Relative difficulties of examinations at GCSE: an application of the Rasch model

Working Paper
31 March 2006

Robert Coe
Curriculum, Evaluation and Management (CEM) Centre
University of Durham, UK

www.cemcentre.org

Introduction

Methods that compare outcomes achieved by the same candidate in different examinations have a number of general advantages and disadvantages in monitoring comparability. On the one hand, by comparing performance in a particular subject with that same candidate's performance in their other subjects, they provide an obvious and superficially compelling control for differences in the general characteristics of individual candidates. These methods generally require no other data to be collected, and no matching with other datasets; hence they may be possible when other methods are not. On the other hand, comparison groups may not be typical of all those who take a particular examination, comparisons between some subjects or examinations may be inappropriate, and the results of such studies may be sensitive to the particular population or subgroup chosen. Any systematic differences in the quality of teaching or general educational provision across examinations could be mistaken for differences in difficulty.

The main methods that have been used are, briefly:

Subject pairs (triples, etc). These methods are simple to do and understand, but seriously limited by the representativeness of candidates taking a particular combination. They can also give inconsistent orderings of 'difficulty' (A>B & B>C, but A<C).

Kelly's (1976) method of concurrent comparisons is a major advance in that it allows the calculation of the grading difficulty of a particular subject to be based on all candidates who have taken that subject along with any other. However, it does require an assumption of unidimensionality (Goldstein and Cresswell, 1996) and estimates of relative difficulty of subjects seem to vary considerably for different subgroups (Sparkes, 2000).

An alternative approach, used here, is to apply the Rasch model to grade outcomes. In this approach, unidimensionality can be tested empirically, rather than simply assumed. The equality of grade intervals, both for a given grade gap across subjects and for intervals within the same subject, can also be tested. Each grade within each subject can be tested for model-fit, as can individual candidates. Identification of misfitting grades and persons provides meaningful and interesting information about them.

Methods

The Rasch model

The Rasch model provides a method for calibrating ordinal data onto an interval scale. Rasch assumes that the difficulty of items and the ability of persons can be measured on the same scale, and that the probability of a person achieving success on a particular item is determined by the difference between their ability and the difficulty of the item. In the Rasch model, these two are related by the logit function, the difference being equal to the log of the odds, and item difficulties and person abilities are estimated in logit units. Rasch's claim to provide an interval scale rests on the fact that the same difference between item difficulty and person ability anywhere on the scale corresponds to the same probability of success. For any two items of different difficulty, different persons will have different probabilities of success, but the odds ratio[1] for each person will be the same regardless of their ability, provided they fit the model.

Rasch analysis uses an iterative procedure to estimate item difficulties and person abilities for a given data set. It allows the fit of the model to be investigated and misfitting items and persons to be identified. For items, good fit with the model indicates that they are unidimensional (i.e. all measuring essentially the same thing) and discriminate appropriately (i.e. more able persons are more likely to be successful). For persons, good fit indicates that their relative probabilities of success on different items are in line with those of others in the population.

The goodness of fit of the Rasch model can be judged from the residuals, the difference between a person's response on a particular item and what would have been predicted by the model. Following Wright (????) these residuals are weighted and standardised for a particular item (or person) in two ways.

'Outfit' is the mean square of the residuals, divided by degrees of freedom, which can be interpreted as an overall measure of how well all responses to that item fit the Rasch model. However, this can be disproportionately influenced by extreme outliers, so another measure, the 'infit', is also used, in which residuals from persons whose ability is outside the range of discrimination for that item are weighted less. Where data fit the model well, subject to normal random error, values of both infit and outfit are expected to be close to 1. Values below 1 indicate better than expected fit, while values above 1 indicate a poor fit to the model. Linacre (2005b) advises that items with values between 0.5 and 1.5 are 'productive for measurement', between 1.5 and 2.0 are 'unproductive for construction of measurement, but not degrading', while infit or outfit greater than 2 'distorts or degrades the measurement system'.

In the context of GCSE examination data, each subject may be thought of as an 'item', although each subject has a number of levels of success (grades). Hence a partial credit model (Masters, ????) can be used, in which the difficulty of each grade within each subject is estimated separately. The current analysis was conducted using WINSTEPS (Linacre, 2005a) in which Masters' partial credit model is estimated using Joint Maximum Likelihood Estimation (Linacre, 2005b).

