Progression from GCSE to A level

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

March 2017

Ofqual/17/6204

Acknowledgements

This report was produced by Paul E. Newton, Qingping He and Beth Black from Ofqual's Strategy, Risk and Research Directorate. Very helpful feedback on an earlier draft was provided by members of Ofqual's Standards Advisory Group, Directorate colleagues, and Tom Benton.

Contents

Background

In a letter dated 13 April 2016, leaders of a number of high profile science organisations[1] wrote to Ofqual's Chair, asking that Ofqual consider its future policy on inter-subject comparability. This was in light of their concern that A level examination standards are not aligned across subject areas and that this is having adverse effects on candidate choice, particularly for the sciences. Their letter was written in response to a wide-ranging inquiry into inter-subject comparability which had been launched by Ofqual in December 2015.[2] They argued as follows:

> 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.

The annex contained three figures, the second of which presented a kind of analysis that has not traditionally figured in debates concerning inter-subject comparability in England: a comparative analysis of progression from GCSE to A level across a range of subjects. Let's call this: Comparative Progression Analysis (CPA). The purpose of the present paper is to explain this analysis, to present in detail the kind of results which it produces, and to prompt debate on the significance of those results for conclusions concerning inter-subject comparability at A level.

CPA concerns the 'progress' made by individual students from their GCSE grade to their A level grade in the same subject area. It considers the distribution of A level grades in a subject, for students who were awarded a particular grade in the same subject at GCSE, to determine whether progression patterns are the same, or different, across subject areas. To interpret outputs from these analyses at face value – in terms of the alignment, or misalignment, of grading standards – it would seem to be necessary to assume: that (groups of) candidates ought (on average) to make the same progress across subject areas, all other things being equal; and that all other things are in fact (more or less) equal.[3]

---

[1] 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.
[2] See https://www.gov.uk/government/collections/inter-subject-comparability-research-documents
[3] In fact, in the final sections of this report, we will question whether these 'reasonable' assumptions are necessary after all; and we will consider, more generally, how best to rationalise the CPA approach.

Contextualising CPA

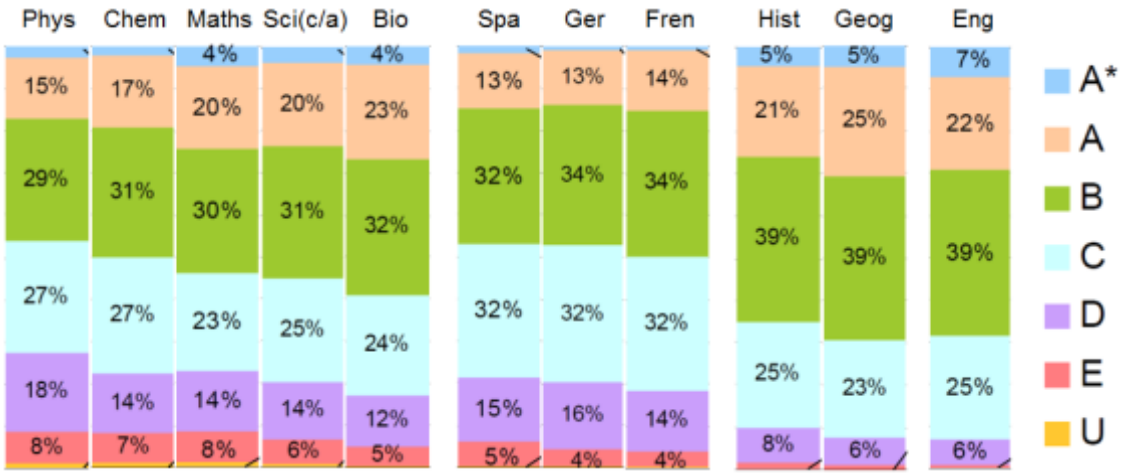
The idea of CPA is not new. For instance, Cambridge Assessment analysed data from the 2005 16+/18+ database in this manner, presenting tables which illustrated "how a candidate could maximise their A-level grades when all other things are equal" (Bell and Emery, 2007, p.3). However, they did not use the tables to draw inferences concerning (a lack of) inter-subject comparability. Quite the reverse, in fact. They challenged the contemporary critique of inter-subject comparability from the Centre for Evaluation and Monitoring in Durham (see Coe, 2008) and they suggested possible explanations for differences in progression patterns across subjects, e.g. differences in motivation between subjects, differences in resources and quality of teaching, differences in uptake for various population subgroups.

The Department for Education analysed results from the National Pupil Database for the cohort who completed Key Stage 4 in 2008, investigating GCSE to A level progression rates for English Baccalaureate subject areas (ESARD, 2012). Although the primary focus of their report was upon progression to A level, per se, it also noted different patterns of results at A level, and even suggested that certain subjects could be seen as more difficult than others:

> It can be seen that over 50% of pupils with a grade A in GCSE physics that go on to A level physics achieve a grade C or lower and, as such, physics could be seen as being more difficult at A level than the other sciences. (ESARD, 2012, p.18)

Figure 1 is a crude cut-and-paste of columns extracted from eleven charts from this report (3.1 to 3.11). Each column represents the distribution of A level grades achieved by GCSE grade A candidates, for a particular subject.

Figure 1. Distribution of A level grades for candidates with GCSE grade A



It illustrates the suggestion, from the above quotation, that physics might be the most difficult science. Following the same logic, it illustrates how history, geography and English all seem to be less difficult than biology, which seems to be the easiest

science at A level. The languages also seem to be amongst the most difficult subjects. Although mentioning the possibility of differential subject difficulty (just once) the report did not develop this suggestion, nor link it to the wider literature on inter-subject comparability.

Cambridge Assessment updated and expanded the ESARD (2012) analysis, using similar methods applied to results from students who were in Year 11 in 2010 (Sutch, 2013). Again, though, it drew no inferences concerning inter-subject comparability, citing the same kind of 'possible explanations' as Bell and Emery (2013).

In summary, although results from similar analyses have previously been published, CPA has never been proposed or evaluated as an approach to investigating inter-subject comparability.

Inter-subject comparability

It seems a little odd that this kind of analysis has not featured previously within inter-subject comparability debates. Within the literature, the most similar approach involves the use of what have previously been called 'value-added' analyses. From this perspective, CPA compares within-subject-area value-added analyses across subjects. Value-added can be calculated in different ways, though; most notably from a baseline of mean GCSE score rather than subject GCSE grade.

Mean GCSE value-added analyses do appear in the literature. In fact, they were a key component of the critique of A level grading standards produced for the School Curriculum and Assessment Authority by Fitz-Gibbon and Vincent (1994). The same analyses were replicated for the Dearing review (Dearing, 1996) and their significance was widely debated at the time (e.g. Goldstein and Cresswell, 1996; Fitz-Gibbon and Vincent, 1997; Newton, 1997).

Even prior to the publication of Fitz-Gibbon and Vincent (1994), statistical techniques for investigating inter-subject comparability had been criticised heavily for a number of reasons. Empirically, they often produced contradictory results when exactly the same analyses were calculated separately for subgroups of the population. For example, a subject which appeared to be harshly graded overall might appear to be very harshly graded when the analysis was re-run only for male students yet not at all harshly graded when re-run only for female students. Theoretically, statistical analyses tend to require the strong assumption that they are tapping into something like 'general academic aptitude' and that the influence of other causal factors on grading standards can safely be ignored.

These two kinds of criticism (empirical and theoretical) are related, as differential subgroup effects can most straightforwardly be explained by the differential influence of other causal factors across subgroups. Imagine, for example, that male students who coasted through science subjects at GCSE realised that they really needed to knuckle down to be successful when they got to A level, and so put in substantially

more effort. Now imagine that this was not true for male students in other subject areas (who continued to coast), nor true for female students in any subject area (who applied similar levels of effort at GCSE and A level). This purely hypothetical example could result in different patterns of value-added across subjects, which might *appear* to suggest differences in grading standards between the sciences and other A level subjects. However, if the same analyses were to be re-run separately for male and female students, we might find no differences in value-added patterns between subject areas for female students, alongside far more extreme differences between subject areas for male students. Since male and female students would all have taken the same A level examinations, the explanation for these gender differences could not be attributed to A level grading standards. Instead, the explanation would lie in different patterns of effort between subsets of candidates, which could not (and should not) be accommodated by 'realigning' standards across subjects.

Conclusions concerning a possible lack of inter-subject comparability at A level, which were drawn from the Fitz-Gibbon and Vincent (1994) mean GCSE value-added analysis, were vulnerable to criticism both empirically and theoretically. Indeed, their report indicated that males achieved higher value-added scores than females across arts, mixed, and science subjects, and that the differential effects were most pronounced for science subjects (Figure 4(iii) on p.29).

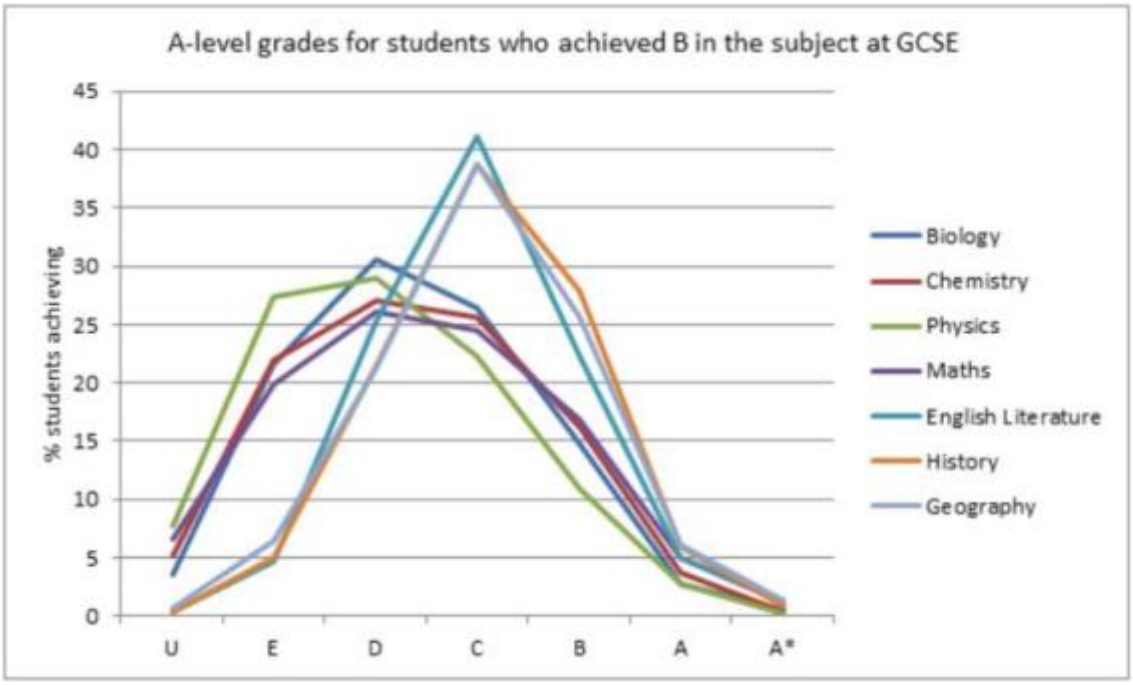
Debate over the interpretability of value-added analyses – and over inter-subject comparability more generally – petered out by the end of the twentieth century. When it was revitalised during the mid-2000s, it was framed principally in terms of results from a new statistical approach, based upon the Rasch method (e.g. Coe, 2008). Indeed, results from Robert Coe's Rasch analyses were reproduced in the aforementioned letter from the science organisations as their first figure.

