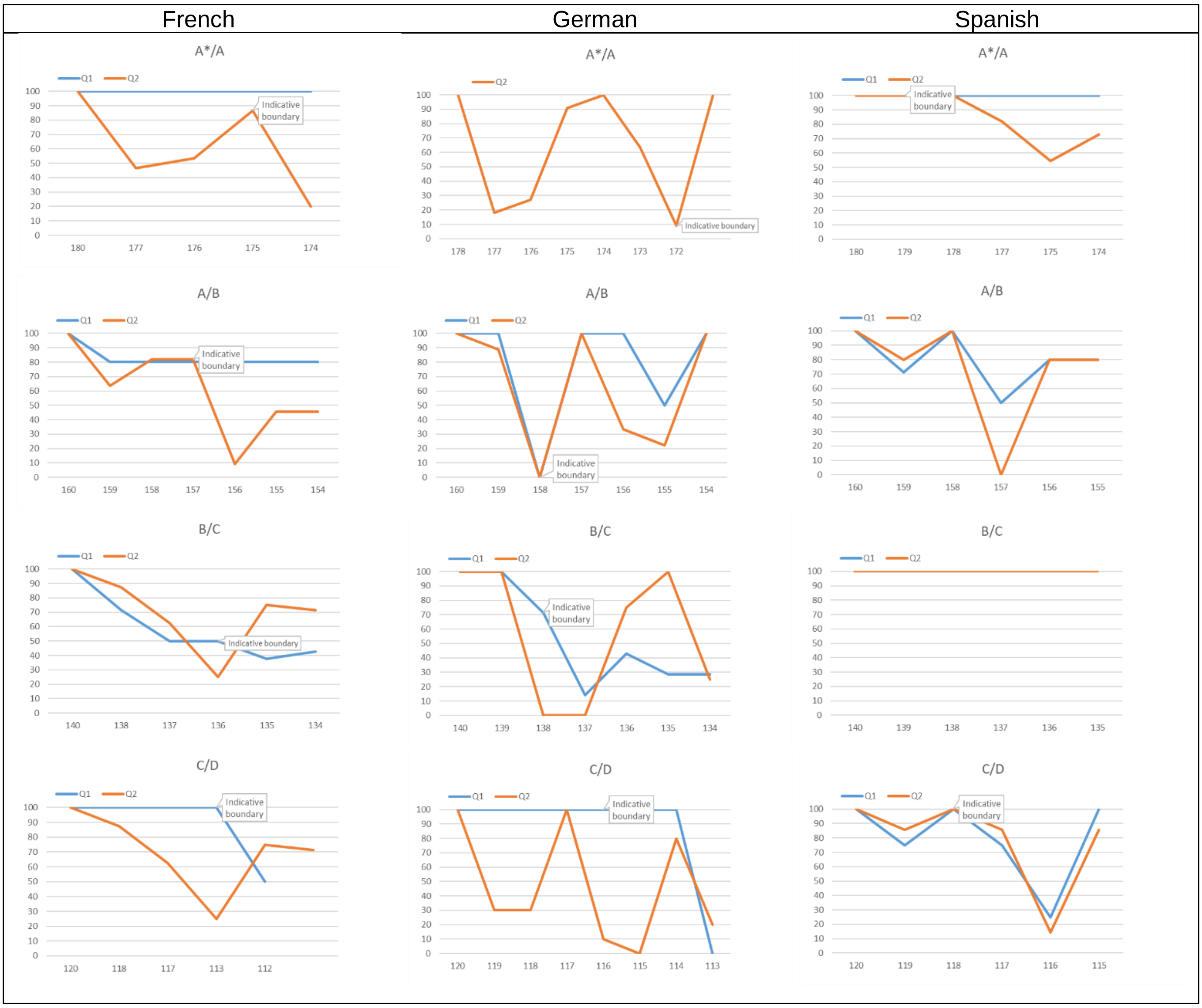


Figure 5. *Percentage of positive endorsements by question in the languages.*

Overall, it can be seen that, in the mark point ranges up to the adjusted grade boundary, the proportion of positive endorsements is either similar across the 2 questions or greater for question 1. The latter pattern suggests that the participants who answered question 1 may have been more ‘lenient’ than those answering question 2. In other words, in some cases they were prepared to admit candidates even though their standard of performance alone may have been seen as less than ideal. This might reflect their awareness that grades are often only one of the criteria for admission to university, and so they were prepared to give ‘benefit of the doubt’ when reviewing scripts in this exercise. Clearly, given a small number of question 1 responses for some grade boundaries, any conclusions and generalisations should be drawn with caution.

Group discussion points

As already mentioned, group discussions were conducted with the participants after each boundary set, as well as at the end of each panel. The panellists were asked about their experience of the script review exercise conducted; about their general perception of the standards of performance in the scripts in each boundary set; whether they felt they could perceive differences between different grades, and within each boundary set; if they felt that a grade boundary adjustment would be acceptable in light of the evidence they had seen; whether this might have a negative impact on university admissions, teaching, etc.

