Cambridge Assessment

Some thoughts on the 'Comparative Progression Analysis' method for investigating inter-subject comparability

Research Report

Tom Benton & Tom Bramley

4 September 2017
(Minor corrections made in November 2017)

Author contact details:

Tom Benton & Tom Bramley
Assessment Research and Development,
Research Division
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG
UK

Benton.t@cambridgeassessment.org.uk
Bramley.t@cambridgeassessment.org.uk

http://www.cambridgeassessment.org.uk

As a department of Cambridge University, Cambridge Assessment is respected and trusted worldwide, managing three world-class examination boards, and maintaining the highest standards in educational assessment and learning. We are a not-for-profit organisation.

How to cite this publication:

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

Table of contents

Summary

This report was originally written as two separate reports which were shared and discussed with Ofqual's Standards Advisory Group as they were considering the issues raised by their investigation of a new approach to inter-subject comparability – namely the 'Comparative Progression Analysis' or CPA method (Ofqual, 2017). We have combined the two reports here for ease of reference.

Our main findings and conclusions are:

1. The CPA method is not really about 'progression', but about conditional grade distributions. Requiring the same conditional grade distributions between GCSE and A level in different subjects is equivalent to requiring that they correlate equally between GCSE and A level. These correlations are a contingent fact about the world and not controllable by exam boards or the regulator by manipulating grade boundaries.
2. Any attempt to explore inter-subject comparability at A level based on progress from GCSE should use data on overall GCSE performance (e.g. mean GCSE grade) rather than purely being restricted to the grade achieved in the specific subject. However, doing this has been criticised in the past and many of the objections are still valid.
3. The CPA analysis itself shows that the larger discrepancies between subjects at A level are at grades C and D. The exam boards only 'set' boundaries at grades A and E – the others are derived by interpolation and extrapolation. Removing some of the larger differences identified by the CPA method might require changing the way intermediate boundaries are set.
4. Using the *overall* (i.e. across all subjects combined) 'progression' relationship from GCSE to A level as a basis it is possible to work backwards to find what changes would be needed to grade boundaries in different subjects to align them to the common CPA pattern. This would result in (for example) lowering the grade boundaries in Physics and raising them in History. Further work would be needed to verify whether either change would be justified in terms of (for example) what universities are saying about the undergraduates they receive.
5. Simulations can show that even if two subjects are equivalent according to CPA for hypothetically defined populations (e.g. the proportion of the GCSE cohort that go on to take A levels), non-random choice of A level subjects and different GCSE-A level correlations in different subjects can create the *appearance* of different CPA for the groups that actually take the A levels in different subjects.
6. Different subjects differ in how they are taught, understood and assessed, and these differences can be reflected in overall (marginal) and conditional grade distributions. Whether these differences should be adjusted for statistically or whether they are part of the inherent nature of different subjects and the way in which they are learned by candidates is a matter for debate.
7. At present, we are nowhere near a scenario where decisions to align subjects could or should be automated purely on the basis of statistical analyses, whether such statistics are derived within subjects over time, between subjects taken simultaneously, or even by some combination of the two. Rather, on any occasion where such an intervention is made we believe it is important to carefully consider the benefits and costs of any action, not only from a statistical perspective, but also in terms of any possible wider implications.

Introduction

At the heart of the CPA method is the idea that more attention could be placed on progress between GCSE and A level *within individual subjects* to explore inter-subject comparability. For example, the fact that candidates who achieved grade B in GCSE Physics tend to go on to achieve no higher than a D at A level is seen as possible evidence that grading standards in this subject are too hard. In contrast, candidates who achieve a B in GCSE History will frequently go on to achieve grade C in History A level.

Such analyses have been performed before. For example, Bell and Emery (2007) provided a fairly comprehensive list of tables detailing the associations between GCSE and A level grades in the same subjects. Similar information is supplied graphically based on GCSEs taken in 2010 by Sutch (2013). However, the idea of using such associations to directly inform decisions is new. In this context this report explores the implications of any such approach a little further.

Is GCSE grade in the same subject the best basis for examining future achievement?

There are a number of options for predicting how we might expect candidates to perform in an A level in a particular subject. Using candidates' GCSE grades in the same subject is only one such possibility. Candidates' mean GCSE grades or achievement in other A level subjects are also frequently considered as reasonable options. The predictive power of each of these possibilities is compared for 14 subjects in Table 2 of Ofqual's report. Interestingly, this showed that GCSE grade in the relevant subject was the predictor with the *lowest* correlation with outcomes of the three examined for 11 out of 14 of the subjects considered. The fact that the chosen measure tends to have a lower correlation with A level outcomes than other alternative actually suggests that it controls for less of the candidate-level factors that may influence outcomes. As such, from a technical perspective, it does not appear to be a good choice for studying inter-subject comparability.

The idea of using subject-specific performance at GCSE to influence A level awarding has been considered before. Specifically, Benton and Lin (2011) considered this in the context of maintaining standards over time and between awarding organisations (within a subject). Their report concluded that "there is little to be gained by any replacement of mean GCSE score as the measure of prior attainment" (page 10) including incorporating data from achievement at GCSE in related subjects. Given that we have (rightly) dismissed the idea of using achievement in a related subject at GCSE to maintain standards between awarding organisations, it would seem odd to return to this concept for the much more challenging task of maintaining standards between subjects.

More thorough analysis looking at the ways in which subject-specific GCSE grades could be used to predict A level performance was conducted by Benton (2015). This research used complex techniques from machine learning to look beyond progress within a given subject between GCSE and A level to consider the extent to which accounting for the exact combination of GCSE subjects candidates had taken and the grades they had achieved helped to predict A level performance. The predictive power of the complex model was compared to a simple model based purely on candidates' mean GCSE grades. This

research concluded that for all the additional effort “for the vast majority of subjects, these improvements [in predictive performance] were small” (page 604). A rare exception was for modern languages (particularly German) where it was found that accounting for the GCSE grade in the relevant language significantly improved predictive power. This finding fits with the results displayed on page 11 of the Ofqual report.

Conceptually we may worry about using a measure such as mean GCSE to explore progress between GCSE and A level due to possible differences in difficulty between the GCSEs taken by different candidates. However, recent analysis (Benton, 2016) has shown that the effect of such differences on measures such as mean GCSE is vanishingly small.

For the reasons above, we believe that from a technical standpoint, any attempt to explore inter-subject comparability at A level based on progress from GCSE should use data on overall GCSE performance (e.g. mean GCSE grade) rather than purely being restricted to the grade achieved in the specific subject. However, despite this, the simplicity of using within-subject GCSE to A level analyses is appealing. For this reason it is worth considering the practical implications of any such approach a little further.

How much difference is there in progress from GCSE to A level between subjects?

Although it included some interesting and compelling graphics, the Ofqual (2017) report did not put any specific figures on the extent to which awarding in some subjects appears to be lenient whilst awarding in others seems harsh. In order to aid later sections of this report, this gap is addressed in this section. As part of this process, the analysis will go beyond Ofqual’s work relating GCSE and A level attainment within each subject and need to impose a practical definition of equal ‘progress’ within each subject.

Analysis was based on A level performance in June 2014 matched to GCSE performance in the same subject in June 2012. All data was drawn from the National Pupil Database (NPD) as supplied by the Department for Education and restricted to pupils in Year 13 in 2014. To begin with, Table 1 shows how grades achieved at GCSE relate to the chances of achieving particular grades at A level in the same subject. For the purposes of this table all GCSE grades below D are combined. Only A level entries from candidates who entered exactly the same subject at GCSE two years earlier are included in this table.

For any A level subject, the information in Table 1 can be used to predict the percentage of candidates we would expect to achieve each grade or above if GCSE-A level progress was constant across subjects – in essence the assumption of the comparative progression analysis provided by Ofqual. Such predictions can then be compared to the actual cumulative percentage of candidates achieving each grade (for candidates with matching GCSE information) to get an idea of the extent to which awarding may possibly be more severe in some subjects than others. The results of such calculations are shown for 12 A level subjects in Figures 1 and 2 for grades A*-B and C-E respectively.