The prima facie challenge

So let us return to this letter, which argued as follows:

> Figure 1 [Rasch] 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 candidates – however, this is less likely if one considers the distribution in output grades for typical B grade candidates embarking on A-levels.
>
> Figure 2 [comparative progression] shows that there are two distinct distributions in candidates 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.



The science organisations have a good point here. In defence of the 'other causal factors' criticism, it is quite likely that they do operate differentially within at least some A level subjects, making it hard to take results from Rasch analyses at face value. For example, it is quite likely that candidates tend to put less time and effort into general studies than into the other A levels which they take; and, if so, then perhaps general studies is not as harshly graded as Rasch analyses seem to suggest. Yet, the distinctness of the subject grouping effect which is evident from the above figure, having controlled for baseline prior attainment at GCSE, does seem to be intuitively persuasive.[4]

This subject grouping effect raises the question of why cohorts of A level students within different subject areas, who start their respective courses from a baseline of 'equivalent' proficiency in those subjects, might end up with substantially different outcomes at the end of their courses. With reference to the above figure, why do GCSE grade B candidates typically end up with around grade D in the sciences at A level while GCSE grade B candidates typically end up with around grade C in the humanities at A level? Perhaps passions are ignited in certain subjects at A level (eg

[4] These analyses control for baseline prior attainment across A level subjects by restricting comparisons to only those candidates who achieved a particular grade at GCSE (e.g. grade B). This assumes that GCSE grades are somehow comparable across subjects. Even if this were not entirely true, it would still not be an unreasonable simplification; particularly as the effects might be more pronounced if results from Rasch analyses at GCSE were to be taken at face value.

English literature) but extinguished in others (eg chemistry)? Perhaps A level teachers are genuinely better in certain subjects (eg history) than in others (eg physics)? Perhaps candidates in certain subjects (eg fine art) are more likely to come from highly effective schools than candidates in others (eg maths)? But how *likely* are these explanations, particularly when the effects in question appear to generalise so consistently across groups of similar subjects? In short, there does seem to be a prima facie challenge for Ofqual to respond to, arising from outputs from CPA.

Analyses

In order to begin responding to this challenge, Ofqual ran a new set of comparative progression analyses, using data from the National Pupil Database. The data related to candidates who were awarded an A level subject grade in 2013. For each A level subject included in the set, all students for whom a prior GCSE grade (in the same subject) was available were included in the analysis. Naturally, this meant that a very large proportion of each A level cohort was included in the analysis.

The samples

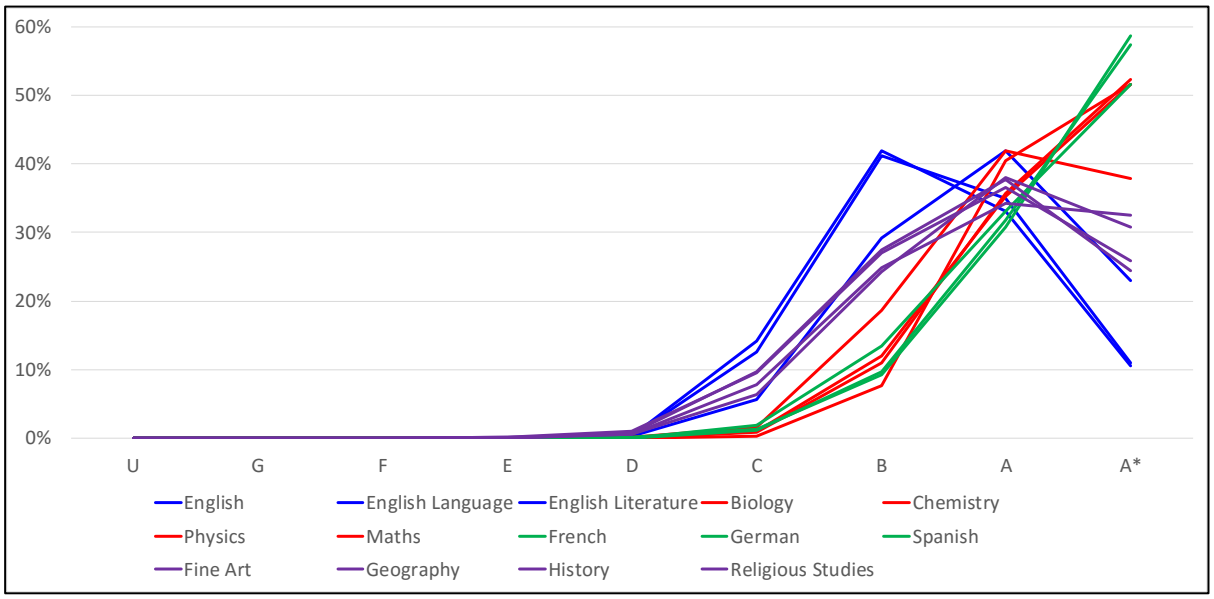
The numbers of candidates involved in the analyses are presented in Table 1.

Table 1. Numbers of candidates involved in the analyses (by prior GCSE grade).

	U	G	F	E	D	C	B	A	A*	TOT
English	0	0	1	2	32	2093	6217	4903	1565	14813
English Language	0	1	0	3	60	2988	9801	8346	2609	23808
English Literature	13	2	2	14	97	2432	12569	18017	9864	43010
Biology	0	1	0	3	21	585	7194	16255	14650	38709
Chemistry	0	1	0	4	7	297	3780	12327	18056	34472
Physics	0	1	0	0	4	214	2745	8053	11777	22794
Maths	4	0	0	5	11	218	5637	29843	38053	73771
French	0	0	1	0	3	105	886	2930	5295	9220
German	0	0	0	0	1	60	443	1091	1698	3293
Spanish	0	0	0	0	2	73	536	1794	3422	5827
Fine Art	0	1	2	2	31	485	1538	2120	2013	6192
Geography	0	1	1	5	202	2639	7491	10310	6673	27322
History	1	2	3	29	381	3978	11302	15274	10826	41796
Religious Studies	0	0	1	15	73	886	3432	5374	4351	14132

It is obvious from Table 1 that very few candidates with less than grade C at GCSE, in their chosen A level subjects, were entered for A level examinations. It seems likely that this is due primarily to school and college curriculum policy decisions, whereby candidates who do not achieve at least grade C in a particular subject at GCSE are not permitted to undertake an A level course in that subject (other than in exceptional circumstances). The highlighting in Table 1 indicates the modal GCSE grade across candidates within each A level subject. When these figures are converted to percentages, cross-subject patterns become more evident, as can be seen from Figure 2.

Figure 2. Distributions of prior GCSE grades across A level subjects (percentages).



Colours have been added to Figure 2 to distinguish four groups of similar subjects: the English (blue); the sciences (red); the languages (green); and the humanities (mauve). Whereas schools and colleges appear to operate a 'minimum C' policy for English and the humanities, they seem to operate a 'minimum B' policy for the sciences and languages.

Another way of reflecting upon the data in this graph is in terms of mean percentages of candidates achieving A or A* at GCSE across subject groups: 52% for the English; 87% for the sciences; 88% for the languages; and 65% for the humanities. In other words, nearly nine out of every ten candidates who sat A level examinations in the sciences and languages had previously achieved either an A or an A* at GCSE. This was true for only two thirds of humanities candidates and one half of English candidates.

The same data are presented slightly differently in Figure 3, in terms of numbers rather than percentages. This helps to illustrate that there are far fewer awards in the languages, in particular. Maths is the clear outlier in the sciences group, with large numbers of candidates who were mainly awarded either A or A* at GCSE.

Figure 4 represents A level outcome data for exactly the same groups of candidates. With the exception of English and English language, the A level grade distributions are reasonably similar across subjects. The languages and sciences tend to award slightly more of the highest grades; but notice that the sciences also tend to award slightly more of the lowest grades.

Figure 3. Distributions of prior GCSE grades across A level subjects (numbers).

