# The effect of subject choice on the apparent relative difficulty of different subjects

**Tom Bramley** Research Division

## Introduction

The work presented here was prompted by a survey carried out by the Office of Qualifications and Examinations Regulation (Ofqual), of opinions on whether the grading of high-stakes academic examinations in England taken at age 16, for General Certificate of Secondary Education (GCSE), and at 18, for General Certificate of Education Advanced level (A level), should attempt in some way to make the different subjects equally 'difficult'. Ofqual published a suite of working papers to inform the debate[1], the second of which (Ofqual, 2015a) was a review of the United Kingdom literature on the topic. In our response to this survey (Cambridge Assessment, 2016), we expressed the view that while:

> *... in a small number of specific cases there may be some reasons for providing decision-makers with an indication of differences in subject difficulty, these are generally substantially outweighed by a much larger number of arguments against taking any of the options outlined by Ofqual to control for inter-subject comparability.* *(p.2)*

One of those arguments – the particular topic of this article – concerns whether it is valid to calculate statistical measures of relative subject difficulty based on the examinee-by-subject matrix containing the grades of each examinee on each subject in a particular examination session (For example, all GCSEs taken in the June 2016 session). There are several different methods of varying complexity that can be used to do this (see Coe, 2007). All of them face the same problems of first defining what is meant by 'difficulty', and second of dealing with the fact that the matrix of data to be analysed contains a large amount of missing data – the grades of examinees on subjects that they did not take. The non-random nature of this missing data (created by the fact that students only choose a subset of the possible subjects) makes the calculation of any statistical adjustment somewhat problematic. It is also likely to make subjects that measure something different to the majority of other subjects appear easier. These two claims are illustrated in this article with a simple example using simulated data.

## Simulated data

Consider a scenario where only four subjects are available: Mathematics, Physics, Chemistry and Art. Assume that in the entire cohort of potential examinees that Mathematics, Physics and Chemistry are highly correlated with each other, but less so with Art – for example with a correlation matrix as in Table 1.

**Table 1: Correlation matrix of scores for simulations (all potential examinees)**

| | *Maths* | *Physics* | *Chemistry* | *Art* |
|---|---|---|---|---|
| Maths | 1.00 | 0.90 | 0.90 | 0.50 |
| Physics | | 1.00 | 0.90 | 0.50 |
| Chemistry | | | 1.00 | 0.50 |
| Art | | | | 1.00 |

The scores of 10,000 examinees were simulated to yield the above correlation matrix (scores in each subject normally distributed with a mean of 0 and a standard deviation (SD) of 1). The scores were converted to grades on an A* to G scale giving a value of 8 to A* and 1 to G, such that the overall distribution was roughly the same in each subject, and reasonably realistic (in fact it matched the national distribution of GCSE Mathematics grades in 2012[2]). Treating the grades as numeric variables, Tables 2 and 3 give the descriptive statistics and correlations among the grades in the different subjects.

**Table 2: Summary of simulated grade distribution (all potential examinees)**

| *Variable* | *N* | *Mean* | *SD* | *Minimum* | *Maximum* |
|---|---|---|---|---|---|
| MathGrade | 10,000 | 4.66 | 1.80 | 0 | 8 |
| PhysGrade | 10,000 | 4.67 | 1.81 | 0 | 8 |
| ChemGrade | 10,000 | 4.67 | 1.80 | 0 | 8 |
| ArtGrade | 10,000 | 4.67 | 1.81 | 0 | 8 |

**Table 3: Correlation matrix of simulated grades (all potential examinees)**

| | *Maths* | *Physics* | *Chemistry* | *Art* |
|---|---|---|---|---|
| Maths | 1.00 | 0.87 | 0.87 | 0.48 |
| Physics | | 1.00 | 0.87 | 0.48 |
| Chemistry | | | 1.00 | 0.48 |
| Art | | | | 1.00 |

We now define 'subject difficulty' statistically such that all these subjects are *by definition equally difficult* because the grade distributions in each of them are the same for the entire cohort of potential examinees.

