Oxford Review of Education
Vol. 34, No. 5, October 2008, pp. 609–636

Comparability of GCSE examinations in different subjects: an application of the Rasch model

Robert Coe*
Durham University, UK

The comparability of examinations in different subjects has been a controversial topic for many years and a number of criticisms have been made of statistical approaches to estimating the 'difficulties' of achieving particular grades in different subjects. This paper argues that if comparability is understood in terms of a linking construct then many of these problems are resolved. The Rasch model was applied to an analysis of data from over 600,000 candidates who took the General Certificate of Secondary Education (GCSE) examinations in England in 2004. Thirty-four GCSE subjects were included in the final model, which estimated the relative difficulty of each grade in each subject. Other subjects failed to fit, as did the fail grade, U. Significant overall differences were found, with some subjects more than a grade harder than others, though the difficulty of a subject varied appreciably for different grades. The gaps between the highest grades were on average twice as big as those between the bottom grades. Differential item functioning (DIF) was found for male and female candidates in some subjects, though it was small in relation to variation across subjects. Implications of these findings for various uses of examination grades are discussed.

Introduction

At first sight it seems a simple enough requirement, and no doubt one that would be widely endorsed by the general public, that the award of the same grade in the same qualification in two different subjects should represent the same standard of achievement. Indeed, something like this naive requirement is enshrined in the code of practice of the regulatory body for external qualifications in England, Wales and Northern Ireland:

> The awarding body's governing council is responsible for setting in place appropriate procedures to ensure that standards are maintained in each subject examined from year to year (including ensuring standards between GCE and VCE qualifications in similar

*CEM Centre, Mountjoy Research Centre 4, Durham University, Stockton Road, Durham DH1 3UZ, UK. Email: r.j.coe@dur.ac.uk

ISSN 0305-4985 (print)/ISSN 1465-3915 (online)/08/050609–28

DOI: 10.1080/03054980801970312

> subjects and between GCSE and GCSE in vocational subjects are aligned), across different specifications within a qualification, and with other awarding bodies. (QCA, 2004a, para 1)

Defining precisely what this means, however, is less straightforward. Indeed, the whole notion of 'standards' is both politically controversial and conceptually problematic (Baird *et al.*, 2000)—an inflammatory combination. In criticising attempts to compare the difficulties of different subjects, some (e.g. Goldstein & Cresswell, 1996) seem to imply that the question can never be resolved. Nevertheless, examination grades in different subjects are treated as interchangeable in contexts such as league tables of school performance or entry to further or higher education. Young people will make decisions about which subjects to study on which their future life chances may depend. Moreover, it is clear that many are unconvinced that the current situation is fair (e.g. Dunford, 2003). It therefore seems important to know whether there is any sense in which it could be meaningful to talk about 'comparability' across different subject examinations.

This issue is discussed in the Introduction to this paper. The main substantive content, however, consists of an analysis, using the Rasch model, of national examinations at age 16 taken by around 600,000 students in England in 2004. Before either, I present a brief outline of existing research on subject difficulties and some of the criticisms that have been made of it.

Controversy over subject difficulties

The use by qualification awarding bodies of statistical comparisons to inform the process of setting grade standards has a long history. Certainly as far back as the 1970s, Subject Pairs Analysis[1] (SPA) was being widely used (Nuttall *et al.*, 1974). Indeed, such methods are still used today, though the emphasis placed on statistical comparisons across subjects in setting thresholds for the award of particular grades in England, Wales and Northern Ireland may be less (Cresswell, 1996; Jones, 2003).

This decline may be traced to the emergence in the 1990s of a number of criticisms of the methods and assumptions of SPA (Willmott, 1995; Alton & Pearson, 1996; Goldstein & Cresswell, 1996; Newton, 1997). Most of these authors were associated with the examination boards themselves and a consensus appeared to emerge among the awarding bodies in England, Wales and Northern Ireland that, as in the subtitle of Newton's (1997) paper, 'statistical techniques do not make the grade'. Instead, the use of examiners' judgements was held to provide a basis for ensuring comparability (Cresswell, 1996).

An alternative perspective came from Fitz-Gibbon and Vincent (1994, 1997) and Dearing (1996) who applied Kelly's (1976) method of concurrent comparisons to A-Level grades. Their findings were that some subjects (in particular, the sciences, mathematics and foreign languages) were indeed more difficult, or at least, more 'severely graded' (Fitz-Gibbon & Vincent, 1994). Meanwhile, in Scotland the Scottish Curriculum Authority was calculating 'National Ratings' based on the same method, to enable consumers of national Standard grade and Higher examinations to compare the ‘difficulty’ of different subjects (Sparkes, 2000). Similar methods have been used for some time in a number of Australian states to adjust examination marks in different subjects when they are combined to form an overall measure that is used to select for entrance to tertiary education (e.g. TISC, 1998). Despite their public criticisms, awarding bodies in England, Wales and Northern Ireland have continued to conduct statistical comparisons to inform the awarding process, favouring methods such as SPA and multilevel models (Newton *et al.*, 2007). Very little published analysis is available, however.

The Rasch model has been used previously in other parts of the world for this kind of analysis, perhaps first by Tognolini and Andrich (1996), and has been used annually in Tasmania since 2000 to equate performance in different subjects for tertiary entrance (TQA, 2000). Rasch has also been used in Cyprus for similar purposes (Lamprianou, 2007).

Criticisms of statistical approaches

There are a number of specific criticisms that have been made of the use of statistical methods to compare the difficulty of different subjects (see, for example, Willmott, 1995; Alton & Pearson, 1996; Goldstein & Cresswell, 1996; Newton, 1997). Space does not permit a detailed review of the arguments and counter-arguments, but it is important to outline some of the key issues here. A more in-depth treatment can be found in Coe (2007).

One of the main arguments against simplistic statistical approaches is that performance is affected by many factors apart from ‘difficulty’. Factors such as the intrinsic interest of the subject, the quality of teaching experienced, extrinsic motivations such as the need for a particular qualification, the candidates’ levels of exam preparation, the amount of curriculum time devoted to it, and many others, could all affect performance, without making that subject more ‘difficult’.

A second argument is that different subjects are essentially incomparable. Clearly examinations in different subjects measure different things, so on what basis can we compare them? It is meaningless to say, for example, that art is easier (or harder) than physics; they are just different.

A third argument is that the groups of students taking particular combinations of subjects (on whom statistical comparisons are based) are not representative of all those who take (or might take) a particular subject. Students who take both physics and sociology, for example, are unlikely to be representative of those taking either subject.

A fourth argument is that if you analyse subject ‘difficulties’ for different subgroups (e.g. males and females) you get quite different results. Hence a judgement about whether one subject is ‘harder’ than another depends very much on who happened to take those subjects. If the characteristics of the entry change, so would the supposed ‘difficulties’.

A fifth argument is that the different methods of estimating the difficulties of different subjects themselves give quite different results.

A final argument is that any attempt to force different subjects to be of equal difficulty would be problematic. The overall abilities of candidates generally reflect the difficulties of subjects, so if the 'harder' subjects were made easier, and the 'easier' subjects made harder, pass rates would be absurdly high in the former and disastrously low in the latter. This would also cause problems during any changeover period.

Many of these are strong and persuasive arguments, and it is clear that one cannot simply claim that one subject is harder than another in any meaningful and defensible way. Indeed, part of the problem here is that the meaning of words like 'harder' in this context is not clear. As Baird (2007) has argued, there are several different conceptions of 'comparability', all of which have significant practical or theoretical limitations.

How can statistical differences be interpreted?

In this paper, I take the view that it may not be possible to define 'standards' or the 'difficulty' of a particular examination in any clear or useful way, but that the observed statistical differences between achievements in different subjects are nevertheless interpretable. I argue that one cannot talk about the validity of a particular method, or even about the assumptions that underpin it, other than by considering particular interpretations or uses of its results. By being clear about how these differences in achievement can be interpreted, and limiting ourselves to those specific interpretations, we can therefore avoid many of the objections that have been made to the use of such methods. In judging such interpretations we must therefore also apply the concept of consequential validity (Messick, 1989) to the examination process itself, and consider the uses to which examination grades are put.