The process of estimating grade difficulties and person abilities in the Rasch model is iterative. Given some estimate of the abilities of the candidates who have taken a particular subject (based on their overall performance in their other subjects), we can examine the relationship between the probability of a particular grade being achieved and the ability of the candidate. We can use some kind of maximum likelihood procedure to select a value for the difficulty of the grade that best explains this pattern of achievement. Having estimated grade difficulties in this way, we can then refine our estimates of candidates' abilities in an exactly analogous way, selecting a value for each person's ability that best explains their pattern of achievement of grades of known difficulty. The process is then repeated using the latest estimates of difficulty and ability, until estimates converge.

Hence the estimate of the difficulty of a particular grade in a particular subject is based on all the candidates who have taken that subject with at least one other. The grade difficulty depends on the relative probabilities of that grade being achieved by candidates of different ability, as determined by their performance in all their subjects and taking into account the different difficulties of all the grades they have gained.

Data

The data used in this analysis were from the national pupil database for pupils in maintained mainstream secondary schools in England who took Key Stage 4 examinations in the summer of 2004. Most of these candidates will have been aged 16 at the time.

The original dataset contained 678,722 pupils, all of whom had been entered for at least one examination. However, a number of examinations had very small entries, making them unsuitable for reliable comparison, so those with fewer than 2,000 entries nationally were dropped from the analysis. A number of pupils (51,024) who had taken only these minority examinations were therefore lost, leaving data from 627,698 pupils and 79 examinations available for this analysis. In the final model with 34 examination subjects included, the analysis was based on the 615,800 candidates who had taken at least two of these 34 subjects.

Included in the dataset were the results of a number of different kinds of KS4 examinations. As well as the traditional GCSE subjects, there were also Vocational and Short Course GCSEs. These qualifications are awarded the same grades as traditional GCSEs (A*, A, B, C, D, E, F, G and U), but are allocated different points by the Qualifications and Curriculum Authority (QCA) to reflect the amounts of time typically spent on these courses. Vocational GCSEs are allocated double points, while Short Course GCSEs receive half the points of traditional GCSEs, for the same grade.

There were also a number of non-GCSE qualifications included in the dataset, such as Full and Part 1 GNVQs at both Foundation and Intermediate level.

Stages in the analysis

As the dataset was large, iterations could be quite slow, so a number of preliminary investigations were conducted with reduced samples. These established that when all subjects were included the model failed to converge within a suitable timescale, and the estimates derived indicated that some subjects were a very poor fit to the model.

The general intention was to arrive at a maximal list of subjects to be included in the model, subject to adequate fit. Two ways to do this were tried, first by starting with all subjects and throwing out those that did not fit well, and then by starting with a small core of well fitting subjects and progressively adding others. In practice the second of these strategies proved to be the most successful in enlarging the pool of subjects included.

Results

Subjects included

The starting point for the analysis was the group of 37 subjects with large entries (over 20,000 candidates). Ten of these subjects had either infit or outfit greater than 1.7 and were removed to produce a core group of 27 well fitting subjects. Other subjects were then progressively added to the set and checked for fit. The final number included was 34, all of which had both infit and

outfit below 2 and at least one grade category with outfit of 1.5 or less. The 34 subjects included, together with their fit statistics, are shown in Table 1.

The estimates of grade difficulty from the model with 27 subjects were compared with those for the same subjects in the full model with 34. Agreement was very close, with a correlation of 0.9999 between the 207 grade estimates in the two models.

Table 1. Fit statistics for the 34 subjects included in the analysis, ordered by Outfit.