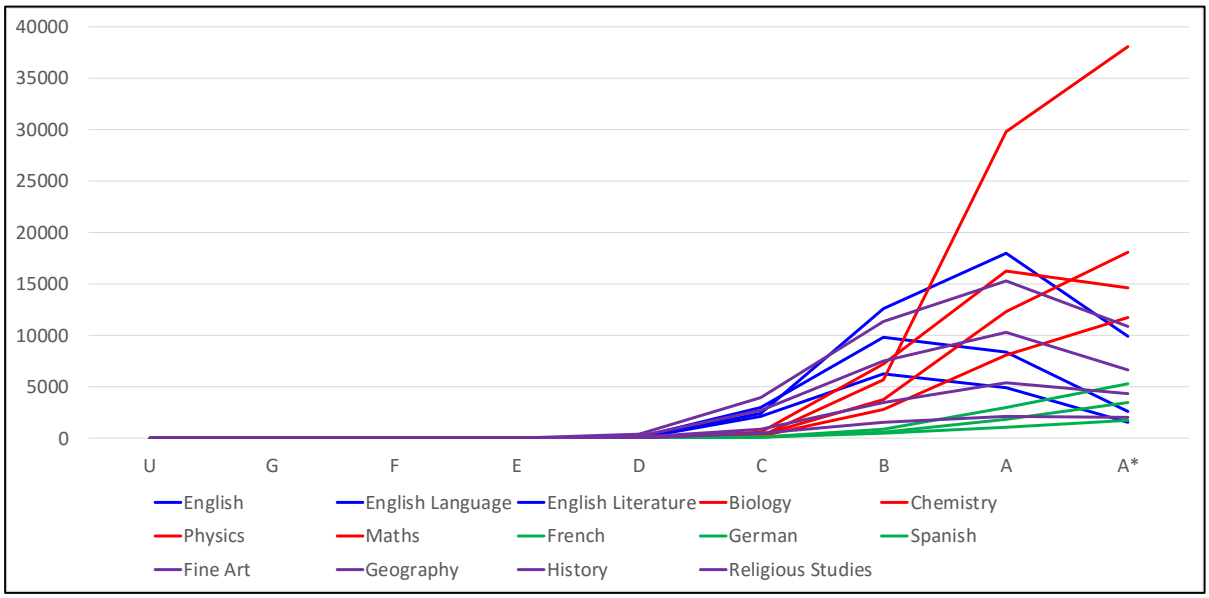
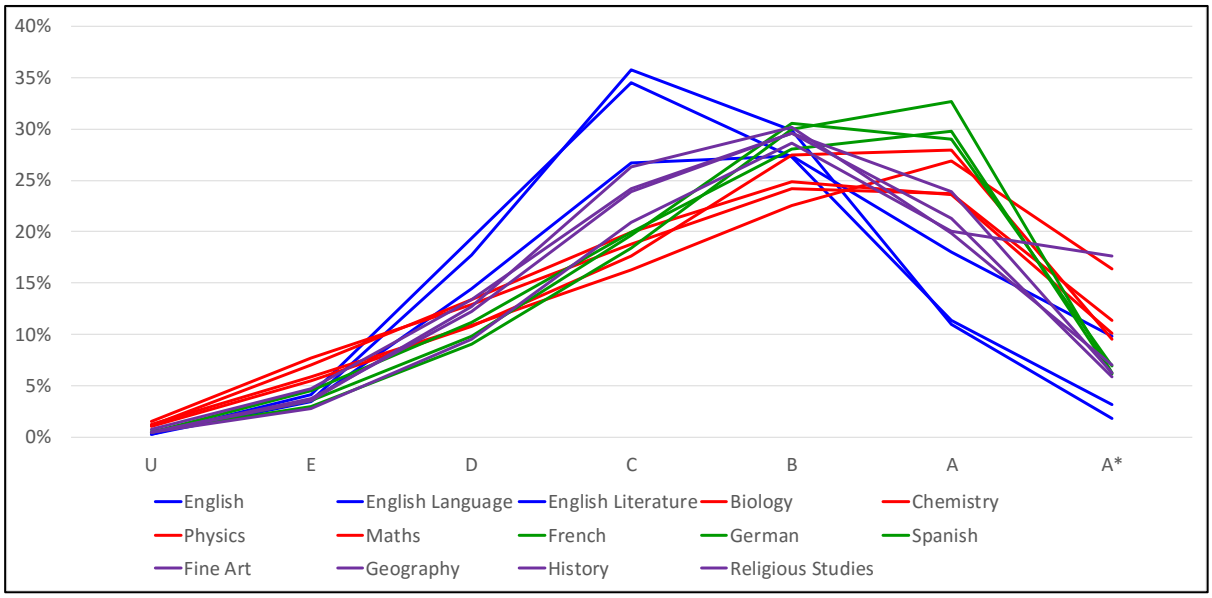


Figure 4. Distributions of A level grades across subjects (percentages).



To explore the data further, correlations were computed to identify relationships between A level subject grades and other variables. These appear in Table 2. The within-subject GCSE to A level correlations were all very similar, ranging from 0.54 to 0.64. There were no obvious trends across subjects. A similar story was true for correlation with mean GCSE grade, ranging from 0.53 to 0.73; as well as for correlation with mean A level grade, ranging from 0.82 to 0.92.

Table 2. Correlations between A level subject grades and other variables.

	Subject GCSE grade	Mean GCSE grade	Mean A level grade
English	0.57	0.66	0.83
English Language	0.57	0.66	0.84
English Literature	0.57	0.73	0.87
Biology	0.60	0.69	0.89
Chemistry	0.57	0.64	0.91
Physics	0.59	0.66	0.92
Maths	0.56	0.59	0.88
French	0.59	0.62	0.84
German	0.64	0.53	0.82
Spanish	0.56	0.53	0.82
Fine Art	0.61	0.59	0.83
Geography	0.62	0.70	0.87
History	0.59	0.69	0.88
Religious Studies	0.54	0.65	0.87

There are a few notable features within Table 2. For instance, the subject GCSE correlation for German is the highest of all (0.64), whereas the mean GCSE correlation for German is the lowest of all (0.53). However, whether much can be inferred from features such as these is unclear; particularly given differences in distributions of prior GCSE grades (leading, for example, to differential range restriction). Bearing in mind that prior GCSE grades were heavily skewed towards the highest grades, particularly for the sciences and languages, it is perhaps surprising that within-subject correlations were as high as those shown in Table 2. On the other hand, even figures this high suggest that only around a third of the variance is 'explained'.

An interesting question arises concerning progress from GCSE to A level across subject areas: what proportion of each GCSE examination cohort progresses to A level in the same subject area? This was not a straightforward question to answer from the data which were analysed for the present report, but 'ballpark' indications can be provided for certain subjects if certain assumptions are made. First, let's assume that the candidates in our samples, for each subject area respectively, comprise more-or-less all of the students who progressed from GCSE in 2011 to A level in 2013. This is an oversimplification, as it was not possible to match data for all candidates, and because some of the A level candidates in our sample would have been awarded GCSE in different years. (The picture is complicated further by the modular structure of these qualifications and their availability across different sessions.) Second, let's assume that the Joint Council for Qualifications (JCQ) statistical releases provide GCSE cohort level data, at subject level, which can be compared fairly (if not absolutely precisely) with A level sample data from the present analysis. Because of a variety of complicating factors (e.g. the split between separate

and combined science at GCSE, the significance of grade C in English and maths and implications for resits and suchlike) this assumption was only made for six subjects from our A level sample, all from the languages and humanities.

Table 3. Subject cohort data (from JCQ statistical releases)

	No. sat GCSE in 2011 (all)	No. sat A level in 2013 (all)	2013 A level as % of 2011 GCSE (all)
French	141472	10249	7.2
Spanish	60773	6923	11.4
German	58382	3999	6.8
History	198316	46927	23.7
Geography	163604	29126	17.8
Religious Studies	199752	19173	9.6

Table 3 is based purely upon published JCQ data, from 2011 and 2013, respectively.[5] The implication from this table is that proportionally more students progress from GCSE to the same subject at A level in history and geography than in the languages or religious studies. For example, nearly one quarter of the GCSE history cohort progresses to an A level in history; whereas not even one tenth of the GCSE French cohort progresses to an A level in French.

Table 4 presents candidate outcomes at GCSE in 2011, for each of the four highest grades, from the full JCQ (GCSE cohort) dataset. Alongside these data, it presents candidate outcomes at GCSE, for the same four grades, from our present (A level sample) dataset. The first column of Table 4 reproduces (from Table 1) the total number of candidates in each subject area from our present (A level sample) dataset. As a percentage of the full 2013 A level cohort, these figures constitute: 90% French; 84% Spanish; 82% German; 89% history; 94% geography; 74% religious studies.

Table 4. Subject sample vs. cohort data

	No. with any GCSE (sample)	No. A* GCSE (sample)	No. A GCSE (sample)	No. B GCSE (sample)	No. C GCSE (sample)	No. A* GCSE in 2011 (all)	No. A GCSE in 2011 (all)	No. B GCSE in 2011 (all)	No. C GCSE in 2011 (all)
French	9220	5295	2930	886	105	14289	23060	30841	33812
Spanish	5827	3422	1794	536	73	8994	10939	12337	12823
German	3293	1698	1091	443	60	5196	9458	14362	15179
History	41796	10826	15274	11302	3978	22608	38473	42043	36094
Geography	27322	6673	10310	7491	2639	17342	29776	32394	35011
Religious Studies	14132	4351	5374	3432	886	23171	40350	45344	37354

[5] Provisional GCSE (Full Course) Results - June 2011 (England Only); Provisional GCE A level Results - June 2013 (England Only).

Finally, Table 5 summarises Table 4, to indicate the likelihood that students with different grades at GCSE progress to A level. Appended to this table are comparable statistics from ESARD (2012). Table 5 suggests, for example, that nearly half of candidates who achieved A* in GCSE history in 2011 progressed to A level, whereas only around one third of candidates who achieved A* in GCSE German in 2011 did so. Although these conclusions are drawn from fairly rough estimates, they closely resemble progression rates which were determined in a more satisfactory fashion by ESARD, for students who sat GCSE in 2008.

Table 5. Estimated progression rate: 2011 GCSE vs. 2008 GCSE

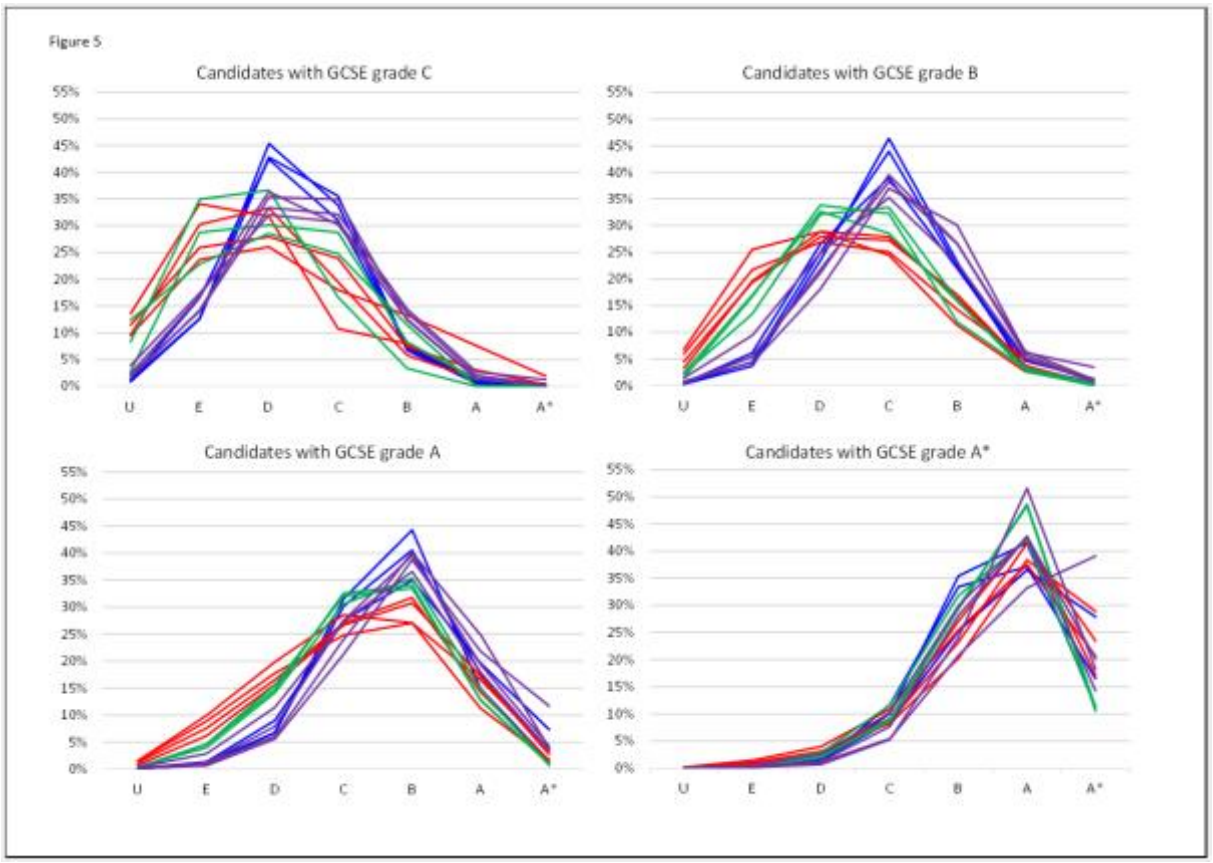
	Progression rate for GCSE A* (2011-)	Progression rate for GCSE A*-A (2011-)	Progression rate for GCSE A*-C (2011-)	Progression rate for GCSE A* (2008-)	Progression rate for GCSE A*-C (2008-)
French	37.1	22.0	9.0	35	9
Spanish	38.0	26.2	12.9	37	14
German	32.7	19.0	7.4	33	8
History	47.9	42.7	29.7	46	28
Geography	38.5	36.0	23.7	35	21
Religious Studies	18.8	15.3	9.6		