The interpretation of these statistical differences draws on Newton's (2005) idea of a 'linking construct'. Newton discusses the extent to which one can interpret scores from two or more tests as comparable in terms of a 'linking construct'. If we want to make a comparison between two examinations, then we can only do this in terms of some common factor, the linking construct. For example, we might acknowledge that examinations in English and mathematics measure two different things, but still believe that they both provide an indication of a candidate's 'general academic ability'. Alternatively, we might interpret what they have in common as being candidates' 'capacity to pass examinations' or their 'level of preparedness for examinations'. If we don't accept that they have anything in common, then there is no basis on which we can compare them.

If a plausible linking construct can be identified, it may be possible to link scores, but 'inferences from linked scores can only be drawn in terms of the linking construct' (Newton, 2005, p. 111). For example, it might be possible to select a specific collection of GCSE examinations for which the linking construct could be 'general academic ability'. If it is accepted that all the subjects in this collection measure (at least to some extent) general academic ability, then we can legitimately compare their outcomes. If we do compare them, then we must interpret these comparisons in terms of our construct of general academic ability. So rather than saying that maths is 'harder' than English we must say that a particular grade in maths indicates a higher level of general academic ability than would the same grade in English.

With this interpretation, the 'difficulty' of a particular examination should not be taken to indicate anything about its level of demand, nor about the subjective experience of candidates, nor about the criteria that must be met for a particular grade to be achieved, nor about the proportion of the cohort who have achieved it. Instead, it simply refers to the relative conversion rate between a given grade in that examination and the construct that links all the examinations being compared. For example, if a group of academic GCSE subjects were considered, we might interpret the linking construct as 'general academic achievement'. Success in any of those subjects could be taken as an indication of general academic achievement, but the conversion rates might vary. In other words, the same grade in two different subjects might correspond to different levels of general academic achievement.

The idea of a linking construct seems to fit particularly well with the application of the Rasch model to the comparison of grades achieved in different subjects. The Rasch model explicitly tests whether a group of subjects can be considered to be measuring a common construct; those that do not fit the model will be identified and can be excluded, leaving a smaller but defensibly 'unidimensional' set. Hence if such a group of examination subjects does exist, the interpretation of the linking construct is relatively unproblematic. The Rasch model has the added advantages that it can estimate the difficulties of each grade in each subject independently, so does not have to make assumptions about the interval nature of any codings used.

A response to the criticisms of statistical approaches

Interpreting subject differences in terms of a linking construct also provides a counter to some of the arguments against the use of statistical methods. For example, the first argument outlined above, that other factors apart from the difficulty of the subject may influence performance, becomes largely irrelevant. If students do better in English because it is more interesting or better taught than maths, we should still treat grades in English as indicating a lower level of our underlying construct of general academic achievement. Moreover, given that many of the uses of examination grades as a selection tool (for higher education or employment) seem to interpret them as an indication of general ability anyway, it does not seem to be too much of a stretch to have to defend this interpretation of differences in attainment across subjects.

The second argument follows suit almost as easily. Whether a particular group of subjects can legitimately be compared becomes essentially an empirical question. If they fit the model adequately, then they can. The third argument is also dismissed if the data can be shown to fit the Rasch model. Estimates of the 'difficulties' of different subjects are independent of the distribution of the sample and depend only on the relative odds of achieving particular grades.

The fourth argument, that relative difficulties vary for different subgroups, is more of a problem. However, the idea that items (i.e. examinations) can be differentially difficult for different groups of persons can be analysed explicitly within the Rasch model using the concept of Differential Item Functioning (DIF). The extent to which GCSE examinations exhibit DIF and how this might affect any estimates of their 'difficulty' will be discussed below.

The fifth argument, that different methods do not agree, also seems potentially problematic. Although it is clear that the various methods give results that are not identical, it is less clear quite how important these differences are. Coe *et al.* (2007) compared the results of using a number of different methods (the Rasch model, Subject Pairs Analysis—weighted and unweighted, Kelly's (1976) method, reference test comparison and multi-level modelling) for estimating relative subject difficulties and concluded that there was broad agreement among different methods, especially when compared with the difference between any of them and the default assumption of comparability.

The sixth and final argument, that equalising the 'difficulty' of different subjects would cause as many problems as it would solve, is not disputed. The argument presented here is that the examination process, including the awarding of grades, should be the business of the awarding bodies. It is clear that there is no single, transparent and simple process that leads to unproblematic comparability of standards (Baird, 2007) and hence the current procedures of combining examiner judgement with statistical evidence may never be substantially improved. Nevertheless, a problem does arise when grades in different subjects are treated as equivalent in contexts such as school league tables or selection to continued education or employment. Either these practices must be outlawed, or, if examinations are to be used in these ways, grades should be equated first.

Methods

Before describing the analysis, it may be helpful to provide a general introduction to the ideas behind the use of the Rasch model.

The Rasch model

The Rasch model (Rasch, 1960/1980; Wright & Stone, 1979) provides a method for calibrating ordinal data onto a scale that is adequate for measurement, with properties such as unidimensionality, linearity, sample and scale independence (Wright, 1997). Unlike other statistical models, Rasch turns the relationship between data and the model upside down. Whereas most statistical modelling attempts to fit a model to existing data, in Rasch the model comes first, since the model embodies the precise requirements for adequate measurement. If data do not fit the model we must reject the data, not the model. Perhaps because of this unconventional approach to modelling, the use of Rasch has been controversial, particularly in the UK, though in many other parts of the world it is widely accepted.

Rasch assumes that the 'difficulty' of items and the 'ability' of persons[2] can be measured on the same scale, and that the probability of a person achieving success on a particular item is entirely 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[3] 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, the model requires 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, their relative probabilities of success on different items must be 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 and Masters (1982) 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 within the target range for that item are prioritised, reducing the effects of outliers that are well out of range. 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.

The interpretation of measures of fit is complex and somewhat controversial (Smith *et al.*, 1998). Some authors advise the use of particular thresholds of acceptable fit (e.g. Linacre, 2005b). Others point out that such rules of thumb are over-simplistic and can often give misleading or inappropriate results, or even that the commonly used fit-indices of infit and outfit are themselves inadequate (Karabatsos, 2000). In this analysis an arbitrary threshold for infit and outfit of 1.7 was used as a starting point for weeding out misfitting subjects. The general approach was to view residual fit statistics as having limited value: they do not guarantee good model fit but can help to indicate poor fit—hence the choice of a relatively tolerant cut-off value. In the final model, Item Characteristic Curves (ICCs) were inspected individually to ensure adequate item fit, and overall model fit was established by a range of indicators (see below).

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 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[4] is estimated using Joint Maximum Likelihood Estimation (Wright & Masters, 1982; Linacre, 2005b).

The partial credit model treats each grade as defining a threshold between those who have achieved that grade (or higher) and those who have achieved a lower grade. Hence it does require an assumption about the order of grades within a subject (e.g. that A is higher than B) but makes no assumptions about the relative sizes of the gaps between them, or about the equivalence of the 'same' grade in different subjects.

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, each time 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 secondary schools in England who took Key Stage 4 (KS4) examinations in the summer of 2004. Most of these candidates will have been aged 16 at the time. Examination results gained by these pupils in earlier years were excluded from the analysis, so that comparisons were based only on examinations taken at the same time.

The original dataset contained 632,273 pupils, all of whom had been entered for at least one examination in 2004. However, a number of examinations had very small entries, making them unsuitable for reliable comparison, so those with fewer than 500 entries nationally were dropped from the analysis. A number of pupils (7250) who had taken only these minority examinations were therefore lost, leaving data from 625,023 pupils and 109 examinations available for this analysis.

In the final model with 34 examination subjects included, but U grades omitted, the analysis was based on the 610,258 candidates who had achieved a G grade or better in at least one of these 34 subjects. This represents 98% of the candidates in the original dataset.

The main examination taken at KS4 in England is the General Certificate of Secondary Education (GCSE). Candidates typically sit examinations in between eight and ten individual subjects. Most courses are largely academic in focus and assessment is primarily by terminal written papers, set and marked by external awarding bodies, though coursework (assessment by teachers) and practical examinations form part of the assessment for many subjects.