Table 1: Overall relationship between GCSE grades and A level achievement in the same subject (combined results across all subjects)

GCSE grade	Percentage of entries achieving each grade or above at A level							Total entries
	U	E	D	C	B	A	A*	
D or below	100.0%	92.2%	68.6%	36.0%	9.6%	1.5%	0.5%	1718
C	100.0%	96.7%	81.0%	46.1%	13.6%	2.0%	0.4%	18565
B	100.0%	97.8%	88.4%	64.6%	29.0%	6.2%	1.1%	70868
A	100.0%	99.0%	94.6%	82.3%	55.9%	21.5%	4.4%	121084
A*	100.0%	99.9%	99.1%	96.3%	86.9%	61.0%	22.6%	106263

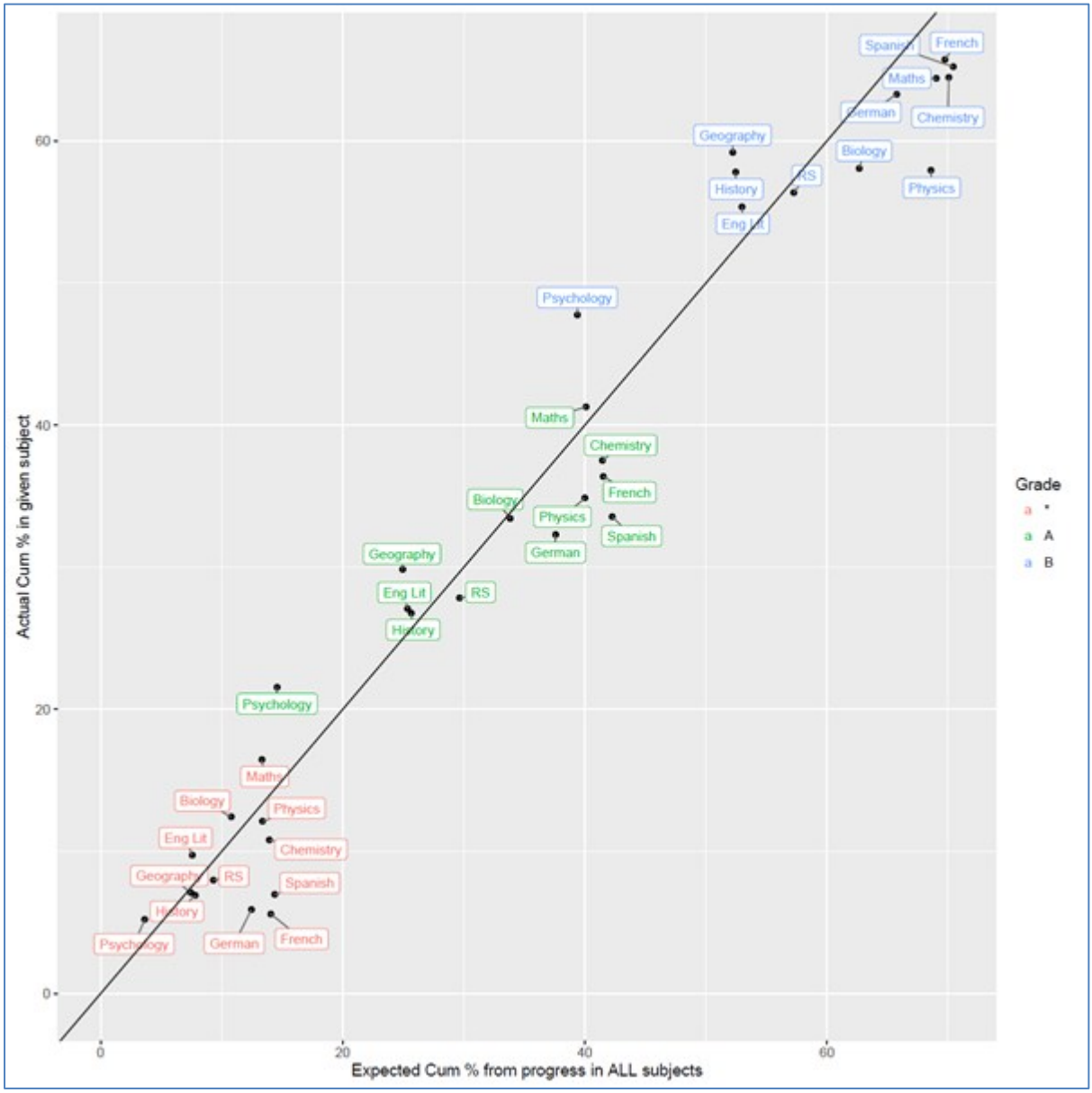

Figure 1: A comparison of the expected grade distributions based on related GCSE performance and the actual grade distribution for twelve A level subjects (grades A*-B)

Although based on different data, Figure 1 confirms many of the findings from the Ofqual report. However, once candidates are analysed overall, the implication in Ofqual's quote from their communication with the science organisations - that there is a clear uniform

influence of severity of grading in sciences (Ofqual, 2017, page 3) - is less obvious. In particular, at grade A (the green labels) Maths and Biology do not appear at all severe compared to the humanities such as English Literature, History and Religious Studies (RS). Furthermore, at grade A, whilst Physics and Chemistry appear generally severe compared to other subjects, the extent of this difference is less than for modern languages. However, larger differences appear at grade B and at lower grades as explored in Figure 2.

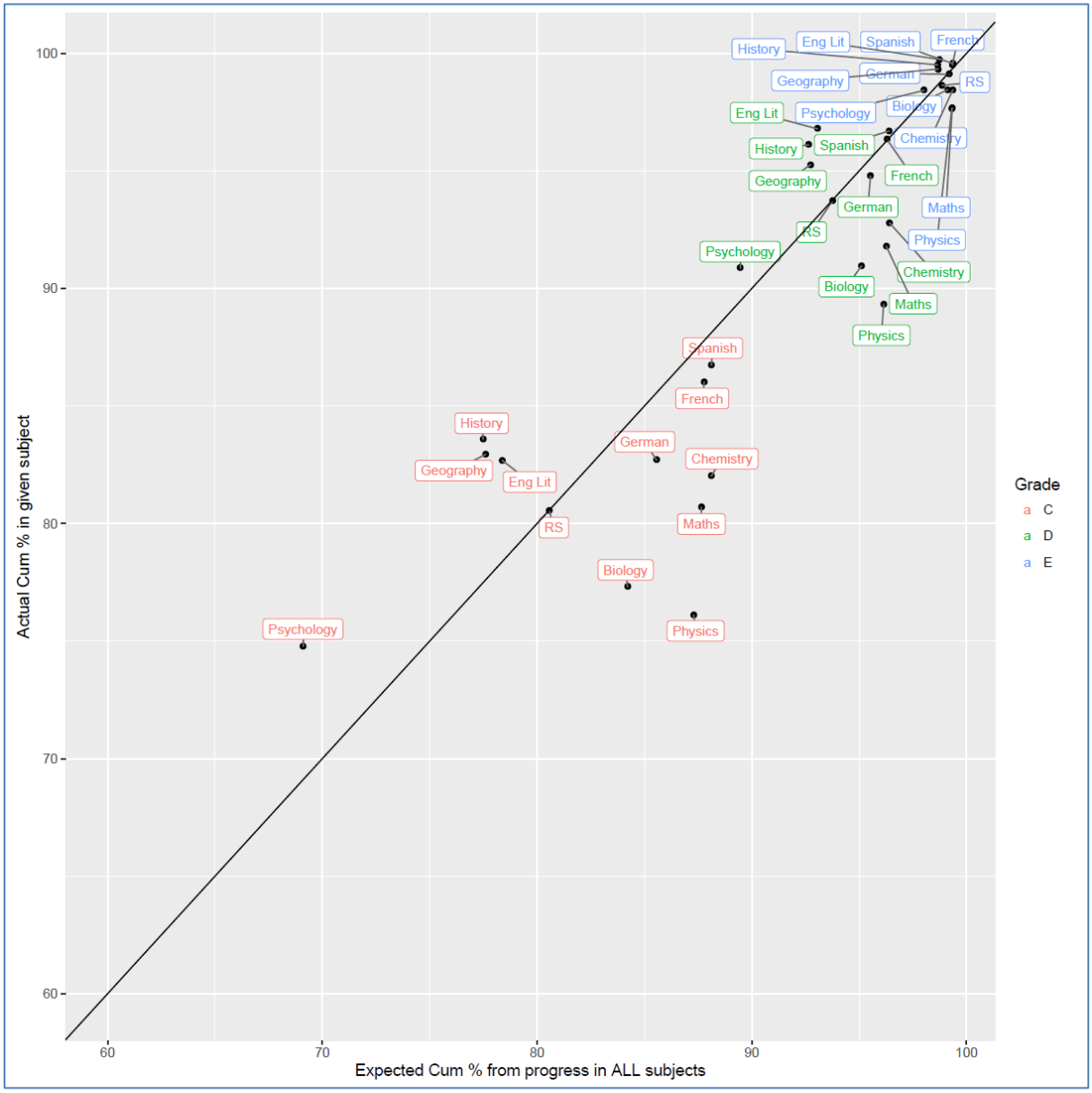

Figure 2: A comparison of the expected grade distributions based on related GCSE performance and the actual grade distribution for twelve A level subjects (grades C-E)

Figure 2 provides particularly worrying results at grade C and D. Due to the fact that the vast majority of candidates achieve grade E or above in all subjects, differences are less visible. This point is important. Within current grading procedures, the only grade boundaries that are actually set by awarding committees are at grades A and E with the remaining boundaries being set to be equally spaced between these. As we have already seen, the differences between subjects at grade A are relatively small. It should be noted that setting

boundaries at grade E can be far more challenging due to the scarcity of candidates near to these boundaries. With this in mind it is interesting that it is in setting this grade boundary where subject differences appear to be emerging.

