

# Ofqual Board

## Paper 58/16

**Date:**

30 November, 2016

**Title:**

A policy position for Ofqual on inter-subject comparability

**Report by:**

Dennis Opposs, Standards Chair

**Responsible Director:**

Michelle Meadows, Executive Director for Strategy, Risk and Research

**Paper for decision**

**Open paper**



**Issue**

1. Ofqual began a public debate on inter-subject comparability in GCSEs and A levels almost a year ago and the Board had a substantive discussion about the issues at its meeting in May. Members are now asked to decide Ofqual's policy position on this matter.

**Recommendation**

2. The Ofqual Board is invited to agree that its policy on inter-subject comparability in GCSEs, AS and A levels should be:
  - 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).

3. The first qualifications that should be examined to see whether an exceptional and compelling case exists are A levels in physics, chemistry and biology, and in French, German and Spanish.

### **Background**

4. The qualifications standards objective in the 2009 Act focuses on us securing “a consistent level of attainment (including over time) between comparable regulated qualifications”. The definition of comparability in the general conditions does not explicitly cover inter-subject comparability. In GCSE and A level terms, it is primarily concerned with us securing reasonable levels of comparability between, for example, AQA history syllabus A and AQA history syllabus B, OCR A level Spanish and Pearson A level Spanish, and between GCSE English literature in 2010 and GCSE English literature in 2015. It is not concerned with making such comparisons as A level biology with A level media studies so that has not been a priority for us.
5. We do, of course, have a duty to take notice of what stakeholders tell us and we are aware, for example, of perceptions that severity of grading in A level German unfairly disadvantages that subject and is a causal factor in the decline in its entries over time.
6. We therefore decided that the time was right to start a public discussion about whether Ofqual should take any action with regard to the comparability of GCSE and A level grade standards between subjects.
7. In December 2015, we began on our website a debate on comparability between subjects in GCSE and A level exams<sup>1</sup>. The intention was that any changes to be implemented would not affect the introduction of the reformed GCSEs and A levels but would take place over a longer timeframe.
8. The launch included a set of six new Ofqual working papers, a video interview to introduce the topic and the publication of an infographic (attached at Annex A) that explains the four policy options we put forward. We also published links to earlier publications - from Ofqual, QCA and others - that help consideration of the topic.
9. At the same time, a survey was opened on our website asking which policy option was closest to people’s preferred position and providing an opportunity for them to explain why. A summary of the outcomes of the survey are attached at Annex B. There were 216 responses. 52 of the replies indicated a subject background, 28 of which were in modern foreign languages.
10. On 4 February 2016 we held a conference for about 100 people in London. The conference programme is attached at Annex C. The main aim of the conference was to provide a range of input and opportunities for discussion to attendees so that they would consider the issue in an informed way before responding to us.

---

<sup>1</sup> See: <https://www.gov.uk/government/collections/inter-subject-comparability-research-documents>

11. Since the start of the year we have carried out some additional research in house to explore the extent to which subject choice is driven by perceptions of difficulty. A summary of the research, including the findings, is attached at Annex D. Students' choices can be based on avoiding subjects where they aren't likely to get their best grade. The main conclusion drawn from the research is that there does seem to be a link between perceptions of subject difficulty and subject entry choices, although perceptions of enjoyment and usefulness appear to interact with, and often supersede, this relationship.
12. Our initiative in beginning the debate also stimulated research elsewhere and this is summarised at Annex E.
13. At its meeting on 18 May, the Board discussed Paper 7/16 which brought Members up to speed with the work on inter-subject comparability and gave them an opportunity to discuss and scrutinise proposals before a final paper was developed for formal decision. The paper sought the Board's views on the arguments put forward in it and how persuasive these were.
14. The main outstanding issue at that time concerned the work that had been started in response to the implications of the letter we received from several high-profile science organisations (annex F). We reported in May that we were replicating and extending the analysis in figure 2 of the letter to see what it may be telling us.

#### **Advice from the Standards Advisory Group**

15. The initial analysis was considered by the Standards Advisory Group (SAG) at its meeting on 1 July. Members generally agreed that the evidence was interesting and required further consideration. The data, based on value added within a subject, was thought to be more challenging to critique than a general value added measure based on mean GCSE. The evidence was considered to be persuasive and not to be dismissed at this stage.
16. We carried out considerably more work on what has become known as the "comparative progression analysis" (CPA) over the summer. A full day meeting to consider the analysis was held on 24 September. Several members of SAG as well as some exam board experts attended.
17. The main paper discussed at that meeting, *Progression from GCSE to A level*, is attached at Annex G. Some of the figures in the paper are reproduced at the end of the annex in forms that may be easier to interpret.
18. The outcomes of that meeting and the papers for it were then considered in detail at the SAG meeting on 14 October.
19. The meeting concluded that the CPA analysis was quite persuasive but that conclusions shouldn't be drawn from this evidence source alone. Some thought that evidence from CPA was not as statistically robust as that from certain other techniques such as mean GCSE value-added analyses (the CPA input variable is a single GCSE, whereas the mean GCSE input variable averages across many GCSEs). There were questions about the assumption

that GCSEs were well aligned across subjects and that teacher quality was taken to be consistent across subjects. There were also concerns that CPA cannot provide evidence for GCSEs nor for A level subjects that do not have GCSEs (such as philosophy).

20. Another paper considered at the SAG meeting, using a new simulation of data from A level science and humanities subjects, suggested that perhaps half of the CPA effect may be caused by a combination of the non-random way in which students choose their A level subjects and the different correlations found between GCSE and A level outcomes in individual subjects. In other words, it seems that a significant part of the effect brought to our attention by the science organisations may be a statistical artefact. It was also noted that A level subjects differ in the extent to which prior GCSE knowledge was a critical requirement, it being much more necessary in, for example, science subjects, than in, for example, humanities.
21. Members suggested that CPA provided one part of a story and should contribute towards a basket of evidence that could be collected in relation to a subject or group of cognate subjects. Where the basket of evidence in a subject was judged to be sufficiently persuasive, Ofqual would have a basis on which to take action.
22. In addition to CPA outcomes, this basket might include, for instance, results from other statistical techniques (for example, subject-pairs analysis and Rasch) and concerns from stakeholder groups (for example, subject associations and HE selectors) as well as contextual data (for example, teacher numbers and quality).
23. SAG has not previously suggested that the statistical and other evidence available is sufficiently strong for Ofqual to support a general recalibration of grade standards across GCSEs and A levels. The most recent analyses do not appear to change that position fundamentally. Even so, there are legitimate concerns about reduced entries in a few subjects that ought to be addressed in some way.
24. Overall SAG agreed that where the basket of evidence provided an exceptional and compelling case in a subject, small changes might be made over a series of years to modify grade standards. However, members considered that making such changes would not necessarily solve some of the problems about which subject communities are concerned. For example, simply adjusting grade boundaries is not likely to be sufficient to address problems in take-up and progression to HE in A level physics. It's worth noting that this is not a subject where there are indications from HE personnel that the A level is "too hard". In addition, even if the A level curriculum is untouched, if grade boundaries are lowered teachers may respond by adjusting their teaching thereby affecting what students learn.
25. SAG's advice is that the policy should be: where Ofqual has a case that it considers compelling, to discuss with other stakeholders implications for, in particular, the curriculum in that subject, before it takes any action to adjust