Comparative Progression Analyses

Figure 5 presents CPA outcomes for candidates who achieved grades C, B, A and A* at GCSE, respectively. So this includes a replication of the second figure from the science organisations' letter. Figure 5 has the same colour coding as Figure 2; the sciences are presented in red; the languages in green; the English in blue; and the humanities in mauve. (Note that Figures 5 to 8 are reproduced in a larger scale at the end of this report.)

For GCSE grade B, Figure 5 replicates the pattern from the science organisations' second figure; extending the analysis to the languages and English, which also cluster by subject group. The graph for GCSE grade C resembles the graph for GCSE grade B; although, for some subjects, the percentages are derived from fairly small numbers and may be correspondingly unreliable. Notice, from both of these two graphs, how grades for the sciences are more widely distributed than grades for the other subjects. This contrast is most stark when the sciences are compared with the English. It is clear that fairly large proportions of GCSE grade B candidates achieve no higher than grade E at A level in the sciences; whereas very few GCSE grade B candidates achieve lower than grade D at A level in the humanities and the English.

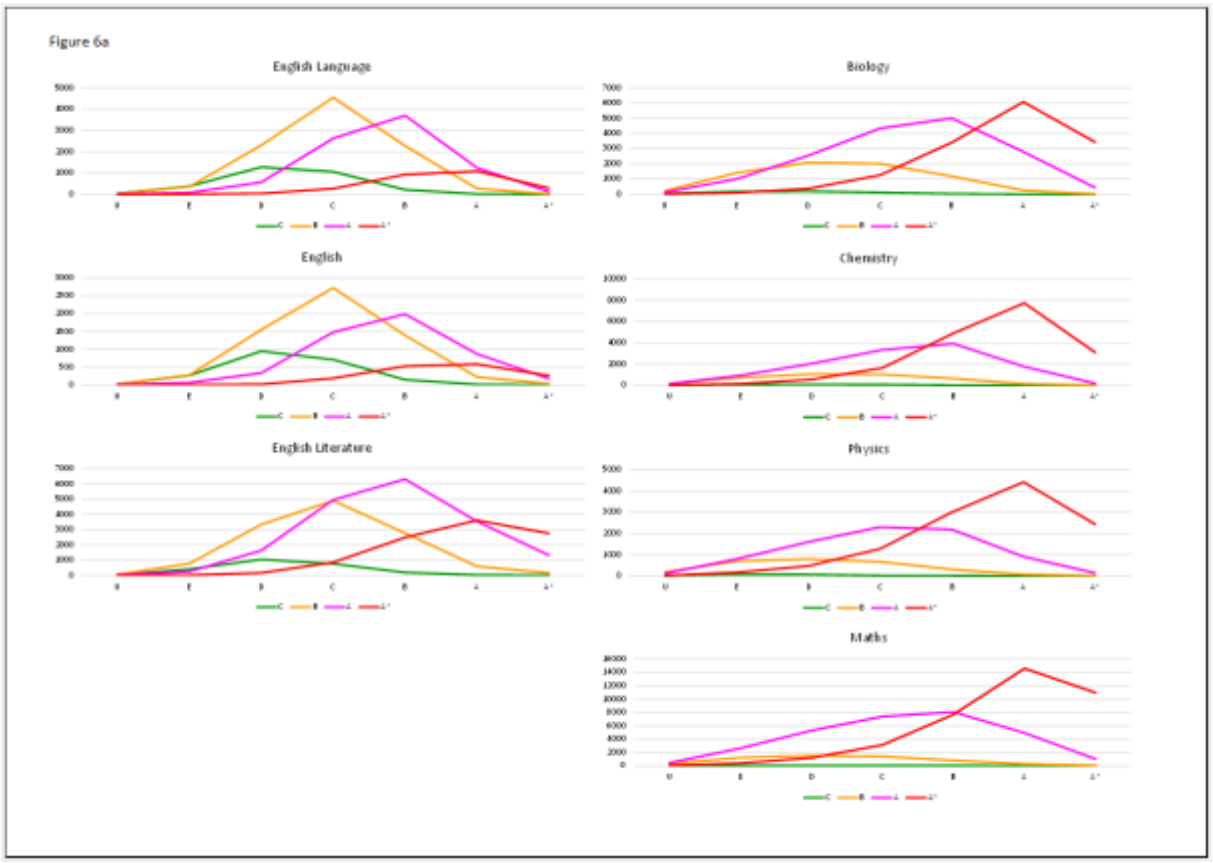
Figure 5. Comparative progression analyses for GCSE grades C, B, A and A*



It is important to notice that the pattern which is clear for GCSE grade B is far less apparent for GCSE grade A. Grade A is an important comparator, because large proportions of candidates begin their A level courses from the GCSE grade A baseline, across all subjects (see Figure 2). Grade A is perhaps more generally representative than GCSE grade B in this respect. From Figure 5, it seems that GCSE grade A candidates typically achieve grade B at A level, across all subject groups, roughly speaking.

The differences between the graphs for GCSE grades A and B represent the kind of differential subgroup effect described above: the same A level grading scales seem, when patterns are taken at face value under various assumptions, to be clearly misaligned when the analyses are run for one subgroup of the A level population (those previously awarded GCSE grade B); whereas the misalignment is far less evident when the analyses are run for another subgroup (those previously awarded GCSE grade A). Remember that we are talking about exactly the same grading scale(s) for both subgroups across all subjects. So this is at least somewhat anomalous. Notice, finally, that the distributional differences are evident across both graphs, with the sciences (and languages) more likely to award grades E and D.

Figure 6a. Numbers of candidates awarded each A level grade (by GCSE C to A*)



Figures 6a and 6b present the same data (as in Figure 5) albeit in terms of numbers rather than percentages and for each subject separately. This time, the colours represent prior GCSE grade: C green; B yellow; A pink; A* red. The subject groups are presented in columns to facilitate comparison: the English and sciences in Figure 6a; the languages and humanities in Figure 6b. Notice how the patterns are very similar within each column. The fourteen pink curves comprise (in toto) the subgroup of GCSE grade A students. Although, as a group, these candidates typically achieve grade B at A level, roughly speaking, the pink peaks are more pronounced for the English and humanities. For the sciences and languages, the pink curves tend to be flatter, such that their typical candidate is probably better described as achieving B-to-C at A level; that is, slightly lower than for the English and humanities.

The fourteen yellow curves comprise (in toto) the subgroup of GCSE grade B students. For the English and humanities, these yellow curves clearly peak at A level grade C. For the sciences and languages, the yellow curves do not clearly peak in the same way, such that their typical candidate is probably better described as achieving C to D at A level, or perhaps even C to E for the sciences. Note that the curves for GCSE grade B students represented far fewer A level candidates for the sciences and languages, on average comprising 12% and 11% of their respective cohorts (compared with 37% and 26% for the English and humanities, respectively).

Figure 6b. Numbers of candidates awarded each A level grade (by GCSE C to A*)