To provide a little more detail on these results, the exact percentages expected to and actually achieving each grade or above are recorded for five subjects in Table 2. For example, this table shows that for Physics, if progress from GCSE to A level was the same as for all subjects then we would expect 87 per cent of (matched) candidates to achieve at least a grade C compared to only 76 per cent who actually do. In contrast, for History, we would only expect 78 per cent of candidates to achieve this grade or above compared to the 84 per cent who actually do. Given these striking differences, these two subjects will be explored further in the next section.

Table 2: A comparison of the expected grade distributions based on related GCSE performance and the actual grade distribution for five A level subjects

A level grade	Cumulative percentage of candidates at each grade in each subject									
	Maths		Biology		Physics		French		History	
	Expected	Actual	Expected	Actual	Expected	Actual	Expected	Actual	Expected	Actual
A*	13.3%	16.4%	10.8%	12.4%	13.4%	12.1%	14.1%	5.6%	7.8%	6.9%
A	40.1%	41.3%	33.8%	33.5%	40.0%	34.9%	41.5%	36.4%	25.7%	26.7%
B	69.0%	64.4%	62.7%	58.1%	68.6%	57.9%	69.7%	65.7%	52.4%	57.8%
C	87.7%	80.7%	84.2%	77.3%	87.3%	76.1%	87.8%	86.0%	77.5%	83.6%
D	96.3%	91.8%	95.1%	91.0%	96.1%	89.3%	96.3%	96.3%	92.6%	96.1%
E	99.3%	97.7%	99.1%	98.5%	99.3%	97.6%	99.3%	99.5%	98.7%	99.5%
U	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

What would happen to grade boundaries if average progress from GCSE to A level grades was fixed across subjects?

Having seen some apparently large differences between Physics and History it is of interest to explore the effect of attempting to align the two subjects statistically, at least in terms of average 'progress', on the actual grade boundaries.

For the purposes of this analysis actual data on the marks achieved by candidates is required. For this reason, analysis is restricted to A levels awarded by OCR[1]. Analysis looked at two OCR A level specifications in History and Physics in June 2015. The necessary data, including matched GCSE information from the National Pupil Database, was available for 4,758 candidates in Physics and 7,652 candidates in History.

Analysis is based upon the total marks achieved in A2 units by candidates in June 2015. For both subjects a total of 200 marks were available on these units. In June 2015, grades were actually awarded based on achievement on a combination of both AS and A2 units. However, in the future awarding will be based on A2 units only. For this reason, we focussed purely on these marks. New grade boundaries were derived based on the total marks on

[1] This data can be matched to GCSE data in the same subjects from all awarding organisations using the NPD.

these A2 units so as to, as close as possible, retain the original A level grade distribution in each subject[2].

The mark distributions on the A2 papers in Physics and History are shown in Figure 3. As can be seen, Physics tends to have a much wider mark distribution (standard deviation=28) than History (standard deviation=21). The red dashed lines show where the grade boundaries would be placed on each specification to preserve the current grade distribution. It can be seen that the locations of grade boundaries are relatively similar across the two subjects. Achieving an A* (the rightmost boundary) requires candidates to achieve a little above 80 per cent of marks (160 out of 200) on each test, achieving a grade A requires 75 per cent of marks (or just less for Physics) and grade E (the leftmost boundary) requires slightly more than 30 per cent of marks.

The blue dotted lines show what would happen if grade boundaries were set so as to match the expected grade distribution based on GCSE attainment in the same subject. These expectations were derived using progress across all A level subjects as recorded in Table 1.

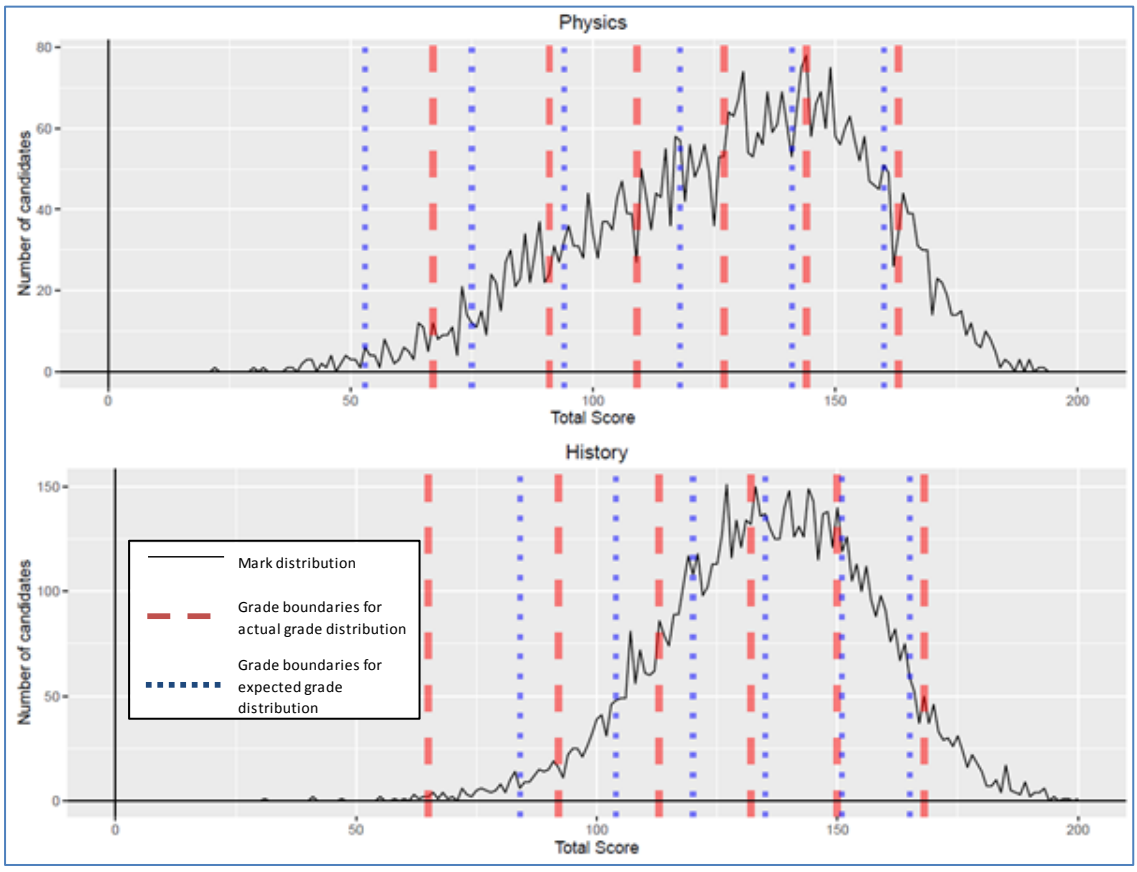

Figure 3: The effect on grade boundaries of applying common predictions across both Physics and History

[2] For the purposes of this analysis we did not require that grade boundaries were evenly spaced. As it turns out, their spacing ends up fairly even in any case. Also note that, for History, candidates could choose from a variety of optional A2 papers. In practice marginally different grade boundaries are set depending upon which combination is chosen. However, these differences are tiny (usually 1 or 2 marks) compared to the changes explored in this section and are ignored for the purposes of this analysis.

Apart from at grades A and A*, directly applying this definition of inter-subject comparability has a dramatic effect on the location of grade boundaries in both subjects. As well as being shown by the blue dotted lines in Figure 3, the original and revised grade boundaries are recorded in Table 3. For Physics, the B to E grade boundaries would end up being lowered by between 9 and 16 marks. In particular this would mean that a grade C would require candidates to achieve less than half of the available marks (only 96 out of 100 required) and a grade E would only require just more than a quarter of marks.

For History, the effect is in the opposite direction. Students would need to secure more than 40 per cent of the marks even to achieve a grade E. Achieving a D would require more than half of the marks. A C grade would now require candidates to achieve at least 60 per cent of the available marks across their A2 assessments.

Table 3: The effect on grade boundaries of applying common predictions across both Physics and History

A level grade	Possible grade boundaries from A2 mark distribution			
	Physics		History	
	From current grade distribution	From proposed grade distribution	From current grade distribution	From proposed grade distribution
A*	163	160	168	165
A	144	141	150	151
B	127	118	132	135
C	109	94	113	120
D	91	75	92	104
E	67	53	65	84

From an educational perspective, the effects above may not be desirable for either subject. The approach would lead to a considerable relaxation of standards in A level Physics. It should be noted that, in contrast, some commentators have suggested that Physics A level is already too easy in terms of preparation for the next stage of education[3]. As such, it is possible that the above grade boundaries would be unacceptable both to awarding committees and to the science organisations raising the issue of inter-subject comparability in the first place. For History, this approach to inter-subject comparability would mean that History candidates would receive no credit at all for achieving the first 40 per cent of marks and that there would be no differentiation in grading between students with marks in this band. Furthermore, the approach would considerably reduce the gaps between grades. Assuming the assessments themselves were unchanged, this would, in turn, reduce the reliability of candidates' final grades.

