On the impact of aligning the difficulty of GCSE subjects on aggregated measures of pupil and school performance

Tom Benton Research Division

Introduction

It is empirically demonstrated that adjusting aggregated measures of either student or school performance to account for the relative difficulty of General Certificate of Secondary Education (GCSE) subjects makes essentially no difference. For either students or schools, the correlation between unadjusted and adjusted measures of performance exceeds 0.998. This indicates that suggested variations in the difficulty of different GCSE subjects do not cause any serious problems either for school accountability, or for summarising the achievement of students at GCSE.

Data source

The analysis in this article is based upon data from the National Pupil Database (NPD) provided by the Department for Education (DfE). In particular the analysis is based upon the GCSE results[1] of all students in Year 11 in England in the academic year 2014/15. Only full GCSE qualifications taken by at least 50 pupils were included within analysis and only each student's best grade in any given subject was retained. Thus the final data set included over 4.5 million GCSE grades from around 600,000 students across 83 GCSE subjects.

Analysis of the impact of adjustments on mean GCSE scores

One simple way to summarise a student's GCSE achievement is to convert their grades to numbers (A*=8, A=7, B=6, C=5, D=4, E=3, F=2, G=1, U=0) and then to calculate their mean grade across all of the GCSEs they took. These summary measures can be averaged across pupils within a school to provide a simple measure of school performance. This section considers the impact of using a particular statistical method to adjust these summary measures.

For the purposes of this analysis, GCSE scores were adjusted using the Kelly method (see Bramley, 2014, for a brief description). This method has been historically used by the Scottish Qualifications Authority (SQA) to rank the difficulty of different Scottish Highers. In essence this method defines a subject as easy if the candidates taking it tend to achieve higher grades in this subject than in their other subjects. On the basis of this assumption, the method is designed to calculate adjustments to grades so that, across the group of pupils taking a particular subject, their mean grade in that subject will equal the average of the mean grades they achieve in all of their other subjects.

The Kelly adjustments for the 83 subjects included in the analysis are shown in Table 1. This shows, for example, that under the assumptions of the method, the "easiest" subject is GCSE Polish. The Kelly rating for this subject is calculated using the fact that the average grade achieved in Polish is 6.9 compared to an average grade of 4.5 across all other GCSEs taken by the same candidates[2]. For this reason the Kelly method suggests an adjustment of subtracting 2.4 from the grades achieved in Polish which (after adjustments have also been made to all other subjects) makes the means match up.

Of course, the fact that so many minor Modern Language GCSEs are amongst those rated "easy" by Table 1 immediately reveals a weakness with the statistical method. It is suspected that many of the candidates taking these subjects are native speakers. For this reason, their tendency to do better in these GCSEs than in others is not necessarily an indication of the GCSEs being easy at all but rather a result of their particularly strong aptitude for the subject. However, notwithstanding this obvious weakness in the Kelly method, it is still of interest to see the impact of applying the Kelly adjustments to all GCSEs on the overall summary measures of achievement.

The mean GCSE score both before and after applying the Kelly adjustments noted in Table 1 were calculated for each pupil in the data set. Across all students, the correlation between these measures was 0.998[3]. To get a measure of overall school performance, the mean of both of these measures was taken across all pupils within each school. The correlation between the school means for the two measures was 0.999 across all 5,236 schools in the analysis, as well as across the 2,928 schools with at least 100 pupils.

To illustrate these findings further, a random sample of 10 schools with between 100 and 200 pupils was selected. The differences between the adjusted and unadjusted measures are illustrated for these schools in Figure 1. The left hand side of the chart compares the measures at pupil level (restricted to students taking at least five GCSEs) whilst the right hand side compares the measures at school level. A line representing equality between the two measures is included in each chart. Within the data used for these charts the correlations between the measures are 0.998 and 0.999 at pupil and school level respectively[4].

At pupil level, there are no very large differences between the measures. In fact there are only five pupils (out of 1,289) where the difference exceeds 0.4 grades and only one where the difference exceeds 0.5. In these cases, the differences are explained by the fact that all five of

1. Since this analysis is based upon the initial unamended version of the NPD, the GCSE results included will not account for changes to students' grades made as part of the Enquiries About Results (EARs) process. Also note that GCSEs taken by this group prior to June 2015 (i.e., early entries) were included within the analysis.
2. Calculations restricted to candidates taking at least two GCSEs.
3. The same value for the correlation was found when analysis was restricted to pupils who had taken at least five GCSEs.
4. Thus matching the correlations reported for the national data.