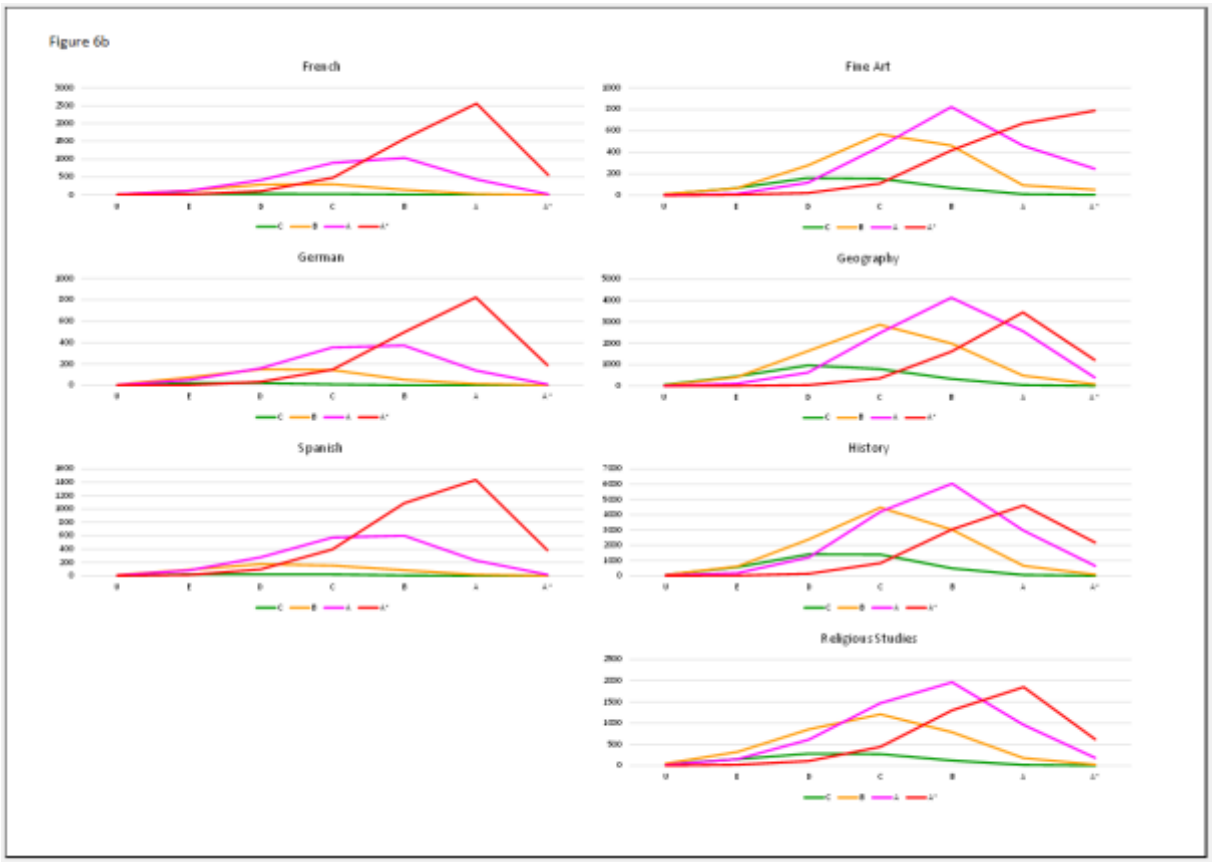
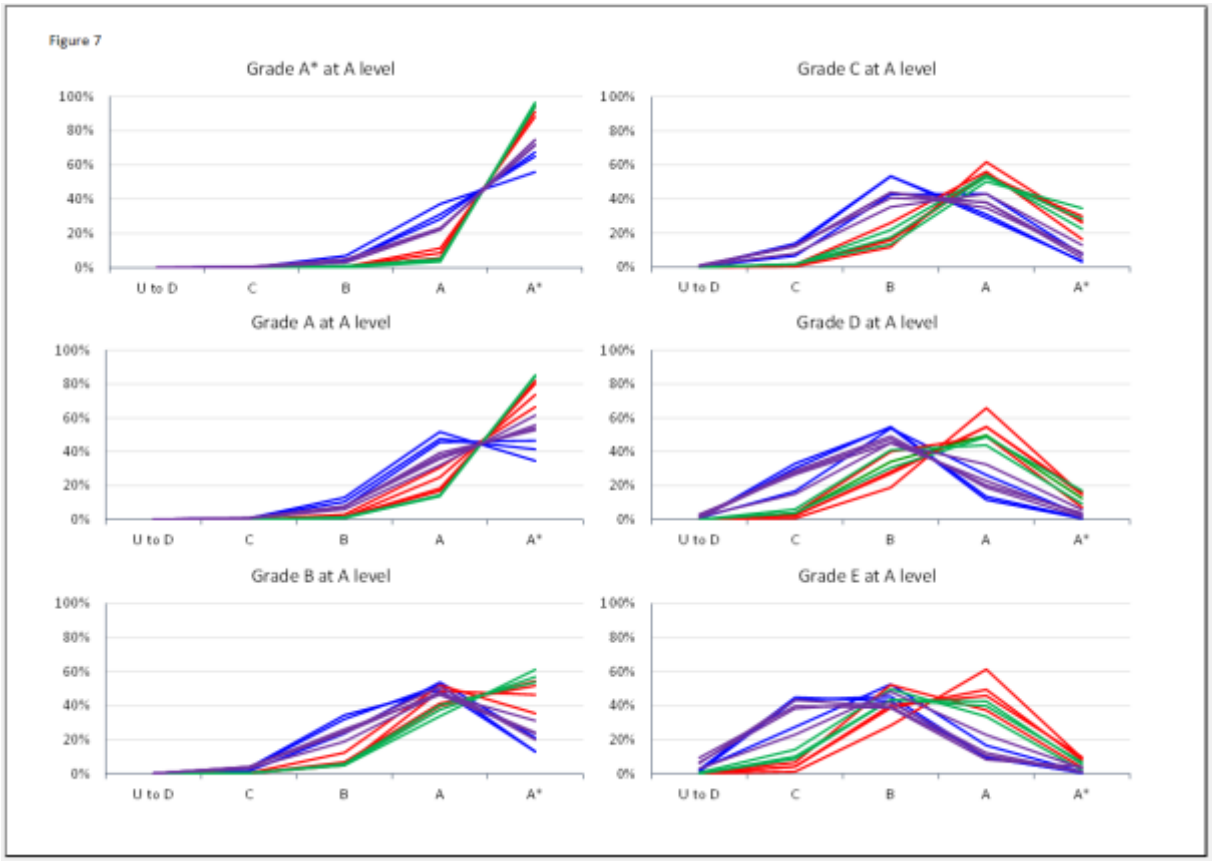


Figure 7 represents, once again, exactly the same comparative progression dataset. This time, though, the lines indicate what percentage of candidates at each A level grade, for each A level subject, had previously been awarded each GCSE grade (for A level grades C to A* respectively). It is analogous to Figure 5, with the axes flipped.

Roughly speaking, for each A level subject grade, the further to the left of each graph a particular subject curve tends to be located, the lower the 'typical' candidate's prior GCSE grade will be. As we would expect, the subject curves creep towards the left (i.e. towards the lower prior GCSE grades) as A level subject grades fall. So the 'full' curves are more evident at the lower A level grades.

At A level grade D, for instance, it is clear that the curves peak at GCSE grade A for the sciences and the languages; whereas, for the English and the humanities, they peak at GCSE grade B. This suggests that English and humanities candidates who achieve the same grade at A level as sciences and languages candidates (grade D) do so from a lower baseline of proficiency in their chosen subject areas. This general pattern is discernible across all A level grades.

Figure 7. Comparative progression analyses for A level grades C, B, A and A*



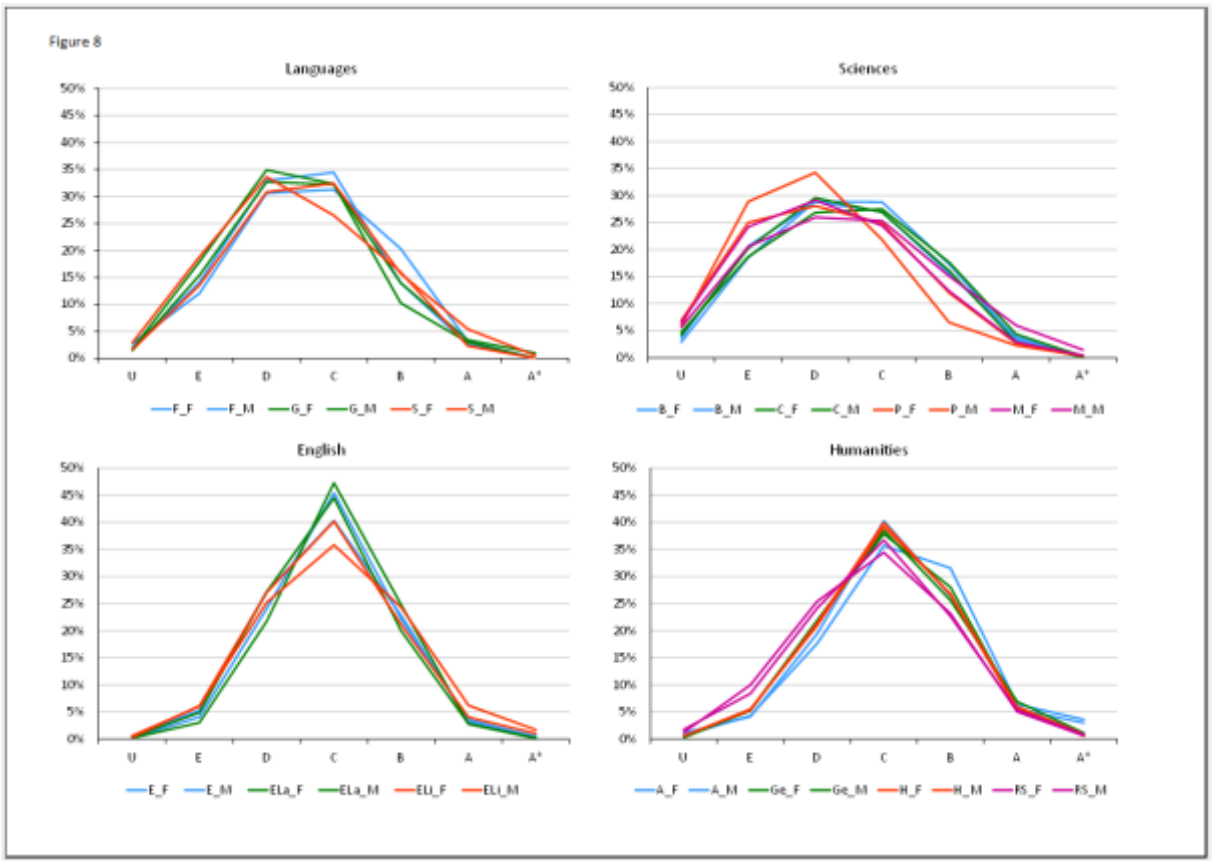
Subgroup analyses

Gender-based subgroup effects are common in statistical analyses of inter-subject comparability. Consequently, the CPA was re-run separately for male and female candidates to explore the possibility of differential subgroup effects.

Figure 8, which presents subgroup analyses, male vs. female, for GCSE grade B students, is representative of an outcome which was fairly consistently replicated across the other GCSE grades; in the sense of providing little evidence of differential subgroup effects.

Incidentally, this figure illustrates more effectively than previous ones the fact that A level grades peak clearly at C for the English and humanities, whilst being less clearly peaked for the languages, and even less so for the sciences. Again, though, these patterns are fairly consistent for male and female candidates.

Figure 8. Subgroup analyses for GCSE grade B students (male vs. female)



Significance

As the first published account of this analysis, the purpose of the present paper is not to draw conclusions from these findings, but to support debate on their significance. The CPA outcomes present a prima facie challenge to the idea that standards are appropriately aligned across A level subject areas, but this challenge is still to be interrogated. The following sections are intended to prompt and to help focus this interrogation. They raise various issues concerning assumptions, findings, action, and theory.

Assumptions

It was suggested that, to interpret outputs from these analyses at face value – in terms of the alignment, or misalignment, of grading standards – we would need to assume that (groups of) candidates ought (on average) to make the same progress across subject areas, all other things being equal; and that all other things are in fact (more or less) equal.

a) Are these assumptions necessary to interpret the outcomes at face value, and is it reasonable to assume that all other things are likely to be (more or less) equal in this context?