[3] See, for example, comments about many Physics degrees now requiring four years of study rather than three in this blog by Tim Oates (http://www.cambridgeassessment.org.uk/insights/just-like-goldilocks-exam-questions-should-be-just-right/).

What would happen to grade boundaries if progress from GSCE grades to A level in Physics was used as a template for History?

The previous section assumed applying a common rate of progress between GCSE and A level across all subjects. However, as mentioned above, this may draw criticism from science organisations who may be unhappy about the prospect of relaxing standards in their own subjects. Instead they may want to ensure that the humanities were as “hard” as Physics.

The impact of assuming that progress from GCSE to A level should be the same in History as it is currently in Physics is explored in Figure 4 and Table 4. For this analysis, a version of Table 1 was created but based only on candidates who took A level Physics (across all awarding organisations) in June 2014. This was then used to derive an expected grade distribution for the OCR A level History specification and, thus, to estimate the appropriate grade boundaries.

The mark distribution and grade boundaries to match the actual grade distribution in Figure 4 are identical to those displayed in Figure 3. However, using progress from GCSE to A level in Physics to set grade boundaries in History leads to even larger amendments to grade boundaries than described earlier. In particular candidates would now need to secure over half the available marks even to achieve grade E. To get a grade D would require a score just 1 mark short of 60 per cent. A situation where over half the mark scale does not lead to any recognition is terms of grades being awarded would be highly unusual. Furthermore, in this assessment, having just 52 marks (out of 200) between candidates who are awarded an A in History and those that are ungraded would reduce the reliability of awarded grades. For both these reasons it is easy to understand why the differences in grade distributions between History and Physics have emerged.

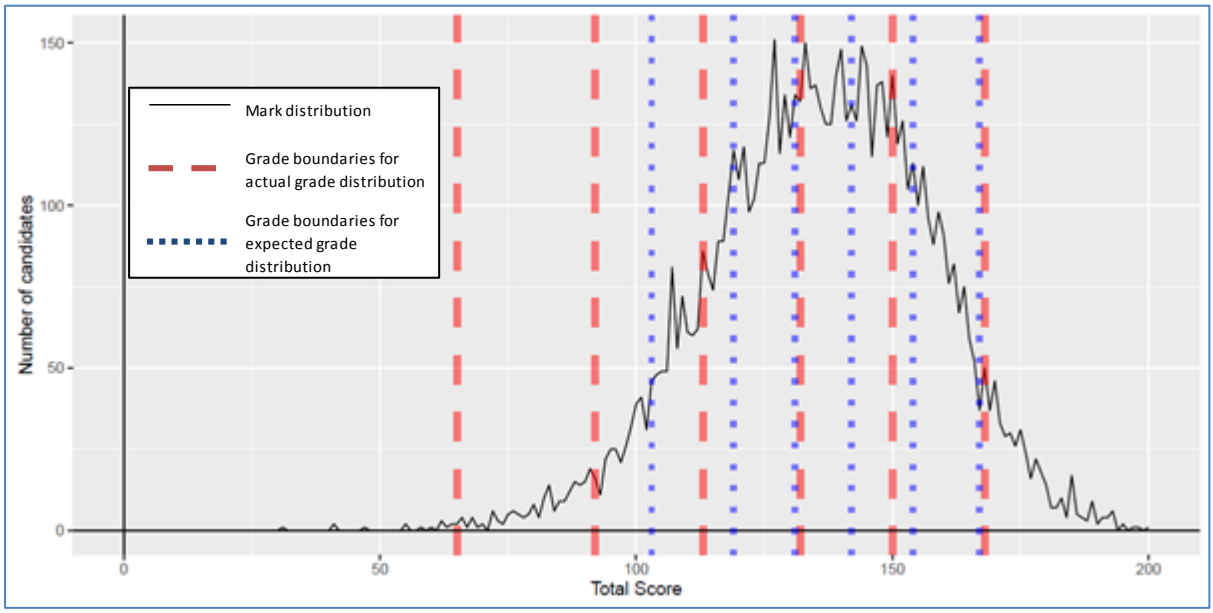

Figure 4: The effect of applying predictions based on progress from GCSE to A level in Physics to grade boundaries in A level History

Table 4: The effect of applying predictions based on progress from GCSE to A level in Physics to grade boundaries in A level History

A level grade	Possible grade boundaries from A2 mark distribution	
	History	
	From current grade distribution	From proposed grade distribution
A*	168	167
A	150	154
B	132	142
C	113	131
D	92	119
E	65	103

Is it misleading to talk about 'progress'?

In this section we argue that describing the new method as 'Comparative Progression Analysis' (CPA) risks sowing the seeds of confusion. To measure progress within a subject, we arguably need two things: a unidimensional attribute, and a unit of measurement. GCSEs and A levels report on different grade scales, so there is no unit of measurement. It is also perhaps debatable whether GCSE attainment and A level attainment in the same subject lie on a unidimensional continuum.

Different subjects by definition measure different things, so even though all GCSEs use the same grade scale and all A levels use the same grade scale, a 'grade' is not a unit of measurement. In contrast, if we consider the physical attributes height and weight (where we do have measurement units) we can see that it is possible to measure the 'progress' in height and weight for a group of candidates from age 16 to age 18. There are meaningful ways in which progress in height could be compared with progress in weight (for example, if most people stopped getting taller at age 16 but continued to get heavier) but they rely on having the same unit to measure height and weight at both ages.

Lacking a unit of measurement, what we are really talking about is bivariate distributions where one variable is the GCSE grade and the other is the A level grade. These are ordered (ordinal) variables that can be thought of as coarse-grained variables created by applying cut-offs to finer-grained variables (aggregate marks, or UMS points). There is nothing in the CPA method that requires a concept of progress[4].

Having defined two variables (GCSE grade and A level grade in the same subject), it then becomes an empirical matter what their bivariate distribution is in a given population. Imagine that the entire population of candidates take History and Physics at both GCSE and A level. The point of this thought-experiment is to remove the complications of unmatched samples and non-random missing data (see Bramley 2016).

If at GCSE the grade distributions in History and Physics were the same, then by most definitions of inter-subject comparability those subjects would have the same grading

[4] This is not to say that pupils make no progress from GCSE to A level.

standards. And likewise at A level if the grade distributions in the two subjects were the same they would have the same grading standards. However, does it necessarily follow that the two subjects must therefore have the same conditional distributions of A level grade for each GCSE grade? Intuitively the answer is 'no' – in which case the CPA method is conceptually flawed inasmuch as it defines equivalent grading standards in terms of equal conditional distributions (mischaracterising these as 'equal progress' or 'equal value-added'). This is illustrated in Tables 5a and 5b below.

Table 5a: Hypothetical bivariate grade distribution in subject with low (zero!) correlation between GCSE and A level grade[5]

		A level grade				
		1	2	3	4	Total
GCSE grade	1	5	5	5	5	20
	2	5	5	5	5	20
	3	5	5	5	5	20
	4	5	5	5	5	20
	5	5	5	5	5	20
	Total	25	25	25	25	100

Table 5b: Hypothetical bivariate grade distribution in subject with medium correlation (0.57) between GCSE and A level grade

		A level grade				
		1	2	3	4	Total
GCSE grade	1	10	6	4	0	20
	2	8	10	2	0	20
	3	3	4	8	5	20
	4	2	3	6	9	20
	5	2	2	5	11	20
	Total	25	25	25	25	100

As can be seen, the subjects have the same grade distributions at GCSE and A level, but the within-grade conditional distributions are different.

In general, there are *(m-1)(n-1)* degrees of freedom in a *m*×*n* cross-table. So even if the marginal distributions are fixed, there is plenty of scope for the cell entries to vary. There is still one degree of freedom in a 2×2 table, so even if GCSEs and A levels were graded pass-fail it would still be possible for two subjects (taken by the same group of candidates) to pass and fail the same overall proportions at GCSE and A level but have different conditional A level pass rates.

[5] For simplicity, the length of both grade scales has been reduced.