grade standards; to improve the quality of assessments where it may be creating detrimental impacts in particular subjects; but to take no coordinated action to align standards across the full range of subjects through grading.

## Analysis

26. Inter-subject comparability is a far thornier area than comparability within subjects or even comparability over time. It is fairly straightforward to describe what we mean if we want to compare AQA geography standards with those in OCR geography, or to compare GCSE maths standards in 2010 with those in 2015. It is hard to explain what it means to compare standards in, say, physics and art. Indeed, some have argued that the *only* way in which students can meaningfully be compared across such disparate subject areas is in relative terms; which would recommend a policy akin to norm-referencing, for example, stipulating the same distribution of grades (at a national level) across all subjects.
27. A common way to conceptualise comparability between subjects involves using a construct called something like “general academic aptitude”. This aptitude is the common thread between subjects. Students with a high “general academic aptitude” are said to be better at learning and they use that advantage irrespective of exactly what subject they are studying be it biology or history or drama. So a student with a high “general academic aptitude” will typically do better in all their subjects than a student with a low “general academic aptitude”.
28. Many who responded to our survey expressed a preference for action based on Rasch-based statistical measures of subject difficulty. Those who indicated their subject background were largely modern linguists so had a primary interest in a subject which has a statistical measure suggesting that it is “difficult” and were looking for grading in that subject to be made more lenient. We can’t, though, be certain what those statistics are measures of or exactly what they are telling us. So, for example, we instinctively don’t believe that it’s harder to get a high grade in general studies than in almost every other A level subject and yet that is what the statistics seem to tell us. There are certainly many universities who do not consider general studies to be the equal of other A levels. We think that the difficulty measure here probably has something to do with students putting less time and effort into general studies than other subjects and so getting a less good grade but we have not been able completely to unpick this effect.
29. Recent research<sup>2</sup> (Annex E) raises the possibility that because correlations between different subjects are unequal (for example, there is a higher correlation between the grades students get in chemistry and biology than between mathematics and art), and because students each choose one particular selection of subjects rather than any other, that subject choice can create spurious differences in the statistical measures of subject difficulty such as Rasch. The differences might be statistical artefacts. All in all, the statistical

---

<sup>2</sup> 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.

measures may not be the panacea to the inter-subject comparability conundrum that some seek.

30. If we can't confidently conceptualise or measure how subjects compare, it would be misleading to make serious changes to try to achieve better inter-subject comparability. There are concerns – and we hear those particularly from some linguists and scientists – but it is not clear how far those concerns will be addressed only by making the adjustments to grading standards that those subject communities may propose. As described in Annex E, since statistical alignment is based on the average of groups of candidates, it may have a much more limited impact at the level of the individual in a selection process or when it comes to school accountability measures.
31. In GCSE and A level French and German, there are real problems of declining entries and issues to do with the assessments, some of which Ofqual has been trying to tackle<sup>3</sup>. It is far from clear, though, that the concerns raised are produced by grading standards rather than by, for example, persistent curriculum changes, exams that don't discriminate at the top end, increasing entries from native speakers and teacher supply issues.
32. When considering whether there is a “compelling case” about grade standards in A level French, German and Spanish, we will factor in Ofqual's present work on native speakers. If that shows that such candidates are typically being awarded the top A level grades despite having unexceptional mean GCSE prior attainment across subjects, we might decide to remove native speakers from the statistical prediction used at the award. The effect of that would be that typical native speakers would still receive high grades but that more non-native speaker candidates would also receive high grades. Overall then, more high grades would be awarded and we might in effect have adjusted the grade standard back to where it was before there were so many native speakers.
33. The findings so far do suggest a difference in performance between native and non-native speakers. However, this does not necessarily mean that there is an issue for standard setting and the next step will be to consider this.
34. We recognise that there are also concerns about whether enough young people in England are choosing to take A level physics. Many would say that action is needed. Yet, if we were to adjust A level physics grading standards so that each student's result is about one grade better than it is now, how far would that go to persuade more students that physics is not just a subject for the “ultra-bright” or to persuade more girls that physics is not “a masculine subject”<sup>4</sup>? In subjects such as physics and further mathematics, a better

---

<sup>3</sup> See, in particular, Ofqual's report “Evaluating the summer 2015 results of A level French, German, and Spanish” at [https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/544636/Evaluating\\_A\\_Level\\_MFLs.pdf](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/544636/Evaluating_A_Level_MFLs.pdf)

<sup>4</sup> These terms come from a recent report, “Tough Choices”, developed by A T Kearney in partnership with Your Life campaign. It is available at: [https://www.atkearney.co.uk/about-us/social-impact/related-publications-detail/-/asset\\_publisher/EVxmHENiBa8V/content/tough-choices/10192](https://www.atkearney.co.uk/about-us/social-impact/related-publications-detail/-/asset_publisher/EVxmHENiBa8V/content/tough-choices/10192)

action than reducing grade standards might be for HEIs to apply differential entry tariffs rather.

35. The position with STEM subjects more generally, where there would be wide agreement that improving take-up would be a good thing, is very complex. So, for example, on statistical (Rasch) measures of difficulty, A level physics and biology are fairly equal yet biology attracts almost twice as many students and a much greater proportion of those are female (60% versus 20%).
36. It is the case that some jurisdictions do make such adjustments but Working Paper 4 and Iasonas Lamprianou's presentation at the 4 February conference both raise questions about whether the benefits have outweighed the public controversies that have followed.