Unsurprisingly, the panellists commented on the difficulty of making holistic judgements in the face of often inconsistent evidence in the scripts, particularly at lower grades, and that some units and questions were not very good discriminators. Whereas a number of panellists raised questions about having to make ‘admissions’ decisions in the absence of any other relevant evidence, there was a recognition that in many institutions the admissions process is centralised and thus more reliant on A level grades for sifting applicants than it might be the case in some selecting universities. Therefore, it was recognised that students who miss or are not predicted to get the required grade might not apply in the first place, or might get a summary rejection from some universities based on the A level grade alone. There were also concerns raised by some representatives about universities raising their grade entry criteria to improve their prestige and position in university league tables, which, in tandem with already low numbers of applicants and perceived high grade standards, risks MFL department closures. Conversely, some panellists pointed out that unconditional offers and offers to study *ab inicio* were also increasingly common and utilised to combat dwindling applications to MFL courses. The latter situation arguably may reduce the importance and relevance normally attached to A level grades in university admissions.

Overall, across the 3 languages, the panellists said that they could confidently differentiate between different boundary sets. In other words, they could see qualitative differences in performances at different grades.

The differences in performance were less clear to them within boundary sets, i.e. between the boundary performance and those just below it. Here, the general view across panels was that they perceived the grade boundaries to be too high (“the papers were one grade lower than they would have expected”). Where they perceived there to be differences between the model scripts and the other scripts in packs, they said that these were often not extreme. The participants across all 3 panels were generally of the opinion that adjusting grade standards would not have a negative impact and that they would have been more than happy to accept students into their institutions who just missed the required grade.

Some comments also chimed with the pattern apparent in the responses for question 1 vs. question 2 (see previous section), as panellists often talked about “seeing flair/potential” in performances below a grade currently required for admission, or that particular students showed they could “make the leap” from sixth form to university despite shortcomings in their scripts/examination performance.

In addition to making their views known during discussions, a number of panellists noted their general views of grade boundary adjustments and other related issues on their response forms. These comments corroborate the main discussion points summarised above. None of the comments written on the forms explicitly disagreed with the need to adjust grade boundaries. We report some of the comments below:

> I felt very strongly that B grade candidates were producing very strong scripts, demonstrating high levels of language competence, and would certainly flourish at Oxford (which currently withdraws our conditional offers if they fall below and Agrade). I'd be much in favour of adjusting the A/B borderline downwards. (French, Oxford)

> Although I would be concerned if these Cs became Bs, I would be a lot more concerned if the current trend (leading pupils not to take French because of the perception that it is more difficult to get a good grade that will impact on their future HE studies choices) continued. It is also to universities to adapt. Otherwise more and more students will start beginners French at university." (French, Newcastle)

> In general, I was surprised that the borderline models were borderline and not higher within the grade, for example the model A*s looked absolutely outstanding to me, not borderline. […] Even where I agreed that the C/D grades were lower than the model, I'm still not sure they deserve a D. (German, Oxford)

> Most candidates, in my opinion, received one grade lower than they deserved. (German, ISMLA)

> There is a serious problem in the marking, which consistently undervalues the work of the candidates. The group consistently thought that marks were too low. […] The impact on the discipline is potentially devastating. (Spanish, King's College London)

> This has been a very revealing, but extremely worrying exercise. I have seen chronic problems with the marking of the essays. Many have been marked too severely (although I say so without having seen the marking criteria or the relative weight given to each element). Even without this, the discrepancies between the "models" and our samples, in which the latter were often superior, is cause for serious concern. Modern languages, I now see, are being attacked from many angles: they have a

> reputation for being tough + the excessively severe marking is reinforcing this. Changing grade boundaries will be a waste of time without tackling examiners' and exam bodies' failures. (Spanish, Aston)

These comments reveal how strongly some of our participants felt about the issue of take-up and severe grading in A level MFLs. This also came through very strongly in the discussions. It should be noted, however, that it is difficult to say based on the current study design whether every participant felt equally strongly about these issues, or entirely agreed with the majority. There is always the risk, widely recognised in the literature, that aspects of group dynamics such as conformity (for example Asch, 1951; Deutsch and Gerard, 1955), polarisation (i.e., adoption of a more extreme position) (Moscovici and Zavalloni, 1969), and, to an extent, ‘group think’ (Baron, 2005) may have created more of a consensus than might have been the case if the discussions were carried out with the participants independently.

Perceptions of grade standard adjustment in the sciences

Overall results

As in the languages,

Figure *6* shows the proportion of positive and negative endorsements by science subject and grade boundary. For A*A boundary in Physics, script review evidence indicates that a small adjustment might be acceptable (to mark point 202), whereas beyond that the evidence is inconsistent. There is evidence that an adjustment might also be acceptable in chemistry, though not to the extent proposed based on the Rach analysis. An adjustment in biology would not appear to be acceptable. At AB boundary, the majority of responses in all 3 subjects suggest acceptability of grade standard adjustment down to about half the width of the mark range based on Rasch analysis. There is little evidence in all 3 subjects that an adjustment would be acceptable at BC and CD boundary.