Tables 6a and b: Hypothetical bivariate grade distributions in two subjects graded pass-fail

		A level		
		Fail	Pass	Total
GCSE	Fail	15	5	20
	Pass	35	45	80
	Total	50	50	100

		A level		
		Fail	Pass	Total
GCSE	Fail	20	0	20
	Pass	30	50	80
	Total	50	50	100

These considerations suggest that it is unreasonable to use the outcomes of CPA analyses (on their own) to support claims of lack of comparability between subjects. Requiring subjects to meet a CPA comparability criterion is in a sense equivalent to assuming or requiring that they correlate equally in the whole population between GCSE and A level – which seems hard to justify. Even if the same group of candidates did take a pair of subjects at GCSE and A level it would not be within the power of the awarding organisations or the regulator to ensure that the CPA criterion was met while at the same time achieving inter-subject comparability in the marginal distributions.

Simulation 1

In an attempt to see whether similar graphs to those in Figure 5 of Ofqual (2017) could arise in conditions where subjects were equally difficult (in terms of overall grade distributions) at GCSE and A level but differed in their correlations, we carried out a small simulation as follows:

- Simulate two bivariate unit normal distributions (N=100,000), one with a high correlation (Subject A, 0.9) and one with a low correlation (Subject B, 0.5). The variables represent GCSE and A level scores (not grades). Then for each subject:
- Assign numerical GCSE grades roughly matching the national cohort distribution of GCSE Maths grades in 2012 by applying cut-offs to the simulated GCSE scores[6].
- Randomly sample 10,000 candidates (i.e. 10% of the GCSE cohort) using stratified sampling with the proportions within each GCSE grade as per Table 1 in Ofqual (2017, p8) for Maths[7].
- Assign numerical A level grades roughly matching the national average distribution of A level grades as per Table 1 in the current paper[8] by applying cut-offs to the simulated A level scores.

Figure 5 shows the A level grade distributions for simulated candidates with GCSE grades from B to A* (6 to 8). It is clear that the conditional grade distributions are different, despite each subject's A level cohort having the same attainment at GCSE, and awarding the same overall proportion of grades at A level.

[6] A* 5.5%, A 15.5%, B 30.2%, C 58.7%, D 77.6%, E 86.7%, F 93.9%, G 98.2%.
[7] 51.6% from those with A*, 40.5% from those with A, 7.6% from those with B, 0.3% from those with C.
[8] A* 9%, A 30%, B 58%, C 81%, D 94%, E 99%.

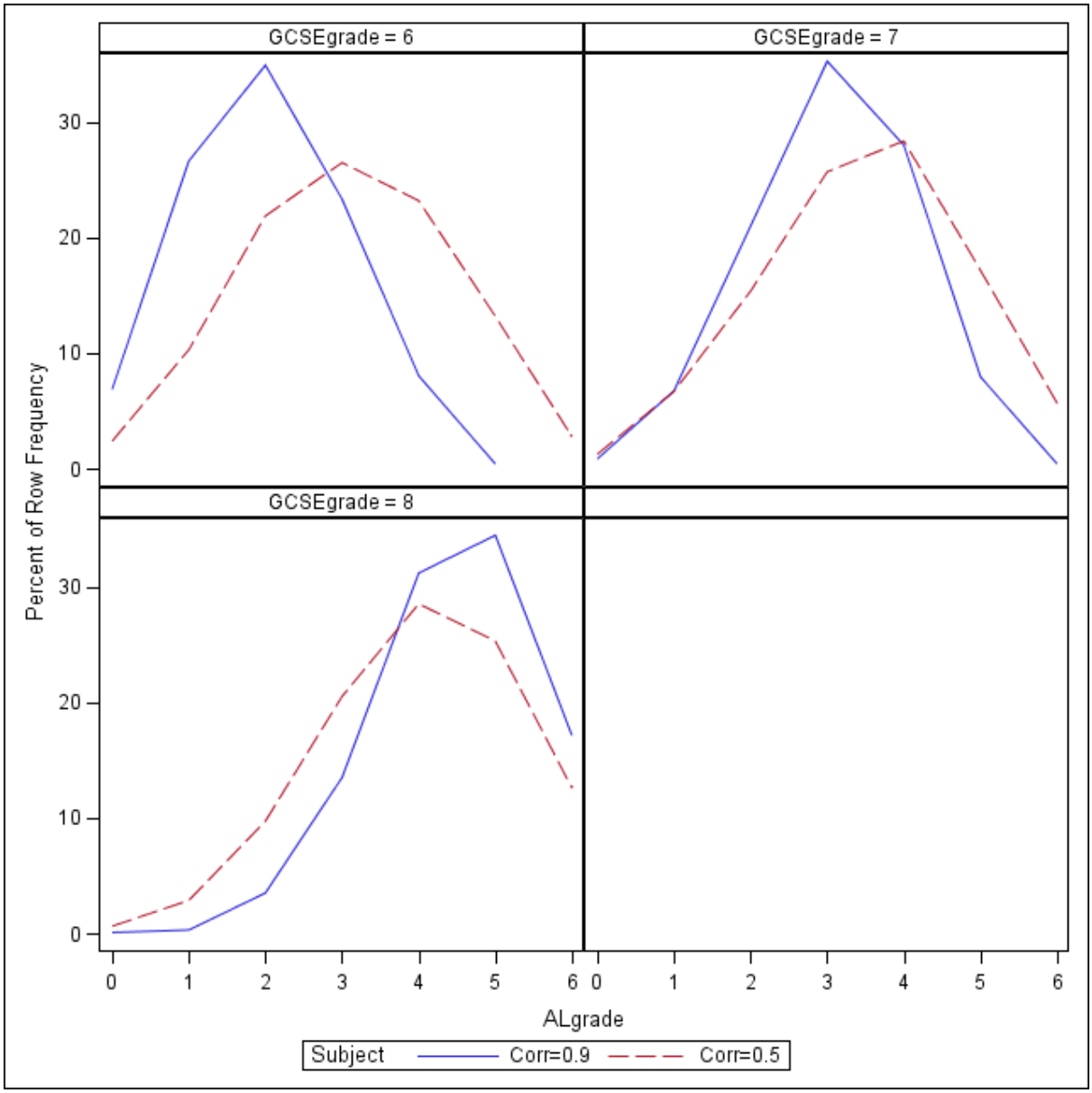

Figure 5: A level grade distributions conditional on GCSE grade (simulated data)

We stress that this simulation does not explain the results reported in Ofqual (2017) because whilst the correlation between GCSE grade and A level grade for the examinees from subject A (with simulated population correlation of 0.9) was 0.615, the same correlation in subject B (with simulated population correlation of 0.5) was only 0.231. The former value is reasonably representative of observed values (Table 2 on p11 of Ofqual 2017), but the latter value is much lower. Furthermore, the simulated GCSE and A level scores were linearly related at population level (because they were simulated to have a bivariate normal distribution), which may not be the case in practice. However, the results do reinforce the point that we should not necessarily expect distributions of A level grades conditional on GCSE grades to be the same in different subjects.

Using conditional means rather than conditional distributions as the basis for comparing subjects

In the previous sections of this report we have argued that an assumption or requirement of the CPA method is that each subject should correlate equally (at the unobservable hyper-population[9] level) between GCSE and A level, and claimed that this is unreasonable. Rather than considering conditional distributions of grades (the CPA method) it is perhaps simpler just to focus on conditional means: that is, for a given grade at GCSE what is the mean grade at A-level? Regression analysis gives one simple way of representing the relationship between GCSE and A level grades in each subject using straight lines.

Figures 6 and 7 use the same set of candidates (those taking A level in June 2014 and GCSE in June 2012)[10]. The divergence of the regression lines shows that candidates with lower GCSE grades in the same subject, or lower mean GCSE grades, achieved on average lower grades at A level in the sciences and modern foreign languages (MFLs) than they did in humanities. The differences between subject groupings are more clearly apparent in the same-subject graph (Figure 6).

These plots make essentially the same point as the CPA analysis but are arguably easier to interpret. Because they are regression lines, they pass through the point [mean(x), mean(y)]. These points are shown as dots. It is noticeable that in Figure 6 the mean GCSE grade in the same subject is often above A, and at a point in the graph where the lines are converging, suggesting that for average candidates and better the differences in A level grades are not too great. But for candidates with a GCSE grade of B in Physics or Psychology the difference between expected scores for a candidate taking A level Physics and one doing A level Psychology is more than a grade (D for the former, C for the latter).

This kind of analysis (using mean GCSE, as in Figure 7) was used (as one of several methods) by Fitz-Gibbon & Vincent (1994) to support claims of lack of subject comparability. It was criticised for a number of reasons by Goldstein & Cresswell (1996), including:

1. Achievement across subjects is multidimensional ("in such a situation, no single reference measure can allow appropriately for achievement in every subject", p439);
2. "The results of reference measure analyses are wholly population dependent" (p439);
3. The relationship between GCSE and A level grades is probably non-linear;
4. Multilevel models should be used (for statistical testing).