Subject	Count	Infit	Outfit	Item-scale correlation
Short GCSE History	199014	1.65	1.64	0.81
GCSE Drama & Theatre Studies	92685	1.52	1.57	0.75
GCSE Spanish	52900	1.53	1.48	0.83
GCSE Information Technology	77122	1.46	1.44	0.80
Short GCSE Religious Education	23067	1.43	1.42	0.79
Voc GCSE Engineering	4443	1.42	1.42	0.74
GCSE Design/Tech & Graphic Prods	97641	1.41	1.42	0.79
GCSE Design/Tech & Resist. materials	103556	1.34	1.37	0.79
GCSE Religious Studies	120604	1.35	1.30	0.83
Voc GCSE Leisure & Tourism	10679	1.34	1.33	0.76
GCSE Sport/Physical Education Studies	121567	1.23	1.28	0.79
GCSE Design/Tech & Textiles Technology	51342	1.21	1.21	0.82
GCSE German	114759	1.21	1.19	0.84
GCSE French	284670	1.15	1.14	0.86
GCSE Latin	8620	1.11	1.02	0.74
GCSE Design/Tech & Food Technology	100630	1.10	1.11	0.84
GCSE Office Technology	24315	1.08	1.09	0.83
GCSE Home Economics: Child Development	28485	1.07	1.07	0.84
GCSE History	200629	1.05	1.03	0.87
Voc GCSE Science	8089	1.01	1.01	0.76
GCSE Media/Film/Television Studies	34271	1.00	1.01	0.84
GCSE Humanities single	15924	0.94	0.93	0.88
GCSE Business Studies	77061	0.91	0.91	0.86
GCSE English Literature	519701	0.84	0.88	0.85
GCSE Statistics	33204	0.85	0.86	0.84
GCSE Sociology	13348	0.86	0.85	0.87
GCSE Mathematics	586471	0.83	0.84	0.88
GCSE Geography	192699	0.80	0.80	0.89
GCSE Physics	42696	0.78	0.77	0.82
GCSE Chemistry	43303	0.76	0.75	0.83
GCSE Science: Single award	49943	0.74	0.76	0.84
GCSE Science: Double Award	471410	0.71	0.72	0.89
GCSE English	587791	0.64	0.69	0.89
GCSE Biology	44121	0.68	0.68	0.84

Overall these 34 subjects seem to have a reasonably good fit to the Rasch model. All item-scale correlations are 0.74 or higher and the overall reliability for the scale is estimated at 0.94[2].

A number of subjects are notable by their absence from this list. Creative subjects, such as GCSE music, art and design, fine art and performing studies, were included in the initial model (37 subjects with entries over 20,000), but did not fit well. In some cases the fit was marginal and they could perhaps have been included, at least for some grades, but the fit was never good.

There were also other classes of Key Stage 4 qualification included in the original data set, but not included in the final list of 34. For example, no GNVQ was able to fit the model. GNVQ Information Technology had a large entry (44,000 candidates) and was included in the initial 37, but had an infit of 3.8 and none of its grade categories (Fail, Pass, Merit, Distinction) fitted either. No other GNVQ had more than 6,000 entries nationally and none was able to be included in any model that converged and provided adequate fit statistics.

Vocational GCSEs were generally poorly fitting, though the 34 does include three vocational GCSEs: science, leisure and tourism, and engineering. Others (information technology, health and social care, and business) failed to fit adequately.

Short course GCSEs also tended not to fit well. Only history and RE were included in the final 34, with IT, RS and French among subjects with substantial short course GCSE entries (over 18,000), but unable to be included.

These 34 subjects may therefore be seen as fitting the model at least reasonably well. Certainly, if their fit were judged to be too low, it would be possible to produce a model with a smaller number of subjects, taken from the bottom end of Table 1. In other words, there is a substantial set of GCSE subjects which can be treated as unidimensional. In general, students who do well on one of them are likely to do well on the others. A single latent construct, general academic ability, forms the basis of all of them.

Misfitting grades

In the preliminary analyses, it became clear that the U grade category was almost always a very poor fit to the model. For example, from the initial group of 37 large entry subjects, only in French was the outfit for U below 2, and then only just at 1.9. It was also apparent that the relatively high estimated difficulty of the U grade was pushing the difficulties of the lower grades upwards, since the partial credit model seeks to preserve the order of categories (grades) within a subject. In other words, candidates awarded U grades were often no less able, as defined by their performance in other subjects, than those awarded G or F, and in some subjects they were

appreciably more able. The model was substantially improved by treating all U grades as missing.

The fact that the U grade fitted so poorly does seem to cast doubt on the widely adopted practice of treating U as the grade below G in allocating a numerical score to it, even if that score is a zero. For example, value-added analyses might be more appropriate with U grades omitted.

Some subjects also had individual grade categories that did not fit the model well. Fit statistics for all grades from the 34 subjects are shown in Table 2. In most cases poor fit was marginal, with outfit values close to the cut off of 1.5. The only grades with outfit above 2 were grade E in Latin (awarded to just 14 candidates nationally) and, perhaps more surprisingly, grade A* in Spanish (outfit = 2.2). A possible explanation for the latter would be if a significant proportion of the 6,000 candidates awarded this grade were native speakers of Spanish whose performance in their other subjects did not match their A* grade in Spanish, but this can be no more than speculation.