As well as the traditional GCSE subjects, the dataset also included Vocational (Applied) 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 in England are allocated different points by the Qualifications and Curriculum Authority (QCA, 2004b) 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. Vocational GCSEs are offered in a more limited range of subjects than conventional GCSEs and the focus is generally more applied and less academic. The majority of the assessment (typically around two-thirds) is conducted by teachers based on coursework, the remainder coming from written examinations. Short Course GCSEs are similar in focus and structure to full GCSEs, but with half the content. For all these examinations candidates are awarded a grade from G to A* or the fail grade, U. The grade is based on overall performance usually across a number of written papers and other assessment tasks, so a fail, U, may be awarded if some parts of the assessment are missing.

There were also a number of non-GCSE qualifications included in the dataset, such as Full and Part 1 General National Vocational Qualifications (GNVQs, focusing on applied and vocational skills, assessed largely by coursework) at both Foundation and Intermediate level, Entry level qualifications, AS levels, etc.

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 1.7 and at least one grade category with outfit of 1.5 or less. These thresholds for acceptable fit are of course arbitrary and changing them would lead to a different set of subjects being included. The justification for these choices is that they are the result of experimenting with different thresholds to produce a set of subjects that included most of those with the largest entries while maintaining acceptable overall model fit (see below). The 34 subjects included in the final model, 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 extremely close, with a correlation of 0.9999 between the 207 grade estimates in the two models. This suggests that adding any further subjects to the model would be unlikely to change the results significantly.

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 0.95.[5] Principal Components Analysis (PCA) of the residuals showed that 83.2% of the variation in examination grades was explained by the single latent trait modelled, with the next biggest contrasting dimension accounting for just 0.9%. This provides a clear endorsement of the unidimensional structure of the data.

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 (the 37 subjects with over 20,000 entries), 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 (43,000 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 6000 entries nationally and none was able to be included in any model that provided adequate fit statistics.

Vocational (applied) GCSEs were mixed in their pattern of fit. Among the 34 are 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 tended not to fit well. Only citizenship and RS were included in the final 34, with IT and PE among subjects with substantial short course GCSE entries (over 10,000), but unable to be included.

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.

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

Subject	Count	Infit	Outfit	Item-scale correlation	Full name
Biology	43134	0.71	0.70	0.87	GCSE Biology
Chemistry	42571	0.74	0.73	0.86	GCSE Chemistry
English	580978	0.73	0.74	0.90	GCSE English
Geography	191094	0.74	0.74	0.90	GCSE Geography
Science Dbl	468586	0.76	0.76	0.90	GCSE Science: Double Award
Sociology	13189	0.76	0.76	0.88	GCSE Sociology
Physics	41910	0.77	0.76	0.85	GCSE Physics
Latin	7685	0.78	0.76	0.80	GCSE Latin
Humanities	15706	0.79	0.79	0.89	GCSE Humanities single
Bus Studs	76595	0.86	0.86	0.87	GCSE Business Studies
History	198899	0.87	0.87	0.88	GCSE History
Science Sgl	48650	0.88	0.87	0.85	GCSE Science: Single award
Maths	571739	0.90	0.90	0.89	GCSE Mathematics
English Lit	514912	0.90	0.92	0.87	GCSE English Literature
Statistics	23628	0.92	0.92	0.85	GCSE Statistics
HE Ch Dev	28352	0.99	0.99	0.84	GCSE Home Economics: Child Development
Media Stds	33314	1.01	1.02	0.84	GCSE Media/Film/Television Studies
Office Tec	23673	1.08	1.09	0.84	GCSE Office Technology
French	277482	1.09	1.09	0.87	GCSE French
VOC Sci	8080	1.11	1.12	0.77	Vocational GCSE Science
RS	115865	1.14	1.12	0.85	GCSE Religious Studies
DT Food T	99972	1.16	1.16	0.84	GCSE Design/Tech & Food Technology
German	112944	1.18	1.17	0.85	GCSE German
VOC Leis&T	10672	1.23	1.26	0.77	Vocational GCSE Leisure & Tourism
DT Textiles	50979	1.26	1.26	0.82	GCSE Design/Tech & Textiles Technology
Spanish	51349	1.32	1.30	0.85	GCSE Spanish
Sport PE	120503	1.32	1.33	0.80	GCSE Sport/Physical Education Studies
SHORT Citiz	22013	1.36	1.37	0.80	Short GCSE Social Science Citizenship
IT	75382	1.37	1.38	0.81	GCSE Information Technology
SHORT RS	195068	1.43	1.44	0.81	Short GCSE Religious Studies
VOC Engin	4443	1.44	1.47	0.74	Vocational GCSE Engineering
DT Graphic	96496	1.46	1.47	0.79	GCSE Design/Tech & Graphic Prods
DT Res Mat	102398	1.48	1.49	0.80	GCSE Design/Tech & Resist. materials
Drama	91398	1.60	1.62	0.76	GCSE Drama & Theatre Studies

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. In the model with 34 subjects there was significant disorder in both grade difficulties and person abilities for this category, with about half the subjects having estimates of the U/G difficulty threshold either higher than or about the same as the G/F threshold. The overall fit of the model was substantially improved by treating all U grades as missing and this is what was done in the final model. Unfortunately, this meant that G then became the bottom category so it was not possible to estimate a difficulty threshold for this grade. Collapsing the U and G categories together would have avoided throwing away a small amount of data, but did not lead to as big an improvement in the fit of the model.

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 1.7; no grades had outfit above 1.9.

Grade difficulties

Rasch estimates of the difficulty of all grades are shown in Table 3 and graphically in Figure 1. Note that the 'difficulty' of a grade in this context refers to the level of general academic ability typically indicated by achieving that grade. In Figure 1, estimates are shown only where 50 or more candidates had achieved that grade, since estimates based on smaller numbers were subject to relatively large errors. Table 3 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 (1.41 logits).[6] The subjects shown have been ranked in order of difficulty using the difference between the difficulty of each grade in that subject and the average for that grade across all subjects, weighted by the number of candidates achieving that grade. Hence the order reflects the relative difficulty of subjects at the grades which are most common.

There are some striking results in Figure 1. Overall, the differences in the level of ability associated with a particular grade across different subjects are substantial. At grade C, for example, Latin is the best part of a grade higher than the next hardest subject, but even the next few subjects (Spanish, French, German) are about a grade higher than those at the other end of the scale (drama, textiles, office technology and English).

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 indicates less ability than the lower grade in others. For example, grade B in German, Spanish or French is about equivalent to an A in child development, textiles or PE. For the lower grades, the overlap seems bigger still, sometimes approaching two grades; a grade F in Spanish, IT or history is almost the same as a D in textiles, PE or drama.

Another interesting observation is that the order of difficulty of different subjects very much depends on which grade is considered. In Figure 1 subjects are ranked by the weighted average relative difficulty across all grades. If the difficulty of one of the middle grades, such as C, B or D, had been used the order would have been similar, though some subjects would have changed a few places. However, if one of the extreme grades F or A* had been used, the order could have changed quite considerably. Indeed the correlation between estimates of difficulty for grade F and A* is just 0.03. Even between A* and C the correlation is only 0.46.