Ofqual (2017) suggested that using performance in the same subject as the reference rather than mean GCSE might get around at least problem 1, and that using simple cross-tabulations of grades rather than fitted statistical models might get around problems 3 and 4. Even if this is accepted (and we have already given some reasons not to, such as the higher correlation of A level grade with mean GCSE than with GCSE subject grade, plus the fact that cross-tabulations of grades are essentially correlations), there is still problem 2 to

[9] This term just means a hypothetical dataset where we know the grades at both GCSE and A level for all candidates in all subjects.

[10] Using data from the National Pupil Database. The numbers of candidates and correlations between subject grade / Mean GCSE and A level are in table A1 in the appendix.

confront. It is not quite clear what is meant by 'wholly population dependent', but if we interpret it as a definitional claim that between-subject comparability could only be properly evaluated if the whole population of A level candidates took A levels in all subjects, then this can only be investigated via simulation.

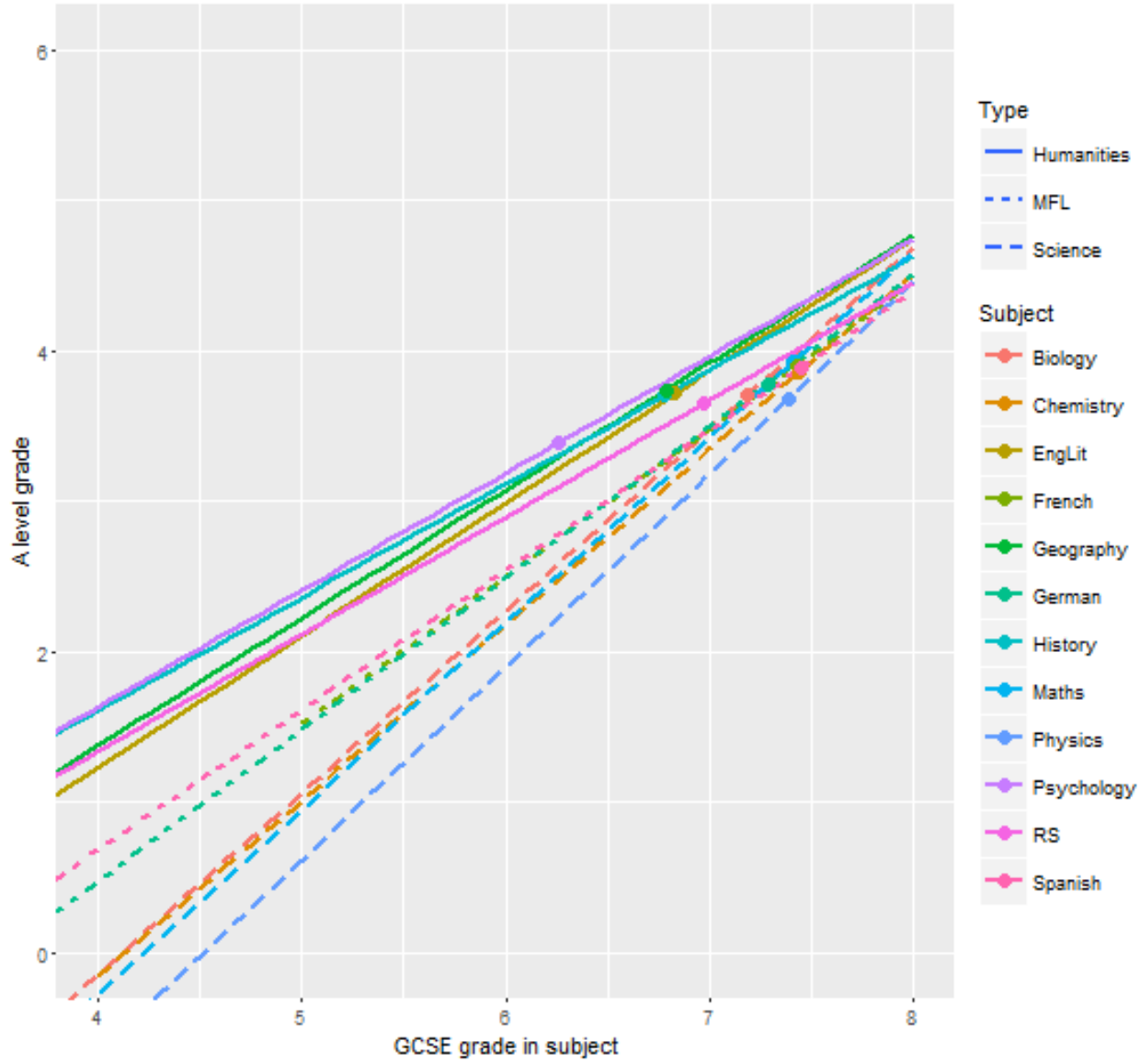

Figure 6: A level grade against GCSE grade in same subject for all A level entries with a GCSE in the same subject (the portion of graph displayed has GCSE grade D or better and A level grade U or better. Dots represent mean grades).

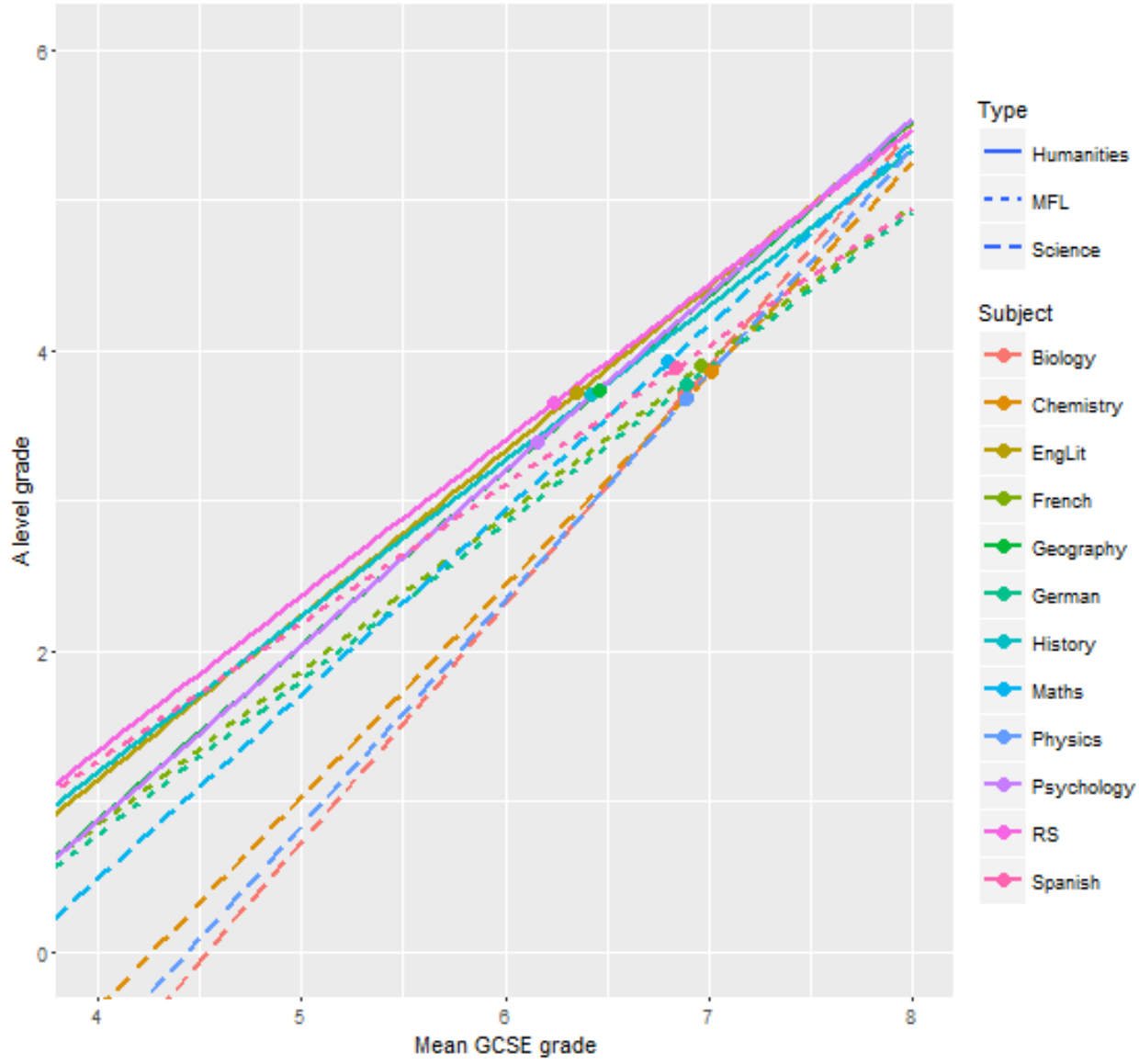

Figure 7: A level grade against mean GCSE grade for all A level entries with a GCSE in the same subject

Simulation 2

Our previous simulation (Simulation 1) showed that differential correlation between GCSE and A level could lead to different CPA results even if the two subjects were equally difficult at GCSE and A level, but that simulation was somewhat unrealistic because we only included one of two factors shown to be relevant to apparent subject difficulty: the correlation between subjects. However, another factor – non-random choice of subject – is also highly relevant, as shown in Bramley (2016). Simulation 2 attempts to improve on Simulation 1 by i) achieving better correspondence between observed and simulated correlations; and ii) incorporating non-random subject choice. The simulation was carried out as follows:

First, we tried to obtain estimates of the correlation between GCSE and A level for a science subject and a humanities subject, starting from the observed correlations reported in Table 2 of Ofqual (2017). Noting the 'restriction of range' for the science subjects arising from the

fact that the A level entry was more selected, with a modal GCSE grade of A* in Table 1 in Ofqual (2017), we applied Thorndike's case 2 correction formula (Thorndike, 1947) to estimate correlations for all A level subjects if the standard deviation of their GCSE grades was the same as that of those taking Maths A level. The results of applying the correction formula are shown in Table A2 in the appendix. After adjustment, Science and MFL subjects had higher correlations than Humanities subjects, justifying using a higher value for Science than Humanities in the simulation.