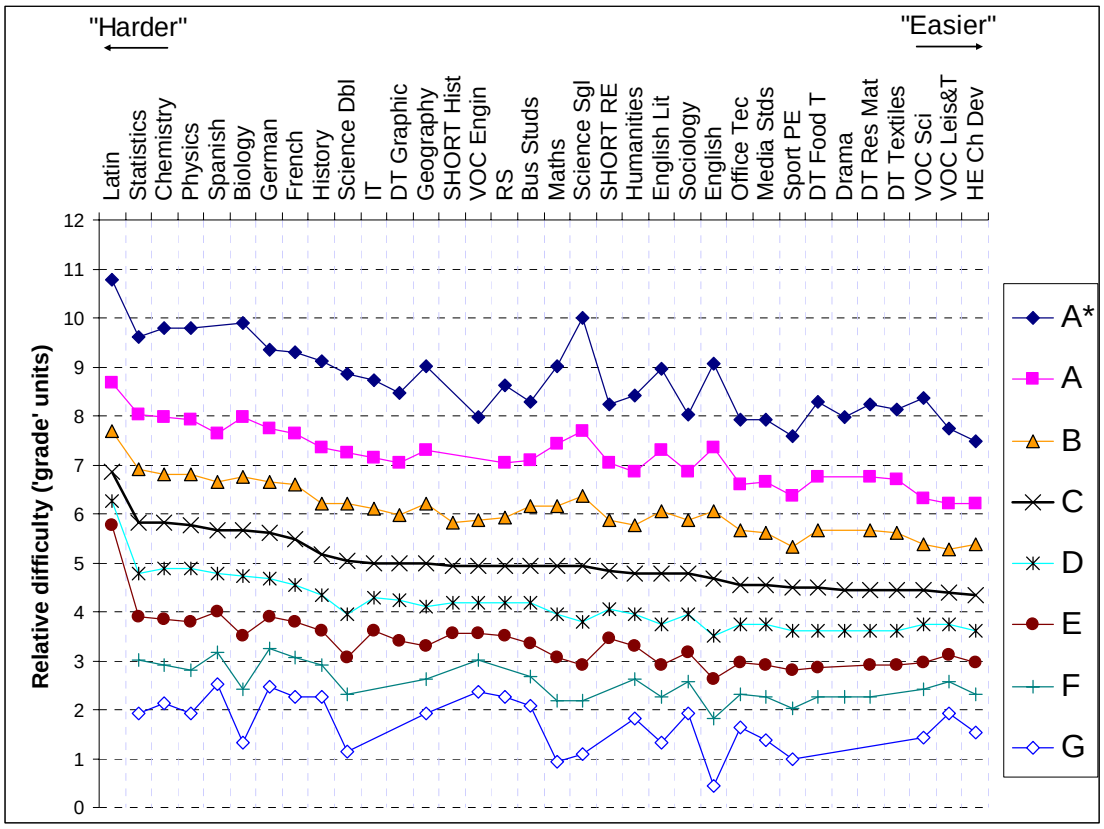
Table 2. Statistics for grades G to A* for 34 subjects