Other statistical analyses of inter-subject comparability have been criticised for relying too much upon the assumption of tapping into something like 'general academic aptitude' and, relatedly, for their inability to support plausible interpretations for subjects that do not correlate well with the majority.

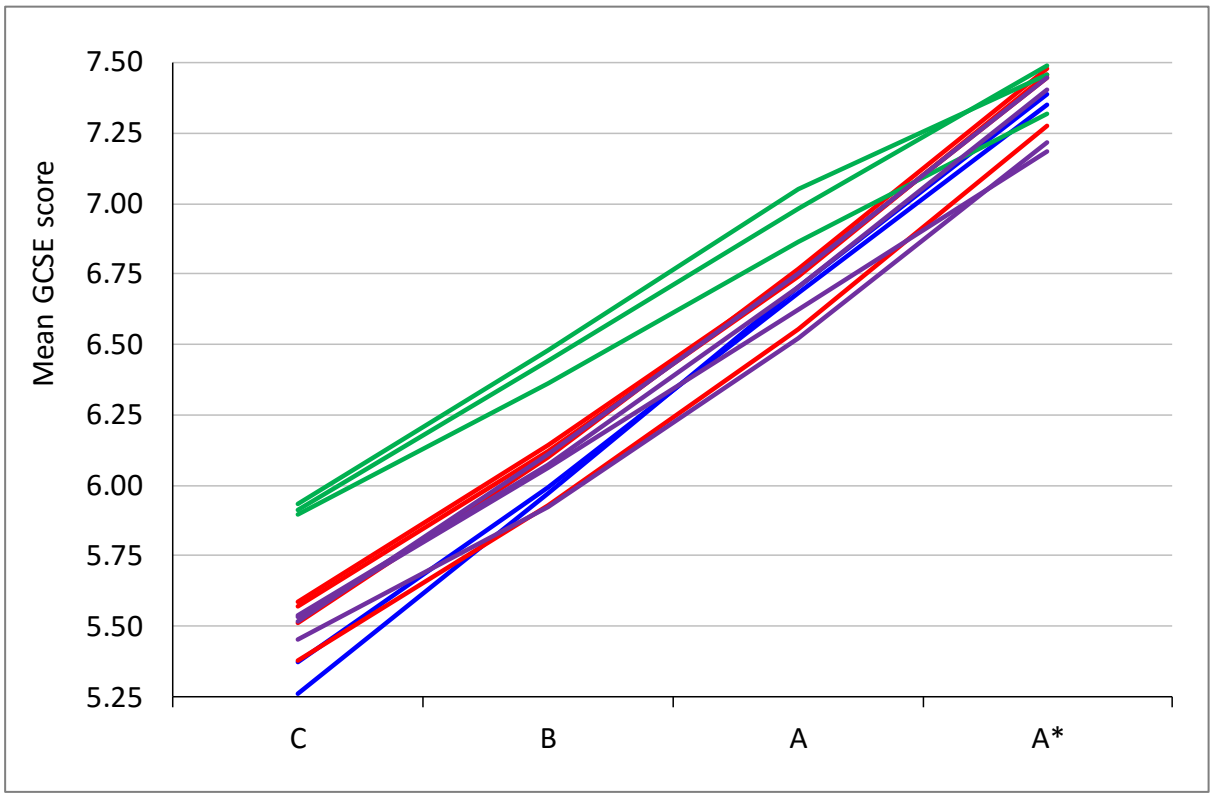
b) Is it likely that these 'new' analyses are more robust to criticisms like these, i.e. that these criticisms do not apply to the same extent, if at all?

The analyses are premised upon the idea that GCSE grades are somehow comparable across subjects. Yet, if the new analyses have any force at A level, then they might hint at similar problems for GCSE.

c) If inter-subject comparability cannot necessarily be assumed at GCSE, then what implications might follow for interpreting CPA outcomes at A level?

In this context, note Figure 9, which, for the candidates in our A level sample, indicates the relationship between their subject GCSE grade and their mean GCSE score. Notice the gulf between grades in the languages vs. grades in the other subject groups: for candidates who achieve grades B or C in the languages, their language grade is typically substantially lower than their mean GCSE score (C is coded as 5 points, B as 6, etc.).

Figure 9. Mean GCSE score for each subject GCSE grade



Finally, recall that the within-subject GCSE to A level correlation coefficients suggested that only around a third of the variance could be 'explained'.

d) Might the magnitude of these coefficients – whether judged to be high or low – have any impact upon our confidence in the analyses?

Findings

The analyses provide strong evidence that schools and colleges tend to operate different progression requirements across subject areas: a 'minimum C' policy for English and the humanities; versus a 'minimum B' policy for the sciences and languages.

e) Could requirements like these impact upon the robustness of A level grading procedures, potentially leading to misalignment?

With the exception of English and English language, the A level grade distributions are reasonably similar across subjects. The languages and sciences tend to award slightly more of the highest grades; but the sciences also tend to award slightly more of the lowest grades.

f) It is clear that A level grades are more widely distributed for the languages and, in particular, the sciences. Why might this be? Is it legitimate? Does it complicate the interpretation of outputs from CPA, or perhaps even undermine them?

Proportionally more students progress from GCSE to the same subject at A level in history and geography than in the languages or religious studies. For example, nearly one quarter of the GCSE history cohort progresses to an A level in history; whereas not even one tenth of the GCSE French cohort progresses to an A level in French. Similarly, nearly half of candidates who achieved A* in GCSE history in 2011 progressed to A level, whereas only around one third of candidates who achieved A* in GCSE German in 2011 did so.

g) Might these differences indicate anything important about A level students in different subject areas? For instance, are A level history students more likely to be the 'cream of the cream' (in history) than A level French students (in French)? If so, then could this affect the legitimacy of conclusions from CPA?

Figure 5 not only replicates patterns from the science organisations' second figure, it also reveals different patterns for the languages and English.

h) Do these additional patterns reinforce, or complicate, recommendations from the science organisations (or both, perhaps)?

However, from Figure 5, the pattern which is clear for GCSE grade B is far less apparent for GCSE grade A. Grade A is an important comparator, because a large proportion of candidates begin their A level courses from the GCSE grade A baseline, across all subjects. As becomes clearer from Figures 6a and 6b, there is still an effect, even for grade A, although it appears to be somewhat smaller.

i) To what extent does this differential subgroup effect (prior grade B versus prior grade A) mean that outputs from CPA cannot be taken at face value?

Distributional differences are evident in Figure 5 – both for prior grade A and prior grade B – with the sciences (and languages) more likely to award grades E and D.

j) Why do these distributional differences occur? And what are their implications for the interpretation of outputs from CPA?

GCSE grade B students represented far smaller proportions of their respective A level examination cohorts for the sciences and languages (12% and 11%) when compared with the English and humanities (37% and 26%).

k) Might this reduce our confidence in inferences from the GCSE grade B analyses (generally) on the assumption that there might be 'different kinds' of grade B student within different subject groups?

The curves in Figure 7 are nothing more than the data from Figure 5 viewed from a different perspective. Not surprisingly, they suggest similar effects.

l) Is Figure 7 any more or less convincing than Figure 5? Does it reveal anything new?

Figure 8 provided little evidence of differential subgroup effects by gender.

m) Why not? Might gender effects become more apparent if the analyses were conducted differently? Might differential subgroup effects emerge if the analyses were broken down differently, eg by school type?

Action

If we were to conclude that CPA outputs do convincingly indicate genuine problems of misalignment at A level, then what actions could be taken?

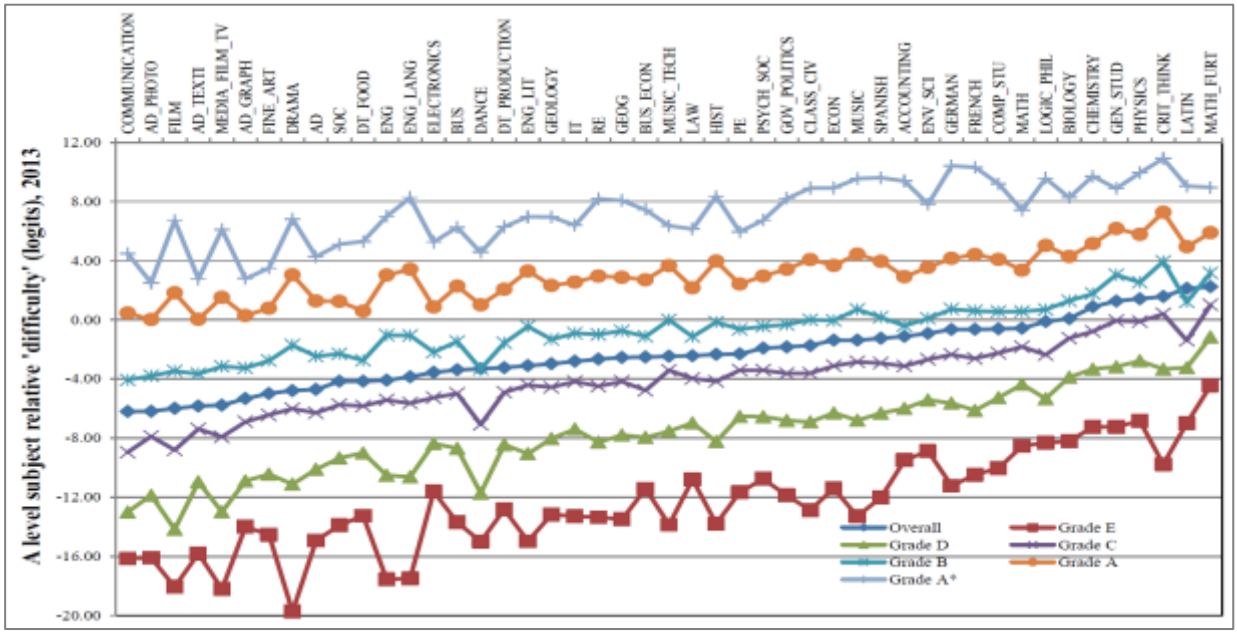
n) Are the outcomes from the analyses, presented above, sufficiently internally consistent to justify taking action? What kind of action might the outcomes recommend?
o) Do the outcomes need to be sufficiently consistent with outcomes from Rasch analyses to justify taking action (see Figure 10)? How much consistency would be sufficient?
p) Because CPA is straightforward and easy to explain, does that make it more appropriate (than, say, Rasch) as a basis for taking action?
q) If we were to conclude that the analyses do indicate genuine problems of misalignment, then would this justify 'special case' realignment, eg just the sciences and languages, or a wholesale recalibration of A level grading standards?

r) If we were to conclude that the analyses do indicate genuine problems of misalignment and a wholesale recalibration, then how would we deal with A level subjects which do not have corresponding GCSEs?

s) Could any action be taken at A level, using GCSE as the baseline, without also (somehow) recalibrating GCSE grading standards?

t) Even if we were totally convinced of the interpretation of CPA outcomes in terms of misalignment of grading standards, would the overall impact of recalibration – costs and benefits weighed against each other – be sufficiently positive to justify action or sufficiently negative to justify not acting?