### **Recommendation**

37. The analysis above suggests both that we should treat statistical measures of subject difficulty with caution and that adjusting grade boundaries in some subjects may not have much impact on take-up. Research (see Annex E) indicates that changing some subject grade standards would only have small effects on performance table rankings<sup>5</sup> or the interchangeability of the grades that individuals are awarded<sup>6</sup>.
38. However, taking no action whatsoever will not help address issues about participation levels in subjects that are seen as important to the nation. Neither is it likely to be seen by subject communities as an adequate response to their concerns about fairness between subjects.
39. We may have a subject where we find that we have a basket of evidence made up of:
  - results from statistical techniques which mainly point in the same direction even if no individual analysis on its own is persuasive, plus
  - concerns from stakeholder groups that indicate educationally damaging consequences, together with
  - contextual data such as figures on teacher supply.
40. That evidence may amount to a potentially compelling case. In such cases, a best first step might be for Ofqual to convene a forum of the main parties – perhaps comprising subject representatives, the DfE (in its role as owner of curriculum content), exam boards and Ofsted - to consider the evidence and discuss what the best answer may be and what effect it might have. That way forward might involve changes to the curriculum for that subject or approaches to teaching (which are beyond Ofqual's remit) as well as adjustments to grade boundaries (which are firmly within it). Whether the action taken should only

---

<sup>5</sup> 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.

<sup>6</sup> AQA research in preparation for publication

involve adjustments to grade boundaries may be questionable given the limited impact that action alone is likely to have.

41. The Ofqual Board is therefore invited to agree that its policy on inter-subject comparability in GCSEs, AS and A levels should be:
  - 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).
42. The first qualifications that should be examined to see whether an exceptional and compelling case exists are A levels in physics, chemistry and biology, and in French, German and Spanish.
43. The initial step should be to clarify how a basket of data for a subject should be put together. In addition, we should determine what criteria should be used to determine what counts as a compelling case. (That might concern an alignment of different evidence sources – statistical, behaviours and public concerns.) To assist that process, the subject communities from schools, colleges and HE should be invited to contribute data and advice.

### **Finance and Resource**

44. Work on improving the quality of assessments in subjects such as A level French is covered by the present business plan. Implementing the proposed policy position will require a considerable amount of senior staff time. There are minimal financial implications.

### **Impact Assessments**

#### Equality Analysis

45. Grade setting focuses on the level of the knowledge, skills and understanding which has been demonstrated in assessments, and does not take account of the particular characteristics of the individual students who have taken those assessments. The grade awarded to each individual student solely reflects the performance of that student in that assessment. To do otherwise would risk introducing different standards in the same qualification for students with protected characteristics and those without. This would not be desirable for students, employers or further and higher education institutions.

We have not identified in respect of the policy proposal in this paper, any potential impacts on students who share any of the protected characteristics.

### Risk Assessment

*This section has been redacted, as its publication would be prejudicial to the effective conduct of public affairs.*

### Regulatory Impact Assessment

46. As no change to the present position is being proposed, there should be no regulatory impact.

### **Timescale**

47. An announcement should be made early in the New Year.

### **Communications**

48. We will need to provide some clear messages, particularly to those subject communities that may be disappointed that we have not completely followed their advice, and to the media. We will consider arranging a specific briefing for such subject communities ahead of announcing our decision. We will also want to make sure that DfE understand our decision in good time.

### **Internal Stakeholders**

49. SRR and GQ directorates.

### **External Stakeholders**

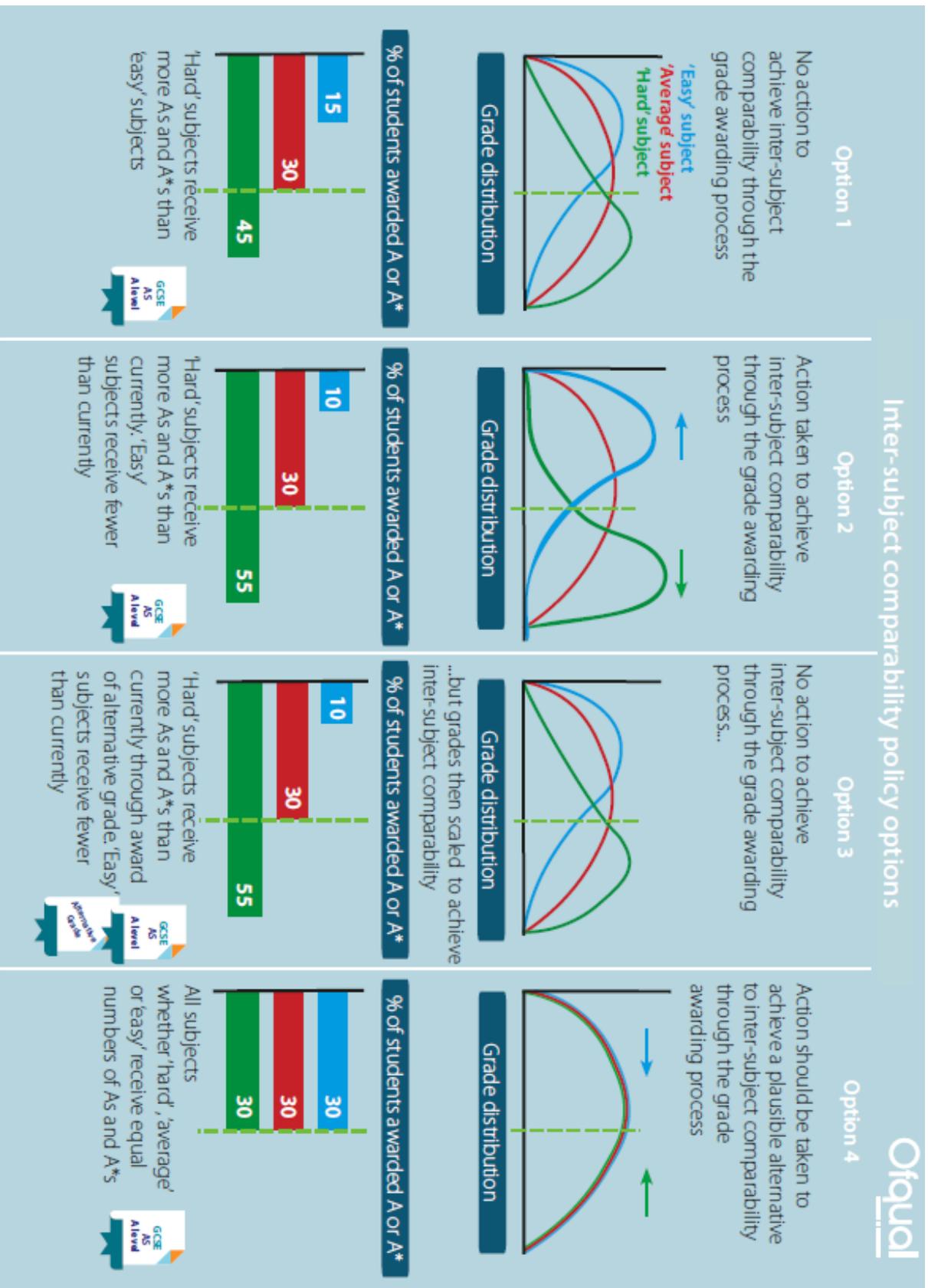
50. Subject communities, particularly groups of linguists and scientists; exam boards; DfE; UCAS and HEIs; headteacher and teacher associations.

