
**Report for SCORE (Science Community Supporting Education), July 2008**

# Relative difficulty of examinations in different subjects

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

CEM Centre, Durham University


www.cemcentre.org

Executive Summary

1. This report reviews the evidence on whether examinations in some subjects can legitimately be described as 'harder' than those in other subjects, and, if so, whether STEM subjects (those that form a foundation for further study in science, technology, engineering and mathematics) are generally more difficult than others.

2. A number of different statistical methods have been used to try to address these questions in the past, including Subject Pairs Analysis (SPA), Kelly's (1976) method, the Rasch model, Reference Tests and value-added (including multilevel) models.

3. Evidence from the existing statistical analyses conducted in the UK appears to show a high level of consistency in their estimates of subject difficulties across methods and over time, though the publicly available evidence is somewhat limited. Nevertheless, there is a clear indication that the STEM subjects are generally harder than other subjects.

4. A separate body of literature has examined the subjective perceptions of difficulty of candidates in different subjects. There is evidence that sciences are often perceived as more difficult.

5. In new analyses conducted for this report, we applied five different statistical methods to two national datasets for England from 2006 examinations, comparing difficulties of 33 A-level subjects and of 34 subjects at GCSE. Agreement across methods was generally high.

6. Analysis conducted by the ALIS project on A-level relative difficulties every year since 1994 shows that these are highly stable over time, and especially so since 2002.

7. When GCSE examination results are analysed separately for different subgroups, such as males and females, there are some differences in the relative difficulties of different subjects. Other subgroup splits, by Free School Meals status and Independent/Maintained school sector, show smaller differences and are more problematic to interpret. At A-level, subgroup differences appear to be much smaller.

8. At A-level, most methods put the range between the easiest and hardest subjects at around two grades, though the use of the Rasch model suggests that this range is a lot larger at the lower grades than at the top, and that grade intervals are far from equal. The STEM subjects are all at the top end of this difficulty range.

9. At GCSE, most methods put the range between the easiest and hardest subjects at around one-and-a-half grades, though one subject (short course IT) is an outlier, so the range is about one grade for the majority of subjects. Again the Rasch model shows that grade intervals are far from equal and the difficulty of

a subject depends very much on which grade is compared. There is a tendency for STEM subjects to be more difficult on average, though this is less marked than at A-level.

10. A number of objections have been made in the past to the simplistic interpretation of statistical differences as indicating subject difficulties, and we discuss these. Although it is possible to argue that statistical differences are meaningless, there are at least three possible interpretations of such differences that we believe are defensible.

11. Given the evidence about the relative difficulties of different subjects, we believe there are three possible options for policy: to leave things as they are; to make grades statistically comparable, or to adjust them for specific uses. These three options are presented and discussed.

# CONTENTS

# *LIST OF TABLES*

# *LIST OF FIGURES*

# *PART I:*
# *INTRODUCTION*

# 1. INTRODUCTION

The controversy over possible differences between subjects with regard to the difficulties of examinations is long standing, and has been the subject of considerable media and public interest. Equally long standing is a perception that science subjects are often the most difficult. More recently, concern over the supply of people with qualifications and skills in 'STEM' subjects (science, technology, engineering and mathematics) has added a new force to the debate over subject difficulties and a new urgency to the concerns over subject difficulties.

As a research area, this topic has received much attention, though much of the work has been conducted by researchers from the awarding bodies and can be found only in internal reports, not publicly available. Nevertheless, enough is available to give a flavour of the results and to show that the debate has been highly controversial. At its heart lie questions about what examinations are for, what we want grades to represent and what we mean by terms such as 'standards', 'difficulty' or 'comparability'. These are all issues with which this report must engage.

This report has been commissioned by SCORE, a group concerned with science education, convened by the Royal Society and its other founding partners, the Association for Science Education, Biosciences Federation, Institute of Biology, Institute of Physics, Royal Society of Chemistry and the Science Council.[1] The original intention of the research, as set out in the tender document, was 'to undertake a rigorous review of available evidence regarding relative difficulties of different subjects as measured by attainment through external testing across the UK'.

## 1.1. A brief historical outline of the controversy over subject difficulties

Going back three decades, the Schools Council published research by Nuttall and colleagues (Nuttall *et al*., 1974) at NFER on comparability of standards between subjects. This reported on analyses of CSE and GCE O-level data for English examination boards. They considered relative subject difficulty by five different methods, and their result showed consistency across the methods and across boards. Science subjects and languages were found to be more difficult than other subjects. However, they stressed that they were raising issues for discussion and there were no firm conclusions with regards to the issue. At the same time, the Welsh exam board WJEC carried out similar subject pairs analyses and got results that supported those of Nuttall *et al*.

Kelly (1975, 1976a, 1976b) published her investigations into 'soft' and 'tough' options in Scottish Higher examinations. She developed the subject pairs methodology from

[1] See http://www.royalsoc.ac.uk/page.asp?id=5216

Nuttall et al. and obtained a similar ordering of subjects for difficulty. She proposed correction factors that would bring grades into line. Subsequently, the Scottish Qualifications Authority (SQA) have calculated correction factors annually and made them available to help interested parties, such as higher education institutions or employers, interpret awarded grades. Kelly also noted variation in the results between subjects over the four years of her study, and particularly between boys and girls, noting that the actual use of a correction factors was thus problematic. She found however that languages, chemistry and physics were consistently more difficult, with biology and mathematics relatively less so.

Newbould (1982, 1983) entered the debate with some results that he believed indicated motivational factors should be considered in the difficulty rankings as well as cognitive ones (measures of ability). Throughout this time, the research department at JMB exam board had been carrying out subject pairs analyses at GCE O-level and A-level and making the information available to awarding committees at the board. The results reported by Forrest and Vickerman (1982) reviewed the results from 1972 to 1980, and despite some variation, the trend was that languages and chemistry, physics and mathematics were found to be harder than other subjects. They did not discuss correction factors, and stressed that they were simply providing research information.

For about a decade then, the interest in comparability between subjects seemed to die down, at least in the published literature. Then, in 1994, Fitz-Gibbon and Vincent's ALIS-based investigation was published which specifically addressed the question "are mathematics and science more difficult than other subjects at A-level?" They carried out their analysis in four different ways, and consistently concluded that these subjects were more difficult. They also found marked differences in the results between boys and girls. Other studies replicated Fitz-Gibbon and Vincent's results. In particular the Dearing Review of 16-19 qualifications, commissioned by SCAA, was taking place around this time. Published in 1996, the study used data available from all the English A-level boards for 1993 to 1995, thus addressing the small and possibly biased sample criticism made of Fitz-Gibbon and Vincent's data, and replicated their results. Alton and Pearson (1996) published similar results from the 16+18+ project funded by the DfEE, although they did note caution in interpretation of the results.

There were many criticisms of Fitz-Gibbon and Vincent's methodology and the assumptions underlying it, particularly from Goldstein and Cresswell (1996). They argued in the 1996 paper, and subsequently Cresswell maintains the argument, that subject pairs and any statistical attempt to analyse inter-subject standards is flawed. The assumption of unidimensionsality, i.e. that exams in the different subjects, even in the same subject, measure the same thing, is criticised. They also argue that stable, representative samples are unattainable. Fitz-Gibbon and Vincent (1997) published a riposte to Goldstein and Cresswell, pointing out they analysed data and as such their result were valid. They defended their position on unidimensionality by noting that a grade A is an A in any subject.

This led to a series of papers that seemed to become more and more philosophical in nature (Cresswell, 1996; Newton, 1997; Wiliam, 1996a, 1996b; Baird et al, 2000) about what standards in various subjects might mean and how they might, or might not, be compared. The conclusion of these authors leans towards viewing standards (i.e. where grade boundaries are drawn) as a social construct, that they are drawn where expert judges (senior examiners in a subject) say they are, and that this is acceptable if society is willing to accept it. Others have pointed out (Sparkes, 2000) that grades are no longer just used for selection purposes for higher education, but also, for example, to compare the performances of schools and colleges.

Recent reviews of quality assurance procedures at QCA have found that these work very well in practice and that QCA is doing a commendable job. It was noted that there are various expectations and demands from different stakeholders of what the examination system should deliver, and that some of these are not realizable in practice. The question is raised (but perhaps not answered) of whether examination results are working for the key purposes for which they are intended. The review by McGaw et al. (2004) for QCA concluded that no examination system has an adequate way to determine whether standards are constant across subjects.

Moreover, it is interesting to note that the Code of Conduct for GCSE and A-level examinations (QCA, 2007) does not explicitly mention the need to ensure comparability across different subjects as part of the requirements for awarding bodies:

> *"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 that standards between GCE and GCE in applied subjects, as well as between GCSE and GCSE in vocational subjects, are aligned), across different specifications within a qualification and with other awarding bodies."* (Para 1.1)

The official position therefore seems to be that the whole notion of comparability across different subjects is too problematic to define, let alone assure, but that the system works about as well as it could do to keep the standard the same for different subjects. Moreover, concern with ensuring the comparability of standards over time appears to have taken precedence over the need to ensure comparability across subjects: the statistical evidence suggests that one cannot do both.

Meanwhile, however, concern about the effect of perceived difficulty of certain subjects on young people's choices to study them has brought a new practical urgency to the question of subject difficulties. In both the Dearing committee report on foreign language teaching (DfES, 2007) and the House of Lords Select Committee on Science and Technology report (House of Lords, 2006), concern was expressed about the potentially off-putting effect of the perceived difficulty of languages and sciences respectively.

The Dearing *Language Review* report (DfES, 2007, para 3.20) discusses the 'widely held view that … that the demands of languages in the GCSE are greater than for the

great majority of subjects', and refers to 'the statistical analysis that appeared to give some support for that view in terms of the level of demand for the award of a grade.' While acknowledging that many factors can affect achievement, the report goes on to conclude that 'we have found strong confirmation of the view that the award of grades is more demanding than for most other subjects', and calls for a 'definitive study' to resolve the matter 'one way or the other'.

The House of Lords (2006) report on *Science Teaching in Schools* is if anything even more forthright than Dearing. While acknowledging that the work of the CEM Centre on A-level difficulties, which it cites, is 'widely if not universally accepted' (para 2.18), it goes on to state that

> *… the "gold standard" of A-levels is now fundamentally compromised. The presumption that an A-level "A" grade represents a fixed level of achievement (embodied in an equal UCAS tariff) is hard to defend. (para 2.24)*
>
> *There is good evidence that students are opting for "easier" A-levels over the sciences and mathematics (para 2.28)*

It seems, therefore, that the government's standard reassurance that 'the DfES and the QCA have always responded to such claims by stating that there is no such thing as an easy or hard A-level' (House of Lords, 2006, para 2.18) is no longer reassuring – if it ever was. Hence the need for this report.

### *1.2. Overview of the report*

This report is divided into four Parts.

Part I consists of this Introduction and Chapter 2, which outlines the different methods that have been used to compare the difficulties of different examinations. Although the focus is on statistical methods, we do also make reference to judgment methods. This Chapter also includes a Section describing the official processes of grade awarding. A final Section previews a discussion of how statistical differences in achievement may be interpreted, in order to put the results presented in Parts II and III in context, though a full discussion of the interpretation of these results is held until Part IV.

Part II summarises the existing evidence about the relative difficulties of different subjects from UK studies. Chapter 3 presents the evidence, with a summary in relation to STEM subjects at the end. Also in Part II, Chapter 4 discusses the existing evidence on the subjective perceptions of difficulty of science subjects.

Part III presents new analysis conducted for this report. Chapter 5 analyses national data for England on A-levels and GCSEs taken in 2006, to investigate the extent to which five different methods are consistent with each other. Chapter 6 uses data from the ALIS project to analyse the consistency of relative difficulties over time. Chapter 7 looks at the question of how much subject difficulties vary for different

subgroups at A-level and GCSE. Chapter 8 summarises all this evidence in relation to the question of whether STEM subjects are more difficult than others.

Part IV discusses the criticisms that have been made of the use of statistical methods to compare subject difficulties and considers how their results may be interpreted (Chapter 9). The final Chapter (10) presents three specific policy options.

# 2. METHODS FOR COMPARING DIFFICULTIES

Many different methods have been used to investigate the comparability of examinations in different subjects. Before reviewing the evidence from the various studies, we briefly outline the different methods. There are broadly two types of approach that have been adopted: statistical and judgement methods. Within each there are a number of varieties, of which the main ones are described briefly below under these two main headings.

A third section of this chapter describes the procedures used by the awarding bodies to award grades to candidates and ensure comparability of standards. A fourth and final section considers the question of how, if at all, statistical differences between levels of performance in different subjects can be interpreted.

## 2.1. Statistical methods

Under this heading we identify five groups of methods. The first three of them (Subject Pairs Analysis, common examinee linear models and latent trait models) may be broadly characterised as 'common examinee methods', as they depend on comparisons among the results achieved by the same candidate in different examinations. These methods have been reviewed by Coe (2007), where more details of the methods, their strengths and weaknesses can be found. The other two (reference tests and 'value-added' methods) depend on comparing the grades achieved in different examinations with those achieved by others who are judged to be similar on the basis of some additional information (such as their performance on a particular reference test). Reference test methods have been reviewed by Murphy (2007) and 'value-added' methods by Schagen and Hutchinson (2007).

### *2.1.1. Subject Pairs Analysis (SPA)*

These methods have been widely used by the examination boards in England, Wales and Northern Ireland, and are perhaps also the simplest conceptually.

The basic version of SPA considers only those candidates who have taken a particular pair of examinations. For each candidate we could determine whether they have achieved the same grade in each or, if not, in which subject they have done better. Simply counting the proportion of candidates in each category would form the basis of a comparison between the two subjects.

A more widely used variation is to compute an average difference in the grades achieved in the two subjects. These methods may be described as 'interval' approaches since any such average will be sensitive to the sizes of the gaps between grades, not just to their order. The conventional way to do this is to convert examination grades into a numerical scale using consecutive integer values (e.g. at

GCSE, U=0, G=1, F=2, …, B=6, A=7, A*=8). For each candidate who took a particular pair of subjects we can calculate the difference between their grade in Subject 1 and Subject 2. The mean of these differences across all candidates is a measure of the difference in 'difficulty' of the pair, in grade units. This method has been widely used to compute pair-wise comparisons (e.g. Forrest and Smith, 1972; Nuttall *et al*., 1974, Ch III). Forrest and Vickerman (1982, p9) note that changing the numerical scale to one that reflects the different mark intervals between grade boundaries produces results which are 'almost indistinguishable from those obtained by the simple method', suggesting that the interval assumption is not too problematic for these methods.

If we are prepared to adopt one of the 'interval' approaches it is relatively straightforward to calculate the mean grade differences for all possible pairs of subjects and to average the mean differences, for each subject separately. So, for example, we can calculate the average difference in the grades achieved in mathematics and every other subject taken with it. An average of these differences will give an estimate of the overall difficulty of mathematics, compared with all the other subjects taken by candidates who also took mathematics. If we do this for all subjects, we arrive at a list of subjects with an estimate of relative difficulty for each. Of course, various weightings may be used in calculating this average. Coe (2007) has described this approach as 'Aggregated Subject Pairs Analysis' (ASPA) to distinguish it from the simple SPA, the main difference being that the samples of candidates for many pairs in SPA are likely to be quite unrepresentative of either subject, whereas in ASPA, the estimate of a subject's difficulty is based on all candidates who took that subject with any other.

### *2.1.2. 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). Under this heading we also include Nuttall *et al.*'s (1974) 'UBMT' method and the use of analysis of variance (ANOVA), neither of which has been used widely.

Kelly's (1976) method overcomes the problem with methods such as ASPA that candidates who take a particular 'hard' subject may be likely to combine it with other 'hard' subjects; and, similarly, 'easy' subjects are more likely to be combined – and hence compared – with other 'easy' subjects. This problem could lead to the extent of the differences between subjects being underestimated.

For example, if a high proportion of those who took chemistry (a relatively 'hard' subject) also took other 'hard' subjects like maths and physics, the average grades they achieved in their other subjects might be quite similar to their grades in chemistry. Methods such as UBMT or ASPA (especially the weighted version) would then estimate chemistry to be of only average difficulty. Kelly's (1976) method essentially uses an iterative procedure to respond to this problem.

Kelly's method begins by comparing the grades achieved by candidates in one subject with their average grades in all their other subjects, and so estimating the difficulty of that subject. This is done for each subject under consideration, using the grades achieved by all candidates who have taken it with at least one other in the set. These 'difficulty estimates' are then used to apply a correction factor to the grades achieved in that subject. So, for example, if chemistry is found to be half a grade more difficult than the average, that half grade is added to the achieved grade for all chemistry examinees. The whole process is then repeated using the 'difficulty corrected' grades in each subject instead of the actual achieved grades, to produce a new estimate of the relative difficulty of these subjects with corrected grades. After a small number of iterations, the corrections shrink to zero and so the estimates of difficulty of each subject converge.

Although it may be conceptually helpful to think of this method as iterative, it can be shown that the result is equivalent to solving a set of linear equations (Kelly, 1976, provides a proof of this, due to Lawley, in an appendix). In practice, solving these equations using a matrix inversion is more efficient than the iterative process for large data sets.

Average Marks Scaling has been used in a number of Australian states for producing aggregated marks from different subjects with different difficulties. Although it does not appear to have been used in the UK, Average Marks Scaling has qualities that make it conceptually interesting and so worth considering in this context. This method differs from the others described so far in that it aims not just to quantify the different difficulties of different subjects, but to rescale their marks onto a common scale, taking account of the spread of abilities of candidates in that subject.

Average marks scaling (AMS) can be thought of as a more sophisticated version of methods such as the UBMT, ANOVA or Kelly's method. It has been applied directly to marks rather than grades, as that is how examination results are generally reported in Australia. AMS corrects both the average mark for a subject and the spread of marks, while preserving the shape of the distribution. AMS could equally well be applied to grades, provided they were coded on a numerical scale. It would then be essentially similar to Kelly's method, but has the advantage that one does not have to assume that the gaps between grades are the same in different subjects; if grades in one subject are relatively compressed, AMS will stretch them out, as well as moving them up or down. However, part of the 'interval' assumption remains as one must still assume that the gaps between grades within each subject are equal.

Marks rescaled by AMS have the following properties (Partis, 1997):

- Within a subject, the order of marks and the shape of the distribution are preserved (i.e. a linear transformation is applied to each).
- The mean scaled score in each subject is equal to the mean scaled score across all subjects taken by all the students in that subject.

- The standard deviation of the scaled marks in each subject is equal to the standard deviation of the unscaled standardized marks across all subjects taken by all students in that subject.

AMS was introduced in Western Australia in 1998, replacing a system in which similar corrections were made on the basis of scores on a reference test, the Australian Scaling Test. It was found that rescaling marks based on the average scores each student had achieved in all their subjects gave results very similar to rescaling based on the reference test, but without the need to sit an additional test (WACC, 1998). Similar scaling approaches are used in a number of other places (e.g. other States in Australia, Cyprus, Fiji, Singapore, British Columbia) to make results in examinations of different difficulty comparable for purposes such as university entrance. However, as Lamprianou (2007) points out, their complexity has sometimes been seen to limit their public acceptability. This issue is discussed later in the context of policy options (see Section 10.1.3, p134).

### *2.1.3. Latent trait models*

The main method of this type is the Rasch model. We know of limited uses of this method (Coe, 2008; Tasmanian Qualifications Authority 2000), but believe it has a number of advantages over the alternatives.

The Rasch model (Rasch, 1960/1980; Wright and Stone, 1979) provides a method for calibrating ordinal data onto an interval scale. 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. It is a requirement of the model that

---

[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 (ie how 'hard' it is to agree with).

[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 $p/(1-p)$ and $q/(1-q)$ respectively. Hence the odds ratio is $[\,p/(1-p)\,] \;/\; [\,q/(1-q)\,]$. The logit function is

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

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

items should be unidimensional (i.e. all measuring essentially the same thing) and discriminate appropriately (i.e. more able persons are more likely to be successful). Unlike other latent trait models, the Rasch model further requires that all items discriminate equally, in other words, the relationship between a person's ability relative to an item and their probability of success on it should be the same for all items. For persons, their relative probabilities of success on different items should be in line with those of others in the population.

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.

In this way the Rasch approach is quite similar to the common examinee linear models described above, though it differs in two important respects. The first is that with Rasch it is possible to estimate the difficulties of each grade in each subject independently, using a 'partial credit' model (Masters, 1982). Hence there is no need to make any kind of interval assumption about the scales on which grades are coded; the Rasch model automatically assigns a value to each grade on a scale which may be said to have the 'interval' property, i.e., the same interval anywhere on the scale denotes the same difference in the probabilities of being achieved. This is a potentially important advantage since to use methods such as Kelly's or ASPA we must assume not only that the intervals between different grades in the same subject are equal, but also that these intervals are the same across all subjects.[4] Given Coe's (2008) finding that the intervals between GCSE grades are far from equal, this may be a significant advantage for the Rasch approach.

The other key difference is that the Rasch model requires the subjects and candidates analyzed to fit a particular model. In this context, fitting the model means that it must be possible to assign ability levels to all persons and difficulty levels to all items

[4] Of course we do not strictly have to assume that they are *equal*, but we have to make some assumption about their relative sizes. Note also that the AMS method requires an assumption about grade intervals within subjects, but not between subjects.

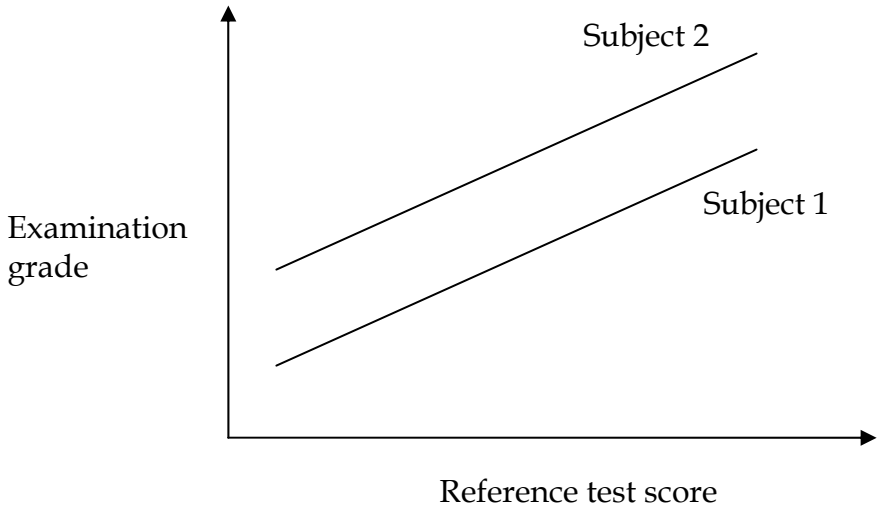
(i.e. subjects and grades) such that when we consider all candidates of a particular level of ability who have taken a particular subject, the proportion of them who achieved a particular grade should be reasonably close to what is predicted by the model. A key requirement for such fit is that both difficulty of items and ability of persons are unidimensional. In other words, there must be essentially just one kind of 'ability' which persons differ in the amount of which they exhibit and which largely accounts for their performance at all grades in all subjects.

If a particular subject, or at least a particular grade in a particular subject, does not fit the model, this will be evident and we can see which grades are not behaving as 'expected'. We can also identify any individuals or groups of candidates who are 'misfits'. The fact that the Rasch model specifically requires a unidimensional concept of ability seems to lend its results to interpretation in terms of the general level of ability represented by a particular achievement.

### *2.1.4. Reference tests*

The fourth group make use of a common reference test, such as an assessment of general ability, taken by all candidates who have taken the examinations we wish to compare. These methods were widely used by the examination boards in the 1970s but seem to have gone out of favour with them since (Murphy, 2007).

*Figure 1: Illustration of the reference test method*




Reference test comparisons 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 is commonly used, with grades in the various

subjects under comparison regressed on the ability test score. This allows the grades achieved by similar (in terms of their general ability) candidates to be compared. An illustration of the approach is shown in Figure 1. Here, candidates of similar ability are achieving generally higher grades in Subject 2 than in Subject 1, so we might conclude that the latter is more difficult (or more severely graded).

The conventional assumption for the use of this approach is that there is a close and equal relationship between the reference test and all the subjects being compared, in other words that the lines are parallel and correlations high (Murphy, 2007). However, Coe *et al.* (2007) have suggested that the results of reference test comparisons may still be interpretable even if these assumptions are not met. In particular, if the reference test is viewed as defining the linking construct (Newton, 2005) in terms of which these examinations are compared, it may be appropriate to compare the level of general ability (or other linking construct) typically indicated by the achievement of a particular grade in a particular examination. In this case, the graph in Figure 1 might be drawn the other way round, with reference test scores regressed on examination grades. With this interpretation, the reference test method is not just a special case of the 'value-added' approach (see below), but is distinctive in its meaning.

One potentially significant advantage of this method over the previous three is that there is no requirement for the examinations being compared to have any common candidates. Hence this method could be used to compare, for example, A-levels and Scottish Highers or International Baccalaureate scores, where few candidates are likely to have taken examinations of more than one type. It has also been used to compare the difficulty of examinations taken in different years.

### *2.1.5. 'Value-added' models*

The fifth and final statistical method is really an extension of the previous one, and the reference test approach could be seen as no more than a special case of the general 'value-added' method. In the general method, the regression model can include additional 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. In principle, many other factors such as motivation or quality of teaching received could be included, though the availability of data on such variables has limited this in practice. Value-added analyses have been widely used by awarding bodies in the UK, particularly since national matched datasets have become available, and the use of multilevel models has become a standard approach (Schagen and Hutchinson, 2007).

A possible multilevel model might be

$$y_{si} = \beta_0 + \sum_c \beta_c x_{ci} + \sum_s \beta_s z_{si} + u_i + e_{si}$$

*Equation 1*

$$u_i \sim \mathrm{N}(0, \sigma_u^2)$$
$$e_{si} \sim \mathrm{N}(0, \sigma_e^2)$$

where

$y_{si}$ is the grade achieved in subject $s$ by candidate $i$

$\beta_0$ is the intercept estimated by the model

$x_{ci}$ are a series of variables representing characteristics of candidate $i$, such as prior achievement, ability, sex, socioeconomic status, etc

$\beta_c$ are a series of fixed-effect coefficients of these characteristics

$z_{si}$ are a set of indicator dummy variables, one for each subject[5], $s$, which take the value 1 if candidate $i$ has taken subject $s$, and 0 otherwise

$\beta_s$ are the fixed-effect coefficients of the $z$s. $\beta_s$ therefore indicates the relative difficulty of subject $s$

$u_i$ is the level 2 (candidate) residual, i.e. the difference between the overall performance of candidate $i$ and what would have been predicted by the model, after taking account of their characteristics, $x_{ci}$ and the difficulties of the subjects taken

$e_{si}$ is the level 1 (examination result) residual, i.e. the difference between the grade achieved by candidate $i$ in subject $s$ and what would have been predicted by the model, given that candidate's characteristics, , and the difficulty, $\beta_s$, of subject $s$

In this model, examination results (level 1) are nested within candidates (level 2). A slightly more complex model might include a third level (school or centre) which would allow for the fact that students in the same school are typically more similar to each other than to those in other schools, and allow school-level residuals to be estimated.

The rationale for the value-added approach is that it presupposes a model of learning in which the learning gains made by students should be predictable from their initial characteristics. If the grading of different examinations is comparable, then

[5] In fact, with $n$ subjects, only $n-1$ variables are required since the final subject is estimated as the default category.

candidates who have made the same learning gains should achieve the same grades, regardless of which examination they take. On this basis, the justification for including particular variables in the model (such as gender or ethnicity) should be made on model-theoretic grounds, not just for opportunistic (the variable happens to be available) or statistical (it correlates with the outcome) reasons.

### *2.2. Judgement methods*

Goldstein and Cresswell (1996) argue that the fundamental problem with purely statistical approaches is that they ignore the educational content of the syllabuses and examinations concerned. Judgement methods rely on the decisions of expert scrutineers, often senior examiners employed by the examination boards, to equate standards based upon their experience (Wiliam, 1996; Adams, 2007). These methods have been widely used to try to compare examinations in the same subject taken at the same time (e.g. different syllabuses within a board or across boards) or in successive years (standards over time). In general they have not been used to try to compare examinations in different subjects, owing to the difficulty of agreeing on what basis one might compare them.

One of the key issues with judgement based methods is that, although it is often possible to make such judgements consistently in terms of overall difficulty it is difficult to decide how much easier or more difficult questions or examinations are. Cresswell (1997) demonstrated that the decisions of examiners are typically correct in identifying when papers are easier or harder from one examination to the next but are generally not correct in estimating how much easier or harder the papers are nor the specific source of those difficulties. Adams (2007) has described as 'the difficulty of comparing an easy task done well and a more difficult task done moderately'. Inevitably, this is a judgement call, but it is clear that different judges may take different positions on this.

A second issue is that such judgements are hard to calibrate across subjects as, within the examination system, expertise tends to be subject specific. If the 'standard' of an examination is held to reside in the skills, knowledge and understanding demonstrated in an observed examination performance judged to be worthy of a particular grade, then we must look for generic, cross-curricular criteria against which to compare two or more different subjects. Such criteria are likely to be either too broad and vague to be precisely operationalisable, or too narrow to be remotely representative of the core of either subject.

Adams (2007) raises a third issue for judging the difficulty of an examination, that of deciding whether a larger syllabus makes the same question more or less difficult. He gives the example of two syllabuses in history where the period studied in Syllabus I is entirely contained within that for Syllabus II. If a question on this period appears on both syllabuses' examinations, is it harder for those who must choose from a wider bank of knowledge (Syllabus II), or for those whose total knowledge is expected to be more limited (Syllabus I)? Or perhaps 'a question is a question' and

the difficulty is the same? Adams notes that when this issue arose in the context of setting grade boundaries for the same level of performance in AS and A-level examinations, 'the collective wisdom of the boards' research officers couldn't agree on this' and it was referred to the regulator.