Figure 10. Reproduction of Figure 4 from Working Paper 3.[6]



Theory

Bearing in mind that there is no generally accepted definition of inter-subject comparability (Newton, 2012), how might the CPA approach be rationalised?

u) What is the nature of the prima facie challenge raised by the CPA approach? Should CPA be evaluated in terms of providing a potential *solution* to a *prediction* problem (ie what is the most appropriate way to predict A level grades)? Or should CPA be evaluated in terms of raising a potential *question* about a *progression* problem (ie why are there different GCSE to A level progression rates across subject areas, if this is not attributable to misalignment of grading standards)? If the former, then perhaps mean GCSE score would be a better baseline than subject GCSE grade after all (ie the

[6] Comparison of the overall subject difficulty and the difficulty at individual grades for the 47 A level subjects from the 2013 exam series. (From He and Stockford, 2015, p.25.)

approach adopted by Fitz-Gibbon and Vincent, 1994). If the latter, then perhaps the 'misalignment' explanation ought to be considered the most parsimonious one until a more plausible explanation can be provided.

v) If outcomes from CPA were deemed to be intuitively plausible, then – in the absence of a generally accepted definition of inter-subject comparability – might it be legitimate simply to choose a definition which was as consistent as possible with the CPA approach? For instance, might it be legitimate to stipulate that inter-subject comparability should be defined in terms of equivalent progress from one educational level to the next? This would imply that (groups of) candidates ought (on average) to make the same progress across subject areas, period, ie without regard to the possibility that other things might not be equal. (And grading standards would be recalibrated across subject areas in order to achieve this.)

w) Are there alternative ways of defining inter-subject comparability, as a basis for understanding the CPA approach? How might they affect the interpretation of outcomes from CPA?

x) Do we need to adopt a particular definition of inter-subject comparability at all? A purely pragmatic alternative might be to propose: that the CPA approach has intuitive plausibility; and that, in the absence of a more plausible approach, it should be acted upon. This would also stipulate that (groups of) candidates ought (on average) to make the same progress across subject areas, period; but as a convenient solution, rather than as a point of principle.

y) What if the nature of progression in learning from GCSE to A level is qualitatively very different across subject areas? Could that provide a justification for progression rate differences, or should any such differences be 'factored out' through the grading process as a point of principle?

z) Similarly, if there were differences in the quality of teaching across subject areas (on average, ie at the national level), then could that provide a justification for progression rate differences, or should any such differences be 'factored out' (on average, ie at a national level) through the grading process as a point of principle?

References

Bell, J.F. and Emery, J.L. (2007). *The Relationship Between A-level Grade and GCSE Grade by Subject. Statistics Report Series No. 7*. Cambridge: University of Cambridge Local Examinations Syndicate.

Coe, R. (2008). Comparability of GCSE examinations in different subjects: an application of the Rasch method. *Oxford Review of Education*, 34 (5), 609-636.

Dearing, R. (1996). *Review of Qualifications for 16–19 Year Olds*. London: School Curriculum and Assessment Authority.

Education Standards Analysis and Research Division (2012). *Subject Progression from GCSE to AS Level and Continuation to A Level. Research Report DFE-RR195*. London: Department for Education.

Fitz-Gibbon, C.T. and Vincent, L. (1994). *Candidates' Performance in Public Examinations in Mathematics and Science*. London: School Curriculum and Assessment Authority.

Fitz-Gibbon, C.T. and Vincent, L. (1997). Difficulties regarding subject difficulties: Developing reasonable explanations for observable data. *Oxford Review of Education*, 23 (3), 291–8.

Goldstein, H. and Cresswell, M.J. (1996). The comparability of different subjects in public examinations: A theoretical and practical critique. *Oxford Review of Education*, 22 (4), 435–42.

He, Q. and Stockford, I. (2015). *Inter-Subject Comparability of Exam Standards in GCSE and A Level* (Working Paper 3. Ofqual/15/5798). Coventry, UK: Office of Qualifications and Examinations Regulation.

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), 433–49.

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.

Sutch, T. (2013). *Progression from GCSE to AS and A level, 2010. Statistics Report Series No. 69*. Cambridge: Cambridge Assessment.

Figure 5

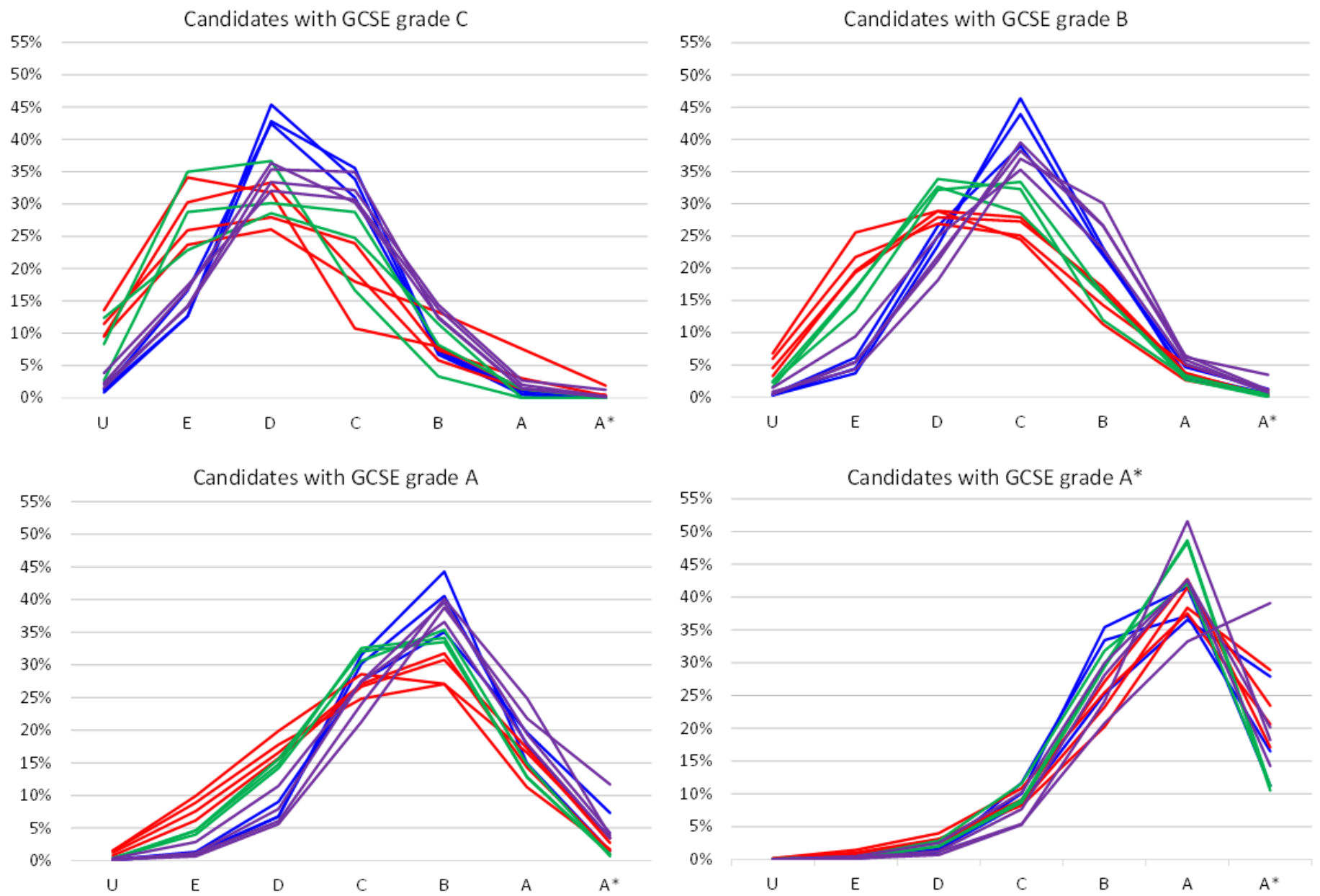

Candidates with GCSE grade C
Candidates with GCSE grade B
Candidates with GCSE grade A
Candidates with GCSE grade A*
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
U
E
D
C
B
A
A*