Table 1: Kelly difficulty ratings for 83 GCSE subjects sorted from lowest ("easiest") to highest ratings

Rank	Subject	Number of candidates	Kelly Rating	Rank	Subject	Number of candidates	Kelly Rating
1	Polish	4,080	-2.38	43	ICT	99,160	-0.05
2	Turkish	1,558	-2.31	44	Dance	11,982	-0.05
3	Portuguese	2,045	-1.87	45	Home Economics: Food	8,623	-0.03
4	Dutch	396	-1.60	46	Physics	123,822	0.01
5	Persian	422	-1.52	47	Science (Core)	371,451	0.01
6	Russian	2,098	-1.28	48	Methods in Mathematics	12,438	0.01
7	Modern Hebrew	441	-1.08	49	Chemistry	124,507	0.02
8	Modern Greek	479	-0.94	50	Biology	127,778	0.03
9	Art & Design (Photography)	22,080	-0.90	51	Electronics	538	0.05
10	Chinese	3,355	-0.82	52	Applications of Mathematics	12,179	0.09
11	Gujarati	597	-0.75	53	Additional Science	304,991	0.09
12	Italian	3,985	-0.74	54	Office Technology	13,969	0.12
13	Urdu	4,209	-0.71	55	D&T Product Design	37,870	0.12
14	Art & Design (3D Studies)	2,156	-0.63	56	Sociology	21,336	0.13
15	Arabic	3,167	-0.63	57	Statistics	51,901	0.14
16	Applied Art & Design	874	-0.58	58	Music	43,519	0.16
17	Home Economics: Textiles	296	-0.55	59	D&T Electronic Products	7,882	0.16
18	Art & Design (Textiles)	7,692	-0.55	60	Geography	211,167	0.21
19	Art & Design	87,940	-0.47	61	D&T Engineering	289	0.23
20	Bengali	897	-0.43	62	Classical Greek	1,191	0.25
21	Art & Design (Fine Art)	51,786	-0.41	63	Other Classical Languages	506	0.28
22	Japanese	865	-0.40	64	D&T Graphic Products	31,779	0.28
23	Art & Design (Graphics)	7,440	-0.37	65	History	228,674	0.28
24	Punjabi	794	-0.34	66	Business Studies: Single	74,023	0.28
25	English Language & Literature	69,086	-0.33	67	Latin	8,297	0.31
26	Film Studies	6,971	-0.31	68	Environmental Science	2,721	0.33
27	D&T Textiles Technology	24,177	-0.31	69	Spanish	85,138	0.33
28	Home Economics: Child Devt	18,096	-0.30	70	D&T Systems & Control	2,976	0.34
29	D&T Food Technology	38,357	-0.28	71	Ancient History	980	0.40
30	Expressive Arts & Performance	3,343	-0.27	72	French	150,486	0.50
31	Media/Film/TV Studies	52,715	-0.24	73	Classical Civilisation	3,937	0.52
32	Performing Arts	6,256	-0.23	74	Psychology	15,961	0.53
33	Drama & Theatre Studies	71,340	-0.15	75	German	52,677	0.54
34	English Literature	407,758	-0.14	76	Economics	9,444	0.56
35	PE/Sports Studies	110,846	-0.14	77	Humanities: Single	8,389	0.57
36	Mathematics	549,695	-0.12	78	Computer Studies/Computing	32,223	0.59
37	Social Science: Citizenship	20,792	-0.12	79	English Studies	720	0.65
38	Geology	638	-0.10	80	General Studies	9,341	0.74
39	Religious Studies	268,738	-0.09	81	Applied Engineering	6,358	0.85
40	Health & Social Care	7,178	-0.08	82	Astronomy	2,320	1.06
41	English Language	307,818	-0.07	83	Law	2,214	1.19
42	D&T Resistant Materials	51,017	-0.06				

these pupils took Polish GCSE. As noted earlier, the statistical method used to calculate subject difficulty may be particularly inappropriate for Minor Language GCSEs and so these adjustments may not be valid in any case. However, more importantly, the analysis shows that, even when such subjects are included, the impact of statistically adjusting the difficulty of subjects is almost zero with the ranking of students remaining largely unaffected.