There are two broad types of judgement methods that have been employed in establishing comparability across examinations:

### *2.2.1. Judgement against an explicit 'standard'*

In the first class of methods, the standard required for the award of a particular grade is made explicit and expert scrutineers must judge whether particular performances in different examinations do or do not meet the standard. These approaches have sometimes been described as 'cross-moderation' methods. Within this class we may identify two approaches: 'identification' and 'ratification'

An 'identification' study seeks to identify 'cut-scores' for grade boundaries by considering a range of scripts. If scrutineers can find one example that is just within the grade and another that is just below it, then the boundary is defined. By contrast, a 'ratification' study begins with scripts that have been judged to be near the relevant grade boundary and asks scrutineers to say whether they agree with that 'standard'. This method has sometimes been described as the 'home and away' approach when it is used to apply one examination board's standards to another's work.

One difficulty with both these types of approach to exemplifying the grade 'standard' is that grade descriptions are themselves problematic. Adams, 2007) points out that they suffer from the problems of 'atomisation', where the desire to describe the criteria precisely leads to a fragmented, reductionist view, the difficulty of accommodating 'compensation', which is the principle common in UK examinations that an inadequate performance in one area can be compensated for by an excellent performance in another, and the fact that grade descriptions generally describe typical, not borderline, performance.

### *2.2.2. Judgement against other scripts*

In recent years an alternative approach has become popular among UK awarding bodies, which avoids the problems of having to explicitly define a 'grade standard'. In the method of 'paired-comparisons', scrutineers are simply given a pair of scripts and asked to judge which is 'better' (Bramley, 2007). The results of multiple judges making these judgements about multiple pairs can be combined using Rasch analysis, putting all the scripts considered onto a single scale of 'quality'. If the scripts are the product of more than one examination and have already been graded with respect to that examination, then it is easy to see how the standards of the respective examinations may be compared.

The paired-comparison method has the advantage that two examinations can be compared without having to make grade descriptions explicit. However, it is still important to clarify exactly on what basis the judgement is made about which script

is 'better'. Issues about the relative value of an easy question done well compared with a hard question done less well, for example, remain. One of the strengths of the method, though, is that it enables the questions of how consistent different scrutineers are in their judgements, and how consistently different scripts are rated, to be investigated. With this method, it is possible to treat the question of whether it is appropriate or meaningful to compare two examinations as an empirical one: if a single latent trait emerges as the unitary construct in terms of which they are being compared then they are comparable, otherwise, not.

A variation on the method of paired-comparisons allows scrutineers to place a larger group of scripts into rank order. A partial credit model can then be applied to create the same kind of scale of 'quality' (Bramley, 2007).

## *2.3. How examination grades are awarded*

### *2.3.1. The grade awarding process in England, Wales and Northern Ireland*

Details of the grade awarding process for examinations in GCSE, GCE, GNVQ and AEA are set out in Section 6 of the Code of Practice for these examinations (QCA, 2007). According to this, grade boundaries 'must be set using professional judgement. The judgement must reflect the quality of candidates' work, informed by the relevant technical and statistical evidence' (6.13). The Code lists twelve different sources of evidence that must be considered by the Awarding Committee for that specification in this process:

> *'Qualitative*
>
> i copies of question papers/tasks and final mark schemes
>
> ii reports from the principal examiner(s)/principal moderator(s) on how the question paper functioned
>
> iii archive scripts and examples of internally assessed work (including, in appropriate subject areas, photographic or videotaped evidence) at the relevant grade boundaries, together with relevant question papers and mark schemes
>
> iv samples of current candidates' work (marked scripts and/or internally assessed material) distributed evenly across key boundary ranges for each component, with enough representing each mark to provide a sound basis for judgement so far as the size of entry and nature of work permit. The material should be selected from a sufficient range of centres where work has been marked/moderated by examiners/moderators whose work is known to be reliable
>
> v any published performance descriptions, grade descriptions and exemplar material, where available

vi any other supporting material (such as marking guides for components where the evidence is of an ephemeral nature)

*Quantitative*

vii technical information – including mark distributions relating to the question papers/tasks and individual questions for the current and previous series, where available

viii information on candidates' performance in at least two previous equivalent series, where available

ix details of significant changes in entry patterns and choices of options

x information on centres' estimated grades for all candidates including:
– qualification-level estimates for linear (including linear unitised) specifications
– unit-level estimates for externally assessed units in all other unitised specifications

xi information about the relationship between component/unit level data and whole-subject performance, where available

*Regulatory authority reports*

xii relevant evidence from the regulatory authorities' monitoring and comparability reports.'

(QCA, 2007, para 6.14)

Following the Awarding Committee, 'The chair of examiners' recommendations must be reviewed by the accountable officer to ensure that grades awarded represent continuity and parity of standards across years, over time and across specifications. In this review, the following evidence must be considered:

i reports from the awarding meeting, including the chair of examiners' recommendations

ii evidence of awarders' professional judgements on the quality of candidates' work within the range considered at the awarding meeting

iii the most complete technical and statistical evidence available, including that outlined in paragraph 6.14 and any generated subsequent to the awarding meeting (for example, information from cognate subjects).' (para 6.22)

We note that the emphasis in this process seems to be on the judgement of examiners, with statistical evidence in a secondary role of providing information to support that judgement. Moreover, the phrase 'where available' suggests that statistical evidence is seen as a desirable, but not essential, part of the process.

However, it seems likely that in practice there will be some variation in the weight that is given to different kinds of evidence.

We also note that there is no explicit mention of comparability across subjects, though 'across specifications' does in theory include this. However, our understanding is that in practice 'across specifications' has been interpreted as meaning across different syllabuses in the same subject or, at most, across 'cognate' subjects.

### *2.3.2. The grade awarding process in Scotland*

The Scottish Qualifications Authority (SQA) sees that it is their responsibility to ensure that a grade a student achieves represents the same standard from one year to the next and also that grades are not easier or harder to achieve across different subjects. Evidence presented at pass mark meetings following examinations is the main way SQA maintains these standards. These meetings result in the production of grade boundaries (the minimum marks required) for A and C grades. Grade boundaries for a B grade are set as the midpoint between the minimum mark for an A grade and C grade. The grade boundaries for a D grade are set as the pass mark for grade C minus 25% of the difference between the pass mark for grade C and grade A. The procedure for setting garde boundaries is briefly summarised here, but more details can be found in SQA (2005).

The grading system is not norm referenced, so there are not a fixed number of passes each year, instead if the paper is judged to present the same challenge and equal difficulty to the previous year then the same grade boundaries will be used.

Around 245 pass mark meetings are held each year, attended by a number of personnel who have expertise in their area and presided over by a Principal Assessor, whose professional judgement is central to the boundary marking process. During the meetings staff need to use professional judgement to decide whether the students and exams are comparable to previous years and whether the subject difficulty is comparables between subjects. Quantitative and qualitative evidence from the following areas are considered when setting grade boundaries:

- Evidence from the exam paper setting process e.g. changes in staff, changes in the structure of the paper
- Evidence from the marking process e.g. differences in the standard of marking, changes in the marking process
- Evidence from exam performance e.g. were there any mistakes in the paper? Did particular questions perform as planned
- Evidence on the candidates e.g. were the type of candidates similar to previous years? Has there been an increase in uptake?
- Quantitative information e.g. teacher estimates on how well they expect students to do, information on candidate population

- Statistical evidence e.g. the frequency distribution of scores in current and previous years, National Ratings (using Kelly method) for the past three years.

The exam process aims for the pass mark for a grade C to be 50% and 70% for a grade A. Taking all the above information into account professional judgement is used to decide whether, and by how much, the grades need to be adjusted to reflect the evidence. All proposals for grade boundaries are backed up with evidence.

The Relative Ratings that SQA produce every year for the Scottish examinations are used to inform the production of future papers. If papers in previous years have come out as difficult on the relative ratings then this will be taken into account in order to maintain equality across the subjects.

### *2.3.3. Raw Scores and Grade Boundaries*

One of the issues that often arises in discussion of the difficulty of an examination is the number of marks required to gain a particular grade. Examples of grades being awarded on the strength of strikingly low percentages of the available marks can be presented in the media as suggesting that the standard cannot be very high.

In fact of course the standard represented by any given percentage of marks achieved depends entirely on the difficulty of the questions and the mark scheme used. If the marks are hard to gain then even a small number of them may indicate a high standard of achievement. Hence we believe that it is not generally possible to say anything about the objective standard of a particular examination or grade from a knowledge of the grade cut-scores, without some knowledge of the questions that were asked, the context in which they were answered and the way marks were allocated to particular responses.

An example may illustrate the problem. One board's examination for AS level physics in 2006 reported that the pass mark for grade E was 40%, but at A2 (the full A-level examination, taken a year later) the equivalent mark was 52%. This certainly gives the impression that an E grade is harder to achieve at A2 than at AS, especially as the material covered in the second half of the course is generally more challenging than the first year's. However, even questions on harder material may be easier than those on easier material; it depends on the demands of the particular question, as well as on the mark scheme. Unless we know how hard the marks were to achieve, we cannot say that the requirement of higher marks necessarily implies greater difficulty. Of course, it may be that in this case A2 was 'harder' than AS, but the pass mark boundaries alone do not prove this.

Nevertheless, it does seem reasonable to infer something about the way an examination will be perceived by a candidate who gains only a small percentage of the marks. Such an examination is likely to be experienced as challenging and the overall experience of taking it may well feel rather negative. For this reason, we deal with the issue of cut-scores for grade boundaries in different examinations in Section 4.2 (p71) under the heading of the subjective perception of difficulty

## *2.4. Interpreting statistical differences in achievement*

Section 2.1 has outlined the main statistical methods for monitoring the comparability of examinations in different subjects. However, the question of exactly what 'comparability' means in this context has been the subject of much debate and some confusion. Other words, such as 'difficulty' or 'standards' are also used in this context and are also the source of controversy and misunderstanding.

Broadly speaking, all the statistical methods for comparing standards across subjects attempt to compare the grades achieved in one subject with grades achieved in other subjects by comparable (or the same) students. These statistical differences may not be completely straightforward to interpret, however. The next two parts of this report (Chapters 3 to 8) present results from a variety of such statistical comparisons, firstly, from existing studies and, secondly, from new research conducted for this report. It is clearly important, therefore, for the reader to have some guidance in how, if at all, any differences between levels of achievement in different subjects may be interpreted.

We reserve a full discussion of the meaning that may legitimately be attached to statistical differences, and the different meanings of terms such as 'difficulty', until Chapter 9, after the presentation of the results. Unfortunately, there is no perfect solution to the problem of having to present a report such as this in some order. While it may seem unsatisfactory to devote over 70 pages to the results of different analyses without first having considered what those results might mean, it is also less than ideal to launch into a discussion of the meaning of 'difficulty' and the various conceptual problems that arise in trying to define and measure it without having some real examples of analysis on which to draw. At the risk of being repetitive, therefore, we present a brief outline of some of the issues and possible interpretations here, in order that the reader may be able to relate them to the results that follow.

### *2.4.1. Problems with statistical comparisons*

If we know that the same students typically achieve a grade less in physics than they do in English, for example, it may seem to be an obvious step to conclude that physics must be harder. However, as several critics have pointed out, there may be a number of reasons why this does not necessarily follow. These arguments are presented in more detail in Section 9.1 (p115) but are outlined briefly here.

The first reason is that differences in achievement could be caused by several factors other than a genuine difference in the standard of the two subjects. Factors 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. This is the problem of 'other factors' (Section 9.1.1).

Second is that the whole idea of comparing two subjects makes sense only if they are in some sense 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 hard to imagine what this common element might be. This is the problem of 'multidimensionality' (Section 9.1.2)

Third is that any statistical comparison can only use those candidates who have actually taken a particular examination to calculate its difficulty. If they are not representative of all those who might possibly take it, then the calculation may not be appropriate. This is the problem of 'unrepresentativeness' (Section 9.1.3).

Fourth is that the statistical differences 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. This is the problem of 'subgroup invariance' Section 9.1.4)

Fifth is that 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 do not always seem to agree, so it is hard to know which estimate (if any) is the 'right' one. This is the problem of 'method inconsistency' (Section 9.1.5).

Sixth is the objection that making all subjects equal in difficulty would cause a number of new problems, including destroying comparability within subjects over time, causing confusion, delaying the awarding process. This is the problem of 'forcing equality' (Section 9.1.6).

### *2.4.2. Interpretations of statistical differences*

Some of the above problems are more serious than others, and we present more extended discussion of their merits in this report (Section 9.2) and elsewhere (Coe, 2007). Fundamentally, our position is that one cannot meaningfully talk about the validity of a particular method *per se* but must talk about the validity of a particular interpretation of a particular method. In other words, whether or not the method is valid depends on how you interpret and use it.

This distinction is important because there are a number of different ways the notion of subject 'difficulties' can be interpreted. Much of the criticism of the use of statistical methods has paid too little attention to the different ways in which their results may be interpreted, and so has inappropriately emphasised irrelevant objections.

Clearly, one can apply any of the methods outlined in Section 2.1 to any examination data. The question is how, if at all, one can then interpret the results. In Section 9.2 (p117) we present four possible interpretations.

The first (Section 9.2.1) is that no valid interpretation is possible. The objections to the use of statistical methods are such that the assumptions are untenable and the results invalid.

The second (9.2.2) is the 'learning gains' interpretation, that statistical differences may indicate the relationship between grades and learning gains in different subjects, provided other factors are taken into account. Precisely what the relevant 'other factors' are, however, may be less clear, and more research is needed before this interpretation can be argued convincingly.

The third (9.2.3) is the 'chances of success' interpretation, that statistical differences may indicate the relative chances of success in different subjects. We have argued that at least a crude form of this interpretation ('a candidate's relative chances of success, on the assumption they are typical of the current entry') does not depend on knowing anything about any 'other factors' that might account for statistical differences, or on some of the assumptions that have been claimed to be a requirement for statistical analyses of comparability (eg unidimensionality).

The fourth (9.2.4) is the 'linking construct' interpretation, that statistical differences may indicate differences in the relationship between grades achieved and some underlying construct such as 'general ability'. This interpretation probably does require examinations to be unidimensional and may also require subgroup invariance, but certainly does not depend on knowledge of any 'other factors' that might account for differences in performance.

Hence we believe that although interpretations of statistical differences are problematic, and it can be argued that statistical methods are invalid, it can also be argued that there are valid interpretations of their results. Indeed, there are three distinct and, we believe, compelling ways in which they can be understood and used.

.

# *PART II*
# *EXISTING EVIDENCE*

# 3. EVIDENCE FROM EXISTING COMPARABILITY STUDIES

In this chapter, we provide a more detailed summary of findings from past subject comparability studies. The summaries are presented in chronological order with different studies presented separately. At the end of this section, we will attempt to put forward some overall findings of this body of work.

## 3.1. Early work (up to the 1980s)

### *3.1.1. Osborn, L. G. (1939) - Relative Difficulty of High School Subjects*

This was an early study, conducted in the United States and did not directly analyse difficulties. We include it largely for historical interest – this issue is not new. Osborn analysed perceptions of difficulty in American High School students. It was noted that difficulty went beyond intellectual requirements of a subject: "it's complex and involves … likes, dislikes, ability, aptitude, teacher's personality". The study found that the hardest subjects for boys were Latin, chemistry, French/Spanish, mathematics, physics, English, history, biology, sociology. For girls, the hardest subjects were chemistry, physics, Latin, French/Spanish, mathematics, biology, English, history, sociology. It was notable that girls perceived science subjects to be harder than boys. Osborn suggested that the laboratory environment may affect attitude to science.

### *3.1.2. Nuttall, D.L., Backhouse, J.K. and Willmott, A.S. (1974) - Comparability of standards between subjects*

This study was published by the Schools Council 1974, but the work was available earlier as an NFER report. It noted that comparability presents researchers with technical and procedural problems of great complexity, so that it is difficult to produce firm, irrefutable evidence. They reported on NFER investigations into CSE and GCE O-level results for 1968. They used five methods of comparability and stated these led to essentially the same results, which indicated that chemistry and physics appeared to be more severely graded. However, this was not consistent across boys and girls when viewed separately. The authors noted that research at the JMB exam board came to the same conclusion regarding physics and chemistry, but added further that all results must be treated with great care.

In the study, they put forward the following much quoted fundamental assumption:

> *"We can see no reason why, if a large group of candidates representative of the population took, for example, both English and Mathematics, their average grades should not be the same"*

However, they found that English and Maths at CSE differed by about half a grade and suggested that this indicated English as being inherently easier than maths.

Nuttall *et al*. used the following five methods in their analysis:

1. Score on an aptitude test and regression against mean grade.

2. Used the aptitude test scores in a more sophisticated regression technique (referred to as a form of structural regression which they called the guideline method).

3. Subject pairs method across ten subjects. This involved:
- Identifying candidates who took a given subject XXX, and the nine other subjects;
- Calculating the mean grade in each pairing of XXX with the other nine;
- Calculate the mean of means for XXX and the other nine;
- The difference is a measure of severity or leniency of subject XXX.

4. The unbiased mean total (UBMT) method, where UBMT for an individual candidate is the mean grade of all other subjects attempted, omitting the one under consideration. For all candidates taking the subject under consideration, the UBMT is the mean grade of all other subjects attempted by those taking the subject under consideration. The measure of severity or leniency is the difference between the UMBT and the mean for the subject under consideration.

5. Analysis of variance, although the authors did not go into detail of how it was done.

Nuttal *et al*. therefore compared results for the five methods, looking specifically at one GCE O-level board. The results are given in the table below; the more positive the values given, the more difficult the subject.

*Table 1: Results from Nuttall et al. (1974)*

| subject | regression | guideline | Subject pair | UBMT | ANOVA |
|---|---|---|---|---|---|
| Art | -0.49 | -0.80 | -0.66 | -0.70 | -0.69 |
| Biology | -0.14 | -0.19 | -0.17 | 0.01 | -0.04 |
| Chemistry | 0.33 | 0.62 | 0.63 | 0.48 | 0.53 |
| Eng language | -0.49 | -0.57 | -0.70 | -0.70 | -0.64 |
| Eng literature | -0.24 | -0.32 | -0.27 | -0.29 | -0.29 |
| French | 0.25 | 0.29 | 0.54 | 0.45 | 0.43 |
| Geography | 0.09 | -0.05 | -0.08 | 0.02 | -0.01 |
| History | 0.16 | 0.04 | 0.21 | 0.25 | 0.21 |
| Mathematics | 0.12 | 0.18 | 0.01 | 0.05 | 0.04 |
| Physics | 0.37 | 0.72 | 0.50 | 0.43 | 0.46 |

The study noted discrepancies between the methods, but more so that there was consistency, particularly with methods 3, 4 and 5 which do not use an external test. They repeated the analysis with three other GCE boards and the general pattern was replicated. The results from Table 1 are also shown in Figure 3 (p65), which illustrates how close the agreement is among the different methods.

Table 2, taken from Nuttall *et al.*, illustrates the ranking of severity when using the ANOVA method for each of four English GCE O'level boards.

*Table 2: Difficulties of ten subjects from four different examination boards*

| subject | board 1 | | board 2 | | board 3 | | board 4 | |
|---|---|---|---|---|---|---|---|---|
| Chemistry | 0.83 | 1 | 0.53 | 1 | 0.71 | 1 | 1.03 | 1 |
| Physics | 0.36 | 3 | 0.46 | 2 | 0.22 | 5 | 0.46 | 3 |
| French | 0.27 | 4 | 0.43 | 3 | 0.63 | 2 | 0.47 | 2 |
| History | 0.17 | 5 | 0.21 | 4 | -0.5 | 3 | 0.18 | 4 |
| Mathematics | -0.09 | 6 | 0.04 | 5 | -0.4 | 7 | -0.47 | 8 |
| Geography | -0.2 | 7 | -0.01 | 6 | -0.54 | 8 | -0.25 | 7 |
| Biology | 0.58 | 2 | -0.04 | 7 | 0.23 | 4 | 0.1 | 5 |
| Eng literature | -0.28 | 8 | -0.29 | 8 | -0.04 | 6 | 0.01 | 6 |
| Eng language | -0.95 | 10 | -0.64 | 9 | -0.67 | 10 | -0.74 | 9 |
| Art | -0.68 | 9 | -0.69 | 10 | -0.63 | 9 | -0.78 | 10 |

The consistency in the rankings is notable, particularly chemistry as the most difficult subject.

The study noted particular assumptions made during the analyses, the validity of which have subsequently been much discussed by other authors.

- Candidates are equally motivated in all subjects;
- The teaching of each subject is equally good;
- The distribution of grades in each subject has the same shape.

They noted that their results were dependent on the acceptance of these assumptions, but the authors believed that these could be justified. However, they stressed that the publication of the study was to raise questions and stimulate discussion on comparability of subjects and not to come to firm conclusions.

### *3.1.3. WJEC reporting on 1971 O-level results*

Using a subject pairs method similar to Nuttall *et al.* (1974), the study concluded similar results. French and chemistry, and to some extent physics, are graded severely; biology and mathematics were close to their reference line or average. The study noted that this was for one year only, but is was different GCE board to Nuttall

*et al.*'s. Again, the report stressed that it was wishing to provoke discussion on the issue.

### *3.1.4. WJEC reporting on 1972 O-level results*

This study repeated the previous years' analysis and got the same results. It noted that pass rates were different for different GCE O-level subjects and asked, "do these various pass rates reflect a different calibre of candidates that enter for these subjects?" To measure this "calibre", the study discussed external testing vs. internal measures, for example like the one used previously where the internal measure were attainment in all other subjects.

The severity rankings given by the report are shown in Table 3

*Table 3: Ranking by relative difficulty of subjects in different years and levels*

| **1971 O'level** | **1972 O'level** | **1971 CSE** | **1972 CSE** |
|---|---|---|---|
| French | French | Mathematics | French |
| Chemistry | Chemistry | French | Mathematics |
| Physics | Physics | Scripture | Chemistry |
| History | History | Physics | Scripture |
| Mathematics | Biology | Biology | English literature |
| Biology | Mathematics | English literature | Physics |
| Art | English literature | Geography | Biology |
| English literature | Art | General science | Geography |
| Geography | English language | History | English |
| English language | Geography | English | Art |
| Scripture | Scripture | Art | History |

The report examined whether any subjects were had been graded "deviantly" out of line. They concluded that generally standards in grading across subjects were within a tolerance of half a grade.

### *3.1.5. Kelly, A. (1976a) - A study of the comparability of external examinations in different subjects: Scottish Higher Examinations 1969-1972*

The aim of the study was to investigate the myth of "soft options" and "tough subjects", that some subjects were easier than others. Kelly proposed a standardisation technique which could be used to " approximate a candidate's grade in a subject to that which would be obtained in the idealised situation in which all candidates took all subjects, and all subjects were marked by the same examiners".

She also noted that "the concept of ability is central to any attempt to standardise subjects", although there might be different abilities associated with different subjects. Her iterative technique, which extended the subject pairs technique of Nuttall *et al.* (1974), she says "compensates for the level of difficulty of an examination paper and the ability of the other candidates taking that paper, against whom an individual is compared."
This led to a correction table where the correction was applied to the mean pass grade for exams in the 5th year in Scotland. Corrections were calculated using 34 subjects, but only the 13 large entry subjects were reported.

*Table 4: Correction values calculated by Kelly (1976a)*

| *Subject* | *1972* | *1971* | *1970* | *1969* | *average* | *1971 boys* | *1971 girls* |
|---|---|---|---|---|---|---|---|
| *Latin* | *0.57* | *0.42* | *0.39* | *0.32* | *0.42* | *0.45* | *0.41* |
| *German* | *0.29* | *0.14* | *0.26* | *0.21* | *0.21* | *0.30* | *0/09* |
| *Chemistry* | *0.9* | *0.09* | *0.07* | *0.17* | *0.10* | *0.02* | *0.25* |
| *Physics* | *0.07* | *0.02* | *0.16* | *0.09* | *0.08* | *-0.04* | *0.26* |
| *French* | *0.13* | *0.02* | *0.08* | *0.04* | *0.07* | *0.2* | *-0.08* |
| *Mathematics0* | *0.0* | *0.06* | *0.16* | *-0.08* | *0.04* | *-0.01* | *0.18* |
| *History* | *0.02* | *0.0* | *0.13* | *0.14* | *0.07* | *-0.03* | *0.02* |
| *Biology* | *0.27* | *0.10* | *0.18* | *-0.09* | *0.11* | *0.04* | *0.16* |
| *Geography* | *-0.05* | *-0.08* | *-0.08* | *-0.08* | *-0.07* | *-0.15* | *0.02* |
| *English* | *-0.12* | *-0.09* | *-0.14* | *-0.06* | *-0.10* | *0.00* | *-0.19* |
| *Art* | *0.04* | *0.08* | *-0.21* | *-0.43* | *-0.13* | *0.02* | *0.10* |
| *Engineering drawing* | *-0.13* | *-0.42* | *-0.33* | *-0.56* | *-0.36* | | |
| *Home management* | *-0.31* | *-0.36* | *-0.33* | *-0.07* | *-0.27* | | |

In terms of grades awarded, there were three pass grades A, B and C to which Kelly gave numerical values of 3, 2 and 1. Thus, using her corrections, 0.42 should be added to the average pass grade in Latin and 0.27 should be deducted from the average pass grade in home management.

What was notable in this analysis was that the rank order of the severities were fairly consistent over the years considered. Also, this was in agreement with previous studies in that languages and sciences were graded more severely, which Kelly interpreted as meaning the examination papers were more difficult. A further analysis with 1971 data indicated there were some large differences in some subjects, notably that chemistry, physics and mathematics were found to beharder for girls than boys. Kelly concluded that it was impossible to standardise the subjects for all candidates and at the same time standardise them for boys and girls separately.

### *3.1.6. Newbould, C.A. (1982) - Subject Preferences, Sex Differences and comparability of Standards.*

Newbould cited Nuttall *et al*. (1974), and others including exam boards, as having evidence as to the differencing difficulties for subjects at O-Level. A possible explanation for this was that exam boards apply different grading standards to different subjects. The premise therefore was that if several candidates take maths

and English and do better on average in English, then maths is more severely graded. However, others have suggested motivation to be an important factor, and therefore a possible explanation for different grades.

Newbould carried out an investigation using O-level data on achievement. The following subject severity values for O-levels 1977 to 79, separated by gender, were calculated. The most severely graded subjects were the one s with the high positive numbers.

*Table 5: Difficulties of O-level subjects from Newbould (1982)*

| Subject | Boys | Girls |
|---|---|---|
| English language | -56 | -87 |
| English literature | 56 | -40 |
| History | -23 | -39 |
| Geography | -27 | -15 |
| Latin | 60 | 5 |
| French | 81 | -7 |
| German | 110 | 47 |
| Spanish | 53 | 17 |
| Mathematics B | -110 | -17 |
| Mathematics A | -89 | 14 |
| Physics | -29 | 65 |
| Chemistry | 16 | 68 |
| Biology | -42 | -10 |

It was found that generally, girls found mathematics and science harder than boys.

Newbould also correlated his severity results with the subject preference results of Duckworth & Entwistle (1974) of GCE candidates. It was found that there was relatively weak correlation (0.3-0.5) between the two sets of results.
The correlation exercise was repeated with subject preference data from Ormerod (1975 – reference required), and a higher correlation (0.8-0.9) was obtained.
Newbould also noted that his severity ordering were similar to those of Nuttall *et al.* (1974) and Kelly (1976).

He also noted that if these analyses have validity then between subject comparability using models based on the cognitive domain may take too simpler a view of the determinants of scholastic achievement; motivational factors may also be important.

### *3.1.7. Newbould, C. A. and Schmidt, C. C. (1983) - Comparison of grades in physics with grades in other subjects: Oxford & Cambridge A-level*

Subject pairs analysis carried out in this study indicated that:

- Physics was considerably more severely graded then biology, mathematics and Nuffield chemistry;
- Physics was considerably more leniently graded than further mathematics;

- Nuffield physics was more severely graded than biology and SMP maths;
- Nuffield physics was more leniently graded than chemistry and mathematics;
- Physics was more severely graded than Nuffield physics.

It should be noted that this study was carried out in an era of project-based assessment involved in Nuffield Physics, SMP maths etc. This had a basic intention of motivating students through making the subjects more interesting and accessible.

The authors stated that the subject parings lend support to the hypothesis that between subject differences are a function of the curriculum (A-level subjects) being followed. A candidate's likelihood of attaining a higher grade in subject x rather than subject y depended more on the orientation of these subjects within the overall curricular package being followed (typically x, y and one other subject). So subject grade comparisons are not, in themselves, infallible guides to between subject standards.

#### *3.1.8. Forrest G. M. and Vickerman C.( 1982) - Standards in GCE subject pairs comparisons 1972-80: JMB exam board*

This study was carried out as an aspect of monitoring standards, involving the routine practice of supplying information to grade award committees. In the introduction, it is stated that "comparability of standards involves in practice a number of extremely complicated issues … measurement of intellectual qualities is not a mechanically precise activity … all examinations involve some degree of imprecision and unreliability". However, they did state that "subject pairs analyses provide an invaluable piece of additional information about comparative standards in examinations."

They give data for O-level and A-level examinations 1972-1980, noting the consistency of grades over time. However, authors did indicate that results indicated a general severity order with languages, chemistry, physics and mathematics being more difficult than other subjects. However, in their conclusion, they stated that "it would be an over simplification of the process involved to suggest that a single factor, such as the previous year's subject pairs comparisons and the observations of a subject committee upon them, could ever override the whole range of evidence and judgements involved."