## Effect of subject choice

We now imagine a situation where each student chooses only two subjects to be examined in, and, for the sake of simplicity, each student

[1] Available online at https://www.gov.uk/government/collections/inter-subject-comparability-research-documents

[2] Cumulative percentage: A* 5.5%, A 15.5%, B 30.2%, C 58.7%, D 77.6%, E 86.7%, F 93.9%, G 98.2%.

chooses their best two subjects (according to the original simulated scores). Tables 4 and 5 show the new descriptive statistics and correlations for the 'observed data'.

**Table 4: Summary of simulated grade distribution (after examinees have chosen their two best subjects)**

| *Variable* | *N* | *Mean* | *SD* | *Minimum* | *Maximum* |
|---|---|---|---|---|---|
| MathGrade | 5,016 | 5.21 | 1.66 | 0 | 8 |
| PhysGrade | 4,956 | 5.19 | 1.65 | 0 | 8 |
| ChemGrade | 5,018 | 5.18 | 1.67 | 0 | 8 |
| ArtGrade | 5,010 | 5.40 | 1.55 | 0 | 8 |

**Table 5: Correlation matrix of simulated grades (after examinees have chosen their two best subjects)**

| | *Maths* | *Physics* | *Chemistry* | *Art* |
|---|---|---|---|---|
| Maths | 1.00 | 0.90 | 0.90 | 0.74 |
| Physics | | 1.00 | 0.90 | 0.76 |
| Chemistry | | | 1.00 | 0.74 |
| Art | | | | 1.00 |

We see from Table 4 that all subjects now appear around half a grade 'easier' (have a higher mean grade) than previously, but that Art is 0.2 of a grade easier than the other three subjects. It is interesting to note from Table 5 that Art is now much more highly correlated with the other subjects (the correlation has risen from 0.48 to 0.75). The effect of subject choice on the grades is easier to see if the six possible subject combinations are considered separately, as in Table 6.

**Table 6: Average grades for each combination of subjects**

| | | *N* | *Mean* | *SD* | *Minimum* | *Maximum* |
|---|---|---|---|---|---|---|
| Maths and Physics | MathGrade | 1,659 | 5.54 | 1.55 | 0 | 8 |
| | PhysGrade | 1,659 | 5.54 | 1.52 | 0 | 8 |
| | ChemGrade | 0 | | | | |
| | ArtGrade | 0 | | | | |
| Maths and Chemistry | MathGrade | 1,672 | 5.54 | 1.53 | 0 | 8 |
| | PhysGrade | 0 | | | | |
| | ChemGrade | 1,672 | 5.54 | 1.53 | 0 | 8 |
| | ArtGrade | 0 | | | | |
| Maths and Art | MathGrade | 1,656 | 4.56 | 1.69 | 0 | 8 |
| | PhysGrade | 0 | | | | |
| | ChemGrade | 0 | | | | |
| | ArtGrade | 1,656 | 5.43 | 1.56 | 0 | 8 |
| Physics and Chemistry | MathGrade | 0 | | | | |
| | PhysGrade | 1,668 | 5.52 | 1.55 | 0 | 8 |
| | ChemGrade | 1,668 | 5.52 | 1.53 | 0 | 8 |
| | ArtGrade | 0 | | | | |
| Physics and Art | MathGrade | 0 | | | | |
| | PhysGrade | 1,668 | 4.52 | 1.68 | 0 | 8 |
| | ChemGrade | 0 | | | | |
| | ArtGrade | 1,668 | 5.40 | 1.55 | 0 | 8 |
| Chemistry and Art | MathGrade | 0 | | | | |
| | PhysGrade | 0 | | | | |
| | ChemGrade | 1,677 | 4.50 | 1.71 | 0 | 8 |
| | ArtGrade | 1,677 | 5.37 | 1.55 | 0 | 8 |

Table 6 shows that the examinees choosing Art have achieved on average 0.8 to 0.9 of a grade better in Art than in the other subject they chose. The 'subject pairs' method of comparing subjects (e.g., Forrest & Smith, 1972; Coe, 2007) would therefore deem Art to be easier than the other three subjects. A more complex method, used in Scotland to calculate the statistical adjustment needed to align the difficulty of different subjects, is Kelly's method (Kelly, 1976; Coe, 2007). The adjustments represent the amount (in grades) that needs to be added or subtracted from each subject such that examinees on average achieve the same grade in that subject than they do on average in their other subjects.