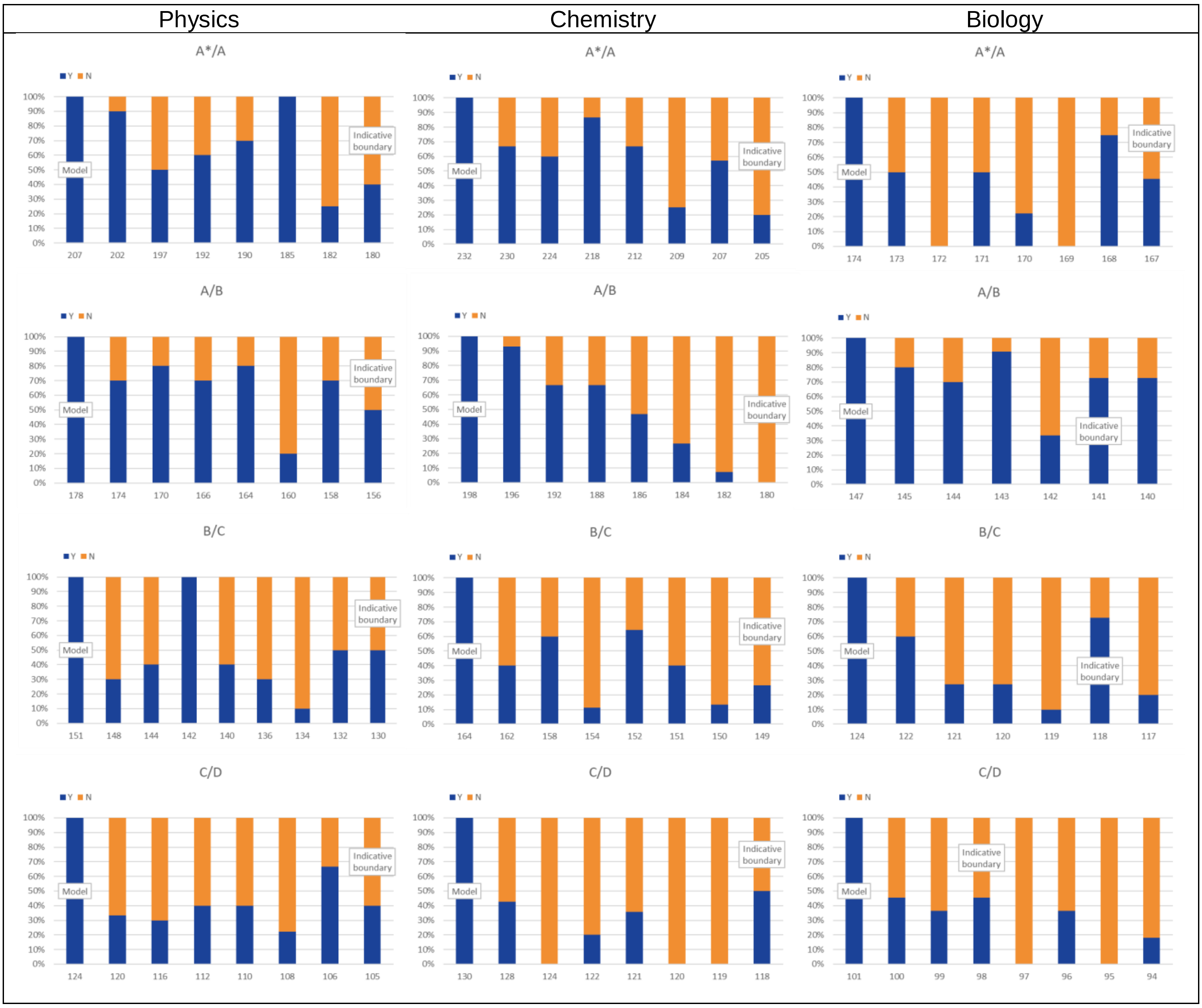


Figure 6. *Proportion of positive and negative endorsements by science subject and grade boundary.*

Overall, similarly to the languages, it can be seen that there is a fair amount of inconsistency in the patterns of responses. Given that gaps between mark points in the sciences are larger than in the languages we may have perhaps expected to see less inconsistency, however the response patterns are fairly noisy here too.

Results by question

Figure *7* shows positive endorsements by question as a percentage of the total number of responses given for each question.