### *3.2. More recent studies (from the 1990s onwards)*

#### *3.2.1. Fitz-Gibbon, C. and Vincent, L. (1994) - Candidates' performance in mathematics and science*

This study addressed the question "is maths/science more difficult at A-level?" The authors defined difficult as "severely graded", which is indicated if attainment is below what would be expected on the basis of adequate statistics. As a result, analyses of ALIS data for 1993 A-level and 1991 GCSE examinations were carried out.

Four methods of comparison were used:

1. Grade pairs – compared maths/science grades with another subject taken. No great difference should be seen on the basis of equal difficulty, however differences were found and also for foreign languages.

2. Corrections factors (as used by Kelly, 1976) - took into account all grades achieved by an individual and examined the proportion of a grade that needed to be added or subtracted in order to equate for difficulty. Physics, chemistry, mathematics and biology (in that order) were shown to be above average in difficulty, together with languages and general studies.

3. Value-added with respect to mean GCSE - students taking maths/science tended to have higher mean GCSE scores and this led to them showing lower value-added progress at A-level. This implied a greater difficulty for these subjects or them being more severely graded at A-level. Girls also showed lower value-added than boys in all the mathematics and science subjects, indicating that girls achieved higher than expected grades at GCSE or lower at A-level.

4. Value added with respect to baseline International Test of Developed Ability. - similar results were obtained as for the previous value-added analysis. Differences in subject difficulties for individuals ranged from a third of a grade up to a whole grade and a quarter. If aggregated over a whole school, it was suggested that the difference could be substantial and may lead to schools and colleges asking students to avoid maths and science subjects. Fitz-Gibbon and Vincent therefore noted a need for further research in the following areas:

- Why the imbalance in some schools and colleges where the take up of maths and science is relatively high, compared to relatively low take up elsewhere?
- What is the long term impact of subject choice?
- What use is made of grade information by admission tutors, employers, careers officers etc?

Based on their analyses, they calculated the correction factors for different subjects shown in Table 6 and Figure 2 (N = number of candidates for which data was available; c.f. = correction factor).

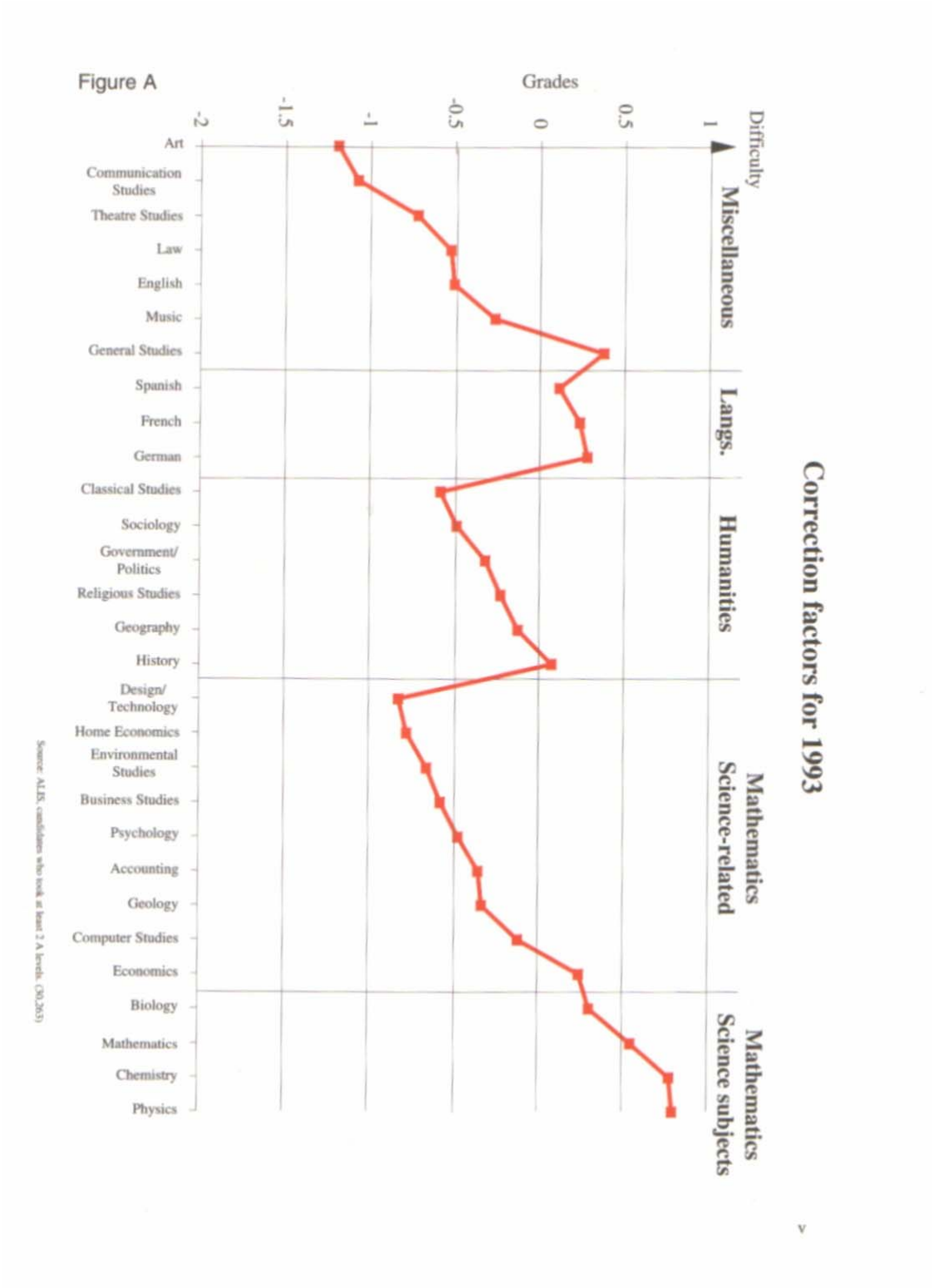
*Table 6: A-level subject difficulties, from Fitz-Gibbon and Vincent (1994)*

| by subject clusters | | |
|---|---|---|
| | N | c.f. |
| Art | 3114 | -1.19 |
| communication studies | 790 | -1.07 |
| theatre studies | 1020 | -0.72 |
| Law | 678 | -0.53 |
| English | 10865 | -0.51 |
| Music | 580 | -0.27 |
| general studies | 6309 | 0.38 |
| Spanish | 573 | 0.11 |
| French | 3766 | 0.23 |
| German | 1500 | 0.28 |
| classical studies | 555 | -0.58 |
| Sociology | 3607 | -0.49 |
| government /politics | 1436 | -0.32 |
| religious studies | 1008 | -0.23 |
| geography | 5728 | -0.12 |
| history | 5839 | 0.07 |
| design technology | 1348 | -0.82 |
| home economics | 463 | -0.78 |
| environmental studies | 244 | -0.66 |
| business studies | 3353 | -0.58 |
| psychology | 1687 | -0.47 |
| accounting | 390 | -0.35 |
| geology | 479 | -0.33 |
| computer studies | 1476 | -0.12 |
| economics | 4343 | 0.24 |
| biology | 5931 | 0.30 |
| mathematics | 8336 | 0.55 |
| chemistry | 5226 | 0.78 |
| physics | 5049 | 0.79 |

| by rank order of correction factor | | |
|---|---|---|
| | N | c.f. |
| physics | 5049 | 0.79 |
| chemistry | 5226 | 0.78 |
| mathematics | 8336 | 0.55 |
| general studies | 6309 | 0.38 |
| biology | 5931 | 0.30 |
| German | 1500 | 0.28 |
| economics | 4343 | 0.24 |
| French | 3766 | 0.23 |
| Spanish | 573 | 0.11 |
| history | 5839 | 0.07 |
| geography | 5728 | -0.12 |
| computer studies | 1476 | -0.12 |
| religious studies | 1008 | -0.23 |
| music | 580 | -0.27 |
| government /politics | 1436 | -0.32 |
| geology | 479 | -0.33 |
| accounting | 390 | -0.35 |
| psychology | 1687 | -0.47 |
| sociology | 3607 | -0.49 |
| English | 10865 | -0.51 |
| law | 678 | -0.53 |
| classical studies | 555 | -0.58 |
| business studies | 3353 | -0.58 |
| environmental studies | 244 | -0.66 |
| theatre studies | 1020 | -0.72 |
| home economics | 463 | -0.78 |
| design technology | 1348 | -0.82 |
| communication studies | 790 | -1.07 |
| art | 3114 | -1.19 |

Plotting these correction factors on a graph:

*Figure 2: A-level subject difficulties, from Fitz-Gibbon and Vincent (1994)*




The study concluded that maths and science subjects were graded more severely than other subject areas.

### 3.2.2. *Alton, A. and Pearson, S. (1996) - Statistical Approaches to Inter subject Comparability: 1994 A-level data for all boards, part of the 16+/18+ project funded by DfEE*

This study investigated several statistical methods for estimating adjustment factors between subjects at A-level and explored the implications of trying to apply such factors in order to align grade distributions. They raised issues of practical limitations and theoretical doubts over validity. Their main conclusions are summarised below:

- Inter-subject relationships are not consistent across time so that adjustment made in any one year would not necessarily be appropriate to subsequent years; entry patterns change and the curriculum changes.
- Inter-subject relationships are not consistent across identifiable subsets (such as boys /girls) so that there would remain major statistical inequalities in outcomes.
- Inter-subject relationships are not consistent across the grade range, so either a very complex system with several adjustment factors would have to be used or one with a single factor which would not be appropriate for certain grades.
- Outcomes of any adjustments would cause great confusion in users, the public and professional bodies.
- Post hoc adjustments change the relationship between judgement and statistics which is at the heart of grading traditions and which form the basis of the SCAA Code of Practice.
- There is a consistency in the findings of the various (previous) studies conducted over a range of examination levels, at different times, and in many geographical regions, a consistency which tends to suggest that there is a loose hierarchy of inter-subject difficulty, which remains fairly constant whatever the context.

By using data from all exam boards, the authors addressed the criticism of unrepresentative data. They used four statistical methods in their analysis:

- Based on prior attainment with mean GCSE grade;
- Subject pairs – weighted according to number of candidates taking the subjects;
- Subject pairs – unweighted;
- Subject triples – unweighted;

The authors noted that there was no reason to say that one method was better than any other. The results of their various analyses are given in the table below:

*Table 7: A-level subject difficulties by different methods, from Alton and Pearson (1996)*

| 1994 A-level Subject | Method | | | | Rank | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean GCSE Grade | Subject pairs weighted | Subject pairs unweighted | Subject triples | | | | |
| Biology | -0.48 | -0.11 | -0.40 | -0.16 | 6 | 9 | 4 | 4.5 |
| Chemistry | -0.68 | -0.36 | -0.40 | -0.43 | 1 | 2 | 3 | 2 |
| Physics | -0.55 | -0.35 | -0.54 | -0.22 | 4 | 3 | 2 | 3 |
| Mathematics | -0.48 | -0.18 | -0.30 | -0.09 | 5 | 8 | 6 | 8 |
| Maths pure | -0.21 | -0.06 | | | | | | |
| Maths applied | 0.10 | 0.04 | | | | | | |
| Maths further | -0.60 | -0.67 | -0.62 | -0.60 | 2 | 1 | 1 | 1 |
| Business studies | 0.51 | 0.36 | 0.35 | 0.61 | 12 | 12 | 12 | 13 |
| Art & design | 0.96 | 0.86 | 0.94 | 0.88 | 14 | 14 | 14 | 14 |
| Geography | 0.01 | 0.09 | 0.03 | 0.32 | 10 | 10 | 10 | 10 |
| History | -0.09 | -0.26 | -0.28 | -0.13 | 9 | 5 | 7 | 6 |
| Economics | -0.16 | -0.19 | -.014 | -0.16 | 8 | 7 | 9 | 4.5 |
| English | 0.55 | 0.56 | 0.51 | 0.56 | 13 | 13 | 13 | 12 |
| Eng literature | 0.37 | 0.33 | 0.32 | 0.35 | 11 | 11 | 11 | 11 |
| French | -0.57 | -0.26 | -0.25 | -0.10 | 3 | 4 | 8 | 7 |

Numerical values for grades were based on A=5, B=4, C= 3, D=2, E=1, so the values in the table are measure in fractions of severity or leniency of grading. The results are consistent with previous studies; mathematics and the sciences, economics and French are more severely graded (more difficult) than the humanities and arts subjects. Comparing boys and girls, the authors observed that in ten of the sixteen subjects the deviation from expectation (based on mean GCSE) for that subject was less than the difference between that value for males and females. Therefore, removing the overall subject deviation would leave a larger one untouched with regards to gender differences.

The authors concluded that it may be argued that the analysis reveals a real phenomenon. However, its persistence across time, examination levels, and cultures may suggest that whatever reality it possesses, it is a reflection more of the complex bundle of factors which influence a student's attainment rather than inter-subject differences in standards. "It is hard to resist the conclusion that any attempt to tinker with standards would lead to more unjustifiable outcomes rather than fewer".

### *3.2.3. Dearing, R. (1996) - Review of qualifications for 16-19 year olds*

A criticism of the Fitz-Gibbon and Vincent analysis was that the schools and colleges were self selecting as they had voluntarily signed up to be part of the ALIS project. Dearing claimed that not only did this result in relatively small samples for some subjects, but also state comprehensive schools and sixth form colleges were over

represented in the data, whereas selective schools, independent schools and further education colleges were under represented. Therefore, another analysis was carried by the School Curriculum, Assessment Authority (SCAA), using the methodology of Fitz-Gibbon and Vincent, drawing on the national database of A-level results for England for 1993,1994 and 1995, thus addressing this criticism.

The results were published by Dearing in his report, and these are reproduced in Table 8. The results are for students who took 3 or more A-levels at age 18.

*Table 8: A-level difficulties over three years, from Dearing (1996)*

| subject | 1993 | | | 1994 | | | 1995 | | | mean |
|---|---|---|---|---|---|---|---|---|---|---|
| | N | c.f. | rk | N | c.f. | rk | N | c.f. | rk | c.f. |
| mathematics | 46323 | 0.90 | 1 | 44748 | 0.38 | 7 | 43981 | 0.75 | 1 | 0.68 |
| chemistry | 26814 | 0.36 | 3 | 27055 | 0.65 | 1 | 28384 | 0.58 | 2 | 0.53 |
| general studies | 46496 | 0.69 | 2 | 45421 | 0.47 | 3 | 48921 | 0.30 | 7 | 0.48 |
| physics | 25435 | 0.32 | 4 | 23774 | 0.63 | 2 | 22888 | 0.24 | 9 | 0.40 |
| French | 20565 | 0.25 | 6 | 19778 | 0.42 | 4 | 19351 | 0.48 | 3 | 0.38 |
| German | 7807 | 0.23 | 8 | 7868 | 0.41 | 6 | 7657 | 0.46 | 4 | 0.37 |
| history | 28970 | 0.25 | 7 | 28456 | 0.41 | 5 | 29001 | 0.35 | 6 | 0.34 |
| economics | 22101 | 0.29 | 5 | 18880 | 0.32 | 8 | 16744 | 0.38 | 5 | 0.33 |
| biology | 27311 | 0.03 | 9 | 29696 | 0.19 | 9 | 31804 | 0.27 | 8 | 0.17 |
| Spanish | 2775 | 0.01 | 10 | 2697 | -0.06 | 10 | 2803 | 0.09 | 11 | 0.01 |
| social science | 26604 | -0.02 | 12 | 29325 | -0.12 | 12 | 31901 | 0.12 | 10 | 0.00 |
| classical studies | 5281 | 0.00 | 11 | 5346 | -0.22 | 15 | 5240 | 0.02 | 12 | -0.07 |
| music | 3238 | -0.05 | 13 | 3436 | -0.26 | 16 | 3626 | -0.15 | 14 | -0.16 |
| geography | 29341 | -0.32 | 15 | 20215 | -0.11 | 11 | 28685 | -0.11 | 13 | -0.19 |
| religious studies | 4654 | -0.19 | 14 | 4792 | -0.13 | 13 | 5190 | -0.28 | 16 | -0.20 |
| other science | 2946 | -0.49 | 17 | 3042 | -0.17 | 14 | 3464 | -0.24 | 15 | -0.30 |
| computer studies | 4453 | -0.49 | 16 | 4461 | -0.37 | 17 | 5000 | -0.46 | 17 | -0.44 |
| communication studies | 7828 | -0.62 | 20 | 8901 | -0.58 | 20 | 10855 | -0.50 | 18 | -0.56 |
| English | 49261 | -0.61 | 19 | 50039 | -0.58 | 19 | 50790 | -0.54 | 19 | -0.58 |
| business studies | 10844 | -0.70 | 21 | 12479 | -0.56 | 18 | 13922 | -0.58 | 20 | -0.61 |
| vocational studies | 1952 | -0.55 | 18 | 1653 | -0.94 | 22 | 1665 | -0.87 | 23 | -0.77 |
| other languages | 1150 | -0.83 | 22 | 1141 | -0.94 | 23 | 1245 | -0.87 | 24 | -0.88 |
| design technology | 5978 | -1.07 | 24 | 6125 | -0.89 | 21 | 3874 | -0.85 | 21 | -0.95 |
| home economics | 1621 | -1.22 | 25 | 1558 | -1.05 | 24 | 1532 | -0.85 | 22 | -1.04 |
| art & design | 15590 | -1.01 | 23 | 15924 | -1.23 | 25 | 16096 | -1.06 | 25 | -1.10 |
| physical education | 1685 | -1.59 | 26 | 2461 | -1.50 | 26 | 3623 | -1.41 | 26 | -1.48 |

The subjects are shown in descending order of difficulty, ordered by a three year weighted average of the correction factors. Ranks in the three years are indicated by

'rk'. It was found that Dearing's analysis confirmed Fitz-Gibbon and Vincent's. However, Dearing noted these results did **not** hold when he did an analysis for those students taking 2 A-levels and all subjects were much more similar.

Dearing also compared subject difficulties with higher education entry requirements as represented by UCAS points. Table 9 shows the average requirements at English universities at the time. (A=10, B=8, C=6, D=4, E=2)

*Table 9: Subject difficulties and HE entry requirements*

| A-level difficulty | | Higher education entry requirements | |
|---|---|---|---|
| Subject | Correction factor | Subject | UCAS points |
| Mathematics | 0.68 | French | 20.5 |
| Chemistry | 0.53 | English | 18.5 |
| Physics | 0.40 | History | 17.9 |
| French | 0.38 | Business studies | 16.3 |
| History | 0.34 | Mathematics | 15.9 |
| Biology | 0.16 | Physics | 15.6 |
| Computer studies | -0.44 | Biology | 15.4 |
| English | -0.58 | Computer studies | 15.3 |
| Business studies | -0.61 | Chemistry | 14.4 |

Dearing noted that the picture changes somewhat if the higher education institutions use the UCAS clearing scheme to fill their course places. In a further analysis, Dearing calculated the number of UCAS points actually scored by students entering degree courses.

*Table 10: UCAS points gained by entrants in different degree courses*

| UCAS points | Subjects |
|---|---|
| more than 20 | mathematics, physics, French |
| 18 –20 | English, history, geography, economics, German, chemistry |
| 16-18 | music, fine art, biology, business and management studies, theology |
| less than 16 | computer studies , design studies |

Dearing noted there were variations on this table across various institutions; he called it the institutional effect, noting certain institutions, for example, some of the former polytechnics (pre 1992) may treat some subjects differently to others.

Dearing's conclusions were therefore as follows:

- There are differences in subject difficulty at A-level. To what extent this is a new phenomenon is not known. These differences may reflect conceptual differences across the subjects themselves, for example physics is not only different in content to geography but requires different skills and almost certainly requires understanding of inherently more difficult concepts (in terms of the general publics' understanding of such concepts).
- There is evidence to show that some of the more difficult subjects at A-level (physics, chemistry, mathematics) are compensated by relatively lower entry requirements to higher education.

- Students starting higher education courses in physics, chemistry and mathematics are amongst the most highly qualified in terms of UCAS points.
- Higher education institutions that mainly offer courses related to the more difficult A-levels tend to recruit students who have achieved higher A-level grades generally.
- In some subjects, such as history and French, higher education institutions appear to make no compensation for relative difficulty of A-level; this may be a function of supply and demand.
- Mathematics and science degree courses are more difficult to fill than some other courses.
- Bridging courses are offered to a much greater degree in mathematics and science related subjects than in others.

### *3.2.4. Pollitt, A. (1996) - The "difficulty" of A-level subjects*

This study investigated whether the relative difficulty rankings which seemed to be established in culturally similar countries ( England, Scotland, New Zealand, South Africa, Australia) would be found in a Pacific Rim country sitting English A-levels. Results from a subject pairs analysis with Mathematics as a referent subject is shown below.

*Table 11: Subject difficulties from a Pacific-Rim country*

| Subject | "difficulty" relative to mathematics |
|---|---|
| Phyics | 1.7 |
| General studies | 1.5 |
| Economics | 0.83 |
| Chemistry | 0.75 |
| Business | 0.30 |
| Biology | 0.22 |
| History | 0.12 |
| Mathematics | 0.00 |
| English | -0.58 |

Thus Physics seemed to be 1.7 grades harder than Mathematics. Therefore, the study in general replicated the "western world" pattern with the exception of mathematics, the Eastern countries performing better at maths. On gender, he found females generally found A-levels to be harder than males. However, the author cautions that subject pairs analysis can be dangerously misleading, offering the simple explanation that in England (or wherever), there are more people good at English than are good at mathematics as one might expect.

### *3.2.5. Patrick, H. (1996) - Comparing Public Examinations Standards over time*

Patrick argued that comparability over long periods (20 years as proposed by SCAA) was impossible because of too many changes in, for example, culture or technology, and that numbers staying on in education (sitting GCSEs, A-levels, going to university) were all rising. She claimed that even 5 years (standard time in board cross moderation exercises) is a long period in which external changes can affect 'standards', i.e. grade allocation.

> *"One of the difficulties faced by those who work in this field is that their results are frequently expected to bear a weight of interpretation far beyond what they can reasonably support".*

However, she noted that it was now in the Code of Practice that archive scripts from one year earlier must be used in the assessment of grade boundaries. She also noted that comparisons over time can indicate trends; e.g. girls rising achievement in GCSE and A-level.

Patrick asserted that the question was not whether standards were rising or falling, but whether they are appropriate. Using maths as an example:

> *"With one lobby claiming that A-level Mathematics is too hard in comparison with other A-level subjects, and another lobby claiming that A-level Mathematics is not hard enough because holders of the qualification are ill-equipped for what is required of them in higher education. These kinds of claims and counter claims do not get us anywhere. Work like that of Sutherland and Pozzi, however, can provide the basis for a more sensible debate. What is A-level Mathematics for? Who is it for? What would it be appropriate for candidates with A-level Mathematics to be able to do in the 1990s? If we could tackle such fundamental questions, and reach agreement on the answers, we would be in a position to design syllabuses and examinations accordingly."* (p.9)

### *3.2.6. Wiliam, D. (1996a) - Meanings and Consequences in Standard Setting*

This was a further philosophical discussion of what are meant by standards. Wiliam rejected norm and criteria referencing in favour of construct referenced. He stated that selection of grade cut off points are arbitrary but not random in that they are based on judgement. He concluded that the validation of standards must include consideration of their consequences as well as their meaning: "where standards exist, they do so by virtue of a shared construct in a community of interpreters, so that all standards are ultimately construct referenced."

### *3.2.7. Wiliam, D. (1996b) - Standards in Examinations: a matter of trust?*

Another discussion of what standards are and who uses them, again putting the construct referenced view. Standards are what a group of people charged with making the judgement say they are. …He quotes Cresswell (1996):

*"Two examinations have comparable standards if candidates for one of them receive the same grades as candidates for the other whose assessed attainments are accorded equivalent value by awarders accepted as competent to make such judgements by all interested certificate users"*

Wiliam therefore argues "that by making central the *value* that is placed on awards, many of the technical difficulties are obviated … the examination system functions only to an extent that the interested certificate users *trust* the judgement of awarders." However, it must be recognised that what he fails to acknowledge is that the awarders' judgment is informed by statistics of both present and past performance. He concludes that "there is no way of establishing comparability of standards other than the professional judgement of a community of experts…. all attempts to define standards or equivalence independently of the social setting in which they are created are bound to fail… we have to *trust* the awarders".

### *3.2.8. Goldstein, H. and Cresswell, M. (1996) - The comparability of different subjects in public examinations: a theoretical and practical critique*

This study criticised attempts at comparability, particularly the search for correction factors. As different examinations have different entry populations, they also bring to the examinations different learning experiences. Also, they are critical of assumed unidimensionality, that exams, even in the same subject, are assumed to be measuring the same thing.

They are particularly critical of Fitz-Gibbon and Vincent (1994). In particular, with subject pairs, one cannot assume that quality of teaching and general education provision have been the same for different subjects. Motivation and interest are also factors to consider. Also there is variation about the mean grade, so any adjustments to grade boundaries are unfair to some candidates. Subject pairs analysis is dependent on stable populations, which is not attainable.

On the use of mean GCSE grades, they argue that achievement across subjects is multidimensional; subjects vary so why should the mean GCSE be a better measure than some other combination of achievements? They also claim that linear regression is misleading, and are critical that a multi-level model approach was not used. Fitz-Gibbon and Vincent also ignored the educational content of the syllabuses and the examinations concerned. They state that "purely statistical procedures which rely upon the information available from a particular cohort of students can never be a valid base for judging, maintaining, or adjusting examination grading standards." They infer that it is too politically sensitive, impractical and ethically indefensible to contemplate any adjustments to grades awarded as a result.

### *3.2.9. Fitz-Gibbon, C. and Vincent, L. (1997) – Difficulties regarding subject difficulties*

Responding to the criticism from Goldstein and Cresswell, Fitz-Gibbon and Vincent stated they analysed data, not philosophise about the meaning of difficulty. However, they did point out that:

*"the term difficult cannot be taken as meaning necessarily or intrinsically difficult; rather subjects are said to be difficult or severely graded if the grades awarded are generally lower than might be reasonably expected on the basis of adequate statistics".*

The authors presented their 1994 results as empirical facts resulting from analysis of data. They point out the same analysis carried out for the Dearing report replicated the results.

They point out they had been asked to analyse data; not consider judgement. They counter the criticism of unidimensionality by pointing out this is how grades are interpreted when it comes UCAS points, an A is an A. Also all grades are independent of subjects in performance tables, which has implications for the take up of "easier subjects". They point this out as an important issue that is raised through their analysis. Similarly, some students may disadvantage themselves in university course and career choice, by taking inappropriate A-levels.

They note that throughout the Western world, there is a phenomenon that more able students are attracted into maths, science and foreign languages, as it is seen in many datasets. They criticised Goldsmith and Cresswell for not acknowledging such a simple explanation. On the need for multi level modelling, they dismissed this as not necessary for data of this nature, as it will make little difference to the results.

### *3.2.10. Newton, P. (1997) - Measuring Comparability of Standards between subjects: why our statistical techniques do not make the grade*

Newton is highly critical of subject pairs analysis as a way of investigating comparability. He states that "these techniques cannot be assumed even to approximate a valid representation of the problem of between subject comparability because they are inappropriate for dealing with the kind of data that our examinations generate".

He says the problem is that when you break down the sample of students into subgroups (boys/girls) you do not always get the same correction factors. He notes any correction factors would also need to vary between types of school, and also argues that previous analyses have shown correction factors are affected by the third / fourth subject taken. Therefore, he believes subject pairs, even as general indicators of subject difficulty, are flawed.

His main criticism is of Nuttall's fundamental assumption that if the sample is large enough, the differences in individual candidates will be ironed out. Newton argues a representative sample is unattainable.
He highlights other assumptions implicit in the subject pairs method and in particular the notion of 'general academic ability'. The basic reason why subject pairs analysis fails is that candidate's performance in public examinations for different subjects are not determined by a unitary underlying "general academic ability", to

the extent that this could provide a meaningful baseline for comparison. He concludes that if there are any genuine differences in standards between subjects, our statistical measures are not capable of determining where they lie.

### *3.2.11. Fowles, D. E. (1998) - The translation of GCE and GCSE grades into numerical values.*

Fowles does not consider subject comparability. Rather, she is critical of coding A=7, B=6 etc., in that it assumes grades form an equal interval scale, whereas the marks distribution in each grade indicate they do not.

### *3.2.12. Baird, J., Cresswell, M. and Newton, P. (2000) - Would the real gold standard please stand up?*

Highlights that "any particular examination script is always a product of the performance of the candidate and the difficulty of the examination and these two factors cannot be disentangled." The paper is a philosophical discussion on what are standards and how standards can be seen from conflicting perspectives. They argue that statistical approaches to examination comparability are not objective as they involve value judgements in the selection of control variables and in the interpretation of the relationships between the controls and the examinations. They defer to the sociological perspective (Wiliam's social construct) of standards, where standards exist only in the act of judgement of grade awarders: "in practice Chief Executives of awarding bodies are the custodians of examination standards in GCE and GCSE". They liken decisions made by awarders and Chief Executives to those of juries and judges.

### *3.2.13. Sparkes, B. (2000) - Subject Comparisons - a Scottish Perspective: Standard Grade 1996 – Highers 1997*

Sums up the debate between Fitz-Gibbon and Vincent, and Goldstein and Cresswell: "whether this phenomenon is because mathematics is inherently more difficult to learn, or because it is less interesting or less motivating than English, or because it is less well taught or resourced, or because its examination is more severely graded, is not known."