	G				F				E				D				C				B				A				A*			
	Count	Rasch measure	Std Error	Outfit	Count	Rasch measure	Std Error	Outfit	Count	Rasch measure	Std Error	Outfit	Count	Rasch measure	Std Error	Outfit	Count	Rasch measure	Std Error	Outfit	Count	Rasch measure	Std Error	Outfit	Count	Rasch measure	Std Error	Outfit	Count	Rasch measure	Std Error	Outfit
Latin	14	0.07	0.04	2.2	37	0.11	0.02	1.8	131	0.19	0.01	1.5	452	0.28	0.01	1.2	965	0.39	0.00	0.9	1488	0.54	0.00	0.6	2989	0.72	0.00	0.8	3301	1.10	0.01	0.9
Statistics	380	-0.51	0.01	0.7	937	-0.31	0.00	0.5	2343	-0.15	0.00	0.8	5560	0.01	0.00	1.0	11597	0.20	0.00	1.1	6149	0.40	0.00	0.8	4615	0.60	0.00	0.8	1801	0.89	0.01	0.9
Chemistry	93	-0.47	0.02	0.8	264	-0.33	0.01	0.7	678	-0.16	0.01	0.6	3113	0.03	0.00	0.8	8407	0.20	0.00	0.7	10617	0.38	0.00	0.6	12043	0.59	0.00	0.7	9023	0.92	0.00	0.9
Physics	60	-0.51	0.03	0.8	157	-0.35	0.01	0.5	582	-0.17	0.01	0.6	2950	0.03	0.00	0.9	8322	0.19	0.00	0.8	10381	0.38	0.00	0.7	12205	0.58	0.00	0.7	8963	0.92	0.00	0.8
Spanish	2495	-0.40	0.00	1.4	4136	-0.28	0.00	1.5	5964	-0.13	0.00	1.4	9058	0.01	0.00	1.2	10745	0.17	0.00	1.0	8047	0.35	0.00	1.1	7068	0.53	0.00	1.5	5986	0.78	0.00	2.2
Biology	113	-0.62	0.03	0.8	281	-0.42	0.01	0.4	747	-0.22	0.01	0.6	2917	0.00	0.00	0.8	8259	0.17	0.00	0.7	11685	0.37	0.00	0.7	12926	0.59	0.00	0.7	8094	0.94	0.00	0.7
German	4679	-0.41	0.00	1.3	8470	-0.27	0.00	1.4	12821	-0.15	0.00	1.2	20866	-0.01	0.00	1.2	30188	0.16	0.00	1.1	18394	0.35	0.00	1.0	12651	0.55	0.00	1.1	7471	0.84	0.00	1.4
French	14913	-0.45	0.00	1.2	27228	-0.30	0.00	1.3	39382	-0.17	0.00	1.2	53144	-0.03	0.00	1.1	60906	0.14	0.00	1.0	40642	0.34	0.00	0.9	29990	0.53	0.00	1.1	20638	0.83	0.00	1.4
History	8265	-0.45	0.00	1.2	13876	-0.33	0.00	1.2	19836	-0.20	0.00	1.0	26447	-0.07	0.00	0.8	37331	0.08	0.00	0.7	41543	0.27	0.00	0.9	36230	0.48	0.00	1.1	18463	0.80	0.00	1.1
Science Dbl	16759	-0.65	0.00	0.7	37769	-0.44	0.00	0.6	64728	-0.30	0.00	0.7	92511	-0.14	0.00	0.7	136695	0.06	0.00	0.8	64166	0.27	0.00	0.7	40750	0.46	0.00	0.7	20713	0.75	0.00	0.9
IT	4103	-0.43	0.00	1.6	6218	-0.31	0.00	1.7	8240	-0.20	0.00	1.5	12223	-0.08	0.00	1.3	19185	0.05	0.00	1.3	14189	0.25	0.00	1.2	9483	0.44	0.00	1.4	3740	0.73	0.00	1.2
DT Graphic	4119	-0.45	0.00	1.8	6441	-0.35	0.00	1.6	11163	-0.24	0.00	1.4	20407	-0.09	0.00	1.4	23530	0.05	0.00	1.3	16809	0.23	0.00	1.3	12619	0.42	0.00	1.4	2789	0.68	0.01	1.2
Geography	6820	-0.51	0.00	0.9	12165	-0.38	0.00	0.8	20041	-0.26	0.00	0.8	30070	-0.11	0.00	0.7	47026	0.05	0.00	0.7	33277	0.27	0.00	0.6	28029	0.47	0.00	0.8	16220	0.78	0.00	1.0
SHORT Hist	13741	-0.43	0.00	1.7	20755	-0.31	0.00	1.8	26578	-0.21	0.00	1.5	31321	-0.10	0.00	1.2	39455	0.04	0.00	1.2	33860	0.20	0.00	1.4	22846	0.37	0.00	1.7	10899	0.61	0.00	1.7
VOC8	576	-0.43	0.01	1.4	807	-0.31	0.01	1.5	932	-0.21	0.01	1.3	910	-0.10	0.01	1.3	743	0.04	0.01	1.3	379	0.21	0.01	1.4	99	0.38	0.03	1.7	8	0.59	0.09	1.5
RS	4403	-0.45	0.00	1.5	7706	-0.33	0.00	1.6	11277	-0.22	0.00	1.4	16396	-0.10	0.00	1.1	22633	0.04	0.00	0.9	25590	0.22	0.00	1.1	22063	0.42	0.00	1.3	11138	0.71	0.00	1.3
Bus Studs	2668	-0.48	0.00	1.0	5169	-0.37	0.00	0.8	8731	-0.25	0.00	0.8	14248	-0.10	0.00	0.9	22852	0.04	0.00	0.8	11789	0.26	0.00	0.7	8132	0.43	0.00	1.0	3584	0.65	0.00	1.2