Paper to be published	Yes
-----------------------	-----

## **Annexes list**

- Annex A** Policy options infographic
- Annex B** Summary survey report
- Annex C** 4 February conference programme
- Annex D** Subject entry choices and perceptions of subject difficulty: are the two linked, and if so, how? An executive summary of the draft Ofqual research report
- Annex E** A summary of recent research on inter-subject comparability
- Annex F** Letter to Ofqual from science organisations
- Annex G** Progression from GCSE to A level

Policy options infographic

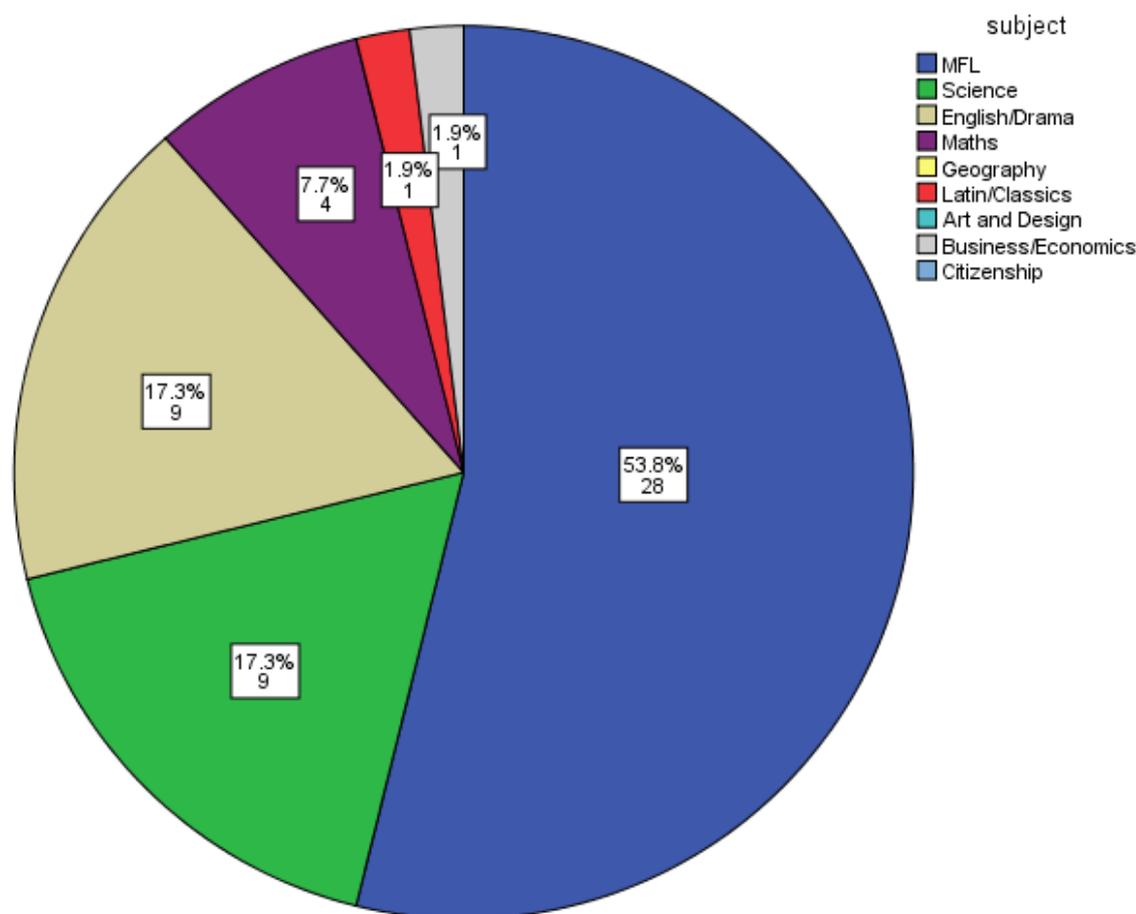


## Summary survey report

### Respondents

There were 216 respondents to the survey. Not all respondents provided details of their position. However, almost 50% of all respondents worked in an education provider, with a fairly even split between teacher, Head of Department/Curriculum Lead and Senior Management (for example, Deputy Head/Head). Other respondents worked in an educational context as a consultant or examiner (5%) or for an educational stakeholder interest group such as a subject interest group or union (4%). A very small number of respondents were from assessment agencies and media organisations.

Some respondents indicated a subject background. These are displayed below. The most common subject represented was modern foreign languages (MFL).