Sparkes points out that Goldstein and Cresswell's criticisms do not answer the concern that students who believe maths, sciences and modern foreign languages to be difficult (what ever that means) may be deterred from taking those subjects.

Sparkes undertook an analysis of Scottish Highers data, noting first that the correction factors developed by Kelly 1976 have been calculated and published annually by the SQA as information available to schools.
He calculated his own correction factors, noting these were similar to those calculated by the SQA using sophisticated software. He was particularly interested in the differences in corrective factors between sub groups of candidates, arguing that

his results for the correction factors were valid, as they had a 95% confidence interval range of 0.2.

He classified candidates into groups according to their GPA (grade points average - mean grade score at standard grade). Although there is variation across the sub groups (for example maths and the sciences are more difficult for less able candidates) the trend for all groups is consistent, and consistent with Fitz-Gibbon and Vincent and others findings. He similarly compared the correction factors for boys and girls, and found results similar to previous studies, i.e. that girls find maths and sciences harder than boys. He noted that girls find mathematics more difficult than boys at all ability levels and that girls find English easier, and he raises the question as to why this is the case. Further analysis into maths and English involved looking at a third and fourth subject studied. Sparkes classified candidates as arts or science candidates according to subjects taken. He noted that the two groups had approximately equal GPAs but the arts candidates found maths and sciences more difficult than the arts subjects and vice versa, suggesting an interest (motivation) factor was present. He suggests his figures underline the observation that "subject difficulty is individual – no subject is 'more difficult' for everyone".

Sparkes compared his results with Fitz-Gibbon and Vincent's and so contrasts England and Scotland. His results indicate that in Scotland, students are not apparently deterred by perceived difficulties of subjects, as maths and science are popular, noting also that students generally take 4 or 5 Highers as opposed to 3 A-levels in England.

Sparkes concluded that his results support those of Fitz-Gibbon and Vincent in that languages, maths and sciences are more difficult subjects. The variation in the sub-groups though leads him also to support Goldstein and Cresswell in that subject difficulty measured in purely statistical terms is unhelpful; it tells us nothing about motivation for and interest in a subject. He cites DES 1980 (reference required) that student subject choice is more influenced by career aspirations than interest in a subject, and suggests the rise or decline in a subjects popularity may have more to do with the national economy and current culture than with that subjects perceived difficulty. He concludes that "difficulties are only a problem for those using examination results for purposes they were not devised for, such as producing school league tables by adding up A-level points or counting the number of subjects passed in Highers."

Sparkes' results for boys and girls are given below.

*Table 12: Subject difficulties for Higher Grade, split by sex*

| Subject | male | female | overall | M-F difference |
|---|---|---|---|---|
| Art & design | 1.44 | 1.79 | 1.64 | -0.35 |
| Biology | -0.22 | -0.52 | -0.38 | 0.30 |
| Chemistry | -0.43 | -1.02 | -0.70 | 0.59 |
| Drama | 0.79 | 1.46 | 1.30 | -0.67 |
| Economics | -0.13 | -0.59 | -0.38 | 0.46 |
| English | 0.09 | 0.40 | 0.26 | -0.31 |
| French | -0.65 | -0.56 | -0.54 | -0.09 |
| Geography | 0.39 | 0.27 | 0.26 | 0.12 |
| History | 0.21 | -0.13 | -0.02 | 0.34 |
| Mathematics | -0.60 | -0.95 | -0.76 | 0.35 |
| Physics | -0.45 | -1.05 | -0.64 | 0.60 |

### *3.2.14. Baker, E., McGaw, B. and Sutherland, S. (2002) - Maintaining GCE A-level Standards*

This study was carried out by an independent panel of expert advisers whose remit was to review the quality assurance procedures that are in place to maintain standards in GCE A-level examinations. Standards were said to involve demands of the specifications and their associated assessment arrangements and the levels of performance required of candidates to gain particular grades.

The panel assessed the quality of the examination systems based on accuracy, validity and fairness. They did not consider inter-subject standards explicitly, but concluded that QCA was doing a commendable job to assure the quality of the A-level examinations. They also stated that there was no scientific way to determine in retrospect whether standards have been maintained.

The study made eight recommendations including :

- QCA should employ a convening function to air issues associated with standards in key areas such a mathematics and science;
- QCA should expand its communication programme to help the public and the profession understand the benefits and limits of its testing programmes and of any modifications being introduced.

### *3.2.15. Jones, B. (2003) - Subject pairs over time: a review of the evidence and the issues*

This paper reviewed the literature for the last 30 years, starting from Nuttall *et al.* in 1974, and noting in a particular that JMB exam board used subject pairs analysis up until 1999 as valuable information that was made available to subject awarding committees at the board.

Jones noted the various criticisms that have been made of subject pairs and the assumptions on which it is based, i.e. inter subject differences might be explained by teaching effects, assessment regime, the multidimensionality of achievement, gender

effects, domain sampling, resourcing, motivation and interest, form of assessment, question difficulty and the distribution of marks.
Jones also notes Willmott's (1995) distinction between factors that make subjects difficult, and factors which do not enable candidates to achieve; it is rarely simple to separate the two.

Jones describes the mid-1990s discussion following the Fitz-Gibbon and Vincent paper, whose results were replicated in the Dearing review, and the criticism of these by Cresswell, Newton and others. He notes in particular that because different correction factors are obtained for different sub groups of candidates, any attempt to level subjects up and down in terms of standards (grades) would be publicly indefensible and educationally indefensible.

Jones summarises the criticism of subject pairs as "since different subjects are designed to measure various subject specific attainments, not a uni-dimensional measure of ability, any statistical method for making comparisons is inappropriate." He goes on to make the case for "expert judgement" about the level of attainment across different subjects but doubt swhether such judges could actually be found.

Jones concludes his discussion with "initially, at least, establishing the relative standards of subjects would need to be an arbitrary, political decision, thus disrupting the main objective of the awarding bodies, maintaining standards over time. Once set, however, subject pairs analysis could be used as one of the ability based indicators, together with, for example, prior attainment, as a proxy ability measure, especially for cognate subjects."

### *3.2.16. McGaw, B., Gipps, C., Godber, R.(2004) - Examination Standards – report of the independent committee to QCA*

This committee was formed as a result of the Tomlinson Enquiry into A-level standards 2002. Amongst their conclusions were :

- No examination system at the school or other level is so tightly or carefully managed.
- Strategies for maintaining standards across time do as well as possible, but there are unrealistic expectations still in play.
- No examination system has found an adequate way to determine whether standards are constant across subjects.

They comment that much of the public discussion of examination results in England is based on the assumption that results are standard referenced with a degree of precision that cannot be delivered. Over the long term, it makes little sense to ask whether examination standards have been maintained since the subjects themselves have changed so much. It would help if different expectations were set, not asking if performance standards are rising or falling over the longer term but asking only if the examinations are making reasonable and appropriate demands of students and if the results work for the key purposes for which they are intended.

On between-subject standards they identify there have been three approaches:

- Investigation into the number of candidates achieving at different grade levels in different subjects, but this is confounded by differences in the nature of candidates in different subjects;
- It can involve a comparison of candidate's prior achievement but that depends on the comparability across subjects of prior achievement;
- It can involve a comparison of performance in a given subject with performance in other subjects taken.

They note this third approach is used in Australia but there are strong assumptions about the comparability of all subjects using this approach. It would be difficult to apply this to A-levels where students typically take two other subjects whereas in Australia they typically take four or five additional subjects.

They also note a fourth qualitative approach involving judgement of the comparability of demands and assessments of students' performance across subjects. However, this depends on finding experts who are sufficiently qualified to make comparisons across at least related subjects. They say initial work suggests that this approach is workable and QCA has initiated some studies to be published 2005.

### *3.2.17. Bramley, T. (2005) - Accessibility, easiness and standards*

This paper includes a discussion of 'difficulty' in the context of maintaining standards through the cut off scores for levels in National Curriculum tests. There is a lot of discussion about the meaning of these words 'standard' and 'difficulty', and he defines difficulty with ability as a latent trait as in Rasch modelling.

Probably of more relevance to the present review is his discussion of accessibility; on changing the wording and presentation of science test items, makes the questions more accessible in that more children get them right. There is an issue here about testing science knowledge and understanding vs. the ability just to read and understand what a question is asking. However, he states that:

> *"it is very difficult to explain these latent trait variables ( intrinsic ability, intrinsic difficulty, motivation, accessibility) without circularity and yet they are so ingrained into the ways of thinking and vocabulary of assessment that practitioners are inclined to talk about them as though everyone knows what they are, and how to measure them … many arguments that draw on these terms are can easily become unsupported assertions or even statements of faith."*

### *3.2.18. Newton PE (2005) Examination standards and the limits of linking*

Newton (2005) considers the arguments about trying to compare the standards in different examinations, particularly when examinations are sufficiently different for the 'linking' to be fairly loose. He discusses the extent to which one can interpret

scores from two or more tests as comparable in terms of a 'linking construct'. 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' (p111). If there is a shared construct, and the same award in each corresponds to the same level of it, then they are comparable.

A problem arises for Newton from the fact that more than one linking construct may be possible, since different linking constructs will imply different definitions of comparability:

> "*Clearly then the different uses to which results may be put hold potentially conflicting implications for approaches to linking standards; more precisely, the different uses imply different definitions of comparability….different users and stakeholders value different linking constructs, wanting to draw different inferences from comparability. …if it were possible to agree upon a single definition for comparability, then why the need to defer to the more nebulous idea of value judgement? The answer seems to reside in a belief that it is not possible to agree upon a single definition: different users wish to use examination results for different purposes, and there are no strong grounds for deciding between those competing values/definitions.*" (p.117)

### *3.2.19. Coe, R. (2008) - Relative difficulties of examinations at GCSE: an application of the Rasch model*

A criticism of the subject pairs based analyses into the comparability of subject difficulties has been that of the assumed unidimensionality of a candidate's ability across the subjects in which they sit examinations. In Coe's approach, the Rasch model was used to test this assumption empirically; the model also allowed grade intervals to be investigated, both for a given grade gap across subjects, and for intervals within the same subject. Results indicated that for a large (34) group of subjects at GCSE, the assumption of unidimensionality does hold up, and that grade gaps are not equal. The analysis used a national data set of GCSE results and equivalent level qualifications from 2004.

In his Rasch model, Coe took subjects as items and the award of a grade as a partial credit. An iterative procedure related grades achieved by candidates in a subject to an ability measure based on their grades awarded in all other subjects taken. Some subjects were found not to fit the model, notably creative studies such as music, art and design and performance studies. Similarly, many GNVQ subjects and vocational GCSE's did not fit the model well, suggesting there was little or no comparability in a sensible way between these and the bulk of GCSE subjects. Coe also noted that the model was substantially improved by omitting grade U from the analyses.

Coe observed in his results that the rank order of subject difficulty varied with the grades, and concluded we cannot really talk about "subject difficulty" in general, but only in relation to particular grades. However, it was notable that the general trend in rank order was similar to that obtained in all previous studies going back to

Nuttall *et al.* in 1974. The individual sciences, chemistry, physics and biology, together with languages, such as Latin, Spanish, German and French are the most difficult subjects. It was also notable that double science, although relatively difficult compared to the 34 subjects, was easier than the individual sciences and also it became relatively easier the lower the grade. A similar phenomenon was observed in mathematics, where at grades A*, A and B it was ranked above 10 in difficulty; this dropped considerably for the lower grades.

### *3.2.20. QCA (2008): Inter-subject comparability studies*

This was a report of four investigations conducted by the Qualifications and Curriculum Authority into the standards of selected subjects at GCSE, AS and A-level. Each study asked subject experts, chosen for their experience of teaching more than one of the relevant subjects, to evaluate 'the demands implied by the syllabus materials within each subject at each level, and a comparison of candidates' work at each level' (p3). Subjects and levels covered are shown in Table 13.

*Table 13: Subjects and levels covered by QCA (2008)*

| *Study* | *Level* | *Subjects* |
|---|---|---|
| 1a | GCSE, AS, A | geography, history |
| 1b | GCSE, AS, A | biology, chemistry, physics (& double science at GCSE) |
| 2a | A | biology, psychology, sociology |
| 2b | A | English literature, history, media studies |

In order to judge the level of demand in each subject, the reviewers were provided with a 'taxonomy of examination demand' which included aspects such as 'complexity, resources, abstractness and strategy' (p13). However, the full taxonomy was not provided in the report, nor was the method by which ratings of these qualities were combined into a numerical scale, nor any evidence for the validity of interpreting ratings on these four elements as a single construct. Following a pilot training exercise in one study (2b) there were 'no significant differences of opinion between reviewers about the particular numerical ratings' (p17), though the report did not say whether these ratings were arrived at independently, how many opportunities for disagreement there were (or even how many reviewers took part) or what level of agreement was reached in the other three studies. The report also noted that part of the problem of judging the demand of an examination was that responses to examination tasks must be converted into marks via a mark scheme, a process that often depends on examiners' interpretations of vague terms such as 'sound' or 'effective' (p17). Even a subject with more demanding questions could be marked more leniently or have a lower threshold or marks for the award of a particular grade.

The second part of the comparability study aimed to address this issue with a review of marked scripts. The review was limited to performance in the written examination section of each subject's assessment and did not include coursework, which was a substantial component in some cases (40% in A-level media studies, for example). Reviewers were provided with a sample of scripts known (according to the original marking) to be at key grade borderlines (grades E and A for AS and A2). They were therefore effectively invited to endorse (or refute) the original grades in full knowledge of the implications of their judgements for conclusions about the comparability of grading. Hence it was impossible to rule out the possibility of a confirmation bias; a stronger design might have asked reviewers to compare the quality of scripts without knowledge of the marks that had been awarded. Given the limitations of the studies, the report notes (p41) that 'only cases where differences in standard were large are reported', though we are not told what the threshold for a difference to be considered 'large' was.

The findings of the four investigations show that there were judged to be some differences in the demand made in different subjects. For study 1a the CRAS scale (complexity, resources, abstractness and strategy) had an interval of two points between the expected ranges for GCSE higher tier and AS-level and another two between AS and A-level, so we may interpret a two point gap as equivalent to about a year's progress for an average student. History was judged to be harder than geography by more than one point at AS and GCSE and just over half a point at A2. In study 1b the science subject were judged to be within half a point at GCSE and one-fifth of a point at AS and A2. However, a different scale appears to have been used in the other studies – unless the results indicate that all the A-levels in study 2a and 2b were easier than the GCSEs in 1a and 1b – so the differences are hard to interpret. Nevertheless, both psychology and sociology appear to be somewhat harder than biology at A2, while media studies was judged rather easier than English literature and history at AS.

More interesting than the numerical differences, though, are the comments about the different assessment characteristics. For example, geography tended to use short, structured questions taken from a well-defined, broad syllabus. In history, by contrast, the syllabus offered more choices, allowing quite narrow focus, and questions were more open-ended, requiring 'essay-style' answers with higher levels of communication skills. The A-level chemistry syllabus reviewed contained a multiple choice element, unlike the other sciences. It was also judged to have insufficient time allocation, making it 'very demanding'. Physics required little extended writing at A-level but required complex mathematical processes. Biology A-level did require extended writing, but the mark scheme did not reward quality of written communication. Biology also offered no choice of questions to candidates, while psychology and sociology did, making them 'less demanding'. Some short-answer questions in sociology were judged to require 'no more than comprehension of a passage' (p31). The demands of English literature were judged to be variable, depending on the particular texts chosen, with some 'particularly inaccessible'. The accessibility and familiarity of texts in media studies provoked disagreement among reviewers about its demands; some felt this made it easier, others that candidates

could be tempted to respond in inappropriately 'colloquial' ways. In media studies candidates were allowed to make use of prepared materials in the examination. All these differences illustrate the difficulties of trying to judge the relative 'demands' of such different examinations. Quite how anyone could equate demands that are so different is far from clear, and a number of comments in the report acknowledge the problems encountered in doing so.

The judgements made in comparing candidates' scripts did not always support the judgements about the demands of the examinations. For example, although when presented with scripts that had been awarded grade E at GCSE, 'reviewers were overwhelmingly of the view that the history candidates performed better than those in geography' (p43), this view was less consistent for grade A candidates and at AS; at A2 geography was judged to be stronger, in contrast to the verdict on the demands of the examination. The comparison of the sciences (study 1b), on the other hand, did endorse judgements of demand, with candidates' work in all subjects judged to be reasonably close in standard. Study 2a did present some discrepancy between the demands of the examinations and the quality of candidates' scripts: at A2 the former had found psychology in line with sociology but both harder than biology, while the script analysis suggested that psychology and biology were in line, but sociology was easier, albeit with some reservations about the confidence that could be placed in this difference. Finally in study 2b, script analysis supported the analysis of demand at AS but at A2 concluded that candidates in media studies were 'less impressive' than those in English, in contrast to the review of examination demands which found them comparable.

The report concludes by stating that its key finding was that 'subjects were generally in line' (p51), which seems odd, given the extent of the differences found and the cautions that surrounded them. These differences appear to be explained by 'differences in approach to assessment rather than standard' (p51). Elsewhere in the report considerable caution is expressed about the validity of comparing subjects that are assessed very differently, and a fair summary might be that although some differences were found, they were not large relative to the degree of uncertainty that would be appropriate given the problems encountered in this kind of approach. The fact that the analyses did not convincingly show that standards were different is not the same, however, as convincingly showing that they were not different.

### *3.2.21. The Scottish Qualification Authority Relative Ratings Data*

The National Ratings produced by SQA are not, and have not been, publicly available on their website. However, interested parties can approach SQA to obtain them. The National ratings are produced annually using the Kelly method. Below we present the 2006 National Ratings produced by SQA:

*Table 14: Standard Grade National Ratings*

| title | National Rating | title | National Rating |
|---|---|---|---|
| Economics | -1.42 | Computing Studies | -0.01 |
| Religious Studies | -0.55 | Administration | 0.01 |
| Mathematics | -0.45 | Classical Studies | 0.09 |
| Accounting & Finance | -0.31 | Modern Studies | 0.09 |
| Contemporary Social Studies | -0.28 | Business Management | 0.16 |
| French | -0.20 | English | 0.23 |
| Biology | -0.18 | Science | 0.26 |
| Geography | -0.18 | Drama | 0.30 |
| Technological Studies | -0.18 | Home Economics | 0.31 |
| Spanish | -0.17 | Craft & Design | 0.32 |
| Chemistry | -0.15 | Art and Design | 0.34 |
| Physics | -0.15 | Gaidhlig | 0.47 |
| Graphic Communication | -0.13 | Gaelic (Learners) | 0.49 |
| Latin | -0.11 | Physical Education | 0.50 |
| German | -0.08 | Music | 0.55 |
| Italian | -0.04 | Social & Vocational Skills | 0.80 |
| History | -0.02 | Urdu | 1.40 |

*Table 15: Intermediate (1) grade National Ratings*

| National Rating | title | National Rating | title |
|---|---|---|---|
| -1.18 | Media Studies | 0.44 | Biology |
| -1.16 | Psychology | 0.58 | Travel and Tourism |
| -0.76 | Italian | 0.59 | Personal and Social Education |
| -0.67 | German | 0.61 | Computing Studies |
| -0.62 | Geography | 0.67 | Administration |
| -0.57 | English | 0.73 | Care |
| -0.57 | Mathematics | 0.74 | Physical Education |
| -0.55 | Accounting | 0.78 | Art and Design |
| -0.42 | Spanish | 0.89 | Woodworking Skills |
| -0.13 | Chemistry | 0.91 | HE: Health and Food Technology |
| -0.1 | French | 0.93 | HE: Lifestyle and Consumer Tech |
| -0.05 | Physics | 0.99 | Hospitality: Practical Cookery |
| 0.08 | History | 1.17 | Drama |
| 0.2 | Music | | |

*Table 16: Higher grade National Ratings*

| National Rating | Title |
|---|---|
| -0.34 | Media Studies |
| -0.30 | Chemistry |
| -0.27 | Psychology |
| -0.26 | Travel and Tourism |
| -0.25 | English |
| -0.24 | Human Biology |
| -0.23 | Mathematics |
| -0.20 | HE: Fashion and Textile Technology |
| -0.19 | Physics |
| -0.17 | Technological Studies |
| -0.16 | Biology |
| -0.13 | Economics |
| -0.12 | Religious, Moral and Phil Stds (new) |
| -0.10 | Computing |
| -0.08 | Philosophy |
| -0.05 | Information Systems |
| -0.02 | Latin |
| -0.01 | Business Management |
| 0.04 | German |
| 0.05 | Accounting |
| 0.05 | Sociology |
| 0.09 | Administration |
| 0.09 | Care |
| 0.13 | History |
| 0.17 | Modern Studies |
| 0.18 | Geography |
| 0.19 | Classical Studies |
| 0.21 | French |
| 0.21 | Italian |
| 0.23 | Geology |
| 0.27 | Drama |
| 0.29 | Early Education and Childcare |
| 0.33 | Religious, Moral and Phil Studies |
| 0.35 | Construction |
| 0.39 | Spanish |
| 0.41 | Product Design |
| 0.50 | Graphic Communication |
| 0.50 | HE: Health and Food Technology |
| 0.52 | Biotechnology |
| 0.56 | Mental Health Care |
| 0.56 | Play in Early Education and Childcare |
| 0.57 | Physical Education |
| 0.58 | Art and Design |
| 0.58 | Photography for the Media |
| 0.61 | Politics |
| 0.65 | Managing Environmental Resources |
| 0.86 | Dance Practice |
| 0.92 | Personal and Social Education |
| 0.98 | Gaelic (Learners) |
| 1.00 | Music |
| 1.12 | Fitness and Exercise |
| 1.14 | HE: Lifestyle and Consumer Tech |
| 1.18 | Gaidhlig |
| 1.21 | Care Practice |
| 1.33 | Design |
| 1.57 | Selling Scheduled Air Travel |
| 1.69 | Sports Coaching Studies |
| 1.97 | Retail Travel |
| 2.19 | Professional Patisserie |
| 2.30 | Hospitality - Professional Cookery |

### *3.3. Summary and synthesis of results*

In the previous sections we have presented summaries of a large number of studies conducted over a long period, using many different methods and datasets. Many studies have considered the question of whether statistical methods are appropriate or helpful in considering questions of inter-subject comparability. We defer discussion of these conceptual issues to the next chapter, where the issue of what, if anything, the statistical differences mean will be considered.

For now we attempt to collect together the empirical evidence about the differences between the grades achieved by candidates who are apparently comparable, in order to judge to what extent there is consistency across the findings and whether they support, on the surface at least, the claim that STEM subjects tend to be more difficult. In considering the statistical findings, we have identified four key research questions:

- Do the different methods give different answers?

- Are relative difficulties consistent over time?
- How much do they vary for different subgroups?
- To the extent that there is consistency in relation to these three issues, do STEM subjects emerge as more difficult?

We address each of these in turn.

### *3.3.1. Do the different methods give different answers?*

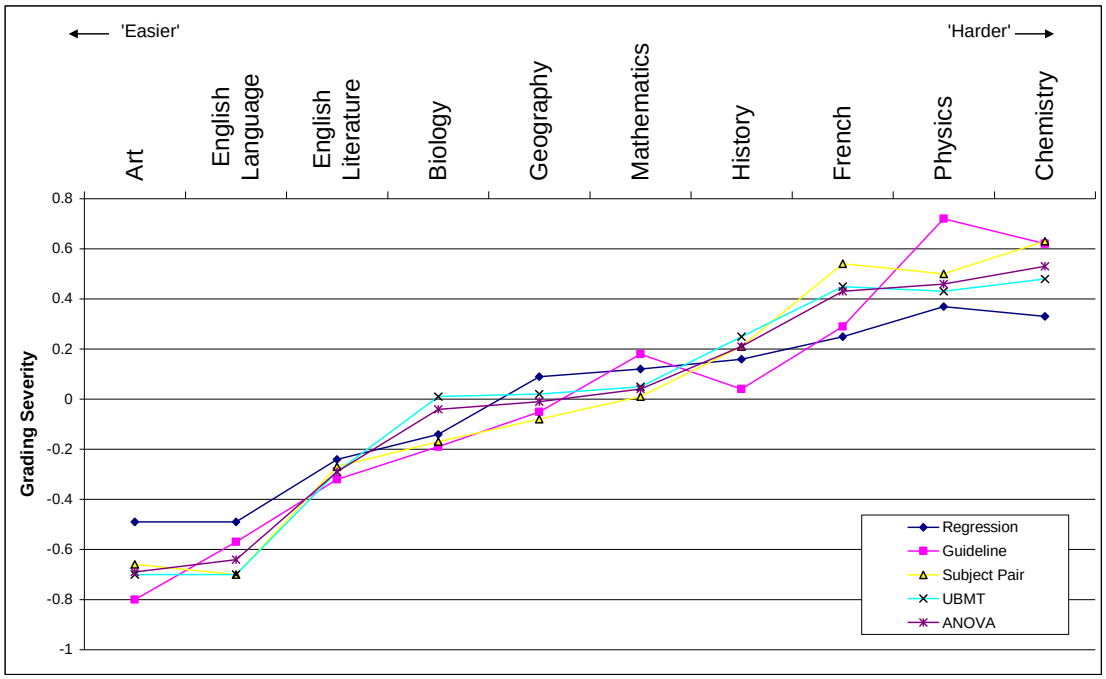
Two studies have specifically addressed this issue, applying different statistical methods to the same dataset. Nuttall *et al.* (1974) used five different methods to analyse the difficulties of ten O level subjects taken in 1968 within one examining board. Correlations among the methods are shown in Table 17 and the actual difficulty estimates are shown graphically in Figure 3.

*Table 17: Correlations among different methods (Nuttall et al., 1974)*

| | Correlations | | | | Standard deviation of subject difficulty estimates |
|---|---|---|---|---|---|
| | Guideline | Subject pair | UBMT | ANOVA | |
| Regression | 0.96 | 0.97 | 0.98 | 0.98 | 0.32 |
| Guideline | | 0.95 | 0.94 | 0.96 | 0.49 |
| Subject Pair | | | 0.98 | 0.99 | 0.47 |
| UBMT | | | | 1.00 | 0.44 |
| ANOVA | | | | | 0.43 |

The average correlation among the five methods is 0.97, suggesting a very high level of agreement. Moreover, with the possible exception of the regression method, for which the spread of difficulty estimates for the ten subjects is a little less, the methods agree well in absolute terms as well as being highly correlated. In no individual subject is the range of difficulty estimates from different methods more than about a third of a grade, compared with one and a half grades range across different subjects; the average difference between each subject's highest and lowest estimate is 0.23 of a grade.

*Figure 3: Agreement across five different methods, from Nuttall et al. (1974)*

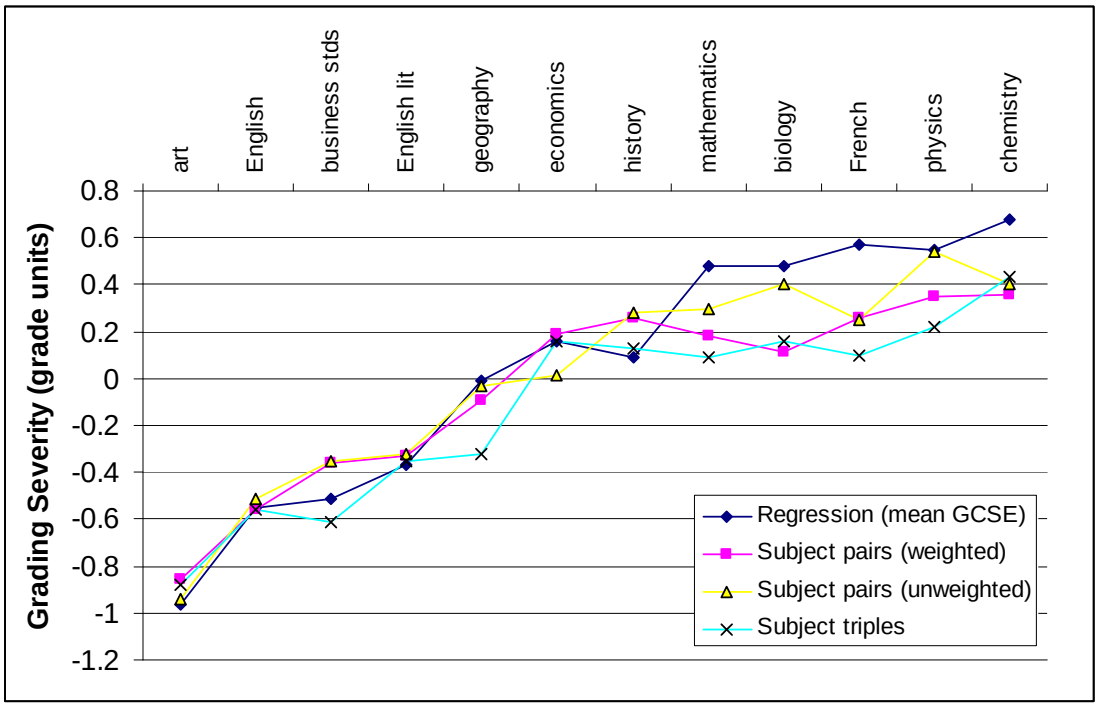



The second study to have used different methods is Alton and Pearson (1996). Their data are from A-levels taken nationally in 1994, for which they compared difficulties of 12 subjects. Correlations and difficulty estimates are shown in Table 18 and Figure 4.