Maths	23419	-0.69	0.00	0.8	54272	-0.46	0.00	0.7	89627	-0.30	0.00	0.8	101594	-0.14	0.00	0.8	132799	0.04	0.00	0.9	110666	0.26	0.00	0.9	51391	0.49	0.00	0.9	28531	0.78	0.00	1.0
Science Sgl	6689	-0.66	0.00	0.9	11886	-0.46	0.00	0.8	12879	-0.33	0.00	0.8	10033	-0.17	0.00	0.8	7204	0.04	0.00	0.7	1328	0.30	0.00	0.5	618	0.54	0.01	0.5	267	0.96	0.02	0.8
SHORT RE	1237	-0.44	0.01	1.8	1933	-0.32	0.00	1.8	3115	-0.23	0.00	1.5	3875	-0.12	0.00	1.2	5439	0.02	0.00	1.2	4559	0.21	0.00	1.2	2296	0.42	0.00	1.2	660	0.64	0.01	1.3
Humanities	1410	-0.53	0.00	1.0	1996	-0.38	0.00	1.0	2443	-0.26	0.00	0.8	2699	-0.14	0.00	0.6	3325	0.01	0.00	0.7	2518	0.19	0.00	0.9	1275	0.39	0.00	1.0	303	0.67	0.01	0.9
English Lit	12470	-0.62	0.00	0.9	27552	-0.45	0.00	0.7	52014	-0.33	0.00	0.7	85775	-0.18	0.00	0.7	135928	0.01	0.00	0.9	115631	0.24	0.00	1.0	70176	0.47	0.00	1.1	22830	0.77	0.00	1.1
Sociology	489	-0.51	0.01	1.0	843	-0.39	0.00	0.9	1414	-0.28	0.00	0.7	2336	-0.14	0.00	0.7	3789	0.01	0.00	0.7	2386	0.21	0.00	0.8	1595	0.39	0.00	0.9	503	0.60	0.01	1.0
English	15280	-0.78	0.00	0.8	34532	-0.53	0.00	0.5	65785	-0.38	0.00	0.5	114478	-0.22	0.00	0.6	152250	-0.01	0.00	0.7	116476	0.24	0.00	0.8	71186	0.48	0.00	0.9	23456	0.79	0.00	0.8
Office Tec	777	-0.56	0.01	1.2	1607	-0.44	0.00	1.0	2843	-0.32	0.00	1.1	4787	-0.18	0.00	1.1	7345	-0.03	0.00	1.0	3327	0.17	0.00	0.9	2575	0.34	0.00	1.1	1103	0.58	0.01	1.3
Media Stds	988	-0.61	0.01	1.2	2086	-0.45	0.00	1.0	3684	-0.33	0.00	1.0	6718	-0.18	0.00	0.9	8613	-0.03	0.00	0.9	6957	0.16	0.00	1.0	4258	0.35	0.00	1.1	1067	0.58	0.01	1.1
Sport PE	1508	-0.68	0.01	1.4	5363	-0.49	0.00	1.0	14699	-0.35	0.00	1.2	29147	-0.20	0.00	1.3	26058	-0.04	0.00	1.3	24472	0.11	0.00	1.3	15350	0.30	0.00	1.4	5290	0.52	0.00	1.3
DT Food T	3689	-0.59	0.00	1.6	7312	-0.45	0.00	1.3	12016	-0.34	0.00	1.2	19572	-0.20	0.00	1.1	26615	-0.04	0.00	1.0	15054	0.17	0.00	0.9	13528	0.37	0.00	1.1	3249	0.65	0.00	1.0
Drama	2022	-0.58	0.01	1.7	4211	-0.45	0.00	1.5	8226	-0.33	0.00	1.6	13706	-0.20	0.00	1.5	22566	-0.05	0.00	1.5	23332	0.14	0.00	1.7	14852	0.34	0.00	1.7	4053	0.59	0.00	1.4
DT Res Mat	4124	-0.61	0.00	1.6	7973	-0.45	0.00	1.5	14091	-0.33	0.00	1.5	23827	-0.20	0.00	1.4	26839	-0.05	0.00	1.4	14195	0.17	0.00	1.2	10090	0.37	0.00	1.3	2944	0.64	0.01	1.2
DT Textiles	1194	-0.58	0.01	1.8	2374	-0.43	0.00	1.6	4358	-0.33	0.00	1.3	8503	-0.20	0.00	1.2	13317	-0.05	0.00	1.1	9627	0.16	0.00	1.0	9406	0.36	0.00	1.2	2738	0.62	0.01	1.1
VOC7	681	-0.60	0.01	1.2	1302	-0.42	0.00	1.1	1866	-0.32	0.00	1.0	2046	-0.18	0.00	1.0	1614	-0.05	0.00	1.0	553	0.12	0.01	0.9	92	0.29	0.02	1.0	9	0.66	0.04	0.4
VOC Leis&T	1483	-0.51	0.01	1.3	1846	-0.39	0.00	1.3	1994	-0.29	0.00	1.1	2090	-0.18	0.00	1.1	1977	-0.06	0.00	1.2	1039	0.10	0.01	1.2	297	0.27	0.01	1.4	20	0.55	0.07	1.1
Art	3167	-0.62	0.01	2.1	7241	-0.45	0.00	2.0	12323	-0.33	0.00	1.8	17670	-0.21	0.00	1.7	33746	-0.07	0.00	1.7	24979	0.12	0.00	1.7	18334	0.33	0.00	1.7	7510	0.56	0.00	1.7
HE Ch Dev	1577	-0.58	0.01	1.2	2982	-0.44	0.00	1.1	4316	-0.32	0.00	1.0	5832	-0.20	0.00	0.9	7410	-0.07	0.00	0.9	3506	0.12	0.00	0.9	2227	0.27	0.00	1.2	714	0.50	0.01	1.2