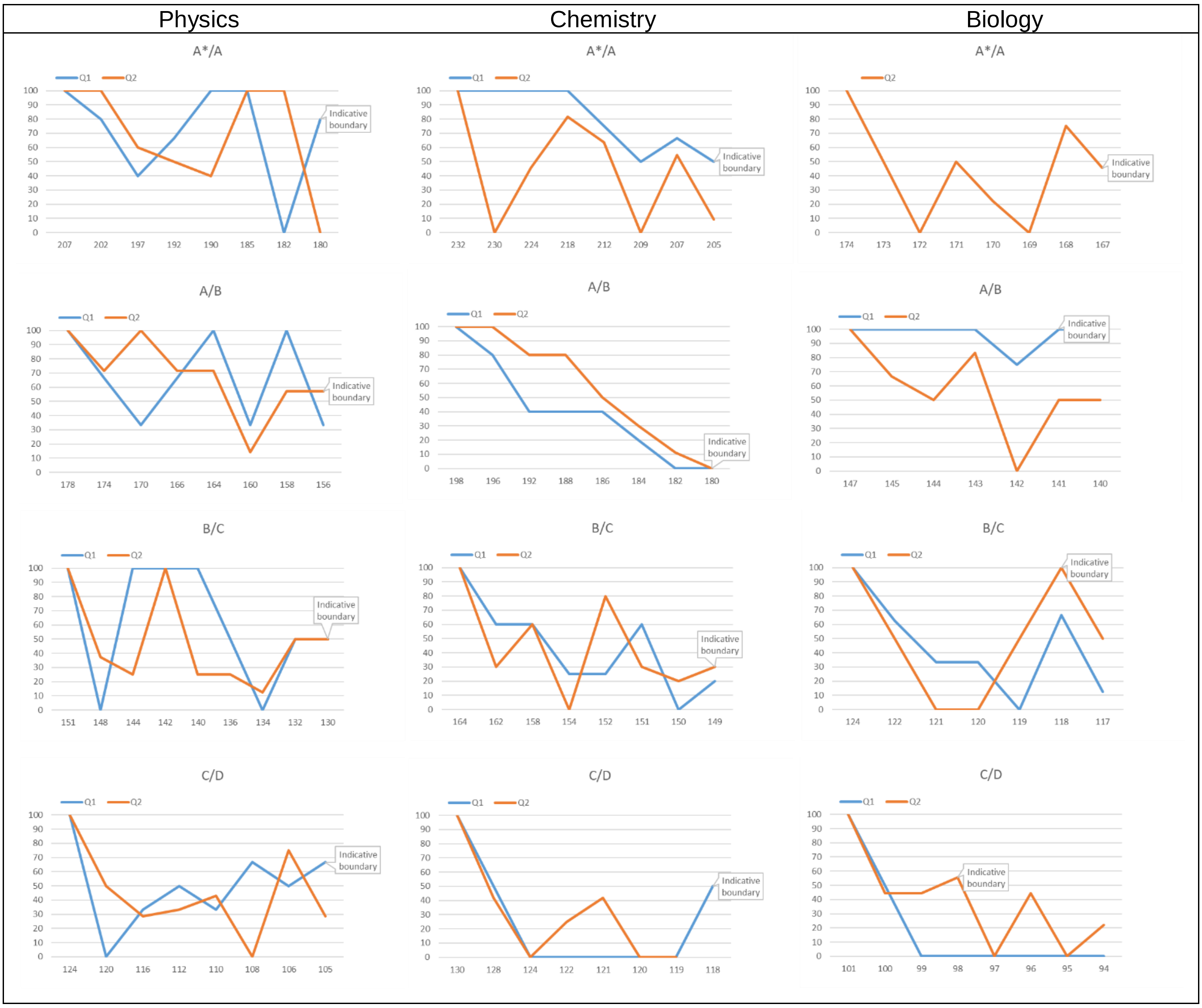


Figure 7. *Proportion of positive endorsements by question in the sciences.*

At A*A boundary in physics there is inconsistency across mark points and to some extent opposite patterns for the 2 questions, although a wider range of mark points was endorsed positively by those answering question 1. The range of positively endorsed mark points for question 1 is also wider in chemistry. No participants answered question 1 at A*A boundary in Biology.

At AB boundary in physics a similar or greater proportion of positive endorsements came from question 1 for most mark points. In chemistry, fewer positive endorsements were made for question 1 throughout the mark point range, and the opposite is true in biology.

At BC boundary, it can be seen that even though, overall, there is not much evidence of acceptability of grade standard change, the responses to question 1 are generally more positive about this in the mark region closer to the original grade boundary in all three sciences (except for mark point 148 in Physics). At CD boundary, there is little evidence across both questions that grade standard adjustment would be acceptable.

Again, given a small number of question 1 responses for some grade boundary sets, any conclusions and generalisations should be drawn with caution.

Group discussion points

Similarly to language panels, group discussions were also conducted with the science panels after reviewing each boundary set, as well as at the end of each panel. Overall comments about difficulties of making judgements based on inconsistent evidence, and regarding the complexities of admissions procedures that cannot be taken into account in this study, were similar to those made by the languages panels. However, there was more variability between different science panels compared to language panels in the general views that came out of the discussions regarding grade standard adjustment.