*Table 18: Correlations among different methods (Alton and Pearson, 1996)*

| | Correlations | | | Standard deviation of subject difficulty estimates |
|---|---|---|---|---|
| | Subject pairs (weighted) | Subject pairs (unweighted) | Subject triples | |
| Regression (mean GCSE) | 0.95 | 0.95 | 0.95 | 0.54 |
| Subject pairs (weighted) | | 0.94 | 0.95 | 0.40 |
| Subject pairs (unweighted) | | | 0.89 | 0.45 |
| Subject triples | | | | 0.41 |

*Figure 4: Agreement across five different methods, from Alton and Pearson (1996)*




Despite the authors' interpretation of these results as showing substantial differences across different methods, the pattern again seems to be of reasonable agreement. The average correlation across methods is 0.94. In this case, unlike in the previous example, the regression method seems to be the one that separates subjects most, but again, the different methods agree reasonably well in absolute terms, especially for those subjects rated easiest. For the subjects at the 'harder' end, the range is a bit more, with, for example, nearly half a grade's difference between the highest and lowest estimate in French.

Fitz-Gibbon and Vincent (1994) also explicitly compared four different methods for estimating relative difficulties. Although they concluded that all methods converged in finding the sciences more difficult than other subjects, they presented data only for the analysis based on Kelly's (1976) method, so it is not possible to quantify the level of agreement across methods.

The answer to the question of how much consistency there is across different methods therefore seems to be that the evidence, such as it is, suggests that there is not too much disagreement among them. Certainly, the difference between choosing one method and another is substantially smaller than the difference between choosing any of them and none. However, it would also be useful to investigate further how well the different methods agree with different datasets.

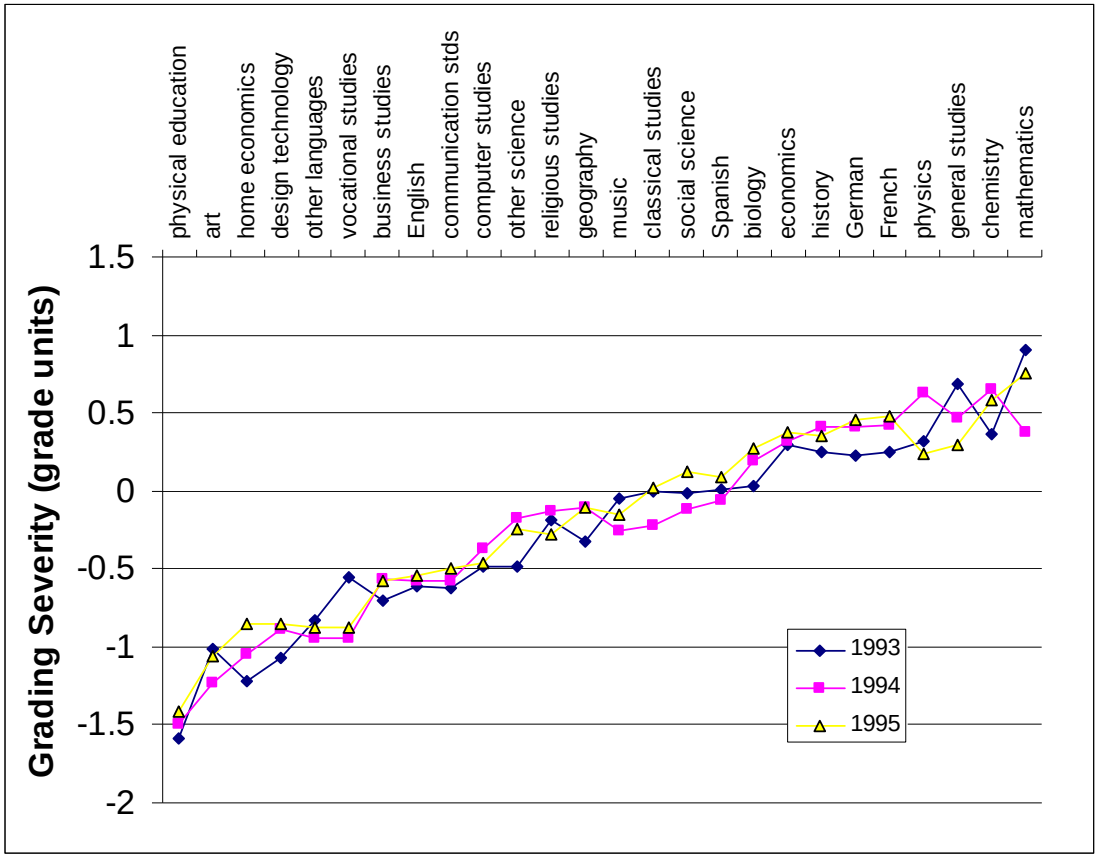
### 3.3.2. *Are relative difficulties consistent over time?*

The evidence available to answer the question of consistency over time is even more limited than for the previous question. Only one analysis that we have found (Dearing, 1996) has estimated difficulties using the same qualification and sample for successive years. This study applied Kelly's (1976a) method to national samples of A-level candidates for 26 subjects in 1993, 1994 and 1995.

*Table 19: Correlations among difficulty estimates in different years, from Dearing (1996)*

| | Correlations | | Standard deviation of subject difficulty estimates |
|---|---|---|---|
| | 1994 | 1995 | |
| 1993 | 0.94 | 0.94 | 0.61 |
| 1994 | | 0.96 | 0.61 |
| 1995 | | | 0.58 |

*Figure 5: Agreement over time, from Dearing (1996)*



Once again, there seems to be a reasonable level of consistency in the statistical relative difficulty of subjects from one year to the next. The average correlation is 0.94. The spread of the 26 subjects is much the same each year and the absolute differences between the highest and lowest estimates for any given subject are small (on average 0.23 grades) compared to the difference between the 'hardest' and 'easiest' subjects (over two grades) in any given year. In other words, the variation in difficulty of different subjects is about ten times the within-subject variation over the three years analysed. Hence the best answer we can give to the question about consistency over time is that estimates of difficulty may well be reasonably stable, though our confidence in this result cannot be strong, given its basis in just one study over three years.

### *3.3.3. How much do they vary for different subgroups?*

A number of writers have pointed to the differences between the estimates of subject difficulties that arise from limiting the analysis to different subgroups (eg Newton, 1997, Sparkes, 2000). For example, if difficulties are calculated separately for male and female candidates the results are different. They suggest that the existence of this difference undermines the validity of interpreting the differences in grades achieved as evidence of differential difficulty since the difficulty of a particular examination should not depend on who happened to take it.

The evidence available to assess the extent to which subgroup difficulties vary is rather limited, however, and, such as it is, does not appear to be very consistent. For example, Newbould (1982) applies subject pairs analysis to 13 O level subjects separately for males and females and produces data with a correlation of just 0.28 between the two sets of difficulties. On the other hand, Sparkes (2000) uses Kelly's (1976a) method with data from Scottish Highers for 11 subjects and generates a correlation of 0.95 between male and female estimates of difficulty.

On the basis of the data available, we do not think it is possible to say how much of a problem the issue of subgroup differences may be.

### *3.3.4. Do STEM subjects emerge as more difficult?*

If we accept that the different statistical methods appear to agree pretty well and that there is reasonable consistency of the estimates of subject difficulty from one year to the next, it seems appropriate to calculate an average difficulty for each subject and compare this for subjects in the STEM (science, technology, engineering and mathematics) group with others.

Before we can answer this, we need to be clear which subjects qualify as 'STEM' and which do not. Splitting subjects into sciences and non-sciences might seem like an obvious division, though for some subjects the classification will be open to argument. However, the concern of SCORE and other groups is about the supply of the next generation of people qualified in the sciences, technology, engineering and mathematics, rather than in the study of those subjects per se. Thus, for example,

although psychology may be a science, A-level psychology is seldom a requirement for further study in any scientific, technological, engineering or mathematical discipline. Equally, although mathematics may not be a science, A-level mathematics is an essential requirement for further study in many of these disciplines. On this basis, we have defined the STEM subjects as the traditional sciences, biology, chemistry and physics, together with mathematics.

*Figure 6: Average difficulties from various A-level studies for STEM and non-STEM subjects*

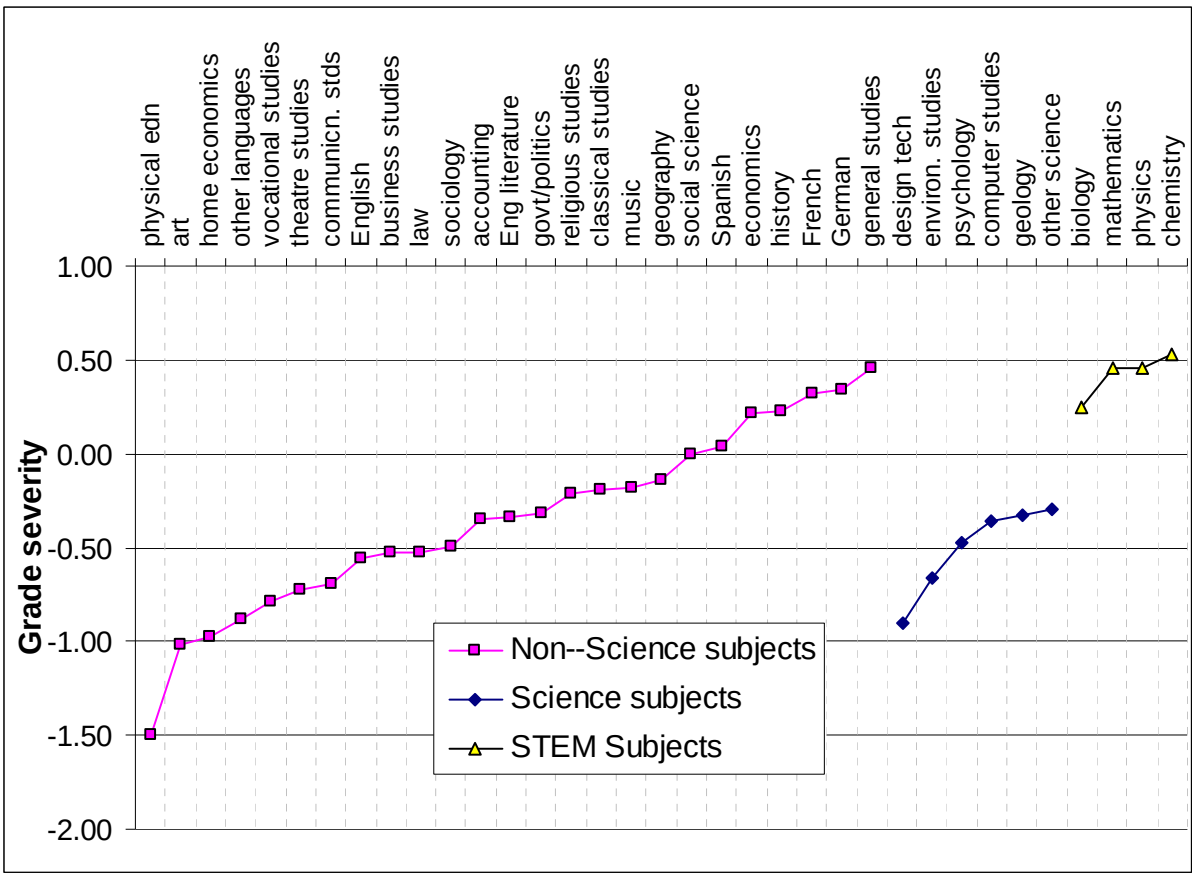



Figure 6 shows data for A-level difficulties averaged across the studies by Fitz-Gibbon and Vincent (1994), Alton and Pearson (1996) and Dearing (1996). From this, it is not clear that the sciences as a whole are more difficult than other A-levels. If we limit our definition of STEM subjects to biology, mathematics, chemistry and physics, then these are certainly among the most difficult of A-level examinations. We should also note that all the data available for this comparison come from studies over ten years old, so the current situation may be different. We present updated analyses of A-level and GCSE data in Chapter 8.

# 4. EVIDENCE ON SUBJECTIVE PERCEPTIONS OF DIFFICULTY

Whether or not science subjects, in particular the physical science subjects, are really more difficult, there is a great deal of research evidence to show that students *perceive* science subjects as more difficult. In this section, we will examine some of this research and why students might have this perception.

## 4.1. Literature on perceptions of difficulty

The issue that students perceive science subjects, in particular physics, as being more difficult has been a recurring theme in research over many years. Duckworth and Entwistle (1974) surveyed 292 Year 5 students (equivalent now to our Year 10) in secondary schools in Lancashire. They found that physics and chemistry were rated as the most difficult subjects whereas English and geography were rated as the easiest. Similarly, Pell (1977) also found in his survey of 180 Year 5 students that physics was rated as the most difficult. Johnson and Bell (1987), in their analysis of the 1984 Assessment Performance Unit (APU) survey, again found that physics and chemistry were rated as the most difficult of the science-related subjects.

More recent studies are consistent with these past findings. Kessels *et al*. (2006) found in their study with 63 older secondary students in Germany that physics was more likely to be associated with difficulty than English. This association was greater for female students compared to male students. In their survey of biology and physics undergraduate students, Spall *et al.* (2003) found that both sets of undergraduates perceived physics as being more difficult than biology. Taking a slightly different view of the problem, Barmby and Defty (2006) found in their analysis of data from the CEM Centre's YELLIS project that Year 10 students had significantly lower expectations for their GCSE grades in physics compared to biology. This contrasted with the fact that this lower overall expectation did not manifest itself in the actual exam grades achieved by the students.

Why then do students have his perception that physics in particular is a more difficult subject? One reason put forward in the research is the mathematical requirements of the subject. Pell (1985) stated that "subject difficulty is related to an over-mathematical approach" (p.131). Spall *et al.* (2003) in their survey of undergraduates also found that physics was seen as a mathematical subject whereas biology was not, implying a possible causal link between mathematical requirements and difficulty. Murphy and Whitelegg (2006) reported a study by Sharp *et al.* (1996) which found that heads of science in schools and colleges considered lack of mathematical knowledge as the main difficulty experienced by students in physics.

Another reason put forward by researchers is that a subject like physics is perceived as being chosen by students who 'do well'. Osborne *et al.* (1998), reporting the findings of a large-scale study by Cheng *et al*. (1995), stated that "the most significant factors affecting uptake of physical sciences were the grades achieved at GCSE in science and mathematics. This suggests that physics and chemistry are only taken by students who do well … the fact that only able pupils do physical sciences reinforces the notion that these are for the intelligent and are therefore difficult" (p.30). This is in agreement with Duckworth and Entwistle (1974), who found that students who chose physics had higher average scores on verbal and mathematical aptitude tests.

### *4.2. Differences in the marks required*

*Table 20: Raw marks and percentages required to achieve grade C GSCE in June 2006 AQA examinations (AQA, 2006)*

| Subject | Maximum mark | Raw mark for C | % required for C |
|---|---|---|---|
| Biology H paper | 135 | 62 | 46 |
| Chemistry H paper | 135 | 54 | 40 |
| Physics H paper | 135 | 48 | 36 |
| English A H paper 1 | 54 | 31 | 57 |
| English A H paper 2 | 54 | 29 | 54 |
| English B H paper 1 | 54 | 27 | 50 |
| English B H paper 2 | 54 | 24 | 44 |
| English Lit A H paper | 66 | 37 | 56 |
| English Lit B H paper | 78 | 45 | 58 |
| French Listening H | 40 | 18 | 45 |
| French Reading H | 45 | 23 | 51 |
| French Speaking H | 40 | 19 | 48 |
| French Writing H | 40 | 17 | 43 |
| Mathematics A Paper 1 I | 100 | 48 | 48 |
| Mathematics A Paper 1 H | 100 | 20 | 20 |
| Mathematics A Paper 2 I | 100 | 45 | 45 |
| Mathematics A Paper 2 H | 100 | 20 | 20 |
| Science Dbl Award 1H | 90 | 37 | 41 |
| Science Dbl Award 2H | 90 | 33 | 37 |
| Science Dbl Award 3H | 90 | 31 | 34 |

One other possible reason that we can put forward is how students actually perform in examinations. We can examine the raw marks obtained by pupils (before scaling) to see if these provide any insight. Using the example of the raw GCSE marks published by AQA for June 2006 examinations (AQA, 2006), we can examine the differences between particular subjects. Table 20 above shows the percentages of raw marks required by students in each examination in order to attain a grade C pass.

In the above table, we have been selective with the subjects and have only included higher tier exam marks (except for mathematics where we have included intermediate tier as well). However, of particularly interest are the three marks for the separate sciences. There is a 10% difference between what is required for a grade C in biology and what is required in physics. Therefore, the questions asked in the physics exam are more difficult because a grade C student can answer less of them correctly. Therefore, we would suggest that there may be a direct link between this inability to answer as many questions and perceived difficulty. This will be particularly true if students are facing these exams or questions from them in school tests or mock exams.

The raw score needed for a particular grade boundary, for example the minimum score needed for a C grade in terms of a percentage, can be considered as an indication of the perceived difficulty of a subject. If students need to get 70% to achieve a C grade in history, for example, and 50% to achieve a C in maths, then it could be argued that the students will perceive the maths paper to be harder as there will be a higher proportion of the exam paper that they get 'incorrect'. Paradoxically, the outside observer may get the opposite impression, concluding that the maths examination must be easier as you only have to get 50% to pass.

### *4.2.1. Analysis of Raw Scores in England and Wales*

In order to see if raw scores vary between exam boards and between subjects we have arbitrarily chosen a sample of STEM and non-STEM subjects and compared the raw score percentages using summer 2006 results obtained from the exam board websites.[6] The results are for full course GCSEs and where more than one syllabus was available we have chosen syllabus A. Where modular and non-modular choices were available we have chosen non-modular. Where given, unweighted marks are provided and where a number of components are shown an unweighted mean is provided.

The results have been split into the results of General papers (where there is no choice of paper difficulty, all students take the same paper), Foundation papers (taken by students who are expected to achieve C grade or below) and Higher papers (taken by students expected to achieve grades A to C). The subject percentages vary less than the percentages for the different levels of exam. Higher level papers require students to get lower percentages in order to obtain a grade C compared to those required to get a grade C in the Foundation papers. However this reflects the differing aims of the papers and the students they are aimed at. When comparing the differences between STEM and non-STEM subjects we can only see a slight difference, mainly in the Higher papers. In the main, non-STEM Higher papers require students to achieve marks in the 50s to be awarded a grade C, compared to mainly marks in the 40s to achieve C grades in STEM subjects. In the sample that we

[6] Grade boundary information can be obtained from the following websites:
AQA: http://www.aqa.org.uk/over/stat.php#grade
Edexcel: http://www.edexcel.org.uk/quals/
OCR: http://www.ocr.org.uk/publications/publications_results.html

have selected mathematics appears to be the subject that requires the lowest grades to obtain a grade C.

*Table 21: A comparison of minimum raw scores needed to achieve GCSE grade C in summer 2006*

| Exam Board | | AQA | | | Edexcel | | | OCR | | | WJEC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Exam Level | | Founndation | General | Higher | Founndation | General | Higher | Founndation | General | Higher | Founndation | General | Higher |
| | Subject | | | | | | | | | | | | |
| Non-STEM subjects | English Language | 55.6% | | 56.3% | 59.0% | | 54.0% | 72.7% | | 58.4% | No data available on | | |
| | English Literature | 54.5% | | 56.9% | 60.0% | | 53.0% | 72.7% | | 56.7% | the website | | |
| | Religious Studies | | 49.8% | | | 46.8% | | | 59.7% | | | | |
| | History | | 47.8% | | | 46.0% | | | 52.8% | | | | |
| | Music | | 45.6% | | | 52.0% | | | 50.8% | | | | |
| | French | 63.9% | | 50.2% | 61.2% | | 42.7% | 66.2% | | 48.5% | | | |
| | Spanish | 59.6% | | 54.2% | error on website | | | 66.2% | | 50.6% | | | |
| STEM subjects | Science (single) | 57.0% | | 42.4% | 47.1% | | 40.9% | 56.7% | | 45.0% | | | |
| | Science (double) | 57.0% | | 41.5% | 51.5% | | 41.1% | 53.6% | | 40.8% | | | |
| | Biology | 58.2% | | 49.1% | 55.6% | | 47.3% | 56.1% | | 42.6% | | | |
| | Chemistry | 56.4% | | 44.2% | 57.4% | | 39.6% | 54.8% | | 40.9% | | | |
| | Physics | 57.0% | | 40.6% | 45.0% | | 38.5% | 54.3% | | 41.7% | | | |
| | ICT | 47.5% | | 45.4% | 51.0% | | 44.0% | 51.3% | | 44.2% | | | |
| | D.T. (Systems & Control) | 71.8% | | 64.5% | 54.0% | | 50.0% | 56.1% | | 45.9% | | | |
| | Mathematics | 48.0% | | 31.1% | 50.4% | | 25.4% | 45.9% | | 30.7% | | | |

### *4.2.2. Analysis of Raw Scores in Scotland*

Similarly to England and Wales SQA produce raw score percentages giving the lower and upper boundary mark for each of the three level at Standard Grade (i.e. Credit, General and Foundation). SQA aims for their lower boundary raw score to be 50% and their upper boundary score to be 70%. In practice these boundaries differ from these percentages slightly as SQA takes into account the difficulty of the exam paper and adjusts the grades so that there is comparability across the subjects. The results for summer 2006 were obtained from the SQA website.[7]

The results show that there is little difference between the percentages required for STEM and Non-STEM subjects. This possibly reflects the fact that SQA makes use of data on Relative Ratings in order to maintain standards across subjects and from year to year (See Section 2.3).

[7] Scottish grade boundary information can be obtained from:
SQA: http://www.sqa.org.uk/sqa/2567.350.html

*Table 22: Comparison of minimum raw scores needed to achieve upper and lower levels at each of the three Standard Grade levels*

| | Level | Credit | | General | | Foundation | |
|---|---|---|---|---|---|---|---|
| | Grade | 1 | 2 | 3 | 4 | 5 | 6 |
| | Subject | | | | | | |
| Non-STEM subjects | English | 64.0% | 44.0% | 50.0% | 38.0% | 56.0% | 36.0% |
| | Religious Studies | 73.6% | 50.0% | 59.7% | 40.3% | 48.3% | 31.7% |
| | History | 73.3% | 50.0% | 62.0% | 46.0% | 65.7% | 42.9% |
| | Music | 73.3% | 55.0% | 62.0% | 48.0% | 60.0% | 48.9% |
| | French | 60.8% | 43.1% | 55.2% | 37.9% | 58.3% | 35.0% |
| | Spanish | 66.7% | 43.1% | 58.6% | 41.4% | 60.0% | 43.3% |
| STEM subjects | Science | 70.0% | 55.0% | 71.3% | 56.3% | 68.3% | 48.3% |
| | Biology | 67.5% | 50.0% | 56.0% | 41.0% | 36.0% | n/a |
| | Chemistry | 76.7% | 55.0% | 73.3% | 58.3% | 50.0% | n/a |
| | Physics | 73.0% | 52.0% | 60.0% | 48.8% | 41.3% | n/a |
| | Computing Studies | 68.1% | 50.0% | 59.7% | 43.1% | 63.9% | 47.2% |
| | Tech Studies | 78.9% | 61.1% | 77.5% | 57.5% | 47.5% | n/a |
| | Mathematics | 72.2% | 48.9% | 76.3% | 55.0% | 65.0% | 46.3% |

*A*

### *4.3. Linking difficulty of science subjects with enrolment*

The research therefore suggests that students may indeed perceive subjects such as physics as being more difficult, though it is not clear how far this may be generalised from available evidence. The question remains, however, what effect this perception is likely to have on the enrolment of students in further courses in subjects like physics?

The picture from past research suggests that the reasons for choosing or not choosing to study further science courses are varied. Pell (1977) did find in his survey of penultimate secondary school students that "those pupils who reject physics after O-level do so mainly because of a perception that the O-level course is uninteresting and difficult" (p. 765). He however also found that A-level students who had chosen physics did so mainly for career reasons. DeBoer (1984) found in his American study that decision to concentrate in science in high school was based on the past performance in science. A slightly later study by DeBoer (1987) showed that intentions to continue on science courses were influenced by expectations of success in that course. This expectation in turn was influenced by their own perceived ability and the difficulty of the course. Garratt (1986) found that perceived difficulty itself was not found to be a factor influencing subject choice at A-level. However, she did find that interest in the subject, followed by past performance and career intentions were important. Crawley and Black (1992) suggested a range of possible influences, including career intentions, wanting to increase knowledge and learn useful information, interest in subjects and fear of failure. Most recently, the study in Australia by Barnes *et al*. (2005), looking specifically at science enrolment differences between male and females, found that "sex differences in enrolment behaviour can

be attributed almost entirely to differences in the perceived career value of courses, differences in how interesting males and females expect to find them and differences in how well they expect to perform" (p.19).

Although the research suggests that the issue of why students choose or do not choose to pursue further study in science subjects is quite complex, we can see that perceived difficulty could play a part. If past performance and expectations of success in subjects do play a part, then students' perceived difficulty of subjects will be related to these. Therefore, whether directly or indirectly, we suggest that perceived difficulty is one of the reasons why students are less likely to study subjects like physics at a higher level.

# *PART III*
# *NEW ANALYSIS*

In the next few Chapters we present new research conducted for this report. Analyses draw on both national data from the National Pupil Database for England and on CEM Centre datasets.

We have used these datasets to compare different methodologies for estimating subject difficulties. These include the different statistical methods described in Section 2.1 (p17). We have also examined consistency over time and for different subgroups. Overall, we have set out to address the four research questions identified in Section 3.3:

- Do the different methods give different answers?
- Are relative difficulties consistent over time?
- How much do they vary for different subgroups?
- To the extent that there is consistency in relation to these three issues, do STEM subjects emerge as more difficult?

# 5. AGREEMENT ACROSS DIFFERENT METHODS

## 5.1. The different methods

The main methods used have been explained in detail in Section 2.1 (p17), but are outlined briefly here.

### *5.1.1. Rasch*

The Rasch model treats the examination in each subject as an instrument for measuring the overall academic achievement of a candidate. In doing so it calibrates the difficulty of each examination onto a common scale. Hence relative difficulties may be interpreted as indicating the correspondence between the grades achieved in a subject and the underlying construct of general academic capacity for achievement. One strength of this method is that it allows the difficulty of each grade in each subject to be estimated independently, though for the purposes of comparing this method with the others, most of which can only estimate an overall difficulty for the subject, only the overall subject difficulty estimate was used.

Although the Rasch model has hardly been used in the UK we have included it in this comparison on theoretical grounds. It is an approach to measurement that allows ordinal data, such as grades, to be put on an interval scale, which is a claim none of the other methods can make. In particular, the fact that the Rasch model does not require us to assume equal intervals between grades makes it worth including, but also suggests that its results may differ from those of the other methods.

### *5.1.2. Subject Pairs Analysis*

Subject pairs analysis (SPA) is conceptually one of the simplest methods. If we take a pair of subjects, such as maths and English, and consider all candidates who have taken both, we can compare the average grades achieved in each. The difference in these average grades provides an indication of the relative difficulties of the two subjects. For example, if the same candidates have achieved on average half a grade better in English than in maths, it suggests that English may be graded a little more leniently.

If we then do the same comparison between maths and each other subject in some set we will have a series of pair-wise comparisons between maths and all the other subjects. An average of these pair-wise comparisons is an indicator of the difficulty of maths, compared with all the other subjects in our set. We can repeat this process for each other subject, so ending up with a relative difficulty estimate for every subject.

SPA has traditionally been one of the most widely used methods by awarding bodies in England, Wales and Northern Ireland, so it seems important to include it in a comparison of methods.

### *5.1.3. Kelly's method*

Kelly's (1976) method can be thought of as a more sophisticated version of SPA. One conceptual problem with SPA is that if candidates who take one hard subject tend also to take other hard subjects with it then the apparent difficulty of that subject may be reduced. Kelly's method avoids this problem by taking each candidate in a particular subject, say maths, and comparing their grade in maths with the average grade they achieved in all their other subjects. The average of this difference for all candidates who took maths provides a first estimate of the relative difficulty of maths, and the same process can provide first estimates of the difficulty of all subjects in a chosen set.

Kelly's method then proceeds to use these difficulty estimates to correct the actual grades achieved in all subjects. If maths was found to be half a grade harder than the average, then half a grade would be added to all maths grades to make them comparable with the other subjects. The process is then repeated, using these corrected grades in place of the actual grades, and estimating a second set of grade corrections for each subject. After a small number of iterations the corrections converge to zero and the corrected grades represent a fair allocation of points to achievements in different subjects. The total correction applied to each subject is taken as the estimate of its difficulty.

Kelly's method was presented in Section 2.1.2 as one of a number of 'common examinee linear models'. It earns its place in this comparison largely because of its regular use in Scotland. It has also been used in some key studies in England (eg Dearing, 1996) and in much of the CEM Centre work. It is the basis of the analysis of the stability of relative subject difficulties over time presented in Chapter 6, so its fit with the other methods needs to be established here.

### *5.1.4. Reference test*

The reference test method uses performance in a test of general achievement or ability as an anchor against which performance in different subjects can be referenced. In the case of comparing subjects at A-level, for example, we may use overall performance at GCSE as the reference test. If we know that candidates who are comparable in terms of their GCSE grades typically achieve higher grades in one subject than another at A-level we may judge the former to be easier.

The reference test method was widely used in comparability studies in the 1970s, but has since fallen out of favour with awarding bodies. Comparisons based on reference tests suffer from a number of limitations, but there are some circumstances where they may be the only method available. For example, if we want to compare the difficulties of examinations that have no common candidates, none of the previously listed methods can be used. Hence it is important to know how well these methods may agree.

### *5.1.5. Value-added methods*

Value-added methods describe a class of approaches that includes the usual implementation of reference test methods, as well as some more sophisticated analyses. Regression models that include a wider range of explanatory variables may be used, so that, for example, our comparison of the A-level grades achieved in two subjects draws on candidates who are similar in not just their GCSE grades but in other characteristics, such as their other prior achievements, their sex, socioeconomic status or the type of school they attend. The use of multilevel models allows the lack of independence between grades achieved by the same candidate or by candidates in the same centre to be incorporated into the model.