Grade difficulties

Rasch estimates of the difficulty of all grades are shown in Table 2 and graphically in Figure 1. In the latter, estimates are shown only where the outfit for that category was 1.5 or less. Table 2 shows difficulties calculated by WINSTEPS in logits. In Figure 1 these have been converted into GCSE grade units, by dividing by the average grade gap across all grades and subjects.

Figure 1. Relative difficulties of grades in 34 GCSE subject, ordered by grade C difficulty.



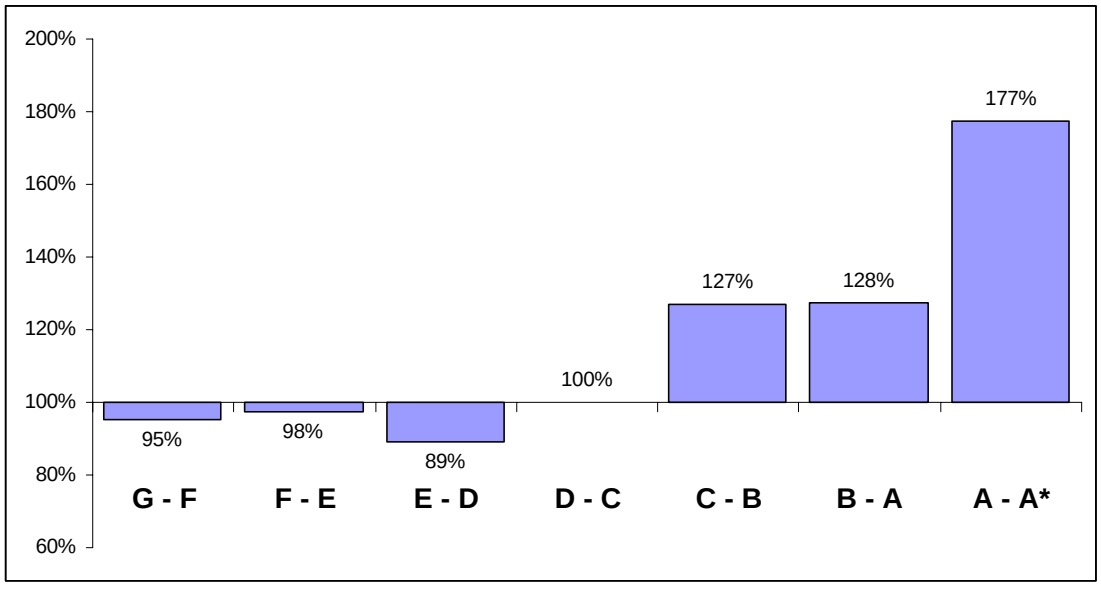
There are some striking results in Figure 1. Overall, the differences in difficulty of a particular grade across different subjects are substantial. At grade C, for example, Latin is about a grade harder than the next hardest subject, but even the next few subjects (statistics, chemistry, physics, Spanish) are about a grade harder than those at the other end of the scale (textiles, vocational science, vocational leisure and tourism, child development).