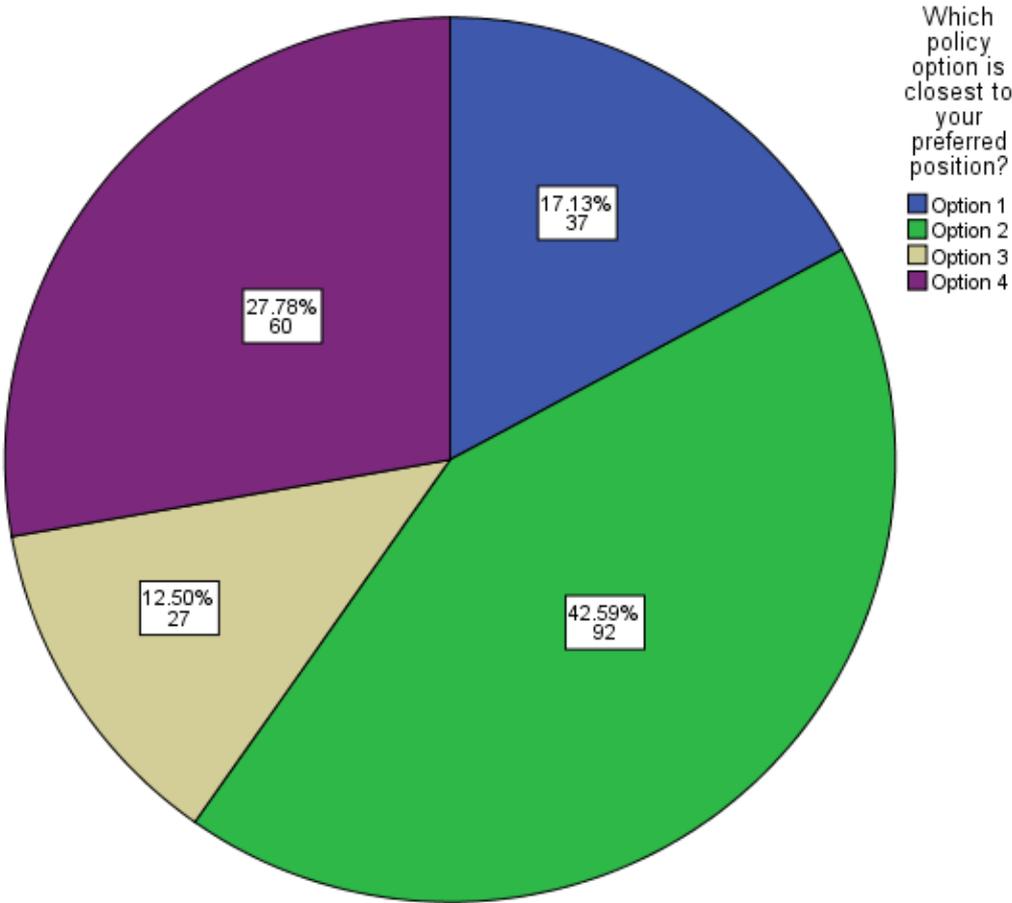


The single most common position represented amongst respondents was “Head of Department for MFL” (n=20). The next most common position was “Teacher of MFL” (n=7) and “Head of Department for English/Drama” (n=7).

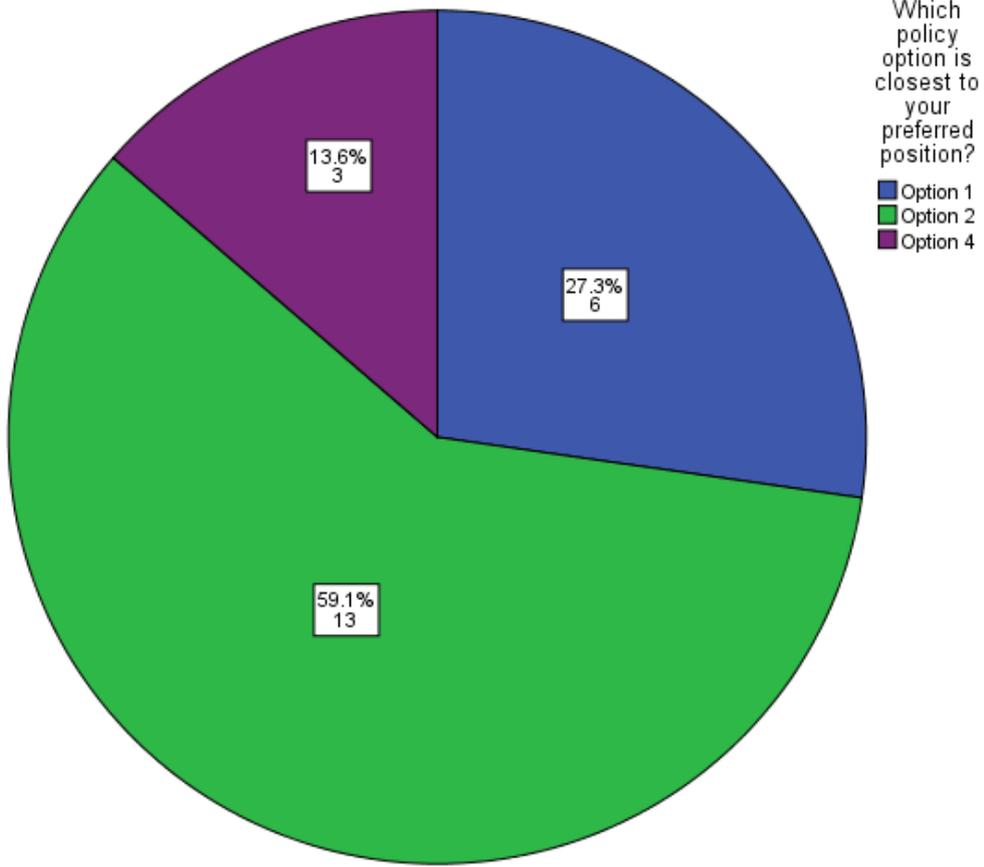
Of the 216 respondents, 135 indicated they were expressing their personal views and 22 that they were expressing those of an organisation or group and 59 did not indicate either.