Table 2. Count and outfit statistics for grades G to A* for 34 subjects

	G		F		E		D		C		B		A		A*	
	Count	Outfit	Count	Outfit	Count	Outfit	Count	Outfit	Count	Outfit	Count	Outfit	Count	Outfit	Count	Outfit
Sport PE	1473	1.74	5306	1.22	14599	1.22	28981	1.36	25796	1.38	24259	1.37	15191	1.35	5242	1.24
DT Textiles	1192	2.13	2359	1.75	4335	1.37	8450	1.31	13244	1.15	9564	1.06	9320	1.15	2709	1.06
Drama	1961	1.98	4158	1.72	8145	1.76	13605	1.63	22299	1.58	22977	1.55	14594	1.60	3968	1.51
Media Stds	931	1.29	2006	1.08	3595	1.00	6540	1.00	8359	0.99	6758	0.95	4174	1.02	1043	1.07
English	15029	1.03	34087	0.72	65291	0.62	113970	0.62	150875	0.68	114868	0.75	69958	0.88	22971	1.07
Office Tec	752	1.23	1562	1.09	2770	1.09	4689	1.12	7127	1.08	3254	0.95	2503	1.11	1072	1.17
DT Food T	3678	1.87	7257	1.46	11955	1.23	19502	1.16	26471	1.04	14952	0.94	13416	1.05	3207	1.00
HE Ch Dev	1576	1.20	2972	1.04	4306	1.00	5806	1.03	7374	0.96	3487	0.82	2209	0.92	709	0.96
English Lit	12385	1.05	27395	0.83	51731	0.79	85392	0.79	135003	0.86	114303	0.92	69192	1.07	22425	1.22
DT Res Mat	4083	2.03	7899	1.75	14003	1.54	23657	1.47	26597	1.41	13959	1.25	9922	1.45	2878	1.27
Sociology	486	0.91	831	0.73	1402	0.66	2306	0.73	3728	0.71	2369	0.76	1578	0.84	498	0.90
Maths	23106	1.00	53744	0.87	89136	0.87	101152	0.83	129347	0.86	107505	0.91	47899	0.94	25916	1.05
RS	4362	1.28	7636	1.28	11127	1.24	16115	1.16	21920	1.08	24322	1.06	20696	1.07	10292	1.03
VOC Sci	681	1.37	1302	1.20	1866	1.09	2046	1.02	1614	1.09	553	0.97	92	1.11	9	0.49
Science Dbl	16710	0.90	37665	0.74	64569	0.72	92190	0.74	136002	0.76	63693	0.74	40370	0.79	20375	0.79
SHORT Citiz	1188	1.86	1868	1.79	3003	1.59	3715	1.41	5141	1.22	4313	1.11	2191	1.06	645	1.19
Science Sgl	6523	1.05	11693	0.90	12710	0.81	9883	0.83	6975	0.83	1276	0.72	482	0.79	123	1.09
Geography	6798	0.88	12118	0.73	19979	0.73	29911	0.76	46665	0.70	32965	0.67	27701	0.74	15972	0.80
Bus Studs	2663	1.03	5156	0.81	8704	0.84	14180	0.90	22707	0.84	11689	0.74	8065	0.88	3551	1.01
Humanities	1391	0.90	1973	0.86	2423	0.79	2664	0.75	3281	0.74	2474	0.73	1245	0.76	300	0.80
DT Graphic	4089	2.16	6415	1.72	11086	1.50	20238	1.52	23298	1.40	16568	1.27	12379	1.34	2696	1.18
VOC Leis&T	1483	1.38	1846	1.35	1994	1.25	2090	1.32	1977	1.20	1039	1.05	297	1.16	20	1.05
SHORT RS	13582	1.59	20469	1.57	26104	1.51	30672	1.44	38813	1.40	33167	1.37	22217	1.38	10516	1.28
History	8243	1.04	13830	0.87	19755	0.84	26328	0.81	37074	0.82	41182	0.85	35808	0.89	18155	0.95
IT	4040	1.68	6145	1.64	8143	1.58	12058	1.50	18798	1.33	13768	1.17	9129	1.22	3558	1.08
Biology	110	1.04	275	0.61	709	0.69	2827	0.84	8064	0.67	11491	0.63	12707	0.68	7920	0.77

Table 2. *(Continued)*

	G		F		E		D		C		B		A		A*	
	Count	Outfit	Count	Outfit	Count	Outfit	Count	Outfit	Count	Outfit	Count	Outfit	Count	Outfit	Count	Outfit
Physics	59	0.99	156	0.68	574	0.79	2904	0.88	8223	0.73	10237	0.69	11989	0.73	8762	0.81
Chemistry	93	0.80	261	0.79	667	0.74	3063	0.84	8296	0.71	10486	0.64	11873	0.68	8837	0.81
VOC Engin	576	1.54	807	1.52	932	1.56	911	1.47	743	1.35	379	1.29	99	1.53	8	1.42
Statistics	242	0.88	695	0.77	1735	0.93	4033	0.93	8493	0.97	4097	0.84	3296	0.96	1166	0.94
French	14734	1.23	27048	1.21	39174	1.13	52736	1.10	60019	1.04	39789	0.94	28312	1.00	17533	1.15
Spanish	2456	1.37	4083	1.29	5890	1.27	8942	1.21	10541	1.16	7874	1.15	6788	1.24	5344	1.90
German	4670	1.43	8450	1.42	12756	1.24	20745	1.22	29921	1.10	18160	1.00	12257	1.00	6735	1.19
Latin	13	1.49	34	1.18	117	0.97	419	0.80	875	0.66	1341	0.63	2688	0.74	2881	0.84

Table 3. Rasch-Thurstone thresholds (difficulty) and standard error statistics for grades F to A* for 34 subjects

	F		E		D		C		B		A		A*	
	Measure	St Err	Measure	St Err	Measure	St Err	Measure	St Err	Measure	St Err	Measure	St Err	Measure	St Err
Sport PE	–5.56	0.03	–3.90	0.02	–2.46	0.01	–0.83	0.01	0.38	0.01	1.88	0.01	3.85	0.02
DT Textiles	–4.50	0.04	–3.25	0.02	–2.29	0.02	–1.09	0.01	0.56	0.01	1.82	0.02	4.28	0.02
Drama	–4.64	0.03	–3.41	0.02	–2.39	0.01	–1.32	0.01	0.24	0.01	2.13	0.01	4.56	0.02
Media Stds	–4.41	0.04	–3.11	0.03	–2.11	0.02	–0.88	0.02	0.57	0.02	2.09	0.02	4.31	0.04
English	–5.18	0.01	–3.57	0.01	–2.33	0.00	–0.90	0.00	0.88	0.00	2.62	0.01	4.90	0.01
Office Tec	–4.28	0.05	–3.06	0.03	–2.09	0.02	–0.91	0.02	0.97	0.02	2.01	0.03	3.75	0.04
DT Food T	–4.50	0.02	–3.15	0.01	–2.13	0.01	–0.90	0.01	0.80	0.01	1.98	0.01	4.53	0.02
HE Ch Dev	–4.18	0.03	–2.89	0.02	–1.94	0.02	–0.90	0.02	0.67	0.02	1.75	0.03	3.51	0.05
English Lit	–4.46	0.01	–3.11	0.01	–2.03	0.01	–0.88	0.00	0.83	0.00	2.62	0.01	4.93	0.01
DT Res Mat	–4.65	0.02	–3.32	0.01	–2.23	0.01	–0.79	0.01	0.99	0.01	2.30	0.01	4.55	0.02
Sociology	–3.60	0.06	–2.54	0.04	–1.69	0.03	–0.70	0.03	0.93	0.03	2.24	0.03	4.15	0.06
Maths	–4.68	0.01	–2.95	0.01	–1.59	0.00	–0.44	0.00	1.13	0.00	2.95	0.01	4.51	0.01
RS	–3.37	0.02	–2.18	0.01	–1.31	0.01	–0.40	0.01	0.74	0.01	2.16	0.01	4.14	0.01
VOC Sci	–4.22	0.05	–2.77	0.03	–1.63	0.03	–0.38	0.03	1.31	0.05	3.14	0.12	5.11	0.36
Science Dbl	–4.43	0.01	–2.88	0.01	–1.68	0.01	–0.46	0.00	1.53	0.00	2.73	0.01	4.31	0.01
SHORT Citiz	–3.44	0.04	–2.32	0.03	–1.38	0.02	–0.47	0.02	0.91	0.02	2.60	0.03	4.59	0.05
Science Sgl	–4.43	0.02	–2.70	0.01	–1.35	0.01	0.04	0.02	2.36	0.03	3.47	0.05	5.18	0.11
Geography	–3.64	0.02	–2.46	0.01	–1.48	0.01	–0.41	0.01	1.21	0.01	2.46	0.01	4.24	0.01
Bus Studs	–3.61	0.02	–2.45	0.02	–1.53	0.01	–0.45	0.01	1.36	0.01	2.49	0.01	4.13	0.02
Humanities	–3.50	0.04	–2.29	0.03	–1.37	0.03	–0.47	0.02	0.85	0.03	2.41	0.04	4.49	0.07
DT Graphic	–3.64	0.02	–2.54	0.01	–1.58	0.01	–0.25	0.01	1.21	0.01	2.55	0.01	5.03	0.02
VOC Leis&T	–3.44	0.04	–2.31	0.03	–1.48	0.03	–0.55	0.03	0.81	0.04	2.55	0.07	5.30	0.25
SHORT RS	–3.22	0.01	–2.03	0.01	–1.15	0.01	–0.29	0.01	0.91	0.01	2.26	0.01	3.96	0.01
History	–3.18	0.01	–1.99	0.01	–1.09	0.01	–0.19	0.01	0.99	0.01	2.43	0.01	4.46	0.01
IT	–3.20	0.02	–2.07	0.02	–1.28	0.01	–0.34	0.01	1.20	0.01	2.62	0.01	4.55	0.02
Biology	–4.38	0.12	–2.95	0.07	–1.89	0.04	–0.40	0.02	1.34	0.02	2.88	0.01	4.91	0.02