GCSE scores in two subjects 'Science' and 'Humanities' were simulated for 500,000 candidates[11] as standard normally distributed variates with a correlation of 0.6. These were assigned grades so as to give the following cumulative distribution: A* 35%, A 75%, B 95%, C and below 100%. A level scores in the same nominal subjects were simulated such that the GCSE – A level correlation for Science was 0.75, and for Humanities 0.55. These were assigned grades so as to give the following cumulative distribution: A* 5%, A 20%, B 40%, C 60%, D 80%, E 90%, U 100%. These distributions were chosen to use realistic round numbers such that the grade distributions after subject choice (see below) would be reasonably close to observed values. The subjects were thus simulated to be equally difficult at GCSE and A level, in the sense that the entire A level population had the same grade distribution in both subjects.

We then implemented two rules to capture the effect of non-random subject choice: i) candidates would choose the subject at A level that they would get the better grade in, preferring Science to Humanities if the A level grades were the same[12]; ii) candidates would drop Science (in favour of a third subject) if their GCSE grade was low (in the simulation this meant dropping out with a probability of 0.5 if they had achieved a B, and with a probability of 0.9 if they had achieved C or below).

These rules resulted in ~27k candidates 'dropping out' of the simulation, leaving ~266k taking Science and ~207k taking Humanities. The 'observed correlation' between GCSE and A level for those remaining was 0.59 for Science and 0.56 for Humanities – close to the observed values in Tables A1 and A2 in the appendix. The GCSE and A level grade distributions for candidates choosing each subject are shown in Tables 7 and 8.

Table 7. Simulated data: A level cumulative grade distributions

	A*	A	B	C	D	E	#cands
Humanities	11.2%	37.0%	63.7%	82.9%	94.2%	97.7%	206633
Science	8.9%	34.7%	62.5%	83.0%	96.2%	99.3%	265714

[11] Simulation 1 simulated the entire GCSE cohort and sampled from it to get the A level cohort. In Simulation 2 the simulated 'population' represents those GCSE candidates who go on to take A levels, not all GCSE candidates. This changes the statistical definition of 'equally difficult at GCSE' to be in terms *only of those who go on to take A levels*.

[12] If they achieved A* in both they were simulated to choose the subject with the higher underlying score. If they achieved U in both they chose at random.

Table 8. Simulated data: GCSE cumulative grade distributions

	A*	A	B	C	#cands
Humanities	37.1%	76.4%	95.4%	100.0%	206633
Science	48.4%	91.7%	99.7%	100.0%	265714

The Humanities candidates achieved higher A level grades despite having worse GCSE grades in the same subject. This is illustrated in Figure 8, which shows a very similar relationship between simulated GCSE grade and A level grade as observed in real data (Figure 6). The differences are not quite as large in absolute terms (note that the y-axis scale is different) but are still more than half a grade for simulated candidates with a grade B at GCSE. The conditional distributions of A level grade for each GCSE grade (i.e. the CPA results) are shown in the appendix.

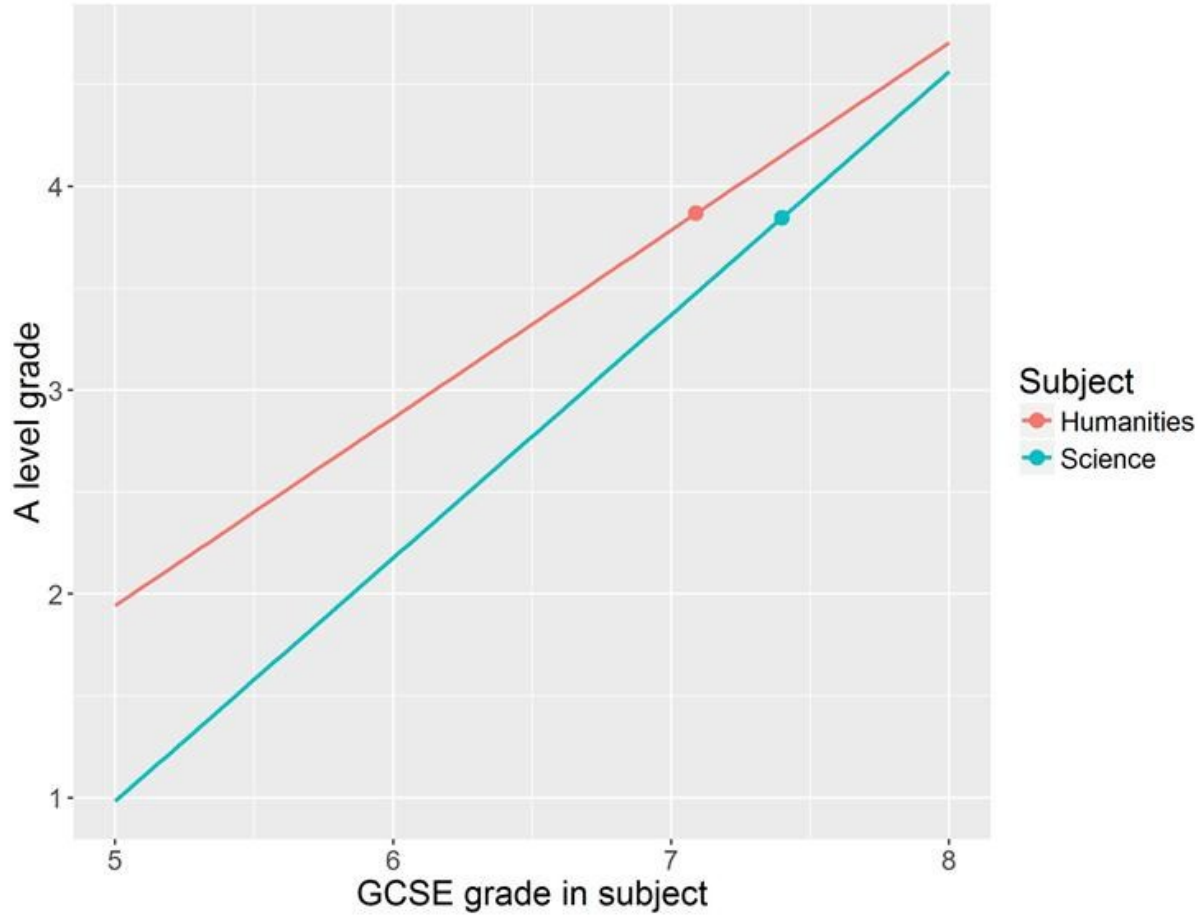

Figure 8. Simulated data. A level grade against GCSE grade in same subject assuming differential correlations and non-random subject choice. Dots represent means.

Again, we are not claiming that this simulation completely explains all the patterns seen in real data. But it is clear that the combination of differential correlation and non-random subject choice can produce apparent differences in subject difficulty when there are none.

In conclusion, A level subjects such as History and Physics are very different. Differences in the way they are taught, understood and assessed are partially reflected in the very different mark distributions for the two subjects. Focussing entirely upon the cumulative percentages of candidates achieving different grades ignores these important distinctions. It is crucial that the effect of any decisions to 'align' different subjects upon the actual location of grade boundaries in terms of marks is considered alongside other sources of evidence. When we look at mark distributions the reason for the discrepancy in History and Physics is clear – candidates achieving extremely low marks are rare in History whilst they occur relatively frequently in Physics. Whether this difference should be adjusted for statistically or whether it is part of the inherent nature of the two subjects and the way in which they are learned by candidates is a matter for debate.