There are two reasons for this: First, it is because the differences between subjects in terms of difficulty are dwarfed by differences in the abilities of pupils across the population. For example, once we account for the number of candidates taking each subject, the standard deviation of the adjustments (Table 1) that will be applied to individual GCSE grades is 0.25. This compares to a standard deviation of 1.6 in the unadjusted mean GCSE scores of pupils. In addition, most pupils take a range of subjects meaning that these adjustments will tend to average out. Secondly, because most pupils will take English, Mathematics and at least one Science GCSE, this ensures some comparability between mean GCSE grades.

On the right hand side of Figure 1 we can see that the rank order of schools is almost entirely preserved regardless of whether adjustments are applied, with the only changes in rank order being amongst schools with extremely similar ratings on both measures[5]. This demonstrates how adjusting for subject difficulty makes essentially no difference to this method of quantifying school performance.

Analysis of the impact of five grade A*-C performance measures

Up to this point this article has only considered measures of performance based on averaging GCSE grades across subjects. However, many school performance measures are based on the percentage of students achieving above some given threshold. Historically, there has been considerable focus upon the percentage of pupils within a school achieving at least five good GCSEs – that is,

5. In fact, to the naked eye only nine data points are visible on the right hand side of the chart due to the fact that the points for 'School 2' and 'School 10' coincide more or less precisely.

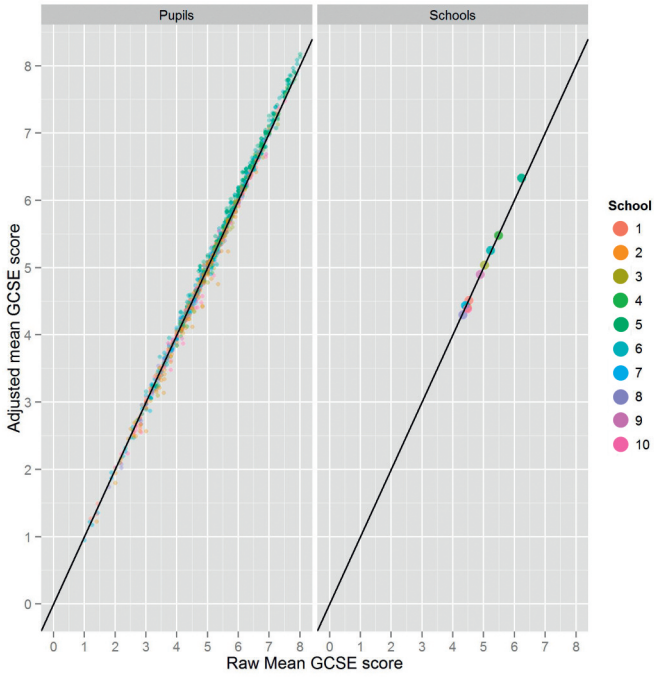

Figure 1: A comparison of adjusted and unadjusted school and pupil GCSE performance measures for a random selection of 10 schools

at grades A*-C[6]. This section estimates what the impact of imposing a statistically defined definition of inter-subject comparability upon GCSEs might be upon this measure.

As the Kelly method only provides adjustments to mean grades, it does not provide an appropriate tool for this analysis. Instead we use an alternative approach: First, we split all pupils into 10 groups (deciles) dependent upon their overall mean GCSE grade. Next, across all subjects combined, we calculate the percentage of GCSEs that are achieved at grade A*-C within each decile. For example, just less than 3% of GCSEs taken by pupils in the lowest decile are awarded grades A*-C compared to 72% for pupils in the fifth decile and 99.96% for the highest decile. Using this information we can predict the percentage of candidates that would be awarded grade A*-C in each subject if the relationship between deciles of achievement and grades awarded was consistent within every subject. This percentage can be compared to the number of candidates that were actually awarded grades A*-C.

Although the NPD does not include a record of the marks achieved by each candidate, it contains sufficient information for us to estimate for each individual pupil the probability that their grade would be awarded at least a grade C if all subjects were adjusted statistically. An example of this is given in Table 2 for GCSE German.