Value-added models, particularly multilevel models, have become increasingly popular among awarding bodies and are widely used to inform the standard-setting process.

## *5.2. National A-level data from 2006*

### *5.2.1. A-level Data*

The dataset analysed here came from the National Pupil Database for England, held by the DfES.[8] It contains results from all the level 3 examinations, including A-levels, taken by candidates at centres in England in 2006. These examination results are also matched to students' earlier achievements at GCSE, Key Stage 3 and 2, and to data from the Pupil Level Annual School Census (PLASC) where available.

The original dataset contained records from 574,000 candidates, of whom 250,000 had taken one or more A-levels. Results from 88 different A-level subjects were available, but some of these had quite small entries. Our analysis was limited to the 33 subjects with at least 5,000 candidates in order to make the results reliable. By limiting the analysis in this way, we did lose a small number of candidates (1,318, ie 0.5% of those with one or more A-level entry) who had taken only the smaller entry subjects.

Information on gender was available for all of the 249,000 students included in our analysis. The vast majority (240,000, ie 96%) also had a matched average GCSE score, with KS3 results available for 84% and KS2 for 87% of the cases. 80% of cases had all three prior attainment scores available. The match with the PLASC data was less satisfactory with over half the cases failing to match, so these data were not included in the analysis.

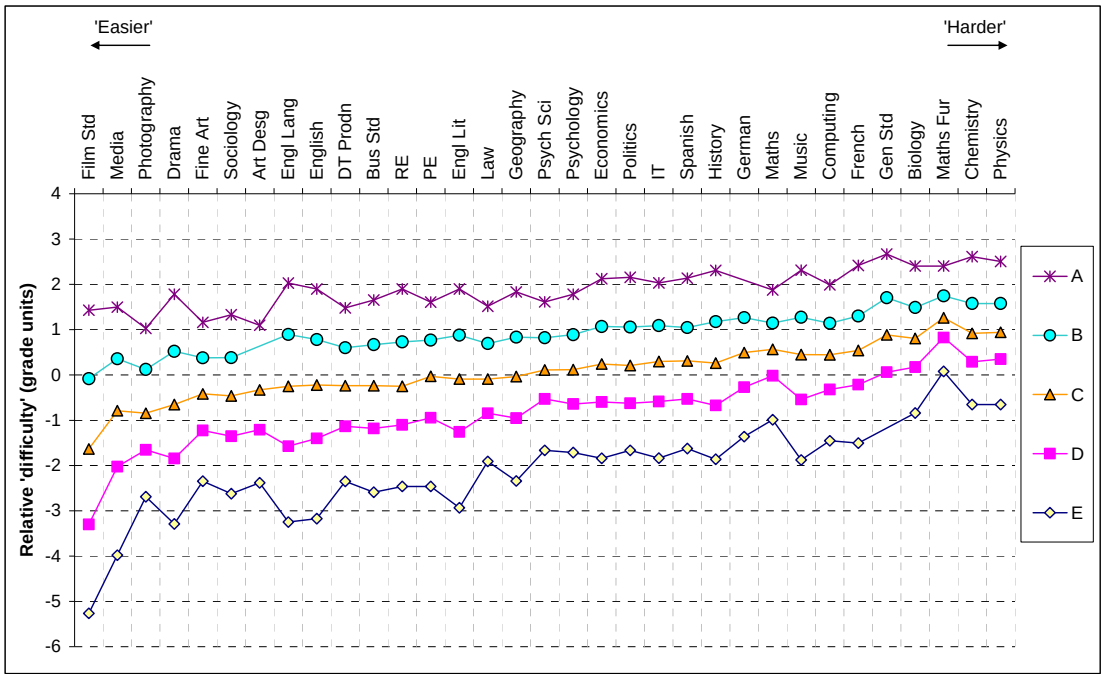
### *5.2.2. Methods*

The five broad methods described in Section 5.1 were applied to this dataset.

[8] The Department for Education and Skills changed into the Department for Children Families and Schools (DCFS) during the writing of this report.

The application of the Rasch model made use of the partial credit model within WINSTEPS, which allows the difficulty of individual grades to be estimated, but the estimate taken as the overall subject difficulty was the item difficulty estimate for each subject. Model fit was good for all subjects (infit and outfit between 0.7 and 1.3) apart from General Studies (infit and outfit both 1.62). However, estimates from General Studies were included in order to be able to compare this method with all the others. Item-measure correlations (i.e. the correlation between a person's grade in a particular subject and the model's estimate of their ability, based on all their grades) were above 0.82 for all 33 subjects. Person reliability was 0.82 and PCA showed 83% of the variance explained by the measures, so there was strong support for the existence of a unidimensional latent construct.

*Figure 7: Rasch estimates of relative difficulty of each grade in each A-level subject*




As the Rasch model estimates difficulties for independently for each grade in each subject, it seems appropriate to report these as well. These estimates are shown graphically in Figure 7. It is clear from these data that the assumption, required by all the other methods, that gaps between grades are equal, is rather questionable.

For example, we can see that the range between the lowest and highest grade within a subject is very variable. Smallest is Further Maths, with hardly more than two 'average' grades between E and A; largest is Film Studies, with almost a seven grade gap. The size of this difference in within-subject range, and the fact that these two subjects are some way from their nearest neighbours in this respect, may help to account for the fact that the Rasch estimates of the difficulties of these two subjects are out of line with estimates from the other methods (see below), all of which make

the assumption that within-subject range is constant. There is also a noticeable tendency for the harder subjects to have the smaller gaps.

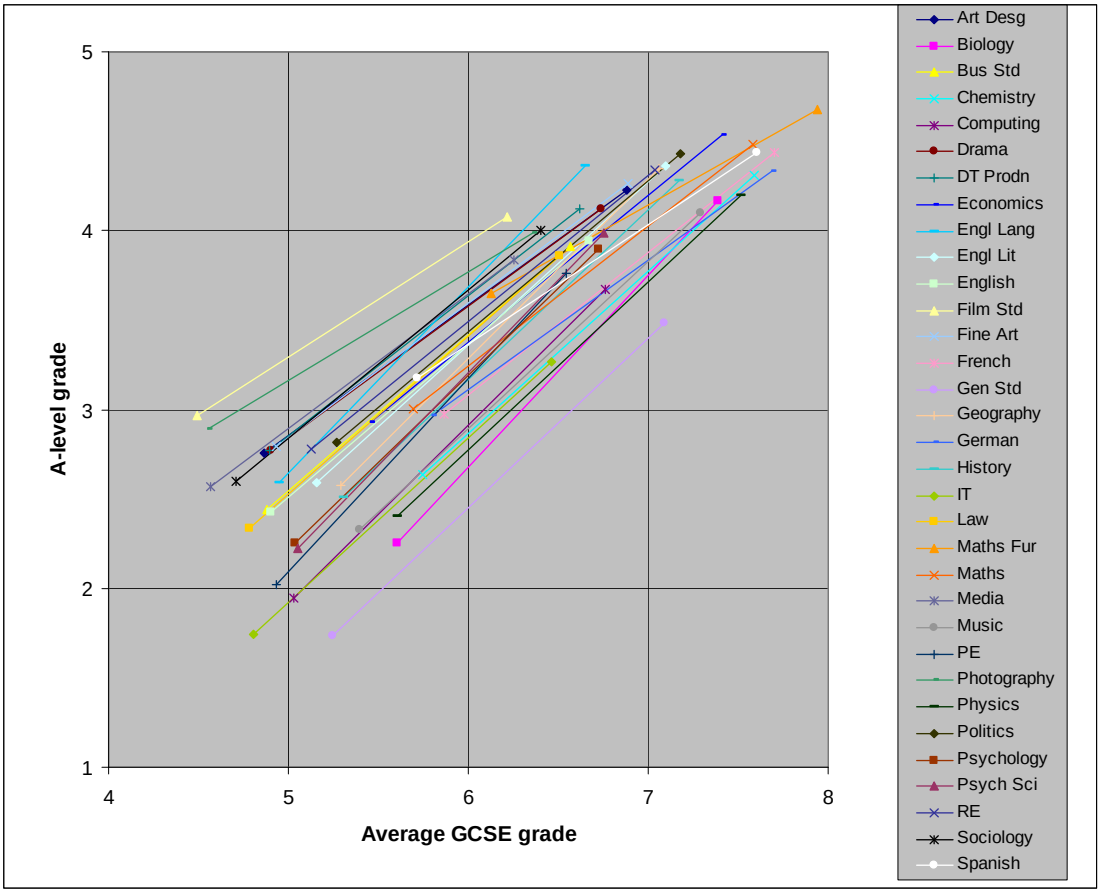
Furthermore, we can see that even within a subject, the sizes of the gaps are often some way from being equal. This leads to some anomalous comparisons of 'subject' difficulty. For example, although for most grades Maths is appreciably harder than English, at grade A English is actually marginally more difficult than Maths. At grade A, Art and Design is about a grade easier than English Language, but for E grade candidates Art and Design is close to a grade harder than English Language. Overall, the correlation between relative difficulties based on grade E and those based on grade A is 0.6, indicating a reasonable level of agreement, but scope for a number of such reversals. There is also a general tendency for the E-D gap to be larger than the others. On average the gap between E and D is nearly 45% bigger than the average gap between D-C, C-B and B-A.

Subject Pairs Analysis was conducted within the statistical analysis program SPSS, using a macro to calculate average differences in subject grades for each pair of subjects. Two different methods were then used to aggregate these. The first treated all subjects as equal and simply calculated the average of the differences between each subject and the other 32. This method is described as 'unweighted SPA'. The second weighted each difference by the number of candidates taking that pair of subjects in calculating this average. This method is described as 'weighted SPA'.

Kelly's method was conducted using an iterative approach, again with an SPSS macro.

The Reference Test method used average GCSE score as the reference test, running a separate regression of A-level grade on average GCSE for each subject. The resulting regression lines, shown for average GCSEs within one standard deviation each side of the mean for that subject, can be seen in Figure 8. It is clear that a student with an average GCSE score of, say, 6 (ie B grades) can typically expect to achieve quite different grades in different subjects. However, the size of the difference depends on the average GCSE score one starts with. In order to compare estimates of subject difficulty from this method with the others, we therefore had to choose a particular average GCSE score at which to calculate 'expected' A-level grades. Although the population mean for all A-level students was 5.6, for a number of subjects with high ability (but relatively small) entries, this value was below their normal range. Hence we used an average that weighted all subjects equally, giving a value of 6.1.

*Figure 8: Regression line segments for A-level grade on average GCSE for each of the 33 subjects*




*GCSE grades coded as A*=8, A=7, B=6, C=5, D=4, etc; A-level grades coded as A=5, B=4, C=3, D=2, E=1, U=0. Regression line segments are shown at the mean of the average GCSE grade for that subject plus or minus one standard deviation.*

Two versions of the Value-Added method were applied. The first fitted an ordinary least squares regression model with A-level grade as the outcome and a dummy variable for each subject, in addition to a set of explanatory variables: average GCSE, KS3 score, KS2 score and sex. This is described as the 'Value-added' model. The equation for this model is:[9]

$$y_{si} = \beta_0 + \sum_c \beta_c x_{ci} + \sum_s \beta_s z_{si} + u_i$$

*Equation 2*

The second model used a 2-level multilevel model with examination results (level 1) nested within candidates (level 2). The same set of explanatory variables was used (average GCSE, KS3 score, KS2 score and sex). This model is described as the 'Multilevel' model and the equation is:[10]

[9] Explanations for the terms in the equation can be found on p23.
[10] Again, see p23 for an explanation of the terms used.

$$y_{si} = \beta_0 + \sum_c \beta_c x_{ci} + \sum_s \beta_s z_{si} + u_i + e_{si}$$

*Equation 3*

$$u_i \sim \mathrm{N}(0, \sigma_u^2)$$
$$e_{si} \sim \mathrm{N}(0, \sigma_e^2)$$

### *5.2.3. Results from A-level analysis*

*Table 23: Estimates of relative difficulty of 33 A-level subjects from seven different methods*

| Subject | Rasch overall (grade units) | SPA (un-weighted) | SPA (weighted) | Kelly | Reference test | Value-added | Multilevel |
|---|---|---|---|---|---|---|---|
| Art Desg | -0.49 | -0.43 | -0.32 | -0.65 | -0.30 | -0.37 | -0.42 |
| Biology | 0.81 | 0.68 | 0.21 | 0.54 | 0.58 | 0.57 | 0.59 |
| Bus Std | -0.34 | -0.23 | -0.25 | -0.41 | -0.14 | -0.21 | -0.26 |
| Chemistry | 0.96 | 0.70 | 0.19 | 0.62 | 0.41 | 0.50 | 0.63 |
| Computing | 0.36 | 0.37 | 0.00 | 0.17 | 0.35 | 0.36 | 0.30 |
| Drama | -0.70 | -0.35 | -0.24 | -0.56 | -0.29 | -0.41 | -0.42 |
| DT Prodn | -0.33 | -0.29 | -0.25 | -0.46 | -0.06 | -0.09 | -0.25 |
| Economics | 0.20 | 0.15 | -0.02 | 0.08 | -0.09 | -0.04 | 0.09 |
| Engl Lang | -0.43 | -0.19 | -0.14 | -0.31 | -0.06 | -0.13 | -0.18 |
| Engl Lit | -0.30 | -0.10 | -0.10 | -0.22 | -0.09 | -0.12 | -0.13 |
| English | -0.43 | -0.25 | -0.13 | -0.35 | -0.09 | -0.17 | -0.20 |
| Film Std | -1.79 | -0.92 | -0.52 | -1.11 | -0.64 | -0.93 | -0.98 |
| Fine Art | -0.50 | -0.43 | -0.32 | -0.65 | -0.31 | -0.38 | -0.45 |
| French | 0.51 | 0.34 | 0.14 | 0.26 | 0.20 | 0.42 | 0.49 |
| Gen Std | 0.87 | 0.78 | 0.52 | 0.66 | 0.82 | 0.81 | 0.80 |
| Geography | -0.13 | -0.08 | -0.17 | -0.20 | -0.02 | -0.04 | -0.05 |
| German | 0.50 | 0.26 | 0.09 | 0.24 | 0.17 | 0.44 | 0.47 |
| History | 0.24 | 0.18 | 0.09 | 0.10 | 0.10 | 0.10 | 0.17 |
| IT | 0.20 | 0.20 | 0.07 | 0.06 | 0.43 | 0.38 | 0.26 |
| Law | -0.13 | -0.20 | -0.15 | -0.32 | -0.13 | -0.22 | -0.19 |
| Maths Fur | 1.27 | 0.25 | 0.18 | 0.63 | -0.28 | 0.19 | 0.47 |
| Maths | 0.52 | 0.26 | -0.07 | 0.20 | 0.04 | 0.21 | 0.26 |
| Media | -1.00 | -0.60 | -0.36 | -0.78 | -0.36 | -0.55 | -0.61 |
| Music | 0.33 | 0.33 | 0.13 | 0.18 | 0.37 | 0.43 | 0.40 |
| PE | -0.21 | -0.12 | -0.21 | -0.30 | 0.09 | 0.10 | 0.00 |
| Photography | -0.82 | -0.72 | -0.37 | -0.95 | -0.47 | -0.74 | -0.79 |
| Physics | 0.95 | 0.75 | 0.22 | 0.65 | 0.50 | 0.60 | 0.66 |
| Politics | 0.23 | 0.10 | 0.03 | 0.03 | -0.16 | -0.16 | -0.06 |
| Psychology | 0.09 | 0.03 | -0.05 | -0.11 | 0.07 | 0.03 | 0.02 |
| Psych Sci | 0.07 | -0.03 | -0.09 | -0.14 | 0.06 | 0.04 | -0.06 |
| RE | -0.24 | -0.12 | -0.16 | -0.27 | -0.21 | -0.24 | -0.23 |
| Sociology | -0.55 | -0.50 | -0.35 | -0.64 | -0.39 | -0.52 | -0.52 |
| Spanish | 0.27 | 0.14 | 0.01 | 0.06 | -0.07 | 0.16 | 0.16 |

Difficulty estimates from seven different methods for each of the 33 A-level subjects are shown in Table 23. Correlations among them are shown in Table 24 and a matrix of the scatterplots of these correlations in Figure 9.

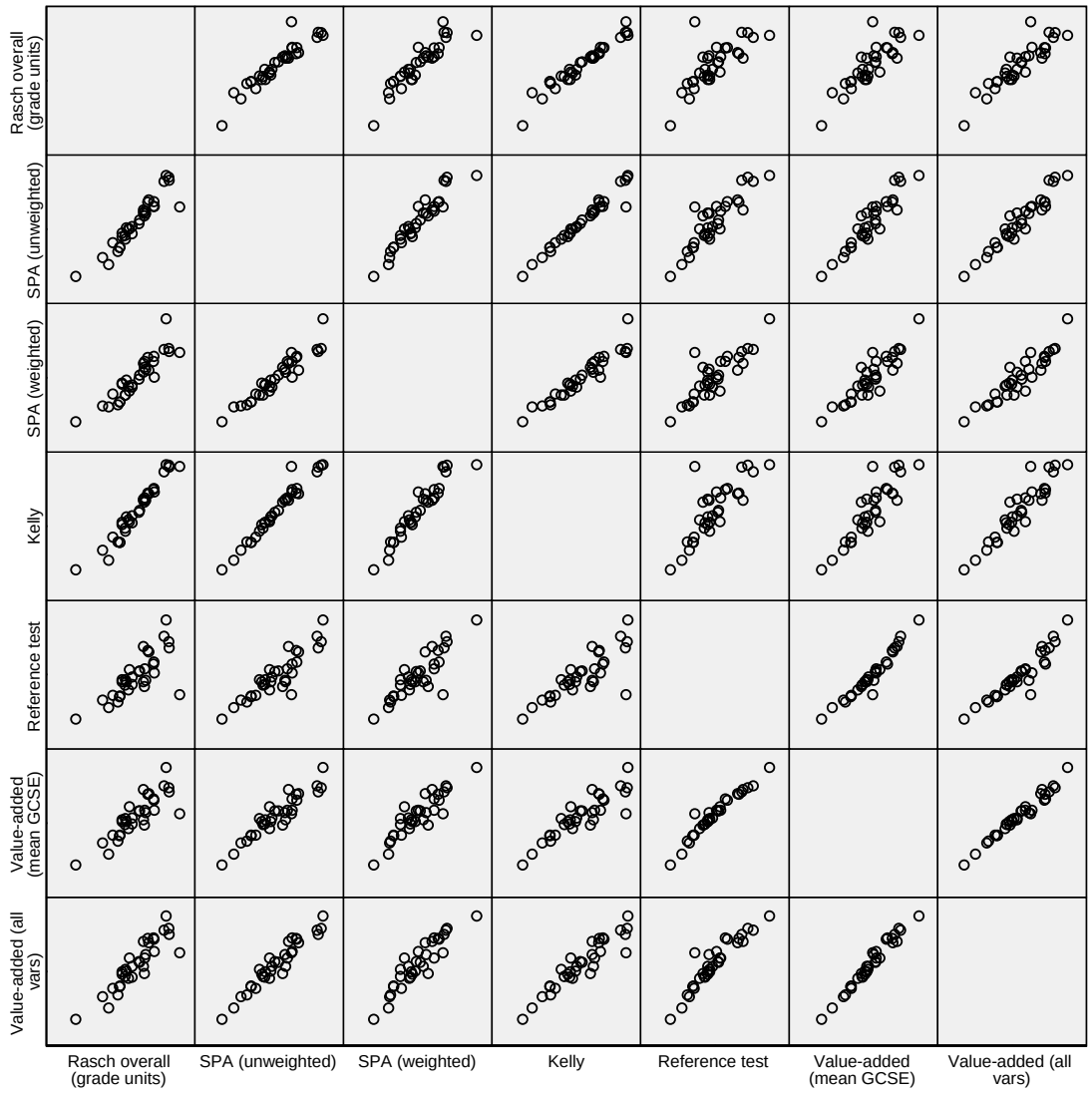
*Table 24: Correlations among difficulty estimates from different methods*

| | SPA (unweighted) | SPA (weighted) | Kelly | Reference test | Value-added | Multilevel model |
|---|---|---|---|---|---|---|
| Rasch | 0.940 | 0.910 | 0.972 | 0.754 | 0.902 | 0.951 |
| SPA (unweighted) | | 0.951 | 0.979 | 0.899 | 0.963 | 0.982 |
| SPA (weighted) | | | 0.955 | 0.856 | 0.923 | 0.953 |
| Kelly | | | | 0.817 | 0.936 | 0.981 |
| Reference test | | | | | 0.939 | 0.884 |
| Value-added | | | | | | 0.949 |

Correlations among the different methods are generally high, with most values in excess of 0.9. One method, reference tests, stands out as having lower correlations with the others. This may be partly a result of the need to choose a specific intercept for this method; subject differences are not constant, but depend on a particular value of average GCSE score.

It is clear that the unweighted Subject Pairs Analysis agrees better with the other methods than the weighted version. This corresponds with a theoretical argument that the unweighted version is to be preferred as a way of comparing subjects (eg Nuttall *et al.*, 1974). Similarly, the Value-added model with all variables included agrees better with other estimates than the less well specified model. Hence in choosing a representative of each method, we have preferred the unweighted version of SPA and the Value-added model with all variables included.

*Figure 9: Matrix of scatterplots of difficulty estimates from the seven methods*

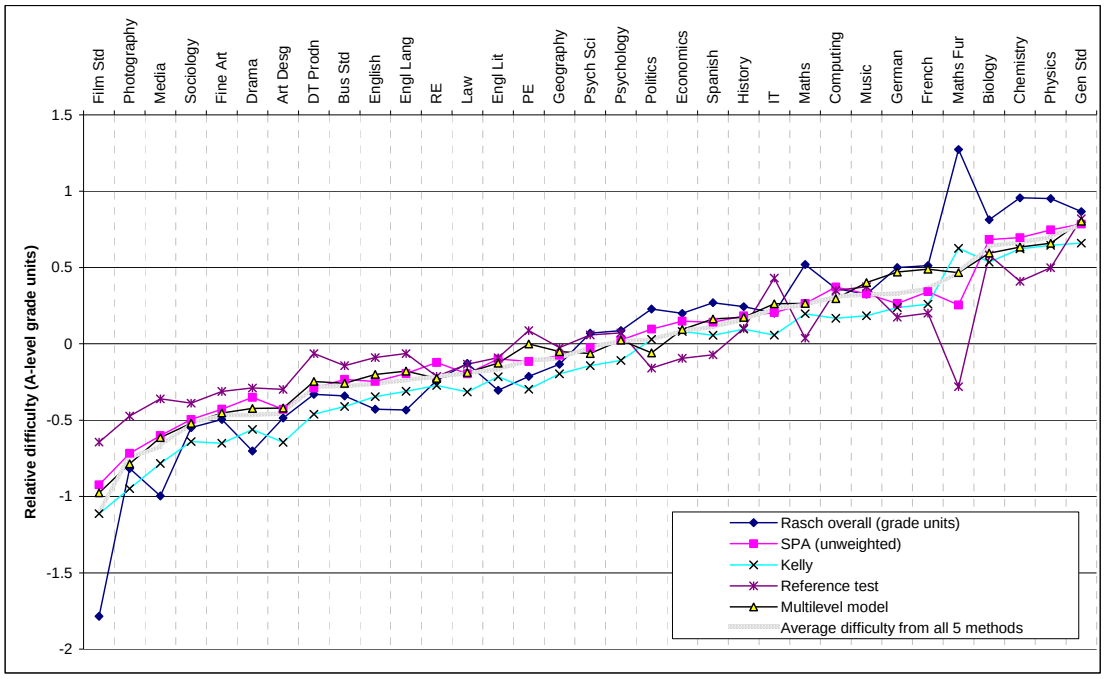



Estimates of subject difficulties from the five preferred methods are shown graphically in Figure 10. With the exception of a small number of subjects where the different methods do not agree so well (Further maths, Film Studies and Media Studies), estimates from the five methods are all within half a grade, and within a third of a grade for the majority of subjects. This compares with a difference of nearly two grades across subjects, averaged across methods. Overall, the average inter-method difference is about 20% of the average inter-subject difference.

The case of Further Maths presents something of a problem, with a difference of over one-and-a-half grades between two of the methods (Rasch and Reference Test). Further Maths is an unusual A-level, with 58% of its 6500 candidates being awarded the top grade, A. Even more extraordinary is the fact that of those who get A in Further Maths, two thirds also get As in all their other A-levels. Overall, therefore, 39% of the candidates who take Further Maths gain no other grade than A in any A-level. This situation inevitably makes it quite difficult to compare the difficulty of

Further Maths with other subjects, since any statistical comparison must be based on differences in the grades that are achieved in different subjects; for a proportion not much less than half, there simply are no differences in the grades gained. In measurement terms, the grading of Further Maths suffers from a significant ceiling effect.

*Figure 10: Relative difficulty estimates of A-level subjects from five different methods, ranked by average difficulty across all methods*



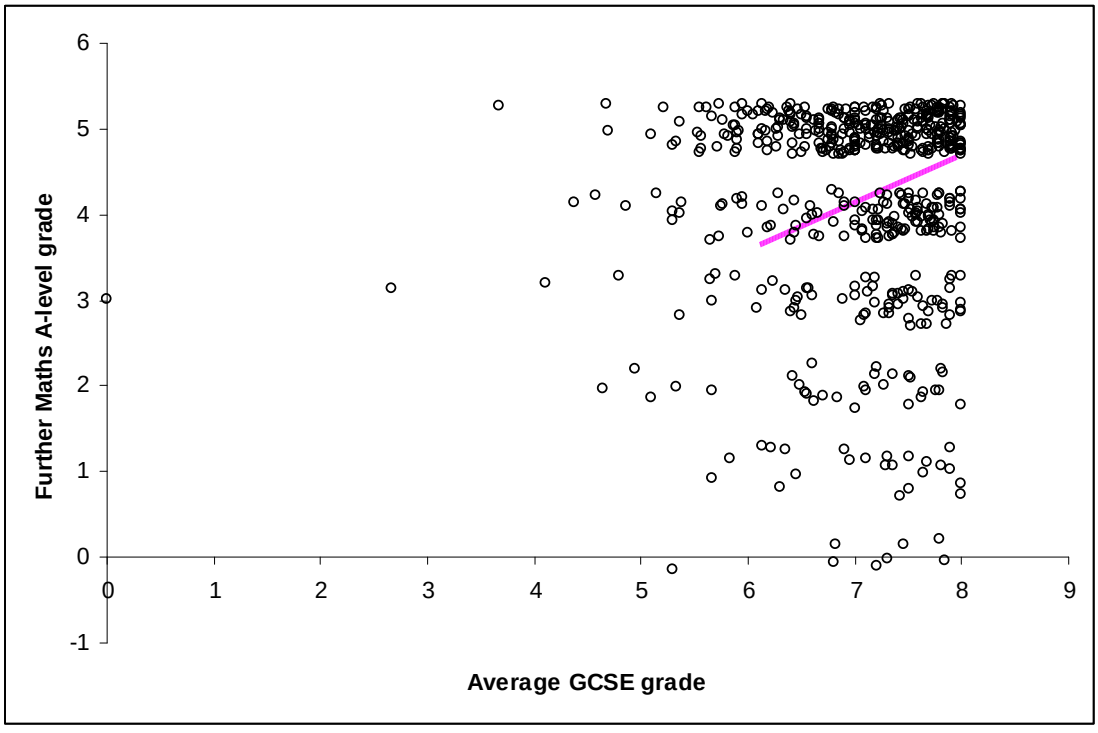

It may be helpful to consider how the different methods deal with this difficulty. For methods such as SPA and Kelly, the results of these candidates with straight As are taken as evidence of the equivalence of the subjects they have taken. Both these methods effectively calculate an average of the differences in the grades achieved between one subject and another. If 39% of the candidates generate a difference of zero, even if there are large differences in grades for the other 61%, the average will be brought down by the zeros. The ceiling on the recognition that can be awarded to the highest levels of performance means that achievements that may be far from equivalent are treated as equivalent in the calculation. A similar problem arises for the Reference Test and Value-Added methods, since the coding of 'A=5' (or any other points score) assigns a value to the top grade that treats all performance awarded that grade as equivalent.

By contrast, in the Rasch model the achievements of those who attain the top grade are not taken as indicating that their ability is at that level, but as providing evidence of ability of at least the level of the grade threshold. In fact, the model cannot estimate the ability of anyone who has achieved an 'extreme score' (i.e all A grades)

and these candidates are effectively ignored in calculating the difficulties of different subjects. Only those who have failed to achieve the top grade in one of more subjects contribute information about relative difficulties. Hence the Rasch model is not so limited by the ceiling on performance in the way that all the other methods are. This may be one reason why the Rasch model gives a wider range of subject difficulties that the other models.

In the case of the Reference Test method the problem is different again. The effect of the ceiling on both Further Maths grade and average GCSE score can be seen in Figure 11, which shows scores on both these variables for a sample of the national entry, together with the regression line segment. The ceiling effect depresses the correlation (0.43) between GCSE and the A-level grade and so flattens the regression line. The average GCSE value of 6.1 that we used to estimate the likely A-level grade, while above the A-level population mean, is well below the main range of candidates in Further Maths. This fact, combined with the flat regression line, leads to the expected grade being higher than is really representative of the difficulty of the subject as a whole. Hence the rather deflated estimate of the difficulty of Further Maths from this method.