**Table 7: Subject difficulty according to Kelly's method (after examinees have chosen their two best subjects)**

| | *N* | *Difficulty* |
|---|---|---|
| Maths | 5,016 | 0.211 |
| Physics | 4,956 | 0.216 |
| Chemistry | 5,018 | 0.219 |
| Art | 5,010 | -0.644 |

We see that Kelly's method has resulted in Mathematics, Physics and Chemistry being 'harder' and Art being 'easier'. Because this is such a simple scenario we can verify the Kelly result by applying it to Table 6. For example, adding 0.219 to the Chemistry mean and subtracting 0.644 from the Art mean of those taking Chemistry and Art gives approximately equal means of 4.72 and 4.73.

If the difficulty adjustments from Kelly's method were applied, when numeric grades in the two subjects were added together (e.g., to form an index of 'general academic ability' like the University and Colleges Admissions Service (UCAS) points score often used by UK universities as part of the student admission process) a student not taking Art would get a boost of ≈ 0.43, whereas a student taking Art would get a reduction of ≈ -0.43. In other words there would appear to be nearly a grade's worth (0.86) of difference between two students with the same raw points score who differed in whether or not they had taken Art. But of course we know from the simulation that (by definition) all the subjects were equally difficult.

## Discussion

In this example, the lower correlation between Art and the other subjects means that there is more 'regression to the mean' – hence for a given score (grade) in Art, the conditional mean score on Mathematics, Physics or Chemistry will be closer to the mean than it would for comparisons of pairs within those three subjects. Because in this simulation examinees are choosing their best subjects, scores on Mathematics, Physics and Chemistry for those examinees for whom Art is one of their top two subjects will be relatively lower (closer to the overall mean) than they are for examinees for whom Art is not one of their top two subjects. Conversely, examinees who are poor at one of Mathematics, Physics and Chemistry are more likely to be poor at the other two than they are to be poor at Art, making Art a more likely best or second best subject. This is illustrated in Figure 1.