**Policy Options – Preferred Positions**

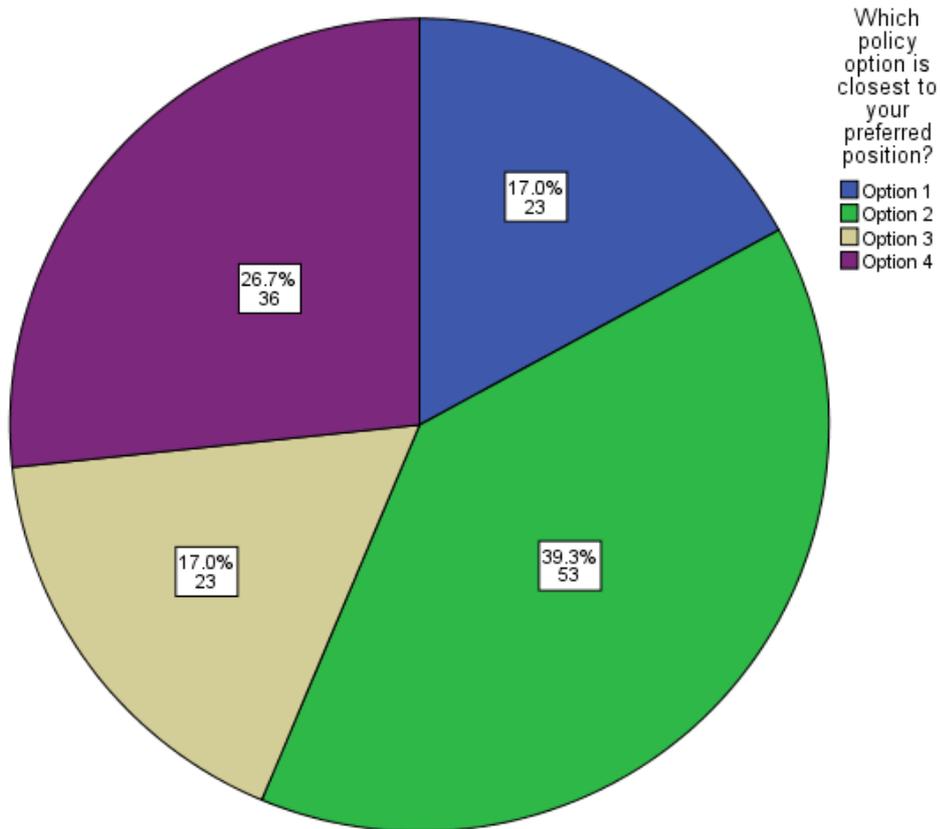
The most common preferred policy option was Option 2 – to achieve inter-subject comparability through the grade awarding process.



*Preferences by respondent type*



'official organisational opinion' respondents



'personal opinion'

*Preferences by subject background*

		subject						Total
		MFL	Science	English/ Drama	Maths	Latin/ Classics	Business/ Economics	
Which policy option is closest to your preferred position?	Option 1	1	0	4	2	0	0	7
	Option 2	22	4	1	2	1	0	30
	Option 3	2	4	0	0	0	0	6
	Option 4	3	1	4	0	0	1	9
Total		28	9	9	4	1	1	52

## Reasons for preferred policy option

Respondents were also asked to supply reasons for their policy options.

These were coded as follows:

1. Congruent with their preferred option
2. Not possible to know whether it is congruent or incongruent
3. Incongruent with their preferred option.
4. Nuancing: Slight modification of preferred option suggested

		reason				Total
		congruent	unclear	incongruent	modification	
Which policy option is closest to your preferred position?	Option 1	30	2	1	4	37
	Option 2	71	9	0	11	91
	Option 3	20	1	2	4	27
	Option 4	18	34	4	3	59
Total		139	46	7	22	214

This indicates that option 4 selectors were less clear on the implications of this option or what it entailed.

Reasons provided for option 1 were noticeably longer than for other options. The mean word count was 80 words compared to 23, 30 and 31 for Options 2, 3 and 4 respectively.

Reasons provided for option 2 usually talked about 'fairness', recognition for hard subjects, and frequently in the context only of MFL or just addressing 'the MFL problem'. In other words, fewer than the 91 respondents probably embrace option 2 in the 'total solution' way for all subjects.

Reasons for option 3 usually talked about it being the least disruptive option or more manageable.

Reasons for option 4 often seemed to allude to 'fairness' between subjects in a way which was unclear, to criterion referencing and to 'equally difficult' to get an A\* in each subject.



**Thursday 4 February 2016**  
**Broadway House**  
**Westminster**

- 10.00      Arrival and registration
- 10.30      **Opening remarks** - Amanda Spielman, Chair, Ofqual
- 10.35      **Perspectives from Ofqual**  
Paul Newton - The very idea of inter-subject comparability  
Dennis Opposs - Why inter-subject comparability matters
- 11.15      **Perspectives from subject communities**  
MFL - Nick Mair, Dulwich College  
Physics - Charles Tracy, Institute of Physics  
English - Jenny Stevens
- 12.15      Lunch
- 1.00        **Perspectives from the education community**  
Geraldine Davies, UCL Academy  
Alison Matthews, Oxford University
- 1.45        **Perspectives from academia**  
Jason Lamprianou, University of Cyprus  
Mike Cresswell  
Robert Coe, Durham University
- 2.45        **Introduction to discussion session** – the policy options
- 2.50        Group discussion
- 3.30        Tea
- 3.50        **Plenary session with panel**  
Suzanne O'Farrell, ASCL  
Jo-Anne Baird, Oxford University  
Gareth Pierce, WJEC
- 4.20        **Summary and next steps**  
Glenys Stacey
- 4.30        Close

### **Subject entry choices and perceptions of subject difficulty: are the two linked, and if so, how? An executive summary of the draft Ofqual research report**

Concerns have been raised about a disparity in the difficulty of different subjects, and that this might be contributing to a lower uptake in certain 'key' subject areas. The purpose of this research was to explore whether teachers' and students' perceptions of subject difficulty might be having an effect on which subjects students choose to study in secondary education, and whether other concerns (eg subject enjoyment or usefulness) might interact with this relationship.

A qualitative research design was chosen to allow for an in-depth exploration of these issues. Interviews and focus groups were held with 49 teachers and 112 students respectively from twelve schools across England. Thematic analysis was performed on coded transcripts and the main drivers of students' and teachers' behaviours were identified.

Teachers agreed that whether a student found a certain subject difficult or not was very much dependent upon each student's individual strengths. Teachers had an influence over students' subject choices both via the setting of policies and by giving advice. Entry criteria policies were often based upon notions of subject difficulty, which served to prevent students from taking subjects they would find too difficult. Some schools also chose not to offer certain subjects because they were seen to be too difficult, again preventing uptake in those areas. Teachers sometimes discouraged students from taking subjects that might be too difficult for them, but stated that this was mostly (although not exclusively) done according to each student's individual strengths, as opposed to any general (ie not person-specific) notions of subject difficulty. However, although subject difficulty was an important consideration for teachers, much of their advice was based upon what each student would enjoy and find useful for future education or employment.