It is, of course, the job of the regulator to ensure public confidence in qualifications. As such, we have no objection if from time to time it is decided that grade boundaries in particular subjects should be deliberately raised or lowered to help improve public trust in comparability. However, such decisions, particularly when it comes to comparisons between different subjects, are complex. At present, we are nowhere near a scenario where such decisions could or should be automated purely on the basis of statistical analyses such as CPA or the techniques that have been suggested in the past, whether this is within subjects over time, between subjects taken simultaneously, or even some combination of the two. Rather, on any occasion where such an intervention is made, we believe it is important to carefully consider the benefits and costs of any action, not only from a statistical perspective, but also in terms of any possible wider implications.

References

Bell, J.F., and Emery, J.L. (2007). *The relationship between A-level grade and GCSE grade by subject.* Cambridge Assessment Statistics Report No. 7. Available from http://www.iaea2008.cambridgeassessment.org.uk/ca/digitalAssets/181132_JB___JE_Statistics_Report_No_7.pdf

Benton, T., and Lin, Y. (2011). *Investigating the relationship between A level results and prior attainment at GCSE*. Coventry: Ofqual. Available from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/488314/2011-09-29-investigating-the-relationship-between-a-level-results-and-prior-attainment-at-gcse.pdf

Benton, T. (2015). Can we do better than using 'mean GCSE grade' to predict future outcomes? An evaluation of Generalised Boosting Models, *Oxford Review of Education*, 41(5), 587-607. http://dx.doi.org/10.1080/03054985.2015.1074563

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.

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.

Fitz-Gibbon, C. T., & Vincent, L. (1994). *Candidates' performance in public examinations in mathematics and science.* A report commissioned by SCAA from the Curriculum, Evaluation and Management Centre, University of Newcastle Upon Tyne. London: School Curriculum and Assessment Authority.

Goldstein, H., & Cresswell, M. (1996). The comparability of different subjects in public examinations: a theoretical and practical critique. *Oxford Review of Education, 22*(4), 435-442.

Ofqual (2017). *Progression from GCSE to A level: Comparative Progression Analysis as a new approach to investigating inter-subject comparability*. Coventry: Ofqual. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/610077/Progression_from_GCSE_to_A_level_-_Comparative_Progression_Analysis_as_a_new_approach_to_investigating_inter-subject_comparability.pdf

Sutch, T. (2013). *Progression from GCSE to AS and A level, 2010*. Cambridge Assessment Statistics Report No. 69. Available from http://www.cambridgeassessment.org.uk/Images/153531-progression-from-gcse-to-as-and-a-level-2010-.pdf

Thorndike, R. L. (1947). *Research problems and techniques*. (Report No. 3). Washington DC: U.S. Government Printing Office.

Appendix

Table A1: Correlation of A level grade (2014) with GCSE subject grade and mean GCSE (2012). Same data as Figures 6 and 7. Source: National Pupil Database, Department for Education.

Subject	Type	Subject grade	Mean GCSE	N
EngLit	Humanities	0.59	0.71	33795
Geography	Humanities	0.60	0.68	24014
History	Humanities	0.59	0.69	36513
Psychology	Humanities	0.54	0.62	2475
RS	Humanities	0.53	0.64	12550
French	MFL	0.57	0.64	6262
German	MFL	0.62	0.65	2468
Spanish	MFL	0.55	0.61	4576
Biology	Science	0.60	0.70	34149
Chemistry	Science	0.55	0.64	30651
Maths	Science	0.54	0.57	35149
Physics	Science	0.58	0.66	20548

Table A2: Original (Ofqual, 2016, Table 2) and adjusted correlations between GCSE and A level

	Original correlation	Adjusted correlation
English	0.57	0.462
English Language	0.57	0.466
English Literature	0.57	0.460
Biology	0.60	0.531
Chemistry	0.57	0.531
Physics	0.59	0.544
*Maths**	*0.56*	*0.560*
French	0.59	0.552
German	0.64	0.570
Spanish	0.56	0.523
Fine Art	0.61	0.458
Geography	0.62	0.473
History	0.59	0.437
Religious Studies	0.54	0.410

* Maths is italicised because it was the basis for the adjustments in the other subjects.

Table A3: Simulated data: A level grade distributions within GCSE grade – Humanities

	A level							
GCSE	A*	A	B	C	D	E	U	#cands
A*	24.1%	38.5%	24.6%	9.9%	2.5%	0.3%	0.1%	76633
A	5.1%	24.2%	32.3%	23.8%	11.2%	2.5%	0.9%	81288
B	1.1%	9.9%	23.1%	28.4%	23.8%	8.6%	5.1%	39206
C	0.1%	2.4%	9.9%	19.3%	30.1%	17.7%	20.5%	9506

Table A4: Simulated data: A level grade distributions within GCSE grade – Science

	A level							
GCSE	A*	A	B	C	D	E	U	#cands
A*	17.4%	40.6%	28.2%	10.8%	2.9%	0.2%	0.0%	128541
A	1.1%	13.8%	30.7%	30.6%	19.7%	3.7%	0.4%	115102
B	0.1%	1.9%	11.4%	25.6%	39.1%	16.8%	5.2%	21329
C	0.0%	0.0%	1.1%	8.5%	31.3%	29.1%	30.1%	742

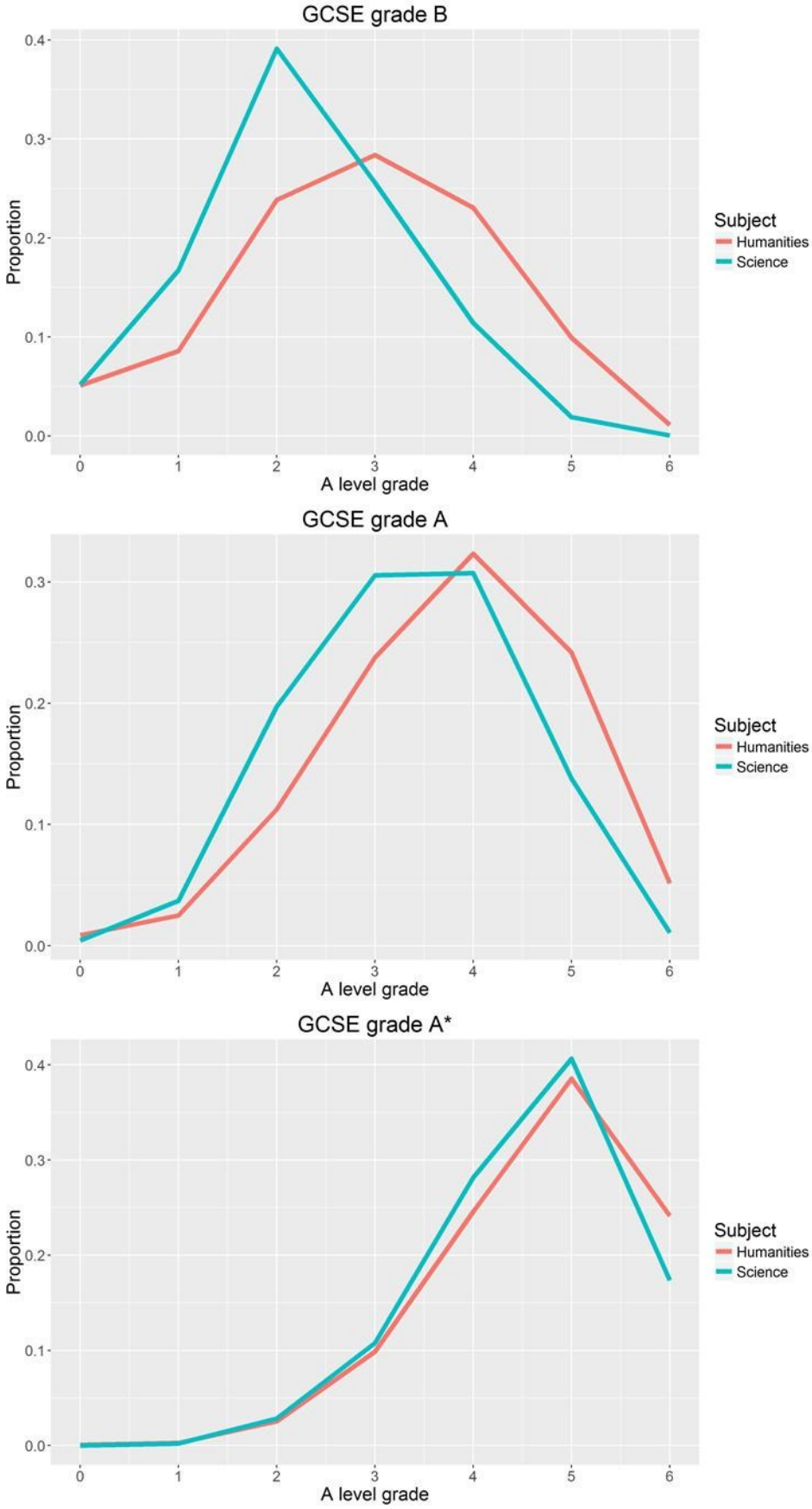

Figure A1: Simulation 2: Comparative Progression