In physics, there were unclear views from the panel regarding whether or not A*A boundary should be adjusted. Some panellists expressed concerns about impact on admissions in both England and Scotland. At AB boundary, the majority expressed concern about adjusting the grade standard and that there was not sufficient evidence in student performance to justify an adjustment, however this was not a unanimous view. The majority view expressed in discussion is to some extent at odds with the pattern that emerged from panellists' responses, where the majority endorsed positive options for both question 1 and question 2, at least to about half way down the mark range reviewed. The panel was also not unanimous regarding an adjustment at BC boundary, though a number of panellists thought that some adjustment would not cause problems for admissions. This is consistent with the pattern of responses to the 2 questions, and with the fact that question 1 had a greater proportion of positive responses than question 2. There was a general consensus in discussion that there should be no adjustment of standard at CD boundary.

In chemistry, the panellists whose institutions require grade A* for admission said they would be uncomfortable if the A grade performances that were seen below the current A* grade boundary were to be 'upgraded' to an A*. This is to some extent at odds with their script review responses to question 1, where the majority of responses suggest that they would accept students with marks about half way down the range to the newly proposed grade boundary. At AB boundary, a general consensus was that an adjustment of the standard would be problematic for admissions. This is supported by the data for question 1 for most of the mark range, though not for question 2. The panellists did not think an adjustment would be appropriate for either BC or CD boundary. The latter is supported by the data collected from script review, however, the positive responses to question 1 for BC boundary suggest that an adjustment of a few marks might be acceptable for admissions in contrast to what was said in discussion.

In biology, the points made in discussion largely accord with the patterns emerging from the responses collected from script review. Namely, there was some support for adjusting the grade standard at AB boundary only (evident in the script review data particularly from the responses to question 1), whereas panellists generally thought that adjusting the grade standard at other grades would not be advisable. Indeed, some comments suggested that the A*A boundary was too low.

A number of panellists (though notably fewer than in MFL panels) noted their views on their response forms too. In contrast to MFL panels, where we did not get any comments explicitly against adjusting grade boundaries, here we did not get any comments explicitly advocating changing the A*A boundary, even though some comments supported changing other boundaries. Some of the comments are reported below:

> In my view, the grades are better than the quality of student work, i.e. A* are As, As are Bs etc. This probably means that the current grade boundaries are just about right. (Physics, Edinburgh)

> I don't think I would want A* to embrace many more students (i.e. those whose knowledge isn't as good as this student's, or who does [sic] not calculate as accurately). (Physics, Imperial College)

> There is not a lot of difference between B, C and D candidates, with some D having quite clear understanding that higher grade candidates cannot show. Happy to lower boundaries (not A* though) to enhance the grade of more candidates. (Physics, Leeds)

> After this exercise I'd be tempted to return to my institution and make an argument to increase our admission grade for chemistry or at least insist that we never take a student with a dropped grade in chemistry (as we occasionally do now). (Chemistry, Kings College London)

> I think it is very important that the A* and A boundaries do not move lower. This would significantly affect our ability to reliably admit the candidates we would want. (Chemistry, York)

Results summary

Tables 7 and 8 summarise the majority views of whether an adjustment to grade standard might be acceptable to our sample of stakeholders in different subjects. We summarise the views emerging from the script review and from the discussions with the panellists in separate columns for each subject.

Tables 7 and 8 present summaries of the script review evidence based on over 60% majority acceptance criterion for languages and sciences respectively. Appendix D contains tables where the script review summary is based on the over 50% majority criterion. For the summaries of discussion views we have used the impressions of the chair and observer of each panel, as evidenced in their notes, of whether a majority of the panellists was of a particular opinion. Hence, the columns summarising the discussions are the same in both tables.

We have used the following categories to summarise the acceptability views from the script reviews:

- Y – adjustment acceptable all the way to and including the indicative grade boundary mark point
- Y part way – adjustment acceptable part way to the indicative grade boundary mark point
- N – adjustment not acceptable (this includes situations where all or all except one mark points in the range between original and indicative grade boundary are not deemed acceptable as a potential new grade boundary)
- Inconsistent – 2 or more mark points in the range between original and indicative grade boundary considered acceptable while others are not, or vice versa.

When summarising the discussion views we have also used the aforementioned categories, except “Y part way” as the discussions were in most cases more general and did not indicate whether approval of adjustment only referred to a portion of mark points presented.

Table 7. *Over 60% majority view of a possible adjustment by language and grade boundary*.

	French		German		Spanish	
	Script review	Discussion	Script review	Discussion	Script review	Discussion
A*A	N	Y	Inconsistent	Y	Y	Y
AB	Y	Y	Y part way	Y	Y	Y
BC	Y part way	Y	Y part way	Y	Y	Y
CD	N	Y	N	Y	Y	Y

Table 8. *Over 60% majority view of a possible adjustment by science subject and grade boundary.*

	Physics		Chemistry		Biology	
	Script review	Discussion	Script review	Discussion	Script review	Discussion
A*A	Y part way	Inconsistent	Y part way	N	N	N
AB	Y part way	N	Y part way	N	Y part way	Y
BC	N	Inconsistent	N	N	N	N
CD	N	N	N	N	N	N

We discuss the patterns apparent in these tables in the following section.