Students also agreed that although some subjects stood out as being more difficult than others, whether or not they found a subject difficult was dependent upon their individual strengths. Students did base their subject choices on perceptions of difficulty, and recognised that they were also sometimes discouraged by their teachers, parents, and friends from choosing subjects that were thought to be too difficult for them. However, as with the teachers, students stated that perceptions of difficulty were not the main basis of their decisions, and focussed more upon enjoyment and usefulness. Importantly, students often stated that they were willing to overlook subject difficulty when they enjoyed it and/or needed it to satisfy their ambitions.

The main conclusion drawn from this research was that there does seem to be a link between perceptions of subject difficulty and subject entry choices, although perceptions of enjoyment and usefulness appear to interact with, and often supersede, this relationship.

### A summary of recent research on inter-subject comparability

1. Tom Bramley of Cambridge Assessment had previously argued<sup>7</sup> that there is a real problem in interpreting the data from all methods of measuring inter-subject comparability as the logic of the approach is based on the unrealisable notion of all students taking all subjects. The same author's recent paper<sup>8</sup> illustrates through simulated A level data that subjects such as art that measure different qualities from the majority of subjects will inevitably appear 'easier' using statistical measures. The paper demonstrates that informed subject choice where there are unequal correlations amongst subjects can create spurious differences in subject difficulty. Support for that finding has come from similar work we have carried out in house using real GCSE candidate data for England.
  
2. At the inter-subject comparability conference on 4 February, Mike Cresswell argued in his presentation, based on his new statistical analysis of simulated A level data, that:
  - the existence of selection processes which use subject grades interchangeably is not a good reason for trying to put subject comparability "right", and that anyway
  - making the sort of statistical alignment of subjects described in Working Paper 3 would do little to improve the quality of selection processes which assume that grades from different subjects are interchangeable. That's because the alignments relate to groups of candidates on average rather than to individuals.
  
3. AQA has since built on that study using a national GCSE results database for England (in preparation for publication). Analyses were undertaken, each based on the original and adjusted grades. At student level, correlations between students' original and adjusted grades were derived. At subject level, tables for a range of subject pairs using both the original and adjusted grades were produced. At school level, a number of average accountability measures were calculated to see how aligning subject grades would affect school rankings.
  
4. At student and school levels, the adjustments in grade boundaries had minimal or no effect on the measures, indicating that the complex task of aligning subjects statistically may not be worthwhile. Moreover, partly due to incomplete data, but mainly due to intentional experimental design, the models and assumptions underpinning the school and student level analyses were

---

<sup>7</sup> Bramley, T. (2014) *Multivariate representations of subject difficulty*. Research Matters (a Cambridge Assessment publication) 18, 42-47

<sup>8</sup> 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.

themselves adjusted. These adjustments were designed to monitor the robustness of the model, and that of the results it produced. In every case, and regardless of how the parameters were adjusted, the effect on the measures was minimal.

5. At subject level, the effect on the paired relationships pre- and post-adjustments varied according to the particular subject pairing. In most cases the effect was small. Sometimes though there was a shift away from the original relationship in one direction or the other (that is, the balance between relative subject leniency/severity shifted). At other times, although overall it remained stable, the original relationship became weaker, with fewer students achieving the same grade in both subjects.
6. This suggests that it may not be possible to make subject grades interchangeable, and may be detrimental in some cases to try to do so.
7. On a similar theme, another new paper<sup>9</sup> from Tom Benton of Cambridge Assessment argues that adjusting aggregated measures of either student or school performance to account for the relative difficulty of GCSE subjects makes little difference. “For either students or schools, the correlation between unadjusted and adjusted measures of performance exceeds 0.998. This indicates that suggested variations in the difficulty of different GCSE subjects do not cause any serious problems either for school accountability or for summarising the achievement of students at GCSE.” The paper is based on the assumption that entries in different subjects would not change and focuses on aggregated measures so isn’t saying that making the grading of GCSE German more lenient might not increase entries in that subject.

---

<sup>9</sup> 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.

## Letter to Ofqual from science organisations



Amanda Spielman  
Spring Place  
Herald Avenue  
Coventry  
CV5 6UB

13 April 2016

Dear Ms Spielman,

We are writing to express our concerns about the continuing problem of variable grading severity of subjects at A-level, the consequent adverse effects on student choice, and the need to take action to address it. We understand that Ofqual is currently considering its position on action and we request that our concerns are put to the Ofqual board when it meets in May.

We are pleased that Ofqual is revisiting the problem and agree with much of the summary contained within the working papers<sup>1</sup>. We also welcome the initial engagement with the community through the recent on-line survey, though note that a more appropriate alternative to option of 'no action' would have been 'some action' -- both for balance and to allow respondents to consider actions other than the three provided. We believe that the problem must be addressed and that there should be a longer, more in-depth consultation to investigate ways of correcting relative grading difficulty and alleviating its effects. We would be glad to contribute, in detail, to such a process.

We disagree with the suggestion in the working papers that the differences in outcomes are the result of a range of factors other than grading severity. The consistency of the grading data suggests that it is far more likely that they result from the same, uniform, influence: severity of grading. Please see the Annex for more detail on our reasoning.

Much work has been undertaken by the community to increase participation in STEM subjects at A-level, and while this has had some positive effect, the underlying problems of grading and perceived difficulty remain. We are concerned that the current inequity in grading is narrowing students' options at A-level, reinforcing gender bias and limiting opportunities for students from lower socio-economic backgrounds. This effect can be summarised through two main mechanisms:

- **Student choice:** A large consideration in student subject choice is what their likely grade is going to be. Currently, there are markedly different likelihoods of getting a given grade in different subjects for the same prior performance, interest, aptitude, teaching quality and application. The perception of difficulty created by this has been seen to have a significant effect on girls' participation in the 'more difficult' subjects particularly.
- **School influence:** There is strong evidence that students are either being debarred from taking those subjects or are self-selecting themselves out of them based on their reputation for difficulty – an effect particularly prevalent in schools in more deprived areas of the country. As a result, there has been a reduction in the range of ability within the entrants to the more severely-graded subjects and, separately, the creation of further barriers to the participation of students from lower socio-economic backgrounds.

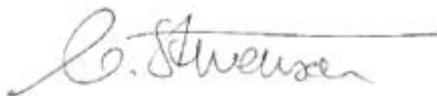
<sup>1</sup> <https://www.gov.uk/government/collections/inter-subject-comparability-research-documents>

It should be noted that the symptoms and consequences are becoming more extreme with time. A decision for no action now will result in the effects becoming more engrained. If action is postponed, not only will more year-groups miss important opportunities, but the problem will be harder to address in the future.