Table 3. *(Continued)*

	F		E		D		C		B		A		A*	
	Measure	St Err	Measure	St Err	Measure	St Err	Measure	St Err	Measure	St Err	Measure	St Err	Measure	St Err
Physics	−3.96	0.16	−2.84	0.09	−1.84	0.05	−0.28	0.02	1.50	0.02	2.85	0.01	4.72	0.02
Chemistry	−3.77	0.12	−2.50	0.07	−1.63	0.04	−0.18	0.02	1.54	0.02	2.90	0.01	4.73	0.02
VOC Engin	−3.13	0.06	−1.90	0.04	−0.92	0.04	0.16	0.05	1.61	0.07	3.49	0.12	6.15	0.39
Statistics	−3.80	0.08	−2.44	0.04	−1.31	0.03	0.04	0.02	2.34	0.02	3.49	0.02	5.56	0.04
French	−3.30	0.01	−1.93	0.01	−0.85	0.01	0.35	0.01	1.82	0.01	3.03	0.01	4.60	0.01
Spanish	−3.00	0.03	−1.73	0.02	−0.74	0.02	0.46	0.01	1.81	0.01	2.91	0.02	4.41	0.02
German	−3.15	0.02	−1.88	0.01	−0.93	0.01	0.25	0.01	1.98	0.01	3.23	0.01	4.87	0.02
Latin	−1.10	0.31	−0.28	0.17	0.46	0.10	1.47	0.06	2.45	0.04	3.36	0.03	5.17	0.03

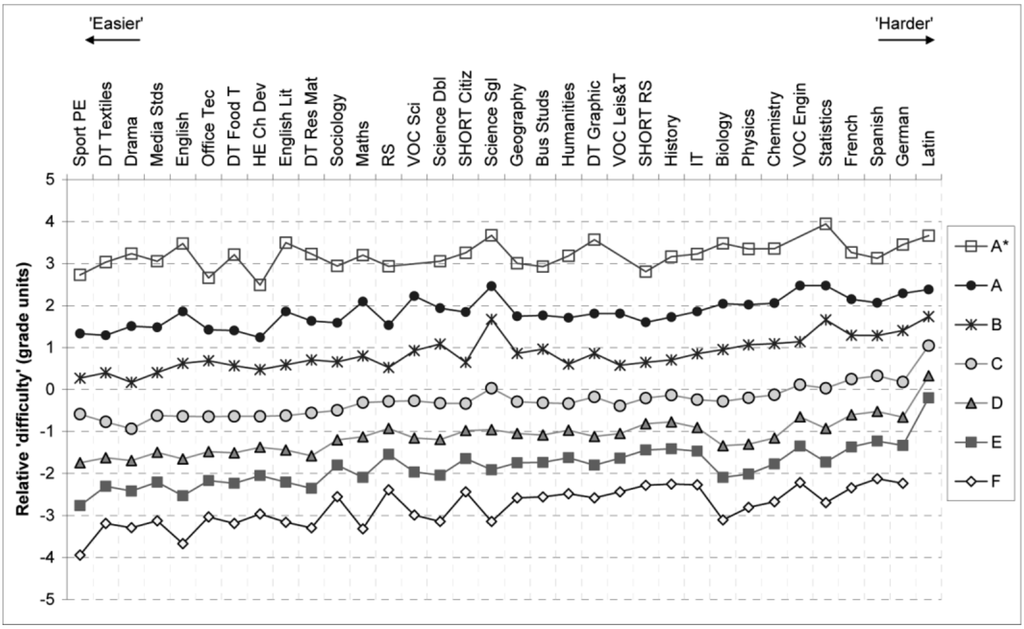

Figure 1. Relative 'difficulties' of achieving each grade in 34 GCSE subjects, ordered by weighted average difficulty

Grade gaps. One of the limitations of most existing statistical analyses of comparability is that calculating differences between grades requires the assumption that the intervals between them are equal. This applies both to intervals between different grades within the same subject, and to intervals between the same grades in different subjects. A strength of the partial credit model is that it calculates separately the difficulty of each grade in each subject, and hence allows the intervals to be estimated directly.

It is clear from Figure 1 that across all subjects the intervals between grades tend to be bigger at the top end than at the bottom. For most subjects, the gaps between F–E, E–D and D–C are roughly similar in size. The C–B and B–A gaps are a little bigger than the gaps between the four lower grades, but the average gap between A–A* is over 50% bigger than the average between all the other grades.

Subgroup comparisons

The possibility of different subgroups performing differently in different subjects was investigated by estimating the differential item functioning (DIF) for each subject, by gender. This analysis takes the overall difficulty of each subject across all grades (without any weighting) and compares how this would have been estimated with populations limited to each subgroup. Information about gender was imported from the Pupil Level Annual School Census (PLASC), which is completed by maintained schools in England, and therefore excludes pupils in independent schools. The analysis included 277,082 female and 276,962 male candidates' results.

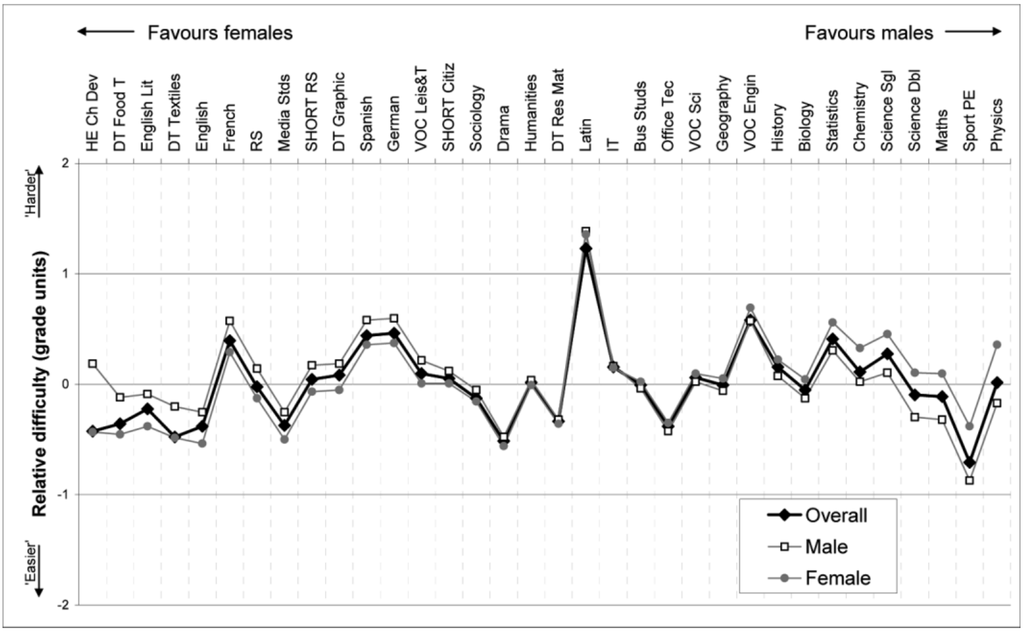

Figure 2. Overall estimate of 'difficulty' for each subject, together with the estimates for males and females separately, ordered by gender difference

Figure 2 shows the overall estimate of 'difficulty' (in average grade units) for each subject, together with the estimates for males and females separately. Subjects are arranged in order of the size of the difference between the two sexes, with subjects that are relatively 'easier' for females on the left.

Although the biggest gender difference is in child development, the number of males taking this subject is low (only 333, about 1% of the entry) so this may not be very reliable. For the subjects with substantial entries for both sexes the biggest difference is in physics which is on average half a grade 'easier' for boys than girls. Although this is a substantial difference, it is much less than the differences among subjects; the overall tendency is for gender-specific estimates of 'difficulty' to agree more than they differ. Across all subjects, male and female estimates of 'difficulty' correlate 0.91 and 0.95 respectively with the overall estimates, and 0.77 with each other.

These differences in relative achievement in these subjects could perhaps be interpreted as a gender bias in the learning and assessment process, provided one remembers that this includes factors such as motivation and effort applied; if girls work harder than boys in English and boys work harder in maths, one could not really blame the *assessment* process if their grades reflect this. An alternative interpretation would be to see performance in terms of two separate dimensions of ability, which we might characterise as 'male-ability' and 'female-ability'. Whatever the reasons though, it may be fair to say that some subjects appear to favour boys (in the sense that they are likely to do relatively better), while others favour girls.

Discussion

A number of issues arise from these results, along with some implications for policy.

Interpreting differential achievement as 'subject difficulties'

We do need to be cautious in interpreting these differences as straightforward differences in difficulty. They reflect the differences between the grades awarded to the students who take that subject and their grades in other subjects and are not necessarily related to the intrinsic or perceived difficulty of the examinations themselves. There could be a number of reasons other than differences in 'difficulty' to explain the phenomenon.

For example, if all the 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 PE or drama, for example. There are other subjects, however, that one might have identified as being typically taken by those who have a special interest or ability in them—for example, music, art or Urdu. Each of these subjects has over 5000 entries and each failed to fit the Rasch model, so their 'difficulties' could not be estimated. A possible explanation for this misfit is that these subjects do indeed draw on special interests or abilities, and hence their grades cannot be interpreted as indicating the same latent trait as the other subjects that were included. In this way, the Rasch model may have automatically filtered out 'specialist' subjects from the comparison.

At the other end of the scale, some subjects may 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. Vocational GCSEs were examined for the first time in 2004, so it is possible that they were still bedding down, making them appear harder than they might in subsequent years. It has been suggested, for example, that a possible reason for maths appearing to be harder than English is that maths is intrinsically less interesting (Goldstein & Cresswell, 1996).

However, as the Rasch model provides separate estimates of difficulty for each grade within a subject, any argument about different levels of motivation or teaching would have to be specific to particular ability ranges. For example, at the top grade (A*), English appears 'harder' than maths, while at A and below English comes out 'easier'. Hence we would have to argue that maths was taught better (or was more interesting) than English to the highest achieving students, but that for lower achievers, English teaching was better. Precisely how one might operationalise quality of teaching across subjects in order to test this hypothesis is not clear, however.

Which subjects can be compared?

Unidimensionality. One of the arguments sometimes made about comparisons of grading in different subjects is that they are fundamentally different and should not be equated. The skills required for, say, English and maths are quite different and are essentially incomparable. One could not really say that either is harder than the other, only that they are different. This argument has a certain superficial appeal, but is challenged by the data presented here.

One of the strengths of the Rasch model is that it allows the assumption of unidimensionality to be tested empirically. If different subjects are highly correlated with each other, in the sense that people who get high grades on one tend also to get high grades on the others, then they may be seen as sufficiently unidimensional. We may infer that a single latent construct, such as general academic ability, forms the basis of all of them. In this case, it would be reasonable to compare the grades achieved in different subjects. On the other hand, if the different subjects fail to fit the Rasch model, then they are not unidimensional and cannot reasonably be compared.

It is clear from the data that a core group of subjects can be considered unidimensional. The 34 subjects included in the analysis may be seen as fitting the model at least reasonably well. Certainly, some grades for some subjects fit less well than others, but 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 top end of Table 1. In other words, there is a substantial set of GCSE subjects which can be treated as unidimensional. Precisely where one would draw the line around this set might be a matter of judgement, but it would be hard to argue that no comparisons were valid.

Model fit and opportunism. From the point of view of conventional statistical modelling it may seem somewhat disingenuous to begin with a set of 109 subjects, throw out those that do not fit the model and then claim that the fact that 34 do fit is evidence of success. However, the point of this investigation has been not to validate the model, which is justified on theoretical grounds, but to investigate whether a possible ‘linking construct’ exists that can be used as a basis for comparing the difficulty of different subjects. For the 34 subjects that do fit we can claim that their grades are broadly measuring a common construct and hence their difficulties can be compared, in terms of the relationship between each grade and that construct. The fact that some subjects or examinations do not fit does not undermine the method; it simply indicates that a comparison of the ‘difficulty’ of those subjects may not be appropriate, since they indicate a different set of skills and abilities from those that fit the model.

Although the Rasch model provides a convenient and theoretically sound way of establishing the comparability of different subject examinations based on a unidimensional latent trait, it remains controversial, at least in the UK. However, it seems likely that the main substantive results of this study are not dependent on the choice of this particular approach, and would be broadly reproduced by other methods. For example, Coe *et al.* (2007) found that agreement among estimates of overall subject difficulty from different methods was generally high, for both GCSE and Advanced Level examinations.

Subjects omitted. The analysis shows that there are some subjects that do not appear to be in line with the others, in the sense that they could be seen as measuring the same latent trait. Among these, for example, are a group of creative or performance subjects, such as GCSE music, art and design, fine art and performing studies, which did not fit the model. It is perhaps not altogether surprising that the skills and abilities required to do well in these subjects are not quite the same as those required for the more conventionally 'academic' subjects that did fit.

Again not too surprising is the fact that Intermediate GNVQs failed to fit. Modes of assessment and intended learning outcomes for these qualifications are generally quite different from typical GCSEs, so it should not be altogether unexpected that they prove not to be measuring quite the same thing. This failure to fit is not a result of the different grading structure—the fact that GNVQs have only three pass grades. If a person's chances of getting a higher grade at GNVQ increased in line with their performance in their GCSEs then they might still fit, but unfortunately they do not.

Perhaps more surprising is the fact that half of the vocational GCSE subjects did fit. Given that the clear intention of these examinations was to provide a different type of qualification for a different group of students (BBC, 2000), and that a number of other vocational and more applied qualifications did not fit, one might have expected that vocational GCSEs would also not fit. That three of them (science, leisure and tourism, engineering) did fit is hard to explain.

The fact that some of these examinations failed to fit the model does raise a question about what it might mean to compare their 'difficulty'. If they are essentially measuring something different from the 34 GCSEs that did fit there is no clear linking construct with which to interpret a comparison. Hence although it may be meaningful to say, for example, that history is 'harder' than sociology, it really has no meaning to ask whether, say, music is harder than either; music is just different, neither 'harder' nor 'easier'.

Implications for using grades for selection

Interpreting grades as evidence for selection decisions. GCSE grades are used in a range of different ways, including to make selection decisions for employment or further education. In these contexts it is arguable that the particular grade achieved in a particular subject is interpreted not so much as direct evidence of specific achievement in that subject, but as an indication of that candidate's general capacity for learning in other contexts. If that is indeed the interpretation then it provides us with a linking construct on which to base our comparisons. For a particular selection decision we might identify a group of subjects that are believed *prima facie* to provide an indication of 'general capacity for learning' in the contexts in which we are interested. Use of the Rasch model would provide a check that these subjects were sufficiently unidimensional for our purpose and would give us an estimate of the level of the latent trait (general capacity for learning) represented by each grade achieved in each subject.

It is worth pointing out here that this interpretation stands even if other reasons might account for the overall differences in achievements in different subjects. For example, suppose candidates in English typically enjoy it more, are taught better, work harder and so genuinely achieve 'more' than candidates in maths; we should still treat grades in English as 'worth less', in terms of their conversion into our linking construct of general capacity for learning, since the advantage they have gained in that subject is unlikely to transfer to the other contexts in which we want to know how well they will do. If our reason for wanting to compare grades in different subjects arises from a need to use them to make selection decisions, then many of the objections and caveats outlined above no longer seem relevant.

We may also note that although only 34 subjects were used in this analysis, we can nevertheless use the model to estimate their ability for 98% of those who took any examination. The majority of subjects may have failed to fit the model or not been included for other reasons, but because most of the popular subjects were included its coverage is still very high. We might argue that because the other subjects do not fit the model, and are hence not providing a measure of the same construct of general academic ability, performance in those examinations should not be taken into account in making selection decisions. On the other hand, in Tasmania where the Rasch model is used to provide a measure of general ability for tertiary entrance selection from performance in examinations with differing difficulties, a method has been devised for including scores from non-fitting subjects in the calculation (TQA, 2004).

The need to equate grades across subjects. When GCSE grades are used to make selection decisions it is clear that there is a problem of lack of comparability. However, it does not necessarily follow that the answer to this problem is to align the grade awarding processes so that grades in different subjects are immediately comparable. For one thing, requiring awarding bodies to do this as part of the grading process could be practically quite complex and would certainly delay the awarding process (Alton & Pearson, 1996). A second objection is that some subjects would no longer discriminate well within the range of abilities of their typical candidates. If all subjects were aligned to a common 'difficulty', then almost half the candidates in Latin would achieve the highest grade, while in child development fewer than one in five would get even the top three grades.

A third, and perhaps more serious, objection is that different definitions of comparability would imply different grading decisions. Two different selection processes might be based on two different linking constructs with two different—but overlapping—sets of subjects.

A fourth difficulty is that such a mechanistic view of comparability rules out any role for judgement in the process. In fact, it is clear that although the use of judgement alone to try to establish comparability is quite unreliable, the use of statistical methods without any accompanying moderating judgment will also lead to some anomalous results (Baird, 2007).

A fifth problem is that even if comparability across subjects could be achieved by such realignment, comparability over time would certainly be compromised by any change.

Given all these problems, it is fortunate that there is actually no need to establish comparability at the point of grading, provided it is clearly acknowledged that the same grade in different subjects may not be 'equivalent' in terms of the level of general academic ability it implies. Of course, this would require a major political shift and is directly against the current requirement of the Code of Practice for England, Wales and Northern Ireland (QCA, 2004a) to ensure 'comparability'. However, what might be seen as politically unacceptable in one context may be perfectly acceptable in another. For example, in Western Australia different subjects are marked independently, but for the purposes of selecting students for tertiary education marks are aligned by 'scaling' to make them comparable. This is explained in a leaflet for general audiences:

> Scaling is needed because students face a choice of subjects. Scaling ensures that students are not disadvantaged by choosing a difficult subject.
>
> In the absence of any sort of inter-subject scaling, students wishing to increase their Tertiary Entrance Rank (TER) might be tempted to enrol only in the easiest subjects. (Why study Calculus when you can get a higher mark in Discrete Mathematics?) Scaling gives students freedom to choose the more challenging subjects. It enables students to make educationally sensible decisions instead of being obliged to choose subjects in which they can score the greatest number of unscaled marks. It ensures a fair treatment of their results. (TISC, 1998)

Implications for using grades to judge school performance

U grade anomaly. The fact that the U grade fitted the model 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. Interestingly, data from the Advanced Level Information System (www.alisproject.org) suggest that at A level, U does in fact behave as the bottom grade, without any such anomaly. In this context it therefore seems ironic that whereas official systems for calculating value-added at GCSE within England have changed points allocations to widen the gap between U and G relative to other grade gaps, at A level, U grades have recently been removed from the calculation.[7]

Big gaps at the top. An interesting finding to emerge from this analysis is that the gaps between grades are significantly larger at the top end. The use of the Rasch model not only allows the difficulty of individual grades to be estimated, but also places these estimates on a genuinely interval scale, hence allowing the sizes of the gaps to be compared. Traditional methods of allocating points to grades generally assume equal intervals between them, although the arbitrariness of such a coding is sometimes acknowledged (e.g. Schagen & Schagen, 2003). The analysis presented here provides a rationale for coding grades numerically in a way that is not arbitrary and which could be used in school effectiveness and other studies that apply regression models to grades.

In practice, the implications of switching to codings based on Rasch estimates of difficulty are not easy to predict. However, it seems likely that by acknowledging that the higher grades are worth a little more than has previously been recognised, our estimates of the performance of relatively high achieving schools would be enhanced, at the expense of those whose absolute performance is lower.

Incentives to take 'easier' subjects. Although I have argued that it is not straightforward to describe one subject as 'easier' than another, the evidence is clear that comparable students are likely to be awarded significantly different grades in different subjects. In any system where the same grade in different subjects is accorded equal value, and where either students have a choice about which subjects to enter or schools have a choice about which to offer, there must be an incentive for candidates to take subjects in which their chances of success are greatest.

Of course, this is not to suggest that a school's decisions about which subjects to offer would be wholly determined by consideration of which subjects are likely to optimise their league table position, or that students will simply choose subjects that are likely to give them the best grades irrespective of factors such as their personal interest or career aspirations. However, it seems hard to believe that such considerations will play no part in the choices that are made. For example, for the majority of students, taking media studies GCSE instead of German is associated with a result that is a whole grade higher. Only for those extremely able candidates, who are pretty much certain to achieve the top grade regardless of subject, is this not the case.

Conclusions

The notion of comparability of different examination subjects is extremely problematic and many researchers have been rightly wary of making over-simplified claims about the 'difficulties' of different subjects. However, it is clear that there are substantial statistical differences in the grades that the same students achieve in different subjects at GCSE and it is not satisfactory simply to say that these are problematic or uninterpretable. For those involved in awarding these qualifications, such an abdication of responsibility is particularly unsatisfactory in a world where grades are widely treated as convertible in contexts such as school league tables and selection to further education or employment. Worse still, such convertibility is often officially endorsed.

To say that one subject is more difficult than another may not be defensible in quite such simple terms. However, if grades in different subjects are interpreted as indicating some common 'linking' construct, such as general academic achievement, then the 'same' grade in a different subject may denote quite a different level of this construct. In this sense, grades in different GCSE subjects are certainly not equivalent: some are harder than others. If grades are to be interpreted in terms of a common construct then it is clear that for some subjects this will be more valid than for others. Although there is a core group of the main ‘academic’ GCSE subjects that can be compared, a number of other subjects do not appear to fit; their grades denote other things and should not be compared. It is also clear that the sizes of the gaps between grades are far from equal and that this presents challenges to some common uses of GCSE grades that make the simpler and more convenient—but false—assumption that grade intervals are equal. A more open acknowledgement of these differences could make our interpretations and uses of examination grades more appropriate and valid.

Notes

1. Subject Pairs Analysis refers to a group of methods that compare the grades achieved by all candidates who took a particular pair of subjects, often aggregating over all possible pairs, to produce an estimate of each subject’s relative difficulty. See Coe (2007) for further explanation.
2. The words ‘difficulty’ and ‘ability’ are used generally in discussing the Rasch model, even when their normal meanings are considerably stretched. For example, in the context of a Likert scale attitude item one may talk about the ‘difficulty’ of an item to mean its tendency to be disagreed with (i.e. how ‘hard’ it is to agree with). The use of these words may initially seem strange to anyone not familiar with the Rasch model. However, I have adopted this convention, partly in order to comply with standard practice, and partly because although the words ‘difficulty’ and ‘ability’ are not quite right for the interpretation intended, I am unable to think of better ones.
3. 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

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

   so the log of the odds ratio is the same as the difference in the two logits, $\text{logit}(p) - \text{logit}(q)$.
4. The partial credit model used here is:

   $$\ln(P_{nij} / P_{ni(j-1)}) = B_n - D_i - F_{ij} = B_n - D_{ij}$$

   where
   P_{nij} is the probability that person n encountering item i is observed in category j;
   B_n is the ability measure of person n;
   D_i is the difficulty measure of item i, the point where the highest and lowest categories of the item are equally probable;
   F_{ij} is the ‘calibration’ measure for item i of category j relative to category j-1, the point where categories j-1 and j are equally probable relative to the measure of the item. (Linacre, 2005b)
   In WINSTEPS the partial credit model is invoked by treating each item as a separate group, using the specification ISGROUPS=0
5. WINSTEPS estimates of reliability are analogous to, but generally underestimates of, internal consistency measures such as Cronbach’s alpha (Linacre, 2005b).
6. There are different ways this might have been done. One of the reviewers of this paper suggested rescaling the logit scale such that the sample of rescaled person ability measures had

the same standard deviation as the set of their mean GCSE scores (i.e. the mean score, for each person, of all subjects they have taken, using A*=8, A=7, etc). The justification for this would be that if we interpret the Rasch measure as an indication of person ability then it makes sense to say that the difference in ability between a person who achieves an average of, say C grades and another who achieves an average of B grades should equal 'one grade' on our recalibrated scale. However, if we emphasise the interpretation of Rasch measure as an indication of grade difficulty, as I have done, then 'one grade' on a recalibrated scale should represent the average difference between the difficulty of adjacent grades, across all grades and all subjects. Fortunately, the two are very close, so it makes little practical difference. The former method produces an 'average grade' interval that is 92% of the size of the latter.

7. The latest (at the time of writing) rationale for the way fail grades are to be treated for value-added purposes at A level is in LSC (2006). An account of the controversy around this issue can be found on BBC News online at http://news.bbc.co.uk/1/hi/education/5134612.stm. The Contextual Value Added (CVA) model used in England is explained at http://www.standards.dfes.gov.uk/performance/word/GuidetoCVA2006.2.doc?version=1

Notes on contributor

Robert Coe is Reader in the School of Education and Director of Secondary Projects in the Curriculum, Evaluation and Management (CEM) Centre at Durham University. His research interests include evaluation, assessment and evidence-based strategies for school improvement.

References

Alton, A. & Pearson, S. (1996) Statistical Approaches to Inter-Subject Comparability. Unpublished UCLES research paper.

Baird, J., Cresswell, M. and Newton, P. (2000) Would the *real* gold standard please step forward? *Research Papers in Education,* 15(2), 213–229.

Baird, J. (2007) Alternative conceptions of comparability, in: P. Newton, J. Baird, H. Goldstein, H. Patrick and P. Tymms (Eds) *Techniques for monitoring the comparability of examination standards* (London, Qualifications and Curriculum Authority).

BBC (2000) Vocational GCSEs target disaffected, BBC News Online, 6 July. Available online at http://news.bbc.co.uk/1/hi/education/821444.stm.

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

Coe, R., Searle, J., Barmby, P., Jones, K. & Higgins, S. (2007) *Relative difficulty of examinations in different subjects.* Report for SCORE (Science Community Partnership Supporting Education) (Durham, Curriculum, Evaluation and Management Centre, Durham University).

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).

Dearing, R. (1996) *Review of qualifications for 16–19 year olds* (London, School Curriculum and Assessment Authority).

Dunford, J. (2003) We do still have to address the issue of equal standards for all A-level subjects, *Guardian Education,* 26 August.

Fitz-Gibbon, C. T. & Vincent, L. (1994) *Candidates' performance in public examinations in mathematics and science* (London, SCAA).

Fitz-Gibbon, C. T. & Vincent, L. (1997) Difficulties regarding subject difficulties: developing reasonable explanations for observable data, *Oxford Review of Education,* 23(3), 291–298.

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.

Jones, B. E. (2003) Subject pairs over time: a review of the evidence and the issues. Unpublished AQA research paper RC/220.

Karabatsos, G. (2000) A critique of Rasch residual fit statistics, *Journal of Applied Measurement,* 1(2), 152–176.

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

Lamprianou, J. (2007) Commentary on Chapter 8, in: P. Newton, J. Baird, H. Goldstein, H. Patrick and P. Tymms (Eds) *Techniques for monitoring the comparability of examination standards* (London, Qualifications and Curriculum Authority).

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

Linacre, J. M. (2005b) *A user's guide to WINSTEPS MINISTEP Rasch-model computer programs* (Chicago, Winsteps.com).

LSC (Learning and Skills Council). (2006) *Treatment of fails in the calculation of value added and distance travelled for the learner achievement tracker* (18 July 2006). Available online at http://readingroom.lsc.gov.uk/lsc/2006/quality/performanceachievement/nattreatmentoffailsinthe calculationofvalueaddedanddistancetravelledmeasures-br-july2006.pdf (accessed 17 November 2006).

Messick, S. (1989) Validity, in: R.L. Linn (Ed.) *Educational measurement* (3rd edn.) (New York, Macmillan).

Newton, P. E. (1997) Measuring comparability of standards between subjects: why our statistical techniques do not make the grade, *British Educational Research Journal,* 23(4), 433–449.

Newton, P. E. (2005) Examination standards and the limits of linking, *Assessment in Education,* 12(2), 105–123.

Newton, P., Baird, J., Goldstein, H., Patrick, H. & Tymms, P. (Eds) (2007) *Techniques for monitoring the comparability of examination standards* (London, Qualifications and Curriculum Authority).

Nuttall, D. L., Backhouse, J. K. & Willmott, A. S. (1974) Comparability of standards between subjects, *Schools Council Examinations Bulletin,* 29.

QCA (Qualifications and Curriculum Authority). (2004a) *GCSE, GCSE in vocational subjects, GCE, VCE, GNVQ and AEA code of practice 2004/5.* Available online at http://www.qca.org.uk/downloads/6295_gcse_vocgcse_gce_vce_gnvq_aea_code-of-practice_04-05.pdf (accessed 30 July 2006).

QCA (Qualifications and Curriculum Authority). (2004b) *Including all approved qualifications in school and college performance indicators.* Available online at http://www.qca.org.uk/14-19/developments/index_including-all-approved.htm (accessed 20 February2007).

Rasch, G. (1960/1980) *Probabilistic models for some intelligence and attainment tests* (Copenhagen, Danish Institute for Educational Research). Expanded edition (1980) with foreword and afterword by B. D. Wright (Chicago, The University of Chicago Press).

Schagen, I. & Schagen, S. (2003) Analysis of national value-added datasets to assess the impact of selection on pupil performance, *British Educational Research Journal,* 29(4), 561–582.

Smith, R. M., Schumacker, R. E. & Bush, M. J. (1998) Using item mean squares to evaluate fit to the Rasch model, *Journal of Outcome Measurment,* 2(1), 66–78.

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

TISC (Tertiary Institutions Service Centre). (1998) *Scaling* (Perth, Curriculum Council, Government of Western Australia). Available online at http://www.curriculum.wa.edu.au/files/pdf/scaling.pdf (accessed 29 September 2006).

Tognolini, J. & Andrich, D. (1996) Analysis of profiles of students applying for entrance to universities, *Applied Measurement in Education,* 9(4), 323–353.

TQA (Tasmanian Qualifications Authority). (2000) *Using Rasch analysis to scale TCE subjects*. Available online at http://www.tqa.tas.gov.au/4DCGI/_WWW_doc/003675/RND01/Rasch_Intro.pdf (accessed 4 September 2007).

TQA (Tasmanian Qualifications Authority). (2004) *How the scaled awards are calculated and used to determine the tertiary entrance score*. Available online at http://www.tqa.tas.gov.au/0477 (accessed 20 September 2007).

Willmott, A. (1995) *A national study of subject grading standards at A-level, Summer 1993*. A report commissioned by the Standing Research Advisory Committee for the GCE A-level examinations, Oxford.

Wright, B. D. (1997) *Measurement for social science and education: history of social science measurement* (Chicago, Institute for Objective Measurement). Available online at http://www.rasch.org/memo62.htm (accessed 4 September 2007).

Wright B. D. & Masters G. N. (1982) *Rating scale analysis* (Chicago, MESA Press).

Wright, B. D. & Stone, M. H. (1979) *Best test design* (Chicago, MESA Press).