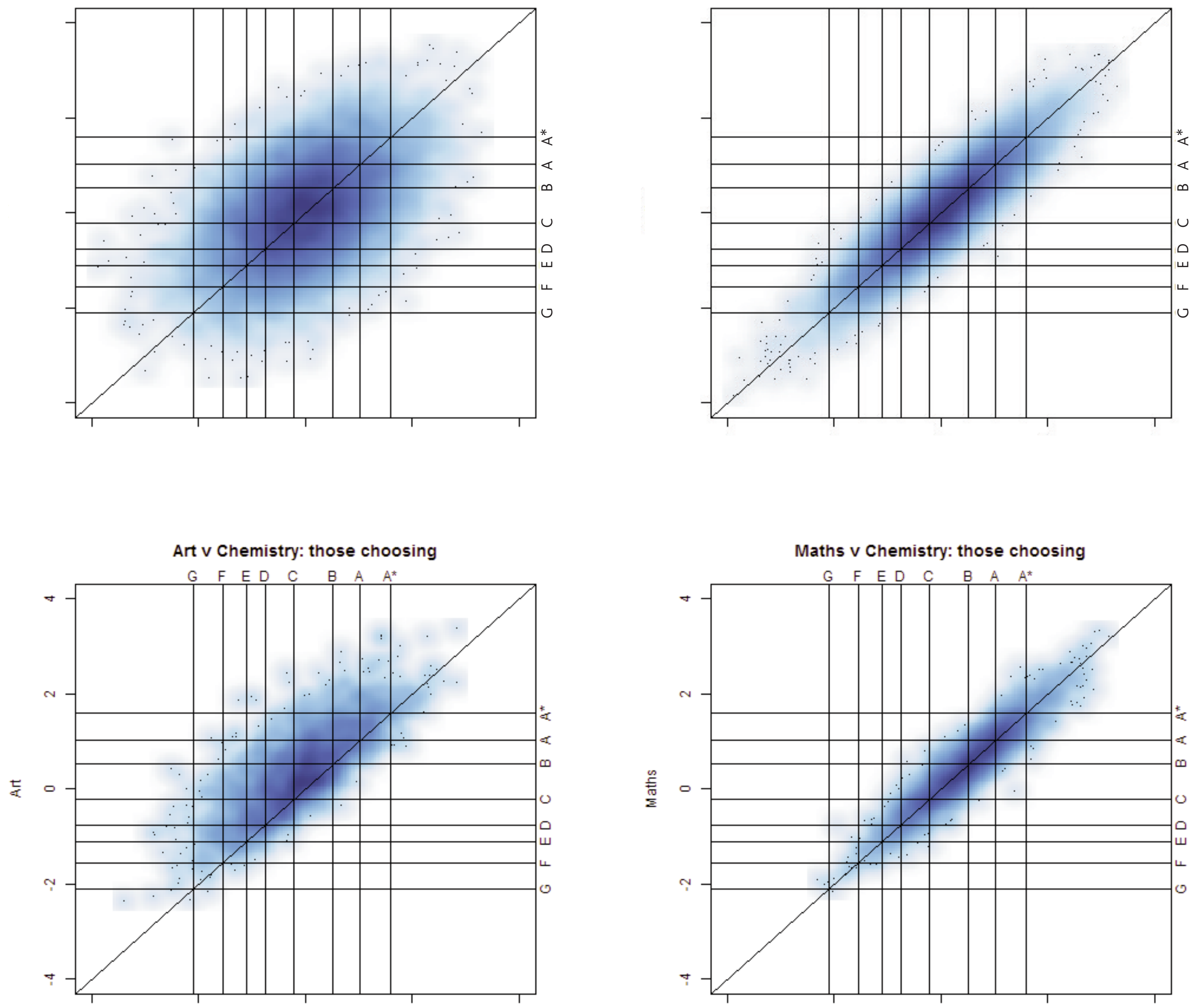

**Figure 1: Effect of choice for those choosing Chemistry and Art compared with those choosing Chemistry and Mathematics (axis values are the original normally distributed simulated scores with mean 0 and SD 1)**

This example highlights the problem in interpreting the data from all methods of measuring inter-subject comparability – the definition of difficulty is based upon a theoretical notion of every examinee having taken every subject. However, in practice examinees (or perhaps in some cases their schools) choose what subjects to take. We do not know, and can never know, what the results would look like if the entire GCSE or A level aged cohort took all the qualifications. So, while it is true that the same overall ranking of subjects by difficulty appears stable across time and even across jurisdictions (as noted in Ofqual [2015a] p.4), all methods for calculating a single adjustment for difficulty are making the same unjustifiable assumption that the 'missing' data (grades on subjects not taken) resembles the data at hand in the relevant way.

This assumption is brought out especially clearly in approaches that use Item Response Theory (IRT), as used, for example, by Coe (2008) and Ofqual (2015b). Here the different subjects have the role of different items (questions) on a single test, and the examinees have a single 'ability' that is supposed to reflect their probability of achieving a given grade in any particular subject. However, the relationship between differences in difficulty between items in a test on the one hand, and differences in difficulty between different academic subjects on the other, is only analogical (Bramley, 2011). Items within a test are usually selected to measure a construct that is explicitly defined via a specification (syllabus). Different academic subjects within a qualification family, such as GCSE, are not designed to measure any particular overall construct connected with the qualification family, so the construct has to be inferred retrospectively as something like 'general academic ability'. However, it is debatable whether there is any underlying ability that can usefully be said to underpin the wide range of subjects on offer at GCSE and A level.