Table 2: Intended percentage of candidates achieving A*-C in GCSE German and cumulative percentage of candidates currently at each grade

Percentage to achieve grade A-C after alignment*	*Percentage of candidates achieving each grade or above*							
	A*	A	B	C	D	E	F	G
85.4	8.0	22.5	45.4	74.2	92.1	97.5	99.3	99.8

6. More recently the main performance measure has been the percentage of pupils achieving at least five A*-C grades *including* English and Mathematics. However, since this more recent measure places a restriction upon which subjects are used, it is of less interest for a piece of research concerned with the impact of differences in subject difficulty.

7. Calculated as 100*(85.4–74.2)/(92.1–74.2).

Predictions based upon performance deciles suggest that, under this definition of subject comparability, 85.4% of candidates should have been awarded grade A*-C. This compares to 74.2% who were actually awarded these grades. The cumulative percentage of candidates awarded each grade is shown in Table 2. Since only 74.2% of candidates achieved grade C or above, any adjustments to grading would leave these candidates within the grade A*-C band. In contrast, 92.1% of candidates achieved grade D or above so that it is clear no candidates with their current grades below D would be reclassified. However, in order to ensure that 85.4% of candidates achieved grade C or above overall it, would be necessary to reclassify some of those candidates who were awarded grade D to grade C. In fact, the top 62.6% of these candidates[7] should be reclassified. On this basis we can say that:

- all candidates currently awarded grades A*-C in German would have a 100% chance of being awarded grades A*-C after statistical aligning of grading standards across subjects;
- all candidates currently awarded grades E, F, G and U in German would have a 0% chance of being awarded grades A*-C after adjustment; and
- candidates currently awarded grade D would have a probability of 62.6% of being awarded grade A*-C after adjustment.

The above calculations were completed for each GCSE subject. Using the probabilities calculated in this way it was possible to calculate the overall probability that each individual student would achieve at least five A*-C grades[8]. By averaging these probabilities across all pupils within a school, it was then possible to estimate the percentage of pupils that would achieve at least five A*-C grades if statistical alignment of GCSE subjects was implemented. This can be compared to the current percentage that actually achieved five A*-C grades.

Figure 2 shows this comparison for all schools with at least 100 pupils. As can be seen, adjusting grading standards to account for any (supposed)

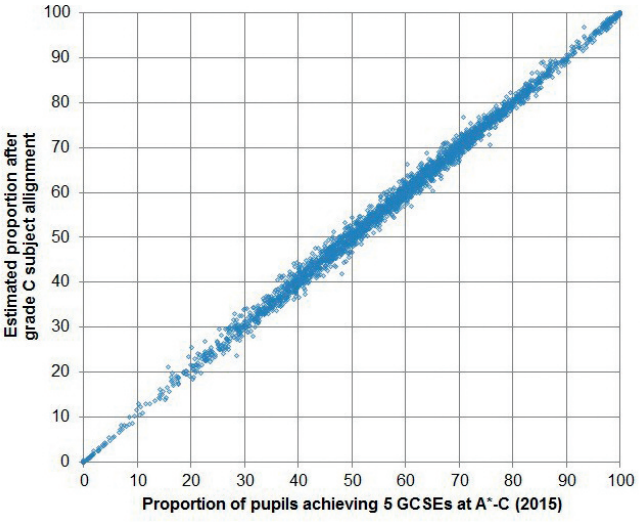

Figure 2: A comparison of adjusted and unadjusted estimates of the percentage of Year 11 pupils in each school who achieve at least 5 A*-C grades at GCSE

8. The process for doing this was fairly complicated and is not described in full in this article. Briefly, it required an assumption of independence between a pupil having their grade adjusted in one subject and having their grade adjusted in another. In essence, this implies an assumption that pupils' marks in different subjects are independent given their grades. Although this assumption is unlikely to hold precisely, given that grades capture nearly all of the useful information in marks, it provides a reasonable starting point. Calculations then treat the number of A*-C grades achieved by each candidate as the sum of independent Bernoulli trials which will (by definition) follow a Poisson binomial distribution.

differential subject difficulty make almost no difference to the ranking of schools. Overall there is a correlation of 0.998 between the original percentage of candidates achieving five A*-C grades and the estimated percentage after adjustments. Furthermore, there are only 8 schools (out of 2,928) where the difference exceeds 5 percentage points and none where it exceeds 10 percentage points. This again indicates that adjustments to grading to account for variations in subject difficulty are unlikely to have any substantial effect upon school performance measures.