*Figure 11: Scatterplot of Further Maths grade and average GCSE*




Although Further Maths is a rather extreme case, a similar ceiling effect occurs in many other subjects at A-level. Overall, almost 10% of all A-level candidates achieve

straight A grades in all their subjects. The percentages of candidates in each A-level subject who gained an A grade in 2006 is shown in Table 25.

*Table 25: Percentages of candidates gaining grade A in each A-level subject*

| | | | |
|---|---|---|---|
| Maths Fur | 58% | Music | 21% |
| Maths | 44% | Psych Sci | 21% |
| German | 37% | Sociology | 21% |
| Spanish | 37% | Law | 20% |
| Economics | 36% | Drama | 19% |
| French | 35% | Psychology | 18% |
| Politics | 32% | DT Prodn | 17% |
| Chemistry | 32% | Bus Std | 17% |
| Art Desg | 31% | Computing | 16% |
| Fine Art | 31% | Film Std | 16% |
| Physics | 30% | English | 15% |
| RE | 27% | PE | 14% |
| Engl Lit | 27% | Engl Lang | 14% |
| Geography | 26% | Media | 13% |
| Biology | 26% | Gen Std | 12% |
| History | 24% | IT | 8% |
| Photography | 23% | | |

#### *5.2.4. Conclusions*

Our overall conclusion from these comparisons is that there seems to be a reasonably high level of agreement across the different methods for A-level data. Certainly, the differences among them are far smaller than the differences in difficulty across the subjects estimated by any individual method. While there may still be reasons to prefer one method to another, and such choices will make a small difference to the results, the argument that the different methods do not agree is not a convincing reason to use none of them.

### *5.3. National GCSE data from 2006*

#### *5.3.1. Data*

The dataset analysed here came from the National Pupil Database for England, held by the DfES.[11] It contains results from all GCSE examinations taken by candidates at

[11] The Department for Education and Skills changed into the Department for Children Families and Schools (DCFS) during the writing of this report.

centres in England in 2006. These examination results are also matched to students' earlier achievements at Key Stage 3 and 2, and to data from the Pupil Level Annual School Census (PLASC) where available.

The original dataset contained records from 698,000 candidates, of whom 635,000 had taken one or more GCSEs. Results from 58 different GCSE subjects were available, but some of these had quite small entries. Our analysis was limited to the 34 subjects with at least 10,000 candidates in order to make the results reliable. By limiting the analysis in this way, we did lose a small number of candidates (1,119, ie 0.2% of those with one or more GCSE entry) who had taken only the smaller entry subjects.

Information on gender was available for all of the 634,000 students included in our analysis. Matched KS3 results were available for 90% and KS2 for 91% of the cases. 86% of cases had both prior attainment scores available. Free School Meal eligibility status was available for 91%.

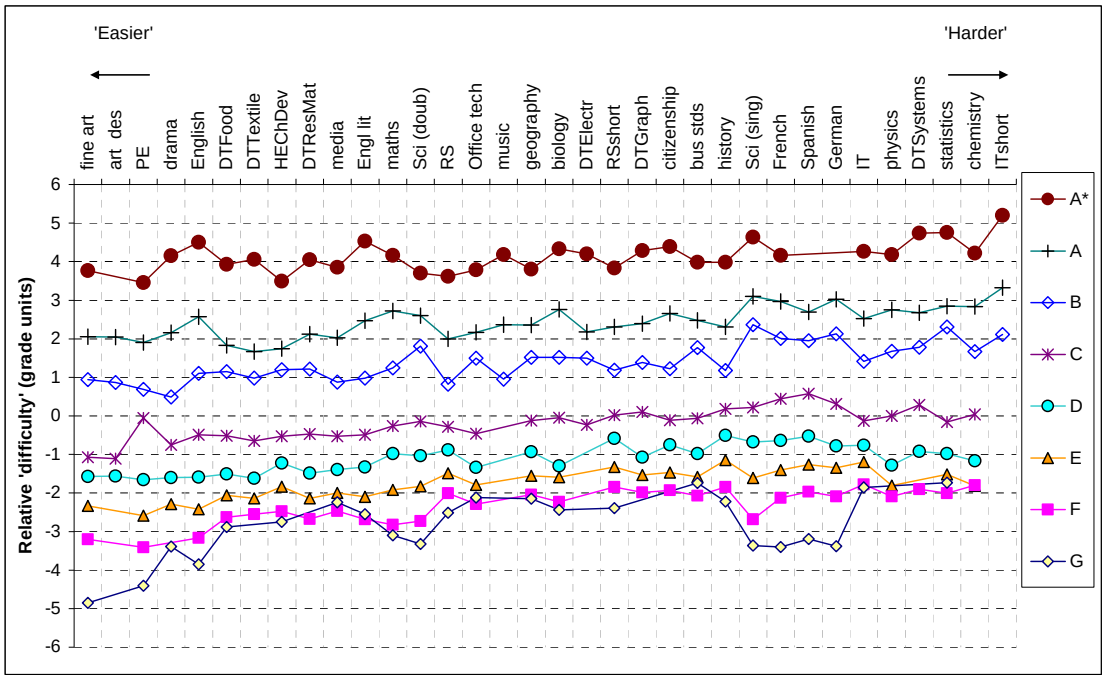
### *5.3.2. Methods*

The five broad methods described in Section 5.1 were applied to this dataset.

The application of the Rasch model made use of the partial credit model within WINSTEPS, which allows the difficulty of individual grades to be estimated, but the estimate taken as the overall subject difficulty was the item difficulty estimate for each subject. Model fit was adequate for most subjects (both INFIT and OUTFIT between 0.6 and 1.5), though five subjects (Art & Design, DT Systems, DT Electronics, Music and Short course IT) did not fit quite as well (INFITs were 1.51, 1.52, 1.56, 1.60 and 1.63, respectively. A previous analysis (Coe, 2008) had found similar results and there is a particular problem with the fit of the lower grades: in many GCSE subjects, candidates who get grade U are actually more able (as indicated by their grades in other subjects) on average than those who get a G, and for some subjects, grades G and F are similarly reversed (see p91, below, for further discussion of this issue).

Despite these anomalies, all grade categories and all the 34 subjects were included in the Rasch analysis in order to be able to compare this method with all the others. Although the fit was not ideal, it was judged to be good enough to allow the analysis to continue. Such a decision is supported by the advice of Linacre (2005) who says that provided INFIT and OUTFIT are below 2.0, items are not detrimental to measurement. Item-measure correlations (i.e. the correlation between a person's grade in a particular subject and the model's estimate of their ability, based on all their grades) were above 0.75 for all 34 subjects. Person reliability was 0.95 and PCA showed 82% of the variance explained by the measures, so there was strong support for the existence of a unidimensional latent construct.

*Figure 12: Rasch estimates of relative difficulty of each grade in each GCSE subject*




*Difficulty estimates are not shown when OUTFIT for that grade exceeded 1.7.*

As the Rasch model estimates difficulties independently for each grade in each subject, it seems appropriate to report these as well. These estimates are shown graphically in Figure 12. It is clear from these data that the assumption that gaps between grades are equal, which is required by all the other methods, is rather questionable.

For GCSEs, the variation in the within-subject range between the lowest and highest grade appears to be smaller than at A-level. Smallest is IT, with a gap of just about six 'average' grades between G and A*, compared with the nominal seven, though several other subjects have similar ranges. Largest is Fine Art, with almost a nine-grade gap.

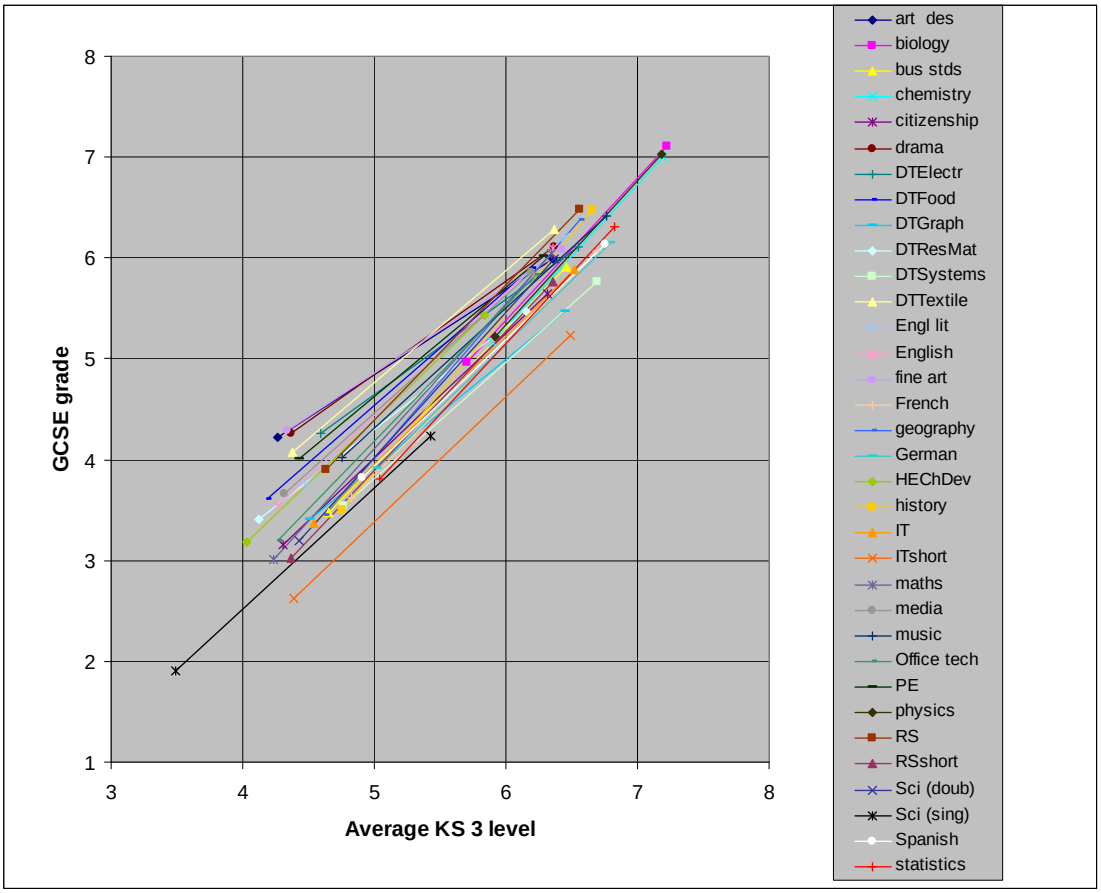
This variation in range, together with apparently random variation in the position of intermediate grades, appears to be enough to lead to some of the same anomalous reversals as were discussed in relation to A-level. For example, the higher grades (A*-C) in French are harder than their equivalents in History: at grade B, the difference is almost a whole grade. However, for the lower grades the direction is reversed: at grade G, History is more than a grade harder than French.

At GCSE there does appear to be more of a problem than at A-level with systematic variation in the size of the gaps at the top and bottom ends, across all subjects. Part of the issue here, as pointed out above, is that in some subjects grade G is actually estimated as being harder than the grade above it, F. What this means is that in these subjects the Rasch model is estimating the 'difficulty' of a particular grade by determining the level of ability at which a person is more likely to be found in that

category than in the grade below. Ability in this context is determined by their performance in all their subjects. If, as here, it turns out that increasing ability is actually associated with lower likelihood of achieving the higher grade then we must conclude that those grades are not really measuring the same thing as grades in all the other subjects.

Whatever the reasons, it is clear from looking at Figure 12 that the Rasch model indicates that the gaps at the lower end of the scale are substantially smaller than those at the top end. In fact the average gap across the bottom four grades (from G to D) is about the same size as the average gap between the highest two (from A to A*). This suggests that the assumption made by all the other models, that the intervals between grades may be treated as equal, is rather problematic.

*Figure 13: Regression line segments forGCSE grade on average KS3 for each of the 34 subjects*




*GCSE grades coded as A*=8, A=7, B=6, C=5, D=4, etc. Regression line segments are shown at the mean of the average KS3 level for that subject plus or minus one standard deviation.*

Subject Pairs Analysis was conducted within the statistical analysis program SPSS, using a macro to calculate average differences in subject grades for each pair of

subjects. Two different methods were then used to aggregate these. The first treated all subjects as equal and simply calculated the average of the differences between each subject and the other 33. This method is described as 'unweighted SPA'. The second weighted each difference by the number of candidates taking that pair of subjects in calculating this average. This method is described as 'weighted SPA'.

Kelly's method was conducted using an iterative approach, again with an SPSS macro.

The Reference Test method used average Key Stage 3 level as the reference test, running a separate regression of GCSE grade on average Key Stage 3 level for each subject. The resulting regression lines, shown for average Key Stage 3 levels within one standard deviation each side of the mean for that subject, can be seen in Figure 13. A student with an average KS3 level of, say, 5 can typically expect to achieve quite different grades in different subjects. However, the size of the difference depends on the average KS3 level one starts with. In order to compare estimates of subject difficulty from this method with the others, we took the population mean of the average KS3 level (5.3) as the point at which to calculate 'expected' GCSE grades.

Unfortunately, with the GCSE dataset we were not able to apply both of the two versions of the Value-Added method that were used in the A-level analysis. The first model was applied exactly as previously. This model, described as the 'Value-added' model, fitted an ordinary least squares regression model with GCSE grade as the outcome and a dummy variable for each subject, in addition to a set of explanatory variables: average KS3 level, average KS2 level, Free School Meals status and sex. The equation for this model is:[12]

$$y_{si} = \beta_0 + \sum_c \beta_c x_{ci} + \sum_s \beta_s z_{si} + u_i$$

*Equation 4*

However, the second ('Multilevel') model could not be used with the GCSE data. The file for this analysis, with one row per examination result and a column for each variable including the 33 subject dummies, contained nearly five million records and over 40 variables. Neither MLwiN nor the 'MIXED' procedure in SPSS could cope with such a large file with the available processing capacity.

### *5.3.3. Results from the GCSE analysis*

Difficulty estimates from seven different methods for each of the 34 GCSE subjects are shown in Table 26. Correlations among them are shown in Table 27 and a matrix of the scatterplots of these correlations in Figure 14.

[12] Explanations for the terms in the equation can be found on p23.

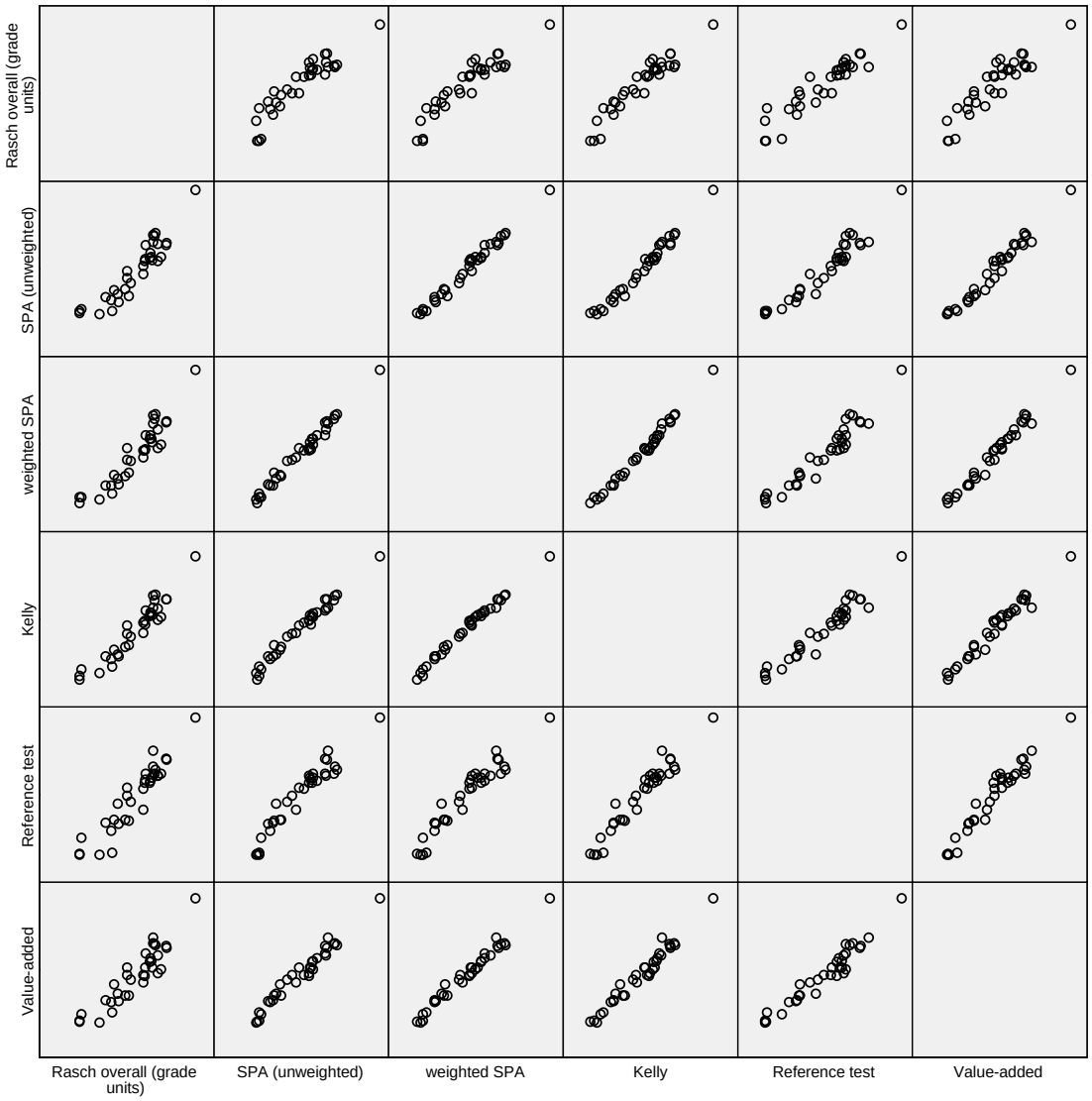
*Table 26: Estimates of relative difficulty of 34 GCSE subjects from six different methods*

| subject | Rasch overall (grade units) | SPA (unweighted) | weighted SPA | Kelly | Reference test | Value-added |
|---|---|---|---|---|---|---|
| art des | -1.20 | -0.43 | -0.38 | -0.49 | -0.58 | -0.50 |
| biology | 0.20 | 0.09 | 0.07 | 0.14 | 0.13 | -0.01 |
| bus stds | 0.34 | 0.13 | 0.17 | 0.22 | 0.18 | 0.15 |
| chemistry | 0.55 | 0.13 | 0.12 | 0.18 | 0.22 | 0.06 |
| citizenship | 0.32 | 0.17 | 0.20 | 0.23 | 0.15 | 0.18 |
| drama | -0.77 | -0.44 | -0.35 | -0.42 | -0.59 | -0.51 |
| DTElectr | 0.21 | 0.11 | 0.08 | 0.10 | 0.16 | 0.08 |
| DTFood | -0.52 | -0.30 | -0.23 | -0.27 | -0.35 | -0.29 |
| DTGraph | 0.31 | 0.12 | 0.17 | 0.19 | 0.13 | 0.14 |
| DTResMat | -0.38 | -0.24 | -0.17 | -0.22 | -0.08 | -0.20 |
| DTSystems | 0.66 | 0.25 | 0.31 | 0.37 | 0.37 | 0.31 |
| DTTextile | -0.50 | -0.41 | -0.30 | -0.35 | -0.57 | -0.40 |
| Engl lit | -0.22 | -0.19 | -0.15 | -0.14 | -0.24 | -0.22 |
| English | -0.64 | -0.27 | -0.23 | -0.24 | -0.27 | -0.27 |
| fine art | -1.20 | -0.41 | -0.33 | -0.45 | -0.59 | -0.49 |
| French | 0.38 | 0.35 | 0.37 | 0.41 | 0.29 | 0.34 |
| geography | 0.17 | 0.04 | 0.07 | 0.12 | 0.07 | 0.00 |
| German | 0.43 | 0.37 | 0.38 | 0.42 | 0.26 | 0.32 |
| HEChDev | -0.46 | -0.20 | -0.14 | -0.17 | -0.24 | -0.10 |
| history | 0.36 | 0.10 | 0.14 | 0.20 | 0.18 | 0.08 |
| IT | 0.48 | 0.26 | 0.25 | 0.27 | 0.20 | 0.21 |
| ITshort | 1.29 | 0.80 | 0.76 | 0.83 | 0.78 | 0.82 |
| maths | -0.18 | -0.08 | -0.01 | 0.00 | 0.00 | 0.00 |
| media | -0.36 | -0.32 | -0.22 | -0.24 | -0.28 | -0.28 |
| music | 0.17 | -0.04 | 0.01 | 0.01 | -0.14 | -0.08 |
| Office tech | -0.10 | -0.13 | -0.02 | -0.03 | -0.06 | -0.05 |
| PE | -1.16 | -0.39 | -0.33 | -0.38 | -0.42 | -0.42 |
| physics | 0.48 | 0.09 | 0.09 | 0.15 | 0.20 | 0.02 |
| RS | -0.14 | -0.26 | -0.12 | -0.12 | -0.25 | -0.22 |
| RSshort | 0.22 | 0.25 | 0.20 | 0.25 | 0.22 | 0.23 |
| Sci (doub) | -0.18 | -0.01 | 0.09 | 0.09 | 0.08 | 0.08 |
| Sci (sing) | 0.38 | 0.28 | 0.30 | 0.28 | 0.45 | 0.40 |
| Spanish | 0.41 | 0.34 | 0.34 | 0.36 | 0.22 | 0.33 |
| statistics | 0.67 | 0.27 | 0.32 | 0.37 | 0.36 | 0.29 |

*Table 27: Correlations among difficulty estimates from different methods*

| | SPA (unweighted) | weighted SPA | Kelly | Reference test | Value-added |
|---|---|---|---|---|---|
| Rasch overall (grade units) | 0.926 | 0.933 | 0.954 | 0.936 | 0.918 |
| SPA (unweighted) | | 0.992 | 0.990 | 0.962 | 0.983 |
| weighted SPA | | | 0.995 | 0.962 | 0.990 |
| Kelly | | | | 0.970 | 0.983 |
| Reference test | | | | | 0.970 |

*Figure 14: Matrix of scatterplots of difficulty estimates from the six methods*

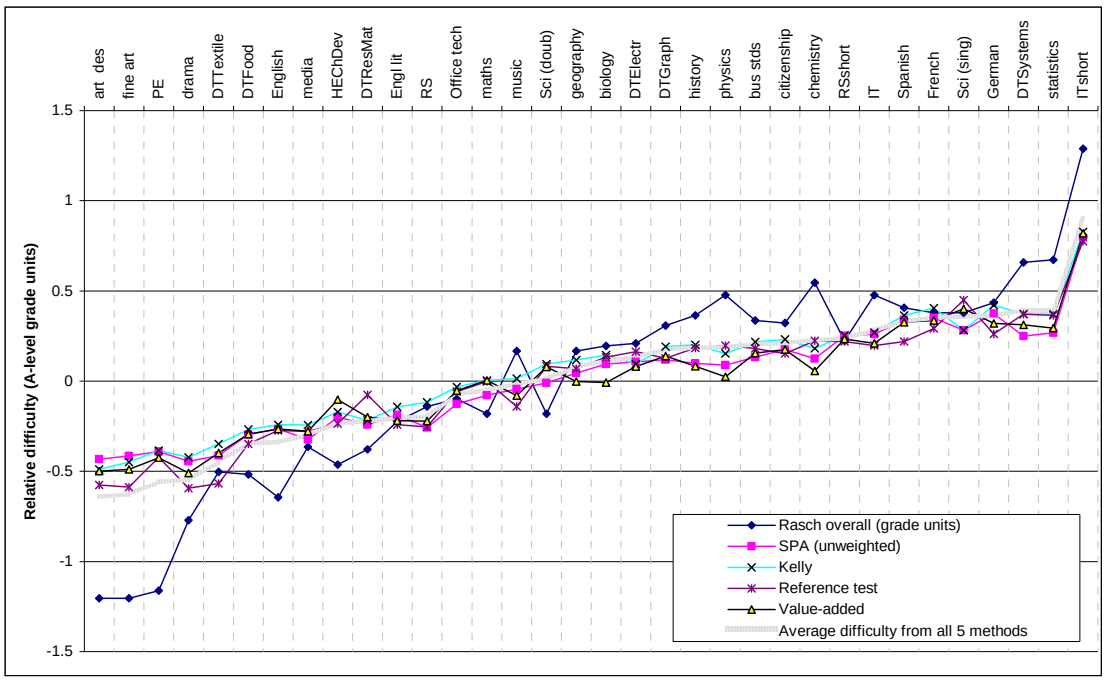



Correlations among the different methods are generally very high, with most values well in excess of 0.9. For four of the methods (the two versions of SPA, Kelly's method and the value-added method) all inter-correlations are above 0.98 and this drops only to 0.97 when the Reference Test method is added to the four. Figure 15

shows a graphical summary of the data in Table 26, and it is clear how close the agreement is between these methods. For these five methods, the difference between the highest and lowest estimate of difficulty is never more than 0.2 of a grade for any subject, and the average difference across all subjects is just 0.1 of a grade. Interestingly, for GCSE data, the unweighted Subject Pairs Analysis and the weighted version are almost identical.

For GCSE subjects, the Rasch method has lower correlations with the others. This may be a result of the significant differences in grade gaps allowed for in this method, but not in the others. Once again, Rasch indicates a wider range between the hardest and easiest subject with about a 2.5 grade gap across subjects, compared with 1.3 for the average of the other methods. This leads to some significant differences in difficulty estimates that may not be apparent from the high correlations. At both ends of the difficulty scale, the Rasch model estimates are some way away from the others. For example, the hardest subject, Short IT, is placed about half a grade harder by Rasch than by the others. At the other end, Rasch makes Art and Design, Fine Art and PE around three-quarters of a grade easier than all the other methods. Such differences are probably large enough to need further investigation and suggest that the choice of one method over another may be important for GCSE subjects, at least if one of them is the Rasch model.

*Figure 15: Relative difficulty estimates of GCSE subjects from five different methods, ranked by average difficulty across all methods*



Finally, we note that Short Course IT is something of an outlier, being rated by all the methods as a grade harder than the next hardest GCSE. It is not clear, however, why this should be.

### *5.3.4. Conclusions*

As at A-level, the different methods for estimating subject difficulties at GCSE are in close agreement. Indeed, with the exception of the Rasch model, the agreement between all the other methods is extremely high. Our conclusion is therefore the same as for A-level difficulties, that debates over which method is to be preferred should not prevent us from acknowledging that there are differences in the difficulties of different subjects.

# 6. CONSISTENCY OVER TIME

## 6.1. Data

The main dataset analysed in relation to this question comes from the ALIS project, which has enabled us to examine relative difficulties of A-level subjects for each year from 1994 to 2006.