Furthermore, subjects that are often taken together by large numbers of examinees (e.g., Mathematics and Sciences) are likely to dominate the retrospective definition of the construct. This presents two issues.

First, many minority subjects will correlate less well with mainstream subjects because there are far fewer common skills, content and understanding between them. Secondly, the self-selection effect is different for minority subjects. Students take minority subjects because they have a particular talent, interest or future requirement for them whereas it could be argued that, although they may also take English, Mathematics, Sciences and History for those reasons, they also take them because they are generally and widely considered good subjects for general progression in Higher Education and employment. This weakens the assumption that there will be a strong relationship between performance in different subjects and means that the subset of students taking minority subjects are often very successful in them.

In summary, when examinees choose subjects that measure something different (from mainstream subjects) on the basis of those examinees' strengths in, and preferences for, those subjects, then it is very likely they will appear easier. Psychometricians have cautioned against the dangers of making statistical adjustments to allow for differences in question difficulty in scenarios where choice of questions is allowed within a single examination (e.g., Wang, Wainer, & Thissen, 1995). It is clearly far more problematic to adjust for differences in difficulty at the subject level.

The simulation reported here resembles A level more closely than GCSE since at A level examinees usually choose three subjects (albeit from a much wider range of possibilities). At GCSE, examinees usually choose eight to ten subjects with Mathematics and English taken by virtually all examinees, and a small subset of other subjects taken by large numbers, meaning that these subjects form an effective 'anchor' setting the scale by which the relative difficulty of less popular subjects is determined. But the above conclusion should still hold: less popular subjects that correlate worse with the anchor will appear easier than they really are, if people choose them based on their ability in those subjects.

Of course, the simulation described here greatly oversimplifies the reality. Not only do examinees have a wider choice of subjects, they do not know beforehand which ones they will score best in, and even if they did they might need to take one of their weaker subjects in order to follow their desired future academic or employment path. The simulation could of course be extended to make it resemble more closely the actual situation at GCSE or A level. One sophisticated approach to this would be that of Korobko, Glas, Bosker, and Luyten (2008) who build statistical models allowing for both multidimensionality (of examinee ability) and non-random subject choice. But the purpose of this very simplified simulation was merely to illustrate the point that multidimensionality and non-random choice of subjects can lead statistical methods for measuring differences in subject difficulty towards the wrong answer. Perhaps the question is where the burden of proof should lie – with those who argue for the use of statistical adjustments to align subjects in terms of difficulty (to show that the example in this article is exaggerated or irrelevant); or with those who argue against (to show that the effect demonstrated here is also likely to apply with more realistic data).

## References

Bramley, T. (2011) Subject difficulty – the analogy with question difficulty. *Research Matters: A Cambridge Assessment publication*. Special Issue 2, 27–33.

Cambridge Assessment (2016). *A Cambridge Assessment response to Ofqual's subject comparability reports*. Cambridge, UK: Cambridge Assessment.

Coe, R. (2007). Common examinee methods. In P. E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. (pp.331–367). London: Qualifications and Curriculum Authority.

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

Forrest, G. M., & Smith, G. A. (1972). *Standards in subjects at the Ordinary level of the GCE, June 1971*. Occasional Publication 34. Manchester: Joint Matriculation Board.

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

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

Ofqual (2015a). *Inter-subject comparability: a review of the technical literature*. *ISC Working Paper 2*. Coventry: Office of Qualifications and Examinations Regulation.

Ofqual (2015b). *Inter-subject comparability of exam standards in GCSE and A Level*. *ISC Working Paper 3*. Coventry: Office of Qualifications and Examinations Regulation.

Wang, X.-b., Wainer, H., & Thissen, D. (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice. *Applied Measurement in Education*, *8*(3), 211–225.