Discussion

The collated script review data provides evidence of the acceptability of some adjustment to most of the grade boundaries in the languages and to somewhat fewer grade boundaries in the sciences. This is to some extent at odds with the evidence from Rasch analysis, which suggested much more of a ‘gap’ in alignment between the sciences and mathematics, than between the languages and geography. This is perhaps not surprising given that the research task is not asking about whether or to what extent there is a difference in grading standards between subjects; it is asking about the acceptability of altering the grading standard within a subject in terms of progression and selection for study in higher education.

More often than not in the science subjects the discussions by the panels indicated a lack of acceptance of any adjustment, even on the 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 grade boundaries in Spanish.

These contrasting patterns of acceptability could be an indication of different perceptions that HE stakeholder communities have of the current situation in these subjects with respect to uptake and progression to HE. There is a strong perception in the MFL community that these subjects are ‘in crisis’, and that this is in no small part due to the severity of the grading standards. Such perception is perhaps less pronounced in the sciences community. This situation has likely provided a backdrop to their decisions in the exercise that the participants were asked to carry out.

It should be noted that a number of methodological issues limit the generalisability of these findings. For both sets of subjects, our samples of HE representatives were small, and, although diverse, not fully representative of the HE sector. In addition, the number of scripts that the participants were able to review in the time available was also relatively small, particularly for the languages and the scripts were only from one specification and board for each subject.

Specifically with respect to the languages, the participants were unable to see the whole candidate work, due to the AS scripts being unavailable. This last issue, however, may not be a major limitation in the context of MFL qualifications since, because of their structure and the nature of languages, similar skills are tested at A2 as at AS. Thus the students' performance at A2 is likely to give a good representation of their overall ability. But it is worth noting that the specifications that the scripts were based on had their final proper sitting in 2017 and have now become 'legacy specifications' following the reform of A level MFL specifications (with new A levels in MFL awarded in summer 2018).

The interpretation of the script review results also hinges on the particular acceptability criterion that we have selected (above 60% acceptability per grade boundary) and could change if a different criterion is selected. The interpretation of the results, and indeed the outcome of the exercise is also linked with whether or not the choice of the reference subjects was appropriate. Determining the reference subjects was a necessary but essentially an arbitrary decision. Even though this decision is open to challenge, it 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 facilitating status in HE admissions; their relative difficulty based on Rasch analysis; and feedback from stakeholders.

The judgemental exercise that the participants were asked to carry out was not straightforward given the inevitably inconsistent evidence that was contained in the scripts and differences between some consecutive scripts of as little as a single mark point. In addition, different judges valued different aspects of scripts, which is perhaps unavoidable given that these subjects at HE have different emphases and specialities. For instance, some MFL representatives valued fluency and flexibility in language more than grammatical knowledge; some valued writing and comprehension skills more than speaking at A level, as they believed that speaking skills can be more easily improved later on with studying abroad for instance. In sciences, participants sometimes had different views about the value of problem-solving skills versus factual knowledge at A level. In addition to this, the participants inevitably would have had somewhat different individual standards – this was evidenced in their discussions of the properties of model scripts for instance. All this could explain some of the apparent inconsistencies in acceptability patterns at script level. However, we believe that, given the questions in this study, it was necessary to have captured this diversity of views within the same subject panels even if they gave a less clear picture of acceptability.

Given the study design, we place more emphasis on the results of the script review exercise and less on the outcomes of panel discussions, even though the latter need to be taken into account to some extent. An issue with the outcomes of the panel discussions is that it is difficult to say whether every participant felt equally strongly or entirely agreed with the majority regarding grading standards and other issues discussed. There is a possibility that aspects of group dynamics such as conformity (e.g. Asch, 1951; Deutsch and Gerard, 1955), polarisation (i.e., adoption of a more extreme position) (Moscovici and Zavalloni, 1969), and, to an extent, 'group think' (Baron, 2005) may have created more of a consensus than might have been the case if the discussions were carried out with the participants independently.

Despite all of this, and with some exceptions, we can see a downward trend of acceptability of the scripts as the marks become lower in most subjects and grade boundary sets. This seems to indicate that, despite the challenges of the task, and the fact that the order of the script was randomised, some genuine judgements were captured in the script review exercise.

In summary, this work provides some evidence from the script review of the acceptability of some grade boundary adjustments in some subjects. This is more clearly the case for modern foreign languages than the sciences. Where this was also the case in the sciences, the discussion by the panels did not always support making an adjustment.

References

Andrich, D. (2015). The problem with the step metaphor for polytomous models for ordinal assessments. *Educational Measurement: Issues and Practice*, 34, 8–14.