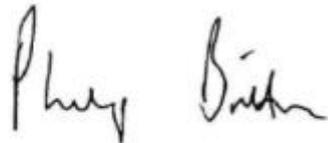
We strongly recommend that the board allocates more time to fully investigate the effects of the problem and find a workable solution.

We would be happy to provide further information or to discuss these issues further with you.

Yours sincerely,



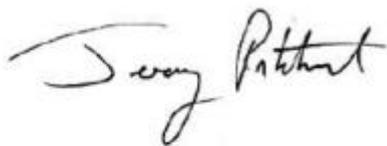
**Corinne Stevenson**  
Chair  
Association for Science Education



**Philip Britton FInstP**  
Vice President (Education)  
Institute of Physics



**Professor Tom McLeish FRS**  
Chair, Education Committee  
Royal Society



**Dr Jeremy Pritchard**  
Chair, Education Training and Policy Committee  
Royal Society of Biology



**Professor Gareth Price FRSC**  
President, Education Division  
Royal Society of Chemistry

## Annex

### An overview of the data of grading difficulty

The available data suggest that there are clear differences in the ways that certain subjects are graded. In particular, science subjects and maths appear to be more severely graded than most other subjects.

The CEM Centre in Durham produced a report in 2008 which looked at relative difficulty in exams in different subjects.<sup>2</sup> Figure 1 below, taken from this report, shows the differences in grade expectation for different subjects allowing for prior attainment.

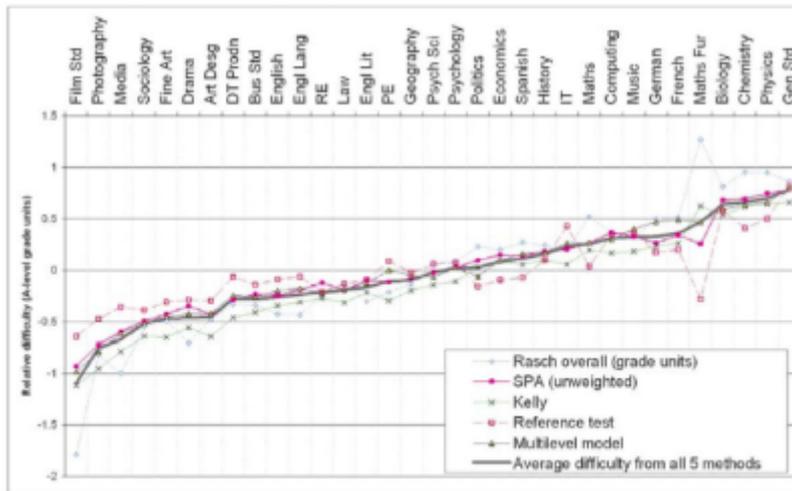


Figure 1

Figure 1 suggests that the sciences and maths are graded more severely than most other subjects. The difference in expected outcome is up to a grade compared with other facilitating subjects and by up to a grade and a half overall. It is possible that other effects could be responsible for this distribution – for example the quality of teaching or the amount of engagements/application of students – however, this is less likely if one considers the distribution in output grades for typical B grade students embarking on A-levels.

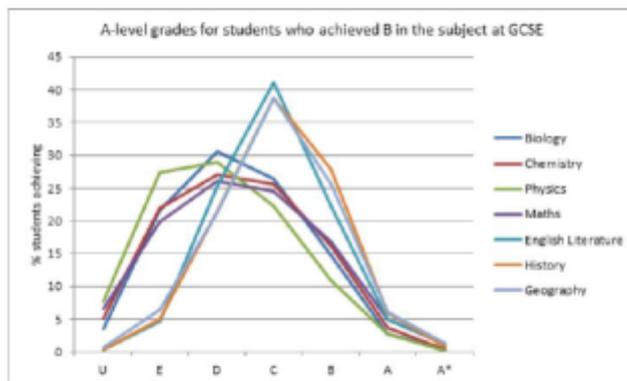


Figure 2

<sup>2</sup> <http://www.score-education.org/media/3194/relativedifficulty.pdf>

Figure 2 shows that there are two distinct distributions in students who received a grade B at GCSE: the sciences and maths follow one general pattern and the other subjects follow another. This suggests that the subjects are systematically graded differently. Considering other explanations, it seems unlikely that there would be such a uniform drop in motivation, quality of teaching (and all other factors suggested) for the sciences and, at the same time, such a uniform retention of all those factors for other subjects.

Evidence exists to suggest that schools apply some kind of filtering and that this filtering is different for the more severely graded subjects.

Figure 3 below shows that there are 720 schools in England with no entries at physics A-level among students with a B grade or less at GCSE. In contrast, there are only 125 schools with the same pattern in history. It seems unlikely that this disparity could have arisen by chance.

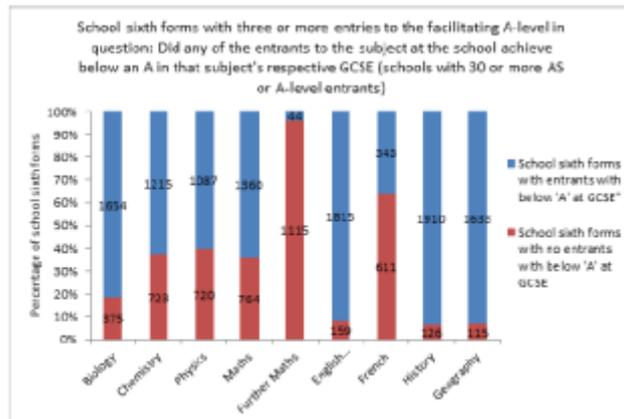


Figure 3

The above data could be explained in a number of ways: that schools have different entry policies for the sciences and maths; that schools are directing students with lower grades away from the sciences and maths (possibly through the use of prediction tools); or that students are selecting themselves out of what are more severely graded subjects. This may, in part, be due to the school's desire to increase their proportion of the highest grades (which, as shown in Figure 2, are more likely in the less severely graded subjects).

In view of the above data, it appears that, in effect, differences in grading severity are resulting in students being denied access to more severely graded subjects because of pressures on schools through performance measures.

## **Progression from GCSE to A level**

Comparative Progression Analysis as a new approach to investigating inter-subject comparability

[This paper has been published on GOV.UK.](#)