Figure 6a

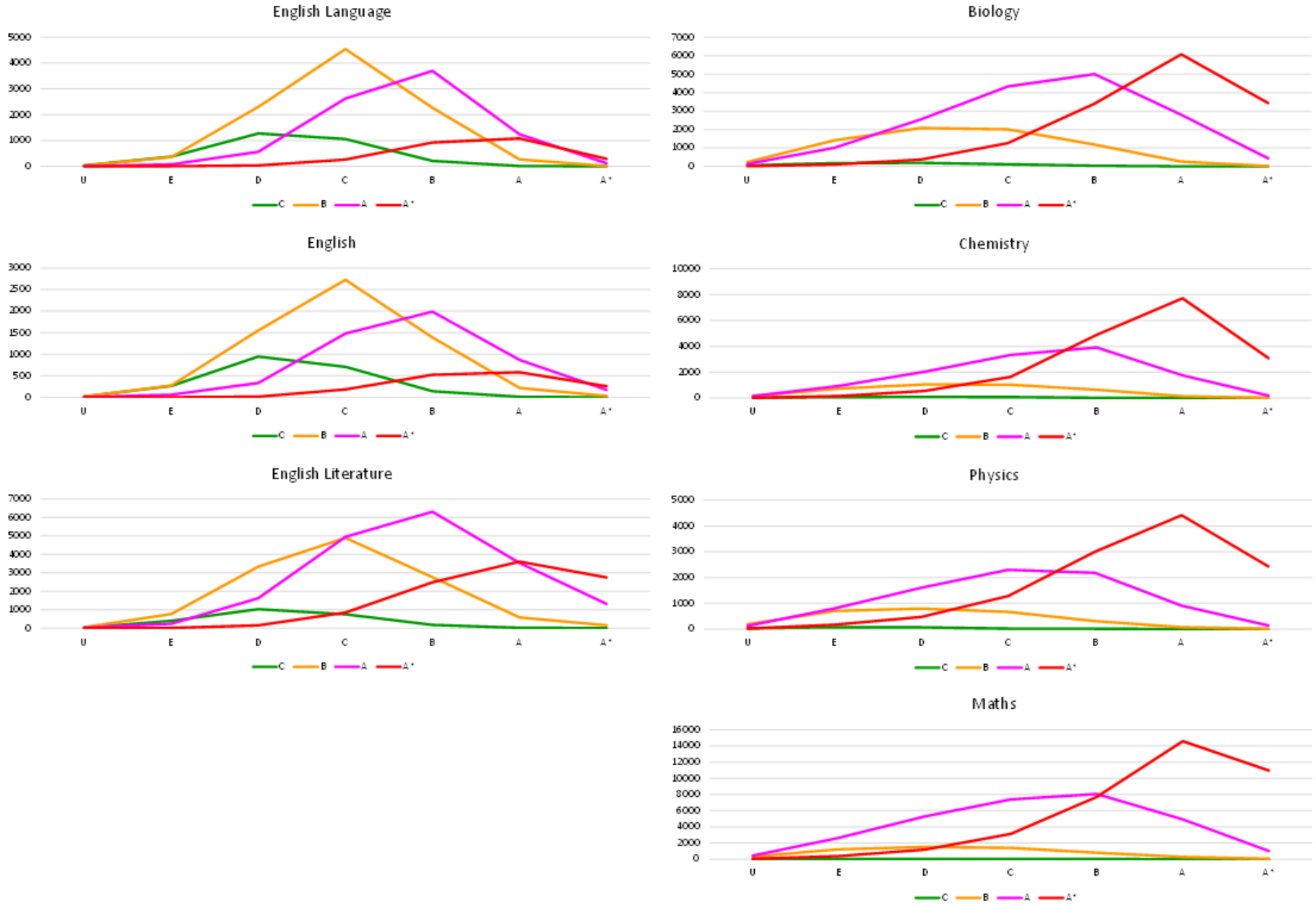

English Language
Biology
English
Chemistry
English Literature
Physics
Maths
U
E
D
C
B
A
A*

Figure 6b


French
Fine Art
German
Geography
Spanish
History
Religious Studies
U
E
D
C
B
A
A*
C
B
A
A*

Figure 7

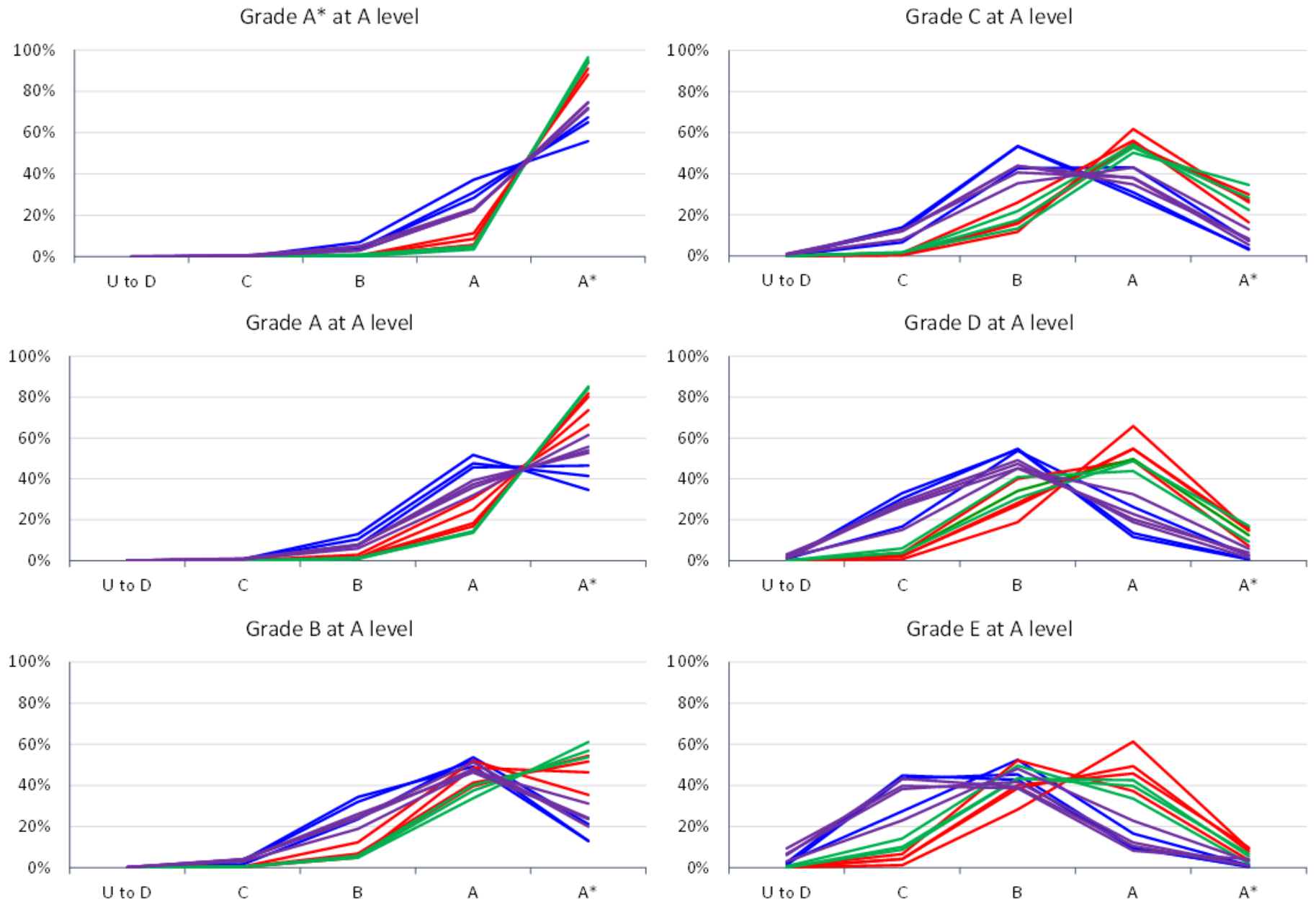
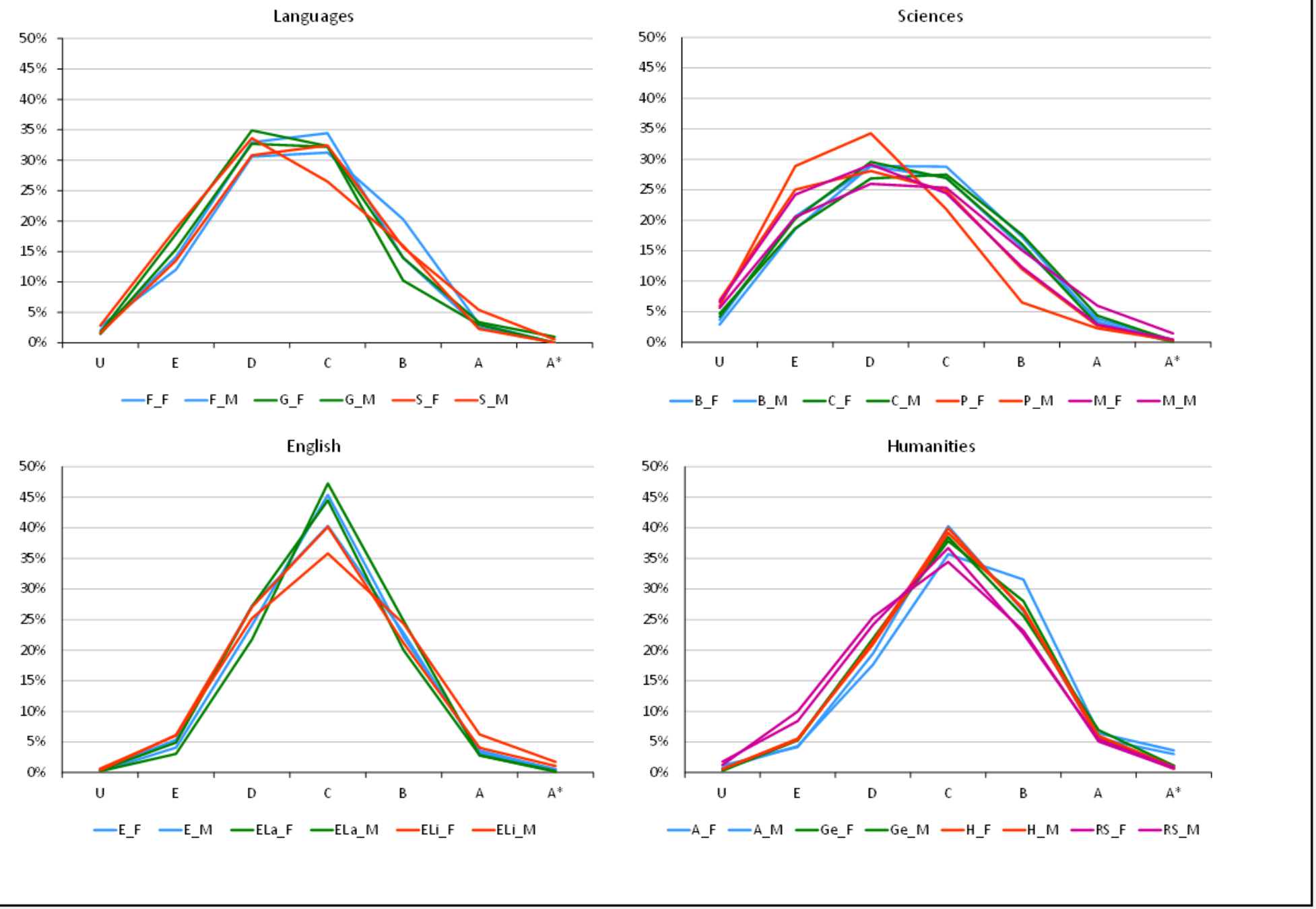

Grade A* at A level
Grade C at A level
Grade A at A level
Grade D at A level
Grade B at A level
Grade E at A level
100%
80%
60%
40%
20%
0%
U to D
C
B
A
A*

Figure 8


Languages
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
U
E
D
C
B
A
A*
F_F
F_M
G_F
G_M
S_F
S_M
Sciences
B_F
B_M
C_F
C_M
P_F
P_M
M_F
M_M
English
E_F
E_M
ELa_F
ELa_M
ELi_F
ELi_M
Humanities
A_F
A_M
Ge_F
Ge_M
H_F
H_M
RS_F
RS_M

We wish to make our publications widely accessible. Please contact us at publications@ofqual.gov.uk if you have any specific accessibility requirements.

This publication is available at www.gov.uk/ofqual.

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation
Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346