A similar pattern can be seen for grades other than C. For every pair of adjacent grades, there is substantial overlap; the higher grade in some subjects is easier than the lower grade in others. For example, grade B in biology, German or French is about equivalent to an A in office technology, media studies or PE, and there are several subjects above biology whose B is harder to get than an A in several others below PE. For the lower grades, the overlap seems bigger still, sometimes approaching two grades; a grade G in Spanish is about the same as an E in English.

Another interesting observation is that the order of difficulty of different subjects very much depends which grade is considered. In Figure 1 subjects are ranked by grade C difficulty. If either of the adjacent grades B or D had been used the order would have been similar, though a few subjects such as English or double science would have changed a few places. However, if the extreme grades G or A* had been used, the order could have changed quite considerably. Indeed, for the 24 subjects that have reliable estimates of difficulty for both grade G and A*, there is no correlation at all (r=0.03) between these values. Even between G and C the correlation is only 0.45. Hence we cannot really talk about 'subject difficulty' in general, but only in relation to a particular grade.

It is also clear from Figure 1 that the intervals between grades tend to be bigger at the top end than at the bottom. Figure 2 shows the average size of the gap between each pair of adjacent grades across all subjects. For most subjects, the gaps between G-F and F-E are close to, though a fraction less than, the gap between D-C. However, the E-D gap seems to be about 10% smaller, while C-B and B-A are just over 25% bigger than D-C. The average gap between A-A* is almost twice as big as the D-C gap.

Figure 2. Relative size of average grade gap, as a percentage of D-C gap



We do need to be cautious in interpreting these differences as straightforward differences in difficulty. They reflect the differences in the grades achieved by the students who take that subject and their grades in other subjects. There could be a number of reasons other than differences in difficulty to explain the phenomenon.

For example, if the only students who enter a particular subject are especially motivated in that subject, then the fact that they do well does not necessarily indicate that it was easier. This might be the case in Drama or vocational subjects, for example.

At the other end of the scale, some subjects may be often not be allocated the same timetable time as others, and hence students may tend to do less well in these subjects than in their others. The GCSE examination itself may be no harder in that subject, but overall students tend to be less well prepared for it. Latin and statistics might be examples of such 'under-timetabled' subjects.

It is also possible that other general factors, such as the quality of teaching, or overall levels of students' interest, motivation and effort, could vary systematically in different subjects.

References

Cresswell, M.J. (1996) Defining, setting and maintaining standards in curriculum-embedded examinations: judgemental and statistical approaches, in H.Goldstein and T. Lewis (Eds) *Assessment: Problems, developments and statistical issues*. Chichester: John Wiley & Sons.

Fitz-Gibbon C.T. and Vincent, L. (1994) Candidates' Performance in Public Examinations in Mathematics and Science. London: SCAA.

Fitz-Gibbon C.T. and Vincent, L. (1997) 'Difficulties regarding subject difficulties: developing reasonable explanations for observable data', *Oxford Review of Education*, 23, 3, 291-298.

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

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

Linacre, J. M. (2005a) WINSTEPS Rasch measurement computer program. Chicago: Winsteps.com

Linacre, J. M. (2005b) A User's Guide to W I N S T E P S M I N I S T E P Rasch-Model Computer Programs. Chicago: Winsteps.com

Sparkes B (2000) 'Subject comparisons - a Scottish perspective', *Oxford Review of Education*, 26 (2): 175-189.

Footnotes

[1] The odds ratio is the ratio of the odds of the two probabilities. In other words if a person has probabilities p and q of success on two items, the odds are $(1 - p)/p$ and $(1 - q)/q$ respectively. Hence the odds ratio is $[(1 - p)/p] / [(1 - q)/q]$. The logit function is

$$\text{logit}(p) = \ln[(1 - p)/p]$$

so the log of the odds ratio is the same as the difference in the two logits, $\text{logit}(p) - \text{logit}(q)$.

[2] WINSTEPS estimates of reliability are analogous to, but generally underestimates of, internal consistency measures such as Cronbach's alpha (Linacre, 2005b).