Bramley, T. (2014). Multivariate representations of subject difficulty. *Research Matters: A Cambridge Assessment Publication*, 18, 42-47.

Bramley, T. (2016). The effect of subject choice on the apparent relative difficulty of different subjects. *Research Matters: A Cambridge Assessment publication*, 22, 23-26.

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

Benton, T. & Bramley, T. (2017). Some thoughts on the '*Comparative Progression Analysis' method for investigating inter-subject comparability*. Cambridge, UK: Cambridge Assessment. http://www.cambridgeassessment.org.uk/Images/416591-some-thoughts-on-the-comparative-progression-analysis-method-for-investigating-inter-subject-comparability.pdf

Coe, R. (2008). Comparability of GCSE examinations in different subjects: An application of the Rasch method. *Oxford Review of Education*, 34(5), 609–36.

Coe, R. (2010). Understanding comparability of examination standards. *Research Papers in Education*, 25(3), 271–284.

Cuff, B. M. P. (2017). Perceptions of subject difficulty and subject choices: Are the *two linked, and if so, how?* Coventry, UK: Ofqual. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/649891/Perceptions_of_subject_difficulty_and_subject_choices.pdf

Dearing, R. & King, L. (2007). *Languages review*. Nottingham, UK: Department for Education and Skills. https://www.languagescompany.com/wp-content/uploads/the-languages-review.pdf

He, Q., Stockford, I. & Meadows, M. (2018). Inter-subject comparability of examination standards in GCSE and GCE in England. *Oxford Review of Education*, Published online: 28 Feb 2018.

He, Q. & Meadows, M. (2018). Using the Rasch model to investigate inter-board comparability of examination standards in GCSE. *Journal of Applied Measurement*, 19(2), 129-147.

Ipsos Mori (2012). Fit for Purpose? The view of the higher education sector, teachers and employers on the suitability of A levels. Coventry, UK: Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/377930/2012-04-03-fit-for-purpose-a-levels.pdf

Laming, D. 2004. *Human judgment.* London: Thomson.

Linacre, J. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16, 878.

Linacre, J. (2015). *Winsteps® Rasch measurement computer program user's guide*. Beaverton, OR: Winsteps.com.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.

Myers, H. (2006). *The 'severe grading' of MFL grades at GCSE and A level*. London: Association for Language Learning.

Kelly, A. (1976). A study of the comparability of external examinations in different subjects. *Research in Education*, 16, 37–63.

Korobko, O. B., Glas, C.A.W., Bosker, R. J., & Luyten, J.W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement*, 45(2), 139–157.

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: part I. *Journal of Applied Measurement*, 4(4), 386-422.

Newton, P. E. (1997). Measuring Comparability of Standards between Subjects: Why Our Statistical Techniques Do Not Make the Grade? *British Educational Research Journal*, 23(4), 433-449.

Newton, P. E. (2012). Making sense of decades of debate on inter-subject comparability in England. *Assessment in Education: Principles, Policy & Practice*, 19(2), 251-273.

Newton, P. E., He, C. & Black, B. (2017). *Progression from GCSE to A level. Comparative Progression Analysis as a new approach to investigating inter-subject comparability*. Coventry, UK: Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/610077/Progression_from_GCSE_to_A_level_-_Comparative_Progression_Analysis_as_a_new_approach_to_investigating_inter-subject_comparability.pdf

Ofqual (2015a). *Comparability of different GCSE and A level subjects in England: An introduction: ISC working paper 1*. Coventry, UK: Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attach ment_data/file/606041/1-comparability-of-different-gcse-and-a-level-subjects-in-england-an-introduction.pdf

Ofqual (2015b). Inter-Subject Comparability: A Review of the Technical Literature. ISC Working Paper 2. Coventry, the Office of Qualifications and Examinations Regulation. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/606043/2-inter-subject-comparability-a-review-of-the-technical-literature.pdf

Ofqual (2016a). *A policy position for Ofqual on inter-subject comparability* (Paper 58/16). Coventry, UK: Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/610111/Board_paper_-_Inter-subject_Comparability.pdf

Ofqual (2016b). Evaluating the summer 2015 results of A level French, German, and Spanish. Coventry, UK: Ofqual. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/544636/Evaluating_A_level_MFLs.pdf

Ofqual (2017). GCE Qualification Level Conditions and Requirements. Coventry, UK: Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/644330/gce-qualification-level-conditions-and-requirements.pdf

Pollitt, A. (1996). The ‘difficulty’ of A level subjects. Report for the University of Cambridge Local Examinations Syndicate. Unpublished.

Raikes, N., Scorey, S. & Shiell, H. (2008, September). *Grading examinations using expert judgements from a diverse pool of judges*. Paper presented at the 34th annual conference of the International Association for Educational Assessment, Cambridge, UK.

Royal Society (2008) *Science and mathematics education 14–19. A ‘state of the nation’ report on the participation and attainment of 14–19 year olds in science and mathematics in the UK, 1996–2007*. London: The Royal Society.