Reference

Bramley, T. (2014) Multivariate representations of subject difficulty. *Research Matters: A Cambridge Assessment publication*, *18*, 42–47. Available online at: http://www.cambridgeassessment.org.uk/Images/174492-research-matters-18-summer-2014.pdf.

Statistical moderation of school-based assessment in GCSEs

Joanna Williamson Research Division

Introduction

School-based assessment (SBA) such as coursework is included in high-stakes qualifications around the world. In the United Kingdom (UK) for example, selected General Certificate of Secondary Education (GCSE) and General Certificate of Education (GCE) Advanced level (A level) examinations include SBA components[1] alongside examination components[2]. Moderation is required in order to address the question of comparability of SBA marks across different centres. Under current procedures for GCSEs and A levels (see Gill, 2015), moderators re-mark a sample of each centre's SBA work. The awarding body uses the relationship between the moderator mark and centre mark (in the re-marked sample) to decide what adjustment, if any, should be applied to that centre's SBA marks.

Statistical moderation is an alternative form of moderation that calibrates and/or monitors the marks of an assessment on the basis of a statistical relationship with another assessment. Its validity depends on the two assessments having a strong relationship in terms of both assessment content and candidate performance, but they need not measure precisely the same construct. In the context of SBA, the most common statistical moderation practice is to calibrate candidate marks on SBA component(s) using marks from the exam component(s) of the same overall assessment. The motivation for statistical moderation is to preserve information about candidates' SBA performance (such as their ranking within the centre) whilst acknowledging that marking may vary between centres. Statistical moderation removes the absolute meaning of SBA marks, and calibrates them to a new scale that is common to all candidates, that is, the exam component.

During recent reforms of GCSEs and A levels, the Office of Qualifications and Examinations Regulation (Ofqual) proposed the use of statistical moderation in GCSE assessment (Ofqual, 2015a). Previous research by Taylor (2005), using results data from the AQA awarding body, found that statistical moderation generally adjusted marks downward, since SBA marks for GSCE and A level were usually higher than exam marks. The study also found that many candidates would have been awarded different grades under statistical moderation, and that there was a disappointing "absence of any pattern, across different specifications" in terms of statistical moderation outcomes (Taylor, 2005, p.51). The present article outlines methods of statistical moderation that are used in jurisdictions around the world, and explores the effect of applying these methods to results data from three Oxford, Cambridge and RSA Examinations (OCR) GCSEs. This involved statistically moderating all SBA components, aggregating SBA marks with exam marks, and then calculating candidates' statistically moderated final grades from these aggregate scores. Analysis focuses on comparing the statistically moderated results to operational results (moderated under existing, non-statistical procedures) in terms of marks, grades, and the rank-order of candidates and centres.

Methods of statistical moderation

Statistical moderation is a form of assessment linking, where "the goal is to put scores from two or more tests on the same scale – *in some sense*." (Kolen & Brennan, 2004, p.423). Given a suitable pair of assessments (e.g., SBA unit and exam unit), there exist multiple ways to statistically moderate. Table 1 shows the methods investigated in this article: the first four methods are variations of linear scaling, the next two are forms of curvilinear scaling and the final method is rank mapping. Of these, the most commonly used method is linear scaling that matches the mean and standard deviation (SD) of SBA marks within each centre to those of the exam marks (Method 2). The three simplest linear methods (1, 2 and 4) and rank mapping (Method 7) were previously investigated by Taylor (2005). Despite different statistical procedures, many of the methods share common outcomes, as summarised in Table 2.

1. Recent qualification reforms have reduced the use of SBA in GCSE assessment (Ofqual, 2015b). Of the 23 'new' GCSEs (9–1) ready for first teaching in September 2015 or 2016, 7 contain SBA components.
2. In GCSE and A level, examination components are always externally set and assessed. They are usually written exams.