*Table 28: Number of candidates in ALIS in each subject from 1994 to 2005*

| | Number of candidates | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **subject** | **1994** | **1995** | **1996** | **1997** | **1998** | **1999** | **2000** | **2001** | **2002** | **2003** | **2004** | **2005** |
| Accounting/Fina | 335 | 402 | 694 | 639 | 577 | 113 | 741 | 868 | 870 | 1,015 | 1,253 | 1,406 |
| Art | | 4,321 | 7,931 | 7,330 | 8,243 | 2,496 | 8,810 | 9,320 | 5,341 | 6,544 | 6,671 | 6,684 |
| Biology | 5,541 | 7,011 | 11,082 | 8,365 | 8,808 | 2,128 | 6,370 | 5,673 | 18,851 | 20,169 | 20,472 | 21,825 |
| Business Studs | | 5,898 | 10,300 | 9,957 | 11,480 | 3,597 | 13,541 | 13,094 | 13,415 | 14,257 | 13,717 | 13,670 |
| Chemistry | | 5,713 | 7,311 | 3,321 | 5,384 | 1,564 | 2,377 | 2,163 | 14,453 | 15,282 | 15,331 | 16,606 |
| Classical Civn | 675 | 952 | 1,962 | 1,753 | 1,569 | 396 | 1,647 | 1,523 | 1,528 | 1,772 | 1,740 | 1,860 |
| Communication S | 1,597 | 2,350 | 1,685 | 1,998 | 1,458 | 320 | 1,329 | 1,482 | 969 | 972 | 1,018 | 940 |
| Computing | 1,536 | 2,188 | 3,539 | 3,475 | 3,918 | 1,117 | 4,515 | 4,634 | 4,329 | 4,350 | 3,868 | 3,504 |
| Design and Tech | 1,533 | 1,677 | 3,313 | 2,969 | 4,195 | 1,247 | 5,034 | 5,178 | 6,029 | 6,945 | 6,993 | 6,986 |
| Economics | 4,152 | 4,201 | 7,553 | 5,823 | 5,637 | 1,605 | 5,856 | 5,670 | 5,969 | 6,358 | 6,147 | 6,554 |
| EnglishLit | 8,967 | 9,118 | 15,939 | 16,604 | 18,191 | 5,280 | 18,681 | 17,968 | 19,521 | 21,012 | 20,913 | 22,109 |
| Environ Science | 327 | 470 | 485 | 651 | 677 | 140 | 674 | 562 | 534 | 583 | 498 | 525 |
| French | 3,991 | 4,721 | 9,184 | 7,437 | 7,116 | 1,987 | 6,190 | 6,271 | 6,224 | 6,465 | 6,252 | 6,136 |
| General Studies | 9,032 | 11,223 | 12,177 | 9,438 | 12,451 | 1,679 | 16,908 | 18,879 | 25,443 | 25,278 | 25,109 | 26,746 |
| Geography | 3,763 | 4,321 | 10,681 | 12,458 | 10,434 | 3,527 | 7,341 | 7,619 | 14,141 | 15,170 | 14,574 | 14,519 |
| Geology | 597 | 747 | 887 | 846 | 761 | 208 | 679 | 742 | 802 | 811 | 769 | 766 |
| German | 1,656 | 1,800 | 3,764 | 3,207 | 3,138 | 956 | 2,901 | 2,870 | 2,715 | 2,813 | 2,617 | 2,565 |
| Government- Pol | 1,938 | 2,133 | 3,478 | 2,994 | 2,750 | 612 | 2,934 | 3,027 | 3,339 | 4,203 | 4,194 | 4,968 |
| Graph Comm | 527 | 530 | 771 | 843 | 396 | 168 | 357 | 378 | 762 | 1,216 | 1,268 | 1,414 |
| History | 6,509 | 7,328 | 14,147 | 12,524 | 12,911 | 3,842 | 13,407 | 13,882 | 16,159 | 18,137 | 18,852 | 19,717 |
| Home Economics | 471 | 486 | 848 | 684 | 581 | 186 | 771 | 753 | 382 | 413 | 371 | 333 |
| Latin | 119 | 123 | 551 | 386 | 436 | 100 | 456 | 452 | 434 | 613 | 616 | 661 |
| Law | 931 | 1,513 | 2,179 | 2,560 | 2,589 | 712 | 3,039 | 2,923 | 4,132 | 4,886 | 5,866 | 6,792 |
| Mathematics | 3,701 | 2,303 | 11,313 | 6,616 | 9,534 | 2,630 | 5,949 | 7,845 | 19,096 | 21,074 | 21,365 | 22,539 |
| Music | 825 | 969 | 2,051 | 1,953 | 2,153 | 678 | 2,337 | 2,363 | 2,449 | 2,870 | 2,892 | 3,153 |
| Photography | 149 | 261 | 536 | 706 | 848 | 155 | 1,050 | 1,170 | 1,392 | 1,826 | 2,343 | 2,534 |
| Physics | 4,398 | 5,233 | 5,380 | 3,495 | 2,669 | 1,002 | 1,550 | 1,615 | 12,468 | 12,790 | 11,860 | 12,118 |
| Psychology | 2,514 | 4,116 | 6,402 | 7,435 | 453 | 105 | 8,397 | 8,780 | 13,289 | 18,550 | 20,794 | 23,128 |
| Religious Studi | 1,075 | 1,409 | 2,837 | 2,613 | 2,716 | 808 | 2,292 | 2,395 | 3,940 | 5,032 | 5,637 | 6,765 |
| Sociology | 4,253 | 5,339 | 7,779 | 7,882 | 7,836 | 2,182 | 8,099 | 7,721 | 8,977 | 10,026 | 10,463 | 11,675 |
| Spanish | 622 | 782 | 1,661 | 1,627 | 1,698 | 539 | 1,843 | 2,020 | 2,096 | 2,400 | 2,532 | 2,567 |
| Theatre St/Perf | 1,499 | 1,847 | 3,557 | 3,750 | 4,420 | 1,403 | 3,321 | 3,197 | 5,789 | 6,530 | 6,642 | 7,058 |

ALIS (the A-level Information System) began in 1983 as a system for helping schools to compare the progress their students have made with that of students in other schools.[13] Currently over 1600 schools and colleges participate in the project, which processes about half of the A-levels taken in the UK. Schools receive value added analysis for individual subject entries, based on simple residual gains when A-level scores are regressed on average GCSE scores, as well as data on a range of student attitudes and perceptions. The numbers of candidates in ALIS in each A-level subject between 1994 and 2005 are shown in Table 28.

## 6.2. Results

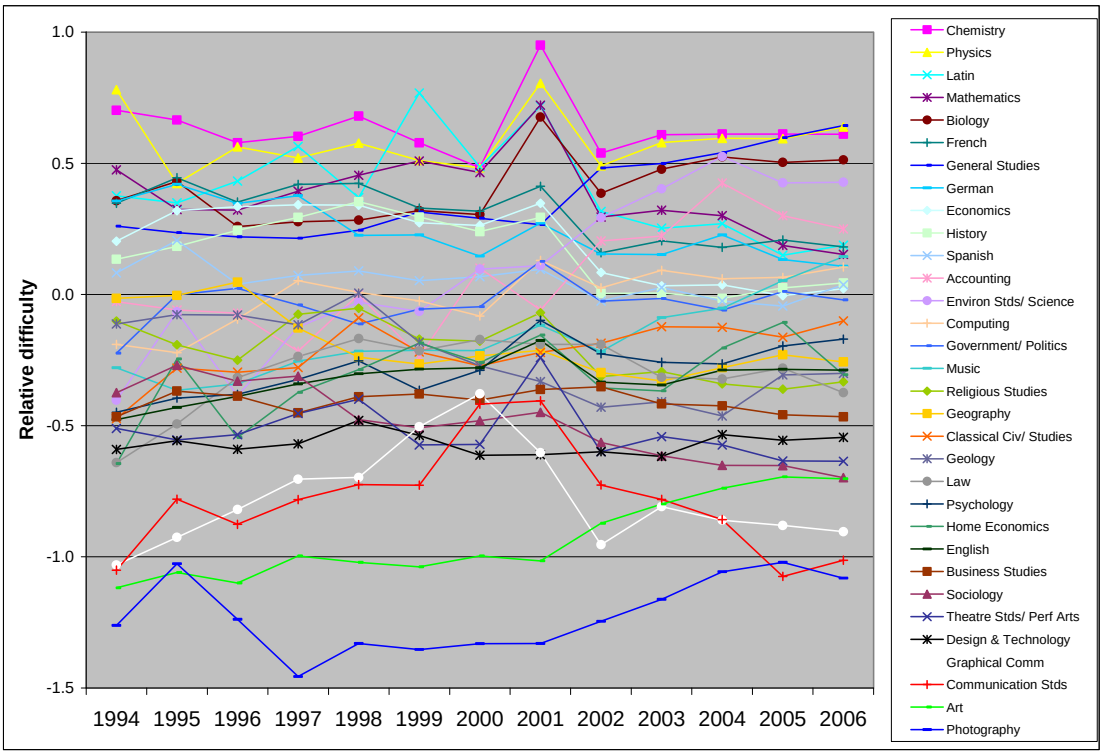
*Table 29: Relative difficulties of 32 A-level subjects from 1994 to 2006*

| | Relative difficulty | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Subject** | **1994** | **1995** | **1996** | **1997** | **1998** | **1999** | **2000** | **2001** | **2002** | **2003** | **2004** | **2005** | **2006** |
| Chemistry | 0.70 | 0.67 | 0.58 | 0.60 | 0.68 | 0.58 | 0.48 | 0.95 | 0.54 | 0.61 | 0.61 | 0.61 | 0.61 |
| Physics | 0.78 | 0.42 | 0.56 | 0.52 | 0.58 | 0.51 | 0.48 | 0.81 | 0.49 | 0.58 | 0.60 | 0.59 | 0.64 |
| Latin | 0.38 | 0.35 | 0.43 | 0.56 | 0.37 | 0.77 | 0.49 | 0.72 | 0.32 | 0.25 | 0.27 | 0.15 | 0.19 |
| Mathematics | 0.48 | 0.32 | 0.32 | 0.39 | 0.45 | 0.51 | 0.46 | 0.72 | 0.29 | 0.32 | 0.30 | 0.19 | 0.15 |
| Biology | 0.36 | 0.43 | 0.26 | 0.28 | 0.28 | 0.32 | 0.30 | 0.68 | 0.39 | 0.48 | 0.52 | 0.50 | 0.51 |
| French | 0.35 | 0.45 | 0.35 | 0.42 | 0.42 | 0.33 | 0.32 | 0.41 | 0.16 | 0.20 | 0.18 | 0.21 | 0.18 |
| General Studies | 0.26 | 0.24 | 0.22 | 0.21 | 0.25 | 0.31 | 0.29 | 0.27 | 0.48 | 0.50 | 0.54 | 0.60 | 0.64 |
| German | 0.36 | 0.42 | 0.35 | 0.38 | 0.23 | 0.23 | 0.15 | 0.27 | 0.15 | 0.15 | 0.23 | 0.13 | 0.11 |
| Economics | 0.20 | 0.32 | 0.33 | 0.34 | 0.34 | 0.27 | 0.26 | 0.35 | 0.08 | 0.03 | 0.04 | 0.00 | 0.02 |
| History | 0.13 | 0.18 | 0.24 | 0.29 | 0.35 | 0.30 | 0.24 | 0.29 | 0.00 | 0.01 | -0.01 | 0.03 | 0.04 |
| Spanish | 0.08 | 0.21 | 0.04 | 0.07 | 0.09 | 0.05 | 0.07 | 0.10 | -0.03 | 0.03 | -0.03 | -0.04 | 0.04 |
| Accounting | -0.03 | -0.06 | -0.07 | -0.22 | 0.00 | -0.24 | 0.10 | -0.06 | 0.20 | 0.22 | 0.43 | 0.30 | 0.25 |
| Environ Stds/ Science | -0.40 | -0.07 | -0.39 | -0.12 | -0.03 | -0.06 | 0.10 | 0.11 | 0.29 | 0.40 | 0.53 | 0.43 | 0.43 |
| Computing | -0.19 | -0.22 | -0.09 | 0.05 | 0.01 | -0.02 | -0.08 | 0.13 | 0.03 | 0.09 | 0.06 | 0.07 | 0.10 |
| Government/ Politics | -0.22 | 0.00 | 0.02 | -0.04 | -0.11 | -0.06 | -0.05 | 0.13 | -0.03 | -0.02 | -0.06 | 0.01 | -0.02 |
| Music | -0.28 | -0.37 | -0.34 | -0.25 | -0.22 | -0.21 | -0.21 | -0.12 | -0.22 | -0.09 | -0.05 | 0.05 | 0.14 |
| Religious Studies | -0.10 | -0.19 | -0.25 | -0.08 | -0.05 | -0.17 | -0.18 | -0.07 | -0.31 | -0.29 | -0.34 | -0.36 | -0.33 |
| Geography | -0.02 | 0.00 | 0.05 | -0.13 | -0.24 | -0.27 | -0.23 | -0.21 | -0.30 | -0.33 | -0.28 | -0.23 | -0.26 |
| Classical Civ/ Studies | -0.47 | -0.28 | -0.30 | -0.28 | -0.09 | -0.22 | -0.27 | -0.22 | -0.19 | -0.12 | -0.12 | -0.16 | -0.10 |
| Geology | -0.11 | -0.08 | -0.08 | -0.12 | 0.01 | -0.18 | -0.27 | -0.33 | -0.43 | -0.41 | -0.46 | -0.31 | -0.30 |
| Law | -0.64 | -0.49 | -0.32 | -0.24 | -0.17 | -0.22 | -0.17 | -0.19 | -0.19 | -0.32 | -0.32 | -0.28 | -0.37 |
| Psychology | -0.45 | -0.40 | -0.38 | -0.32 | -0.25 | -0.37 | -0.29 | -0.10 | -0.23 | -0.26 | -0.26 | -0.20 | -0.17 |
| Home Economics | -0.65 | -0.25 | -0.55 | -0.37 | -0.29 | -0.19 | -0.26 | -0.16 | -0.36 | -0.37 | -0.20 | -0.11 | -0.31 |
| English | -0.48 | -0.43 | -0.39 | -0.34 | -0.30 | -0.28 | -0.28 | -0.18 | -0.33 | -0.35 | -0.29 | -0.29 | -0.29 |
| Business Studies | -0.47 | -0.37 | -0.39 | -0.45 | -0.39 | -0.38 | -0.40 | -0.36 | -0.35 | -0.42 | -0.42 | -0.46 | -0.47 |
| Sociology | -0.37 | -0.27 | -0.33 | -0.31 | -0.48 | -0.51 | -0.48 | -0.45 | -0.56 | -0.62 | -0.65 | -0.65 | -0.70 |
| Theatre Stds/ Perf Arts | -0.51 | -0.55 | -0.53 | -0.45 | -0.40 | -0.57 | -0.57 | -0.24 | -0.60 | -0.54 | -0.57 | -0.63 | -0.64 |
| Design & Technology | -0.59 | -0.56 | -0.59 | -0.57 | -0.48 | -0.54 | -0.61 | -0.61 | -0.60 | -0.62 | -0.53 | -0.56 | -0.54 |
| Graphical Comm | -1.03 | -0.93 | -0.82 | -0.70 | -0.70 | -0.50 | -0.38 | -0.60 | -0.95 | -0.81 | -0.86 | -0.88 | -0.90 |
| Communication Stds | -1.05 | -0.78 | -0.88 | -0.78 | -0.73 | -0.73 | -0.42 | -0.41 | -0.73 | -0.78 | -0.86 | -1.07 | -1.01 |
| Art | -1.12 | -1.06 | -1.10 | -1.00 | -1.02 | -1.04 | -1.00 | -1.02 | -0.87 | -0.80 | -0.74 | -0.70 | -0.70 |
| Photography | -1.26 | -1.03 | -1.24 | -1.46 | -1.33 | -1.35 | -1.33 | -1.33 | -1.25 | -1.16 | -1.06 | -1.02 | -1.08 |

[13] See www.alisproject.org

Since 1994 ALIS has applied Kelly's (1976) method to the results from each year's A-levels and reported 'relative ratings' to show the relative difficulties of A-level subjects.[14] This method was explained in detail in Section 2.1.2 (p18). Results are shown in Table 29 and graphically in Figure 16. Although the latter is quite complex with such a large number of subjects shown, it can be seen that most of the main subjects appear to be relatively stable for most years.

*Figure 16: Stability of relative A-level difficulties over time*



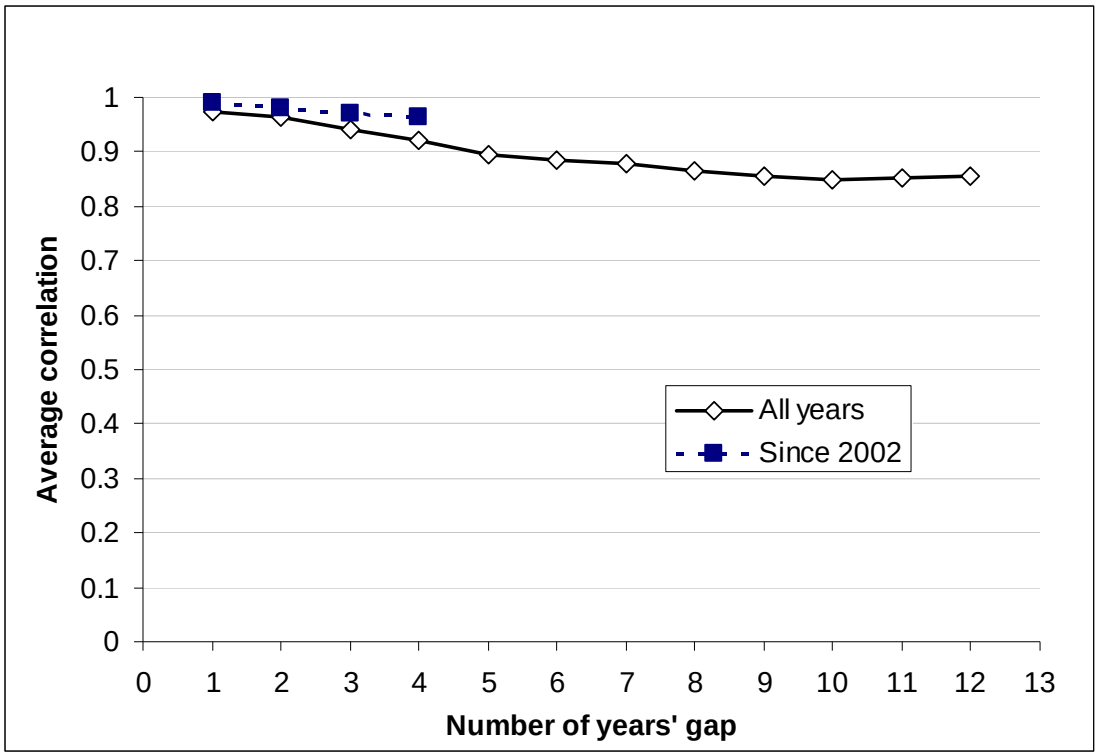

It can also be seen that for many subjects there was something of a blip in 2001, with difficulties returning to normal in 2002 and remaining fairly stable since then. Correlations among the difficulty estimates for subjects across different years are shown in Table 30. For consecutive years, correlations range from 0.93 (between 2001 and 2002) upwards, with an average of 0.97. As the gap between the two years being compared increases, correlations fall gradually, levelling off at around 0.85 when the gap is nine or more years.

[14] This analysis has been conducted by Dr Paul Skinner for the ALIS projects. See http://www.alisproject.org/Documents/Alis/Research/A-Level%20Subject%20Difficulties.pdf

*Table 30: Correlations among subject difficulty estimates in different years*

| | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1994 | = | 0.97 | 0.98 | 0.95 | 0.95 | 0.92 | 0.90 | 0.91 | 0.89 | 0.88 | 0.85 | 0.85 | 0.85 |
| 1995 | 0.97 | = | 0.97 | 0.95 | 0.95 | 0.93 | 0.91 | 0.92 | 0.91 | 0.89 | 0.86 | 0.86 | 0.86 |
| 1996 | 0.98 | 0.97 | = | 0.98 | 0.96 | 0.95 | 0.92 | 0.92 | 0.90 | 0.87 | 0.83 | 0.83 | 0.84 |
| 1997 | 0.95 | 0.95 | 0.98 | = | 0.98 | 0.98 | 0.95 | 0.96 | 0.91 | 0.89 | 0.85 | 0.84 | 0.85 |
| 1998 | 0.95 | 0.95 | 0.96 | 0.98 | = | 0.97 | 0.96 | 0.96 | 0.93 | 0.92 | 0.88 | 0.87 | 0.88 |
| 1999 | 0.92 | 0.93 | 0.95 | 0.98 | 0.97 | = | 0.97 | 0.96 | 0.91 | 0.90 | 0.86 | 0.85 | 0.85 |
| 2000 | 0.90 | 0.91 | 0.92 | 0.95 | 0.96 | 0.97 | = | 0.96 | 0.94 | 0.92 | 0.89 | 0.86 | 0.86 |
| 2001 | 0.91 | 0.92 | 0.92 | 0.96 | 0.96 | 0.96 | 0.96 | = | 0.93 | 0.92 | 0.88 | 0.86 | 0.86 |
| 2002 | 0.89 | 0.91 | 0.90 | 0.91 | 0.93 | 0.91 | 0.94 | 0.93 | = | 0.99 | 0.98 | 0.96 | 0.96 |
| 2003 | 0.88 | 0.89 | 0.87 | 0.89 | 0.92 | 0.90 | 0.92 | 0.92 | 0.99 | = | 0.99 | 0.98 | 0.98 |
| 2004 | 0.85 | 0.86 | 0.83 | 0.85 | 0.88 | 0.86 | 0.89 | 0.88 | 0.98 | 0.99 | = | 0.99 | 0.98 |
| 2005 | 0.85 | 0.86 | 0.83 | 0.84 | 0.87 | 0.85 | 0.86 | 0.86 | 0.96 | 0.98 | 0.99 | = | 0.99 |
| 2006 | 0.85 | 0.86 | 0.84 | 0.85 | 0.88 | 0.85 | 0.86 | 0.86 | 0.96 | 0.98 | 0.98 | 0.99 | = |

*Figure 17: Average correlations between difficulty estimates over different time intervals*




The average correlation for all pairs of years with a given time interval between them is shown graphically in Figure 17. The graph also shows the average correlation between years, limited to the years from 2002 to 2006 when the pattern appears more stable. For these more recent examinations since the introduction of Curriculum 2000,

the correlations are indeed higher, suggesting that relative difficulties have become even more stable than they were previously. The average of the four correlations across a one-year gap is 0.99, falling only as far as 0.96 when the gap is four years. It is also worth pointing out that even higher inter-year correlations can be achieved by removing some of the smaller subjects from this analysis.

### *6.3. Conclusions*

Our overall interpretation of these data is that relative subject difficulties at A-level are quite stable, even over a decade. In the period since 2002, relative difficulties have been particularly stable. Hence, at least at A-level, it does not appear to be too crucial which particular year one takes the data from.

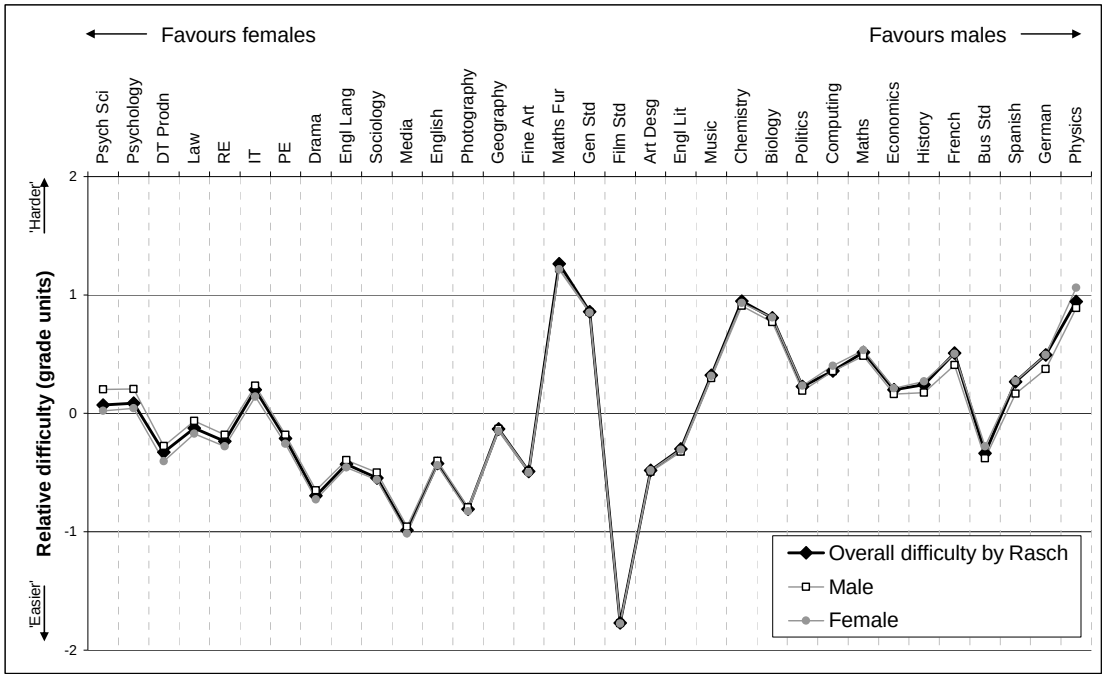
# 7. VARIATION FOR DIFFERENT SUBGROUPS

Our third main research question concerns how much the estimates of relative difficulty of different subjects vary for different subgroups. We present analyses of the 2006 A-level and GCSE national datasets for England. Given the broad agreement that was found across methods in Chapter 5, we present just two of the methods applied to each dataset, the Rasch model and Kelly's method. Difficulty estimates are calculated separately for male and female candidates to illustrate the process of investigating subgroup invariance. In addition, two other subgroup divisions have been used for the GCSE dataset, splitting the data by Free School Meals status and by Independent/Maintained sector.

## 7.1. A-level 2006 data

### 7.1.1. Differential difficulty by gender

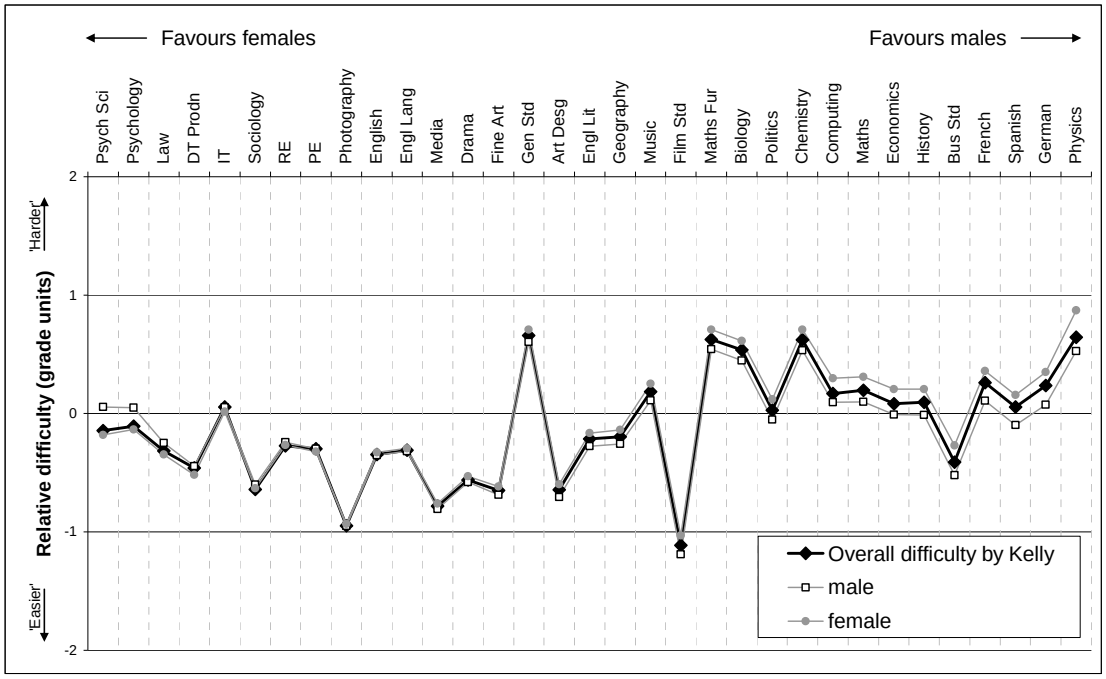
*Figure 18: A-level subject difficulties, by gender, using Rasch*




Results of the Rasch gender analysis of the A-level data are shown in Figure 18, and the analysis for the same data using Kelly's method in Figure 19. These graphs show the overall difficulty for each subject, as reported above, together with the relative difficulties when the analysis was restricted to each gender. Subjects are listed in order of the size of the difference between the estimates for males and females: for

subjects on the left, the difficulty was greater for males than females, so they may be said to 'favour females'; for those on the right of the graph, the difficulty was greater for females, hence those subjects 'favour males'.

*Figure 19: A-level subject difficulties, by gender, using Kelly's method*




For A-level subjects, both the Rasch and Kelly methods are in agreement that there is very little difference between difficulties for males and females. In the Rasch analysis the difference between male and female difficulties for every subject is less than 0.2 of a grade and the mean absolute difference is 0.07 of a grade, ie just 2% of the overall difference between the hardest and easiest subject. The correlation between male and female difficulties for the 33 subjects is 0.99.

For Kelly's method the agreement is not quite as perfect as for Rasch, though still very high. The difference between difficulties for the two sexes is below 0.2 of a grade for the vast majority of subjects, but four subjects (French, Spanish, German and Physics) favour males by nearer to 0.3 of a grade. However, this is still not a large difference; the mean absolute difference is 0.14 of a grade, 8% of the overall difference between the hardest and easiest subject. The correlation between male and female estimates of difficulties is 0.97.

## 7.2. GCSE 2006 data

### 7.2.1. *Differential difficulty by gender*

The same analyses were conducted on the GCSE dataset, and the results are presented as for the A-level data, in Figure 20 and Figure 21.

*Figure 20: GCSE subject difficulties, by gender, using Rasch*

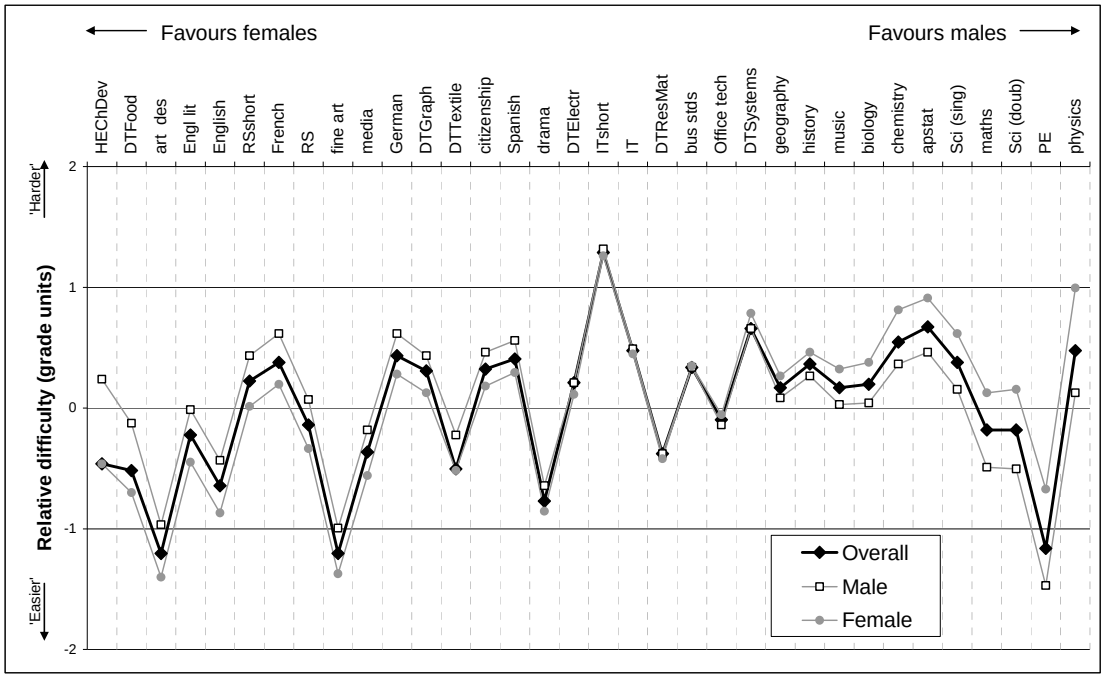