Scharaschkin, A., and Baird, J. (2000). The Effects of Consistency of Performance on A Level Examiners’ Judgements of Standards. *British Educational Research Journal*, 26, 343-357.

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370. https://www.rasch.org/rmt/rmt83b.htm

Appendix A: List of participating institutions

This list includes all participating institutions in alphabetical order (within each subject), excluding Ofqual subject specialists as these were not representing any institutions and worked in a consulting function.

French	German	Spanish
Cardiff University	Aston University	ISMLA
Newcastle University	ISMLA	Aston University
Oxford Brookes University	Newcastle University	Cardiff University
The Association for Language Learning	University of Bristol	King's College London
University of Edinburgh	University of Cambridge	University of Exeter
University of Oxford	University of Hull	University of Hull
University of Portsmouth	University of Liverpool	University of Liverpool
University of Sheffield	University of Oxford	University of Manchester
University of Southampton	University of Warwick	University of Nottingham
University of St Andrews	University of York	University of Oxford
University of Warwick		University of York
University of York		
York St John University		

Physics	Chemistry	Biology
Durham University	Cardiff University	King's College London
Imperial College London	De Montfort University	Royal Society of Biology
Institute of Physics	Keele University	St George's, University of London
University of Edinburgh	King's College London	University of East Anglia
University of Hull	London Metropolitan University	University of Lincoln
University of Leeds	Newcastle University	University of Liverpool
University of Leicester	Queen Mary University of London	University of Portsmouth
University of Oxford	The Royal Society of Chemistry	University of Reading
	University College London	University of St Andrews
	University Of Brighton	University of Surrey
	University of Edinburgh	
	University of Oxford	
	University of York	

Appendix B: Order of script presentation

French		German		Spanish		Physics		Chemistry		Biology	
A*A	180	A*A	180	A*A	180	A*A	207	A*A	232	A*A	174
	180		180		180		207		232		174
	175		176		179		202		224		171
	177		172		177		190		212		167
	174		174		178		180		205		169
	176		178		174		197		207		170
	178		175		175		185		218		172
	179		173	AB	160		192		209		173
	173		177		160		182		230		168
AB	160	AB	160		159	AB	178	AB	198	AB	147
	160		160		155		178		198		147
	157		159		156		156		180		143
	158		158		158		164		186		141
	156		156		157		170		192		144
	155		154	BC	140		174		184		140
	159		157		140		160		188		145
	154		155		139		166		182		142
BC	140	BC	140		136		158		196	BC	124
	140		140		135	BC	151	BC	164		124
	138		139		137		151		164		118
	136		138		138		144		152		121
	135		136	CD	120		136		158		117
	137		134		120		130		149		120
	134		137		119		134		151		122
	139		135		118		140		150		119
CD	120	CD	120		117		142		162	CD	101
	120		120		116		132		154		101
	113		116		115		148	CD	130		100
	117		119			CD	110		118		99
	114		117				116		124		96
	119		118				105		121		97
	118		114				120		128		95
	112		115				108		122		94
	116		113				106		120		
							112		119		

Appendix C: Examples of response forms

Q1

A*/A

NAME (Please print):

Please record your decisions by circling one of the options and noting any comments in the table below:

Candidate No.	Deserving of admission?	Comments
23393		
28520		
26001	**Y Y? N?**	
25921	**Y Y? N?**	
20117	**Y Y? N?**	
25604	**Y Y? N?**	
28556	**Y Y? N?**	

Q2

A*/A

NAME (Please print):

Please record your decisions by circling one of the options and noting any comments in the table below:

Candidate No.	How does each script compare to the model?	Comments
23393		
28520		
26001	**BETTER SIMILAR WORSE**	
25921	**BETTER SIMILAR WORSE**	
20117	**BETTER SIMILAR WORSE**	
25604	**BETTER SIMILAR WORSE**	
28556	**BETTER SIMILAR WORSE**	
26086	**BETTER SIMILAR WORSE**	

Appendix D: Summary of over 50% majority views

Table 9. *Over 50% majority view of a possible adjustment by language and grade boundary.*

	French		German		Spanish	
	Script review	Discussion	Script review	Discussion	Script review	Discussion
A*A	Inconsistent	Y	Inconsistent	Y	Y	Y
AB	Y	Y	Y part way	Y	Y	Y
BC	Y part way	Y	Y part way	Y	Y	Y
CD	N	Y	N	Y	Y	Y

Table 10. *Over 50% majority view of a possible adjustment by science subject and grade boundary.*

	Physics		Chemistry		Biology	
	Script review	Discussion	Script review	Discussion	Script review	Discussion
A*A	Y part way	Inconsistent	Y part way	N	N	N
AB	Y part way	N	Y part way	N	Y part way	Y
BC	N	Inconsistent	Inconsistent	N	Y part way	N
CD	N	N	N	N	N	N

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/18/6450/4**