Summary Report, July 2008

Relative difficulty of examinations in different subjects

Robert Coe, Jeff Searle, Patrick Barmby, Karen Jones, Steve Higgins

CEM Centre, Durham University

This document presents a summary of a larger research report which can be found at www.cemcentre.org/SCORE2008report.pdf. The study was commissioned by SCORE (Science Community Representing Education) to address the following research questions:

- How (if at all) can the relative difficulty of different examinations be defined?
- What existing studies have presented evidence about relative difficulties of UK examinations in different subjects?
- To what extent do the results of existing studies converge?

and by conducting new analysis of GCSE and A-level examination data for England:

- How well do different statistical methods agree?
- How consistent are differences over time?
- How much do estimates of relative difficulty vary for different subgroups?

Finally, integrating existing and new research:

- Are STEM (science, technology, engineering and mathematics) subjects generally more difficult than others?
- What are the practical implications of any differences in difficulty?
- What policy options are there?

Main conclusions from the study

Summary of statistical differences

Overall, there are substantial differences in the average grades achieved by the same (or comparable) candidates in examinations in different subjects, at both GCSE and A-level. Although the choice of which statistical method to use can make some difference to this, the inter-method differences are always much smaller than the overall differences in difficulty of different subjects by any method.

Differences between subjects appear to be extremely stable over time. Restricting the analysis to particular subgroups (eg by gender, Free School Meals eligibility or sector) can make some difference to estimates of difficulty, but again these differences are always much smaller than the overall differences in difficulty of different subjects.

> *"... there are substantial differences in the average grades achieved by the same (or comparable) candidates in examinations in different subjects. ..."*

Given this level of consistency across methods, time and subgroups, it seems appropriate to examine the relative difficulties of science and non-science subjects. At A-level, the STEM subjects are not just more difficult on average than the non-sciences, they are without exception among the hardest of all A-levels. The other sciences are also generally more difficult than the non-sciences, though the difference is not as extreme. At GCSE the STEM subjects are a little more difficult on average than the non-sciences, though the difference is less than at A-level.

> *" At A-level, the STEM subjects are not just more difficult on average than the non-sciences, they are without exception among the hardest of all A-levels."*

Responses to evidence of statistical differences

That there are statistical differences in the grades achieved in different subjects is beyond question. What is less clear is how these differences may legitimately be interpreted. We have identified four possible interpretations:

- *Statistical differences are meaningless*. The criticisms (outlined below) of the various approaches used to identify such differences are sufficiently serious to render any valid interpretation of the differences impossible. It makes no sense to say that one subject is harder than another.

- *Grades in different subjects reflect different levels of achievement*. Provided we have taken *all* relevant factors into account (such as ability, prior attainment, gender, socioeconomic status, attitudes, motivation, quality of teaching, time spent, etc), statistical differences indicate that equivalent learning gains correspond to different grades in different subjects. To say one subject is harder than another means a student would have to demonstrate more learning in it to get the same grade.

- *Candidates' chances of success are different in different subjects*. Provided characteristics of the candidate (as in the previous interpretation, but excluding any subject-specific factors like motivation, teaching quality) have been taken into account, statistical differences indicate the relative chances of success of equivalent candidates in different subjects. To say one subject is harder than another means that, other things being equal, a student is likely to get a lower grade in it.

- *Grades in different subjects reflect different levels of ability*. Statistical differences may indicate differences in the relationship between grades achieved and some underlying construct such as 'general academic ability'. To say one subject is harder than another means that the same grade in it indicates a higher level of general ability.

"To say one subject is harder than another means that the same grade in it indicates a higher level of general ability."

Policy implications

We have identified three possible policy responses to the issue of subject difficulty.

- *Do nothing*. Some have argued that interpreting statistical differences as 'difficulty' is too problematic; any changes would cause as many problems as they would solve and the current situation is largely acceptable. We are unconvinced by these arguments.

- *Make all subjects the same standard*. We could equate the standards of grades in different subjects to make them statistically comparable. Although this would be transparent and fair for using grades for selection, it would create a number of perverse anomalies: some subjects would be far too hard for the candidates that currently take them, others would be far too easy; comparability over time would be lost. For these reasons, this option seems unattractive.

- *Change the way grades are used*. This would involve introducing some kind of scaling so that some grades are acknowledged to be worth more than others for certain purposes. Existing statistical differences would continue, but a fair 'conversion rate' would be applied whenever grades in different subjects were to be treated as equivalent, for example in league tables or UCAS tariffs. Although there are some difficulties with this option, it appears to us to be the best way forward.

"Existing statistical differences would continue, but a fair 'conversion rate' would be applied whenever grades in different subjects were to be treated as equivalent, for example in league tables or UCAS tariffs."

Review of existing work on comparability of subject examinations

Summary of findings from existing studies

We reviewed 29 different studies of cross-subject comparability, conducted in the UK since the 1970s, most of them containing some empirical evidence about differences in difficulty.

Two previous studies have compared different statistical methods for estimating subject difficulties. Although one of them interpreted the results as showing substantial differences, data from both suggest that agreement across methods is good and that between-method differences are much smaller than between-subject differences by any method.

Only one previous study has compared difficulty estimates (at A-level) calculated in different years. Variation over time was about one tenth of the size of the difference between subjects, suggesting that relative difficulties are reasonably stable over time.

Two studies have reported differences in relative difficulty for different subgroups, both comparing male and female candidates. In one, differences were large, in the other, small, so the evidence is inconclusive.

Three existing studies provide estimates of relative difficulties of enough subjects at A-level to allow STEM subjects (biology, chemistry, physics and mathematics) to be compared with others. From this it is clear that the STEM subjects are among the most difficult, though the inclusion of other sciences (e.g. design technology, psychology, geology) makes the comparison more equal.

Evidence on perceptions of difficulty

Whether or not they are objectively more difficult, there appears to be a widely held perception that science subjects, and in particular physics, are more difficult than others. Perceived difficulty may be one of the reasons for students to choose not to continue to study some STEM subjects, though it is hard to untangle the real reasons behind people's choices and the evidence is mixed.

Evidence from new analysis conducted for this study

We analysed data from two different datasets of national cohorts in England. The first contained the results of 250,000 candidates who took A-level examinations in 2006; the second contained the results of 635,000 candidates who took GCSEs in the same year.

How well do different methods agree?

We applied five different statistical methods for comparing the grades achieved by comparable (or the same) candidates in different subjects. These were subject pairs analysis, Kelly's method, the Rasch model, the reference test method and value-added (multilevel) models. For some methods, more than one version was applied.

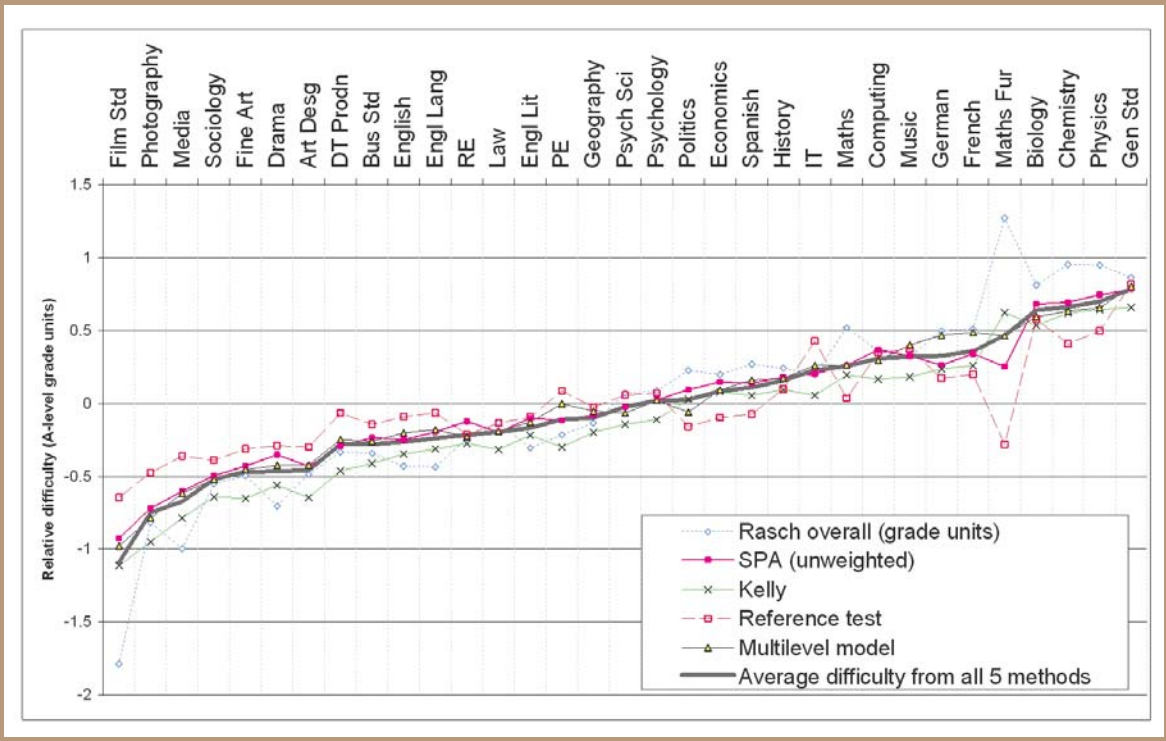
"We applied five different statistical methods for comparing the grades achieved by comparable (or the same) candidates in different subjects."

At A-level, agreement across methods was high, though a few subjects generated different estimates. With the exception of further mathematics, film studies and media studies, estimates within subjects from the five methods were all within half a grade, and within a third of a grade for the majority of subjects.

"At A-level, agreement across methods was high, though a few subjects generated different estimates."

This compared with a difference of nearly two grades across subjects. Overall, the average inter-method difference for a given subject was about 20% of the average inter-subject difference for a particular method. A comparison of the difficulty estimates for A-levels using the different methods is shown in Figure 1.

Figure 1 : Estimates of A-level subject difficulties by different methods



The Rasch model gave the largest separation across subjects, with a range of three grades between easiest and hardest, while other methods put this range between 1.5 and 1.8 grades. Despite this absolute difference, correlations between Rasch and the other methods were above 0.90, apart from with the reference test method which did not correlate as well with any of the other methods (correlations with the reference test method were between 0.75 and 0.90).

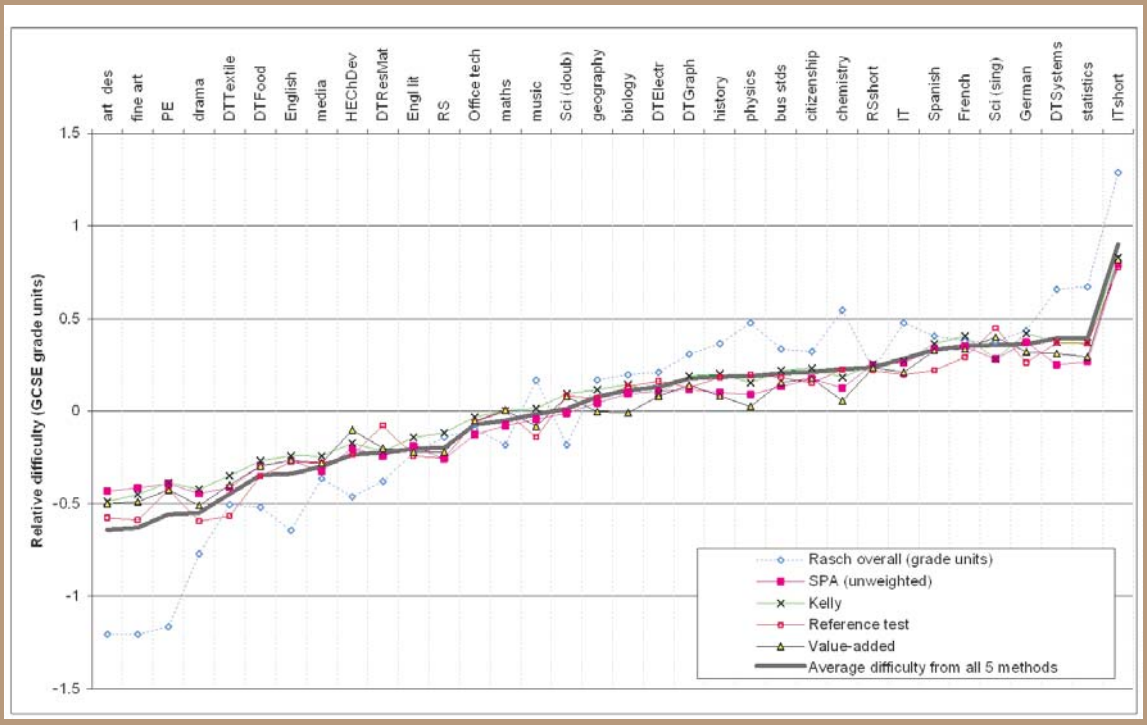
"At GCSE, there was extremely good agreement across methods."

At GCSE, there was extremely good agreement across methods. For all the methods apart from the Rasch model, correlations between subject difficulty estimates were all in excess of 0.96; the difference between the highest and lowest estimate of difficulty was never more than 0.2 of a grade for any subject, and the average difference across all subjects was just 0.1 of a grade. This compares with a difference of 1.3 grades between the hardest and easiest subject, averaged across all these methods. The average inter-method difference was about 10% of the average inter-subject difference.

In the Rasch model the separation between subjects was greater than in the other methods, with the gap between hardest and easiest rising to 2.5 grades. This led to some larger differences in the absolute difficulties at the extremes, though the correlation between Rasch and the other methods was still high (0.92 or above). Despite the fact that the Rasch model is slightly out of line with the others, there are some technical reasons for preferring it.

A comparison of the difficulty estimates for GCSEs using the different methods is shown in Figure 2 .

Figure 2 : Estimates of GCSE subject difficulties by different methods



How consistent are differences over time?

This analysis used data from the Alis (A-level Information System, www.alisproject.org) project for A-levels taken between 1994 and 2005 in 32 subjects. Kelly's method was used to estimate relative difficulties.

The average correlation between difficulties in successive years was 0.97. As the gap between the two years being compared increased, correlations fell gradually, levelling off at around 0.85 when the gap was nine or more years. Since 2002, relative difficulties have been even more stable. The average of the four correlations across a one-year gap was 0.99, falling only as far as 0.96 when the gap was four years. Hence, at least at A-level, relative difficulties change little from year to year.

"Hence, at least at A-level, relative difficulties change little from year to year."

How much do subgroups vary?

The national A-level and GCSE datasets were analysed again, but this time restricted to particular subgroups to investigate how much the relative difficulties varied. Both Kelly's method and the Rasch model were used.

"For A-level subjects, split by gender, both the Rasch and Kelly methods were in agreement that there was very little difference between difficulties for males and females"

For A-level subjects, split by gender, both the Rasch and Kelly methods were in agreement that there was very little difference between difficulties for males and females. In the Rasch analysis the difference between male and female difficulties for every subject was less than 0.2 of a grade. The correlation between male and female difficulties for the 33 subjects was 0.99. For Kelly's method the maximum difference was 0.3 of a grade and the correlation 0.97. Even by this method the mean absolute difference between subject difficulties for the two genders (0.14) was only 8% of the overall difference across subjects.

For GCSEs, the gender difference was somewhat bigger than at A-level. From the Rasch analysis, three subjects had as much as half a grade's difference between male and female difficulties (child development, in favour of females, and PE and physics, in favour of males), and a further four subjects had differences greater than 0.3 of a grade (food technology, favouring females; mathematics, single and double science favouring males). Overall, the correlation between male and female estimates of difficulty for the 34 subjects was 0.77. Kelly's method gave similar gender differences, but as the overall range between the hardest and easiest subjects was smaller for Kelly's method than for Rasch the correlation between male and female difficulties was correspondingly lower at 0.66.

"For GCSEs, the gender difference was somewhat bigger than at A-level."

Two further splits of the GCSE dataset were made, by Free School Meals status and school sector (independent/maintained), and difficulties calculated for each subgroup using Kelly's method. In the former, differences were small, with just four subjects showing more than 0.2 grade's difference in difficulty (fine art and art and design, easier for those with FSM; history and geography, relatively hard for those with FSM). Split by school sector, differences were a little larger, but still averaged only 16% of the inter-subject range. For both these splits the subgroups were very unequal in size and in their overall levels of achievement.

Methods that have been used to compare difficulties

A number of different methods have been used to try to determine the relative difficulties of examinations in different subjects. Here we summarise the different methods and their limitations.

We identified five distinct types of *statistical* methods:

- *Subject Pairs Analysis*. For a given pair of subjects, grades achieved by candidates who took both are compared. These pair-wise comparisons may be aggregated to provide an index of relative difficulty for each subject. This method has been widely used by awarding bodies in England.
- *Common examinee linear models*. These methods effectively compute the relative difficulties of different subjects from a matrix of examination by candidate results. In practice the calculation amounts to the solution of a set of linear simultaneous equations. These approaches have not been used much by the GCE and GCSE awarding bodies, but have been widely used in Scotland (Kelly's method) and Australia (Average Marks Scaling).
- *Latent trait models*. The Rasch model specifies the probability that a candidate will achieve a particular grade in a given subject, given their ability and its difficulty. Values of ability and difficulty are then estimated iteratively in order to best fit the observed data on the candidate's achievement.
- *Reference tests*. These make use of a common test, usually some kind of general ability test, as an anchor or reference against which performance in different subjects can be judged. A regression model allows the grades achieved by similar (in terms of their general ability) candidates in different subjects to be compared.
- *Value-added models*. The regression (often multilevel) model can include any explanatory variables that help to explain variation in examination performance, such as a candidate's prior attainment, gender, socioeconomic status, type of school attended, etc.

Two further *judgement* methods have been applied to comparisons of subjects that are sufficiently similar.

- *Judgement against an explicit 'standard'*. Cross-moderation studies compare candidates' scripts from different examinations with an explicit description of the 'grade standard' in order to identify or ratify a specific grade boundary mark for each. This approach has been the preferred method of awarding bodies and the regulator in England until recently.
- *Paired comparisons*. Pairs of scripts are compared and scrutineers asked to judge which is better, without knowledge of the marks they have been awarded.

Criticisms of these methods

Even if statistical methods show that candidates typically achieve higher grades in one subject than another it does not necessarily follow that the former is easier, for the following reasons:

1. *Other factors*. Differences in achievement could be caused by several factors other than a genuine difference in the standard of the two subjects. Factors that vary between subjects such as the motivation of candidates, the quality of teaching, the intrinsic interestingness of the subjects, and many others, could all account for differential levels of achievement.

2. *Multidimensionality*. Comparing two subjects makes sense only if they are in some way comparable. For this they must have something in common in terms of which they can be compared. For some pairs of subjects, for example physics and art, it can be argued that it is hard to imagine what this common element might be.

3. *Unrepresentativeness*. If those candidates who have actually taken a particular examination are not representative of all those who might possibly take it, then statistical comparisons may not generalise beyond that group. If the characteristics of the candidates were to change, so would the difficulty.

4. *Subgroup invariance*. Statistical differences in apparent difficulty between subjects may vary for different subgroups. We might find, for example, that maths appears to be harder than English for girls, but the difficulty is reversed for boys. Such subgroup variation undermines the simple interpretation of these differences as difficulty.

5. *Method inconsistency*. There are many different methods by which one can compare the relative difficulties of different subjects and no convincing *a priori* reason to prefer one to the others. Unfortunately, their results have not always seemed to agree, so it is hard to know which estimate (if any) is the 'right' one.

6. *Forcing equality*. Making all subjects equal in difficulty would cause a number of new problems, including destroying comparability within subjects over time, causing confusion and delaying the awarding process.

Judgement methods are also limited, for the following reasons:

1. *Breadth of criteria*. If standards are defined by criteria, these must be broad enough to allow different subjects to be compared, which makes them likely to be somewhat imprecise.

2. *Crediting responses to different levels of demand*. Comparing, for example, a good answer to an easy question with a partial answer to a harder question proves to be problematic.

3. *Crediting different types of performance*. Various other differences arise in the style of assessment in different examinations. Comparing, for example, levels of knowledge demonstrated in essay and multiple choice questions is very difficult.

4. *Statistics underpin judgements anyway*. Unless an 'expert' who had never seen any candidates attempt any examination questions could judge the difficulty purely from knowledge of the subject, then any judgements made may well be dependent on 'statistical' evidence.

5. *Interpretation and context*. Criteria and standards have to be interpreted by judges within a context. Judges are unlikely to have all the relevant information about the context in which the examination performance was given, or to be able to take account of it if they have.

6. *Aggregating judgements*. Examination grades typically depend on performance in a range of tasks with different criteria, so a comparison must specify some method for aggregating these judgements. The resulting grade is unlikely to correspond to a single criterion.

" At A-level, the STEM subjects are not just more difficult on average than the non-sciences, they are without exception among the hardest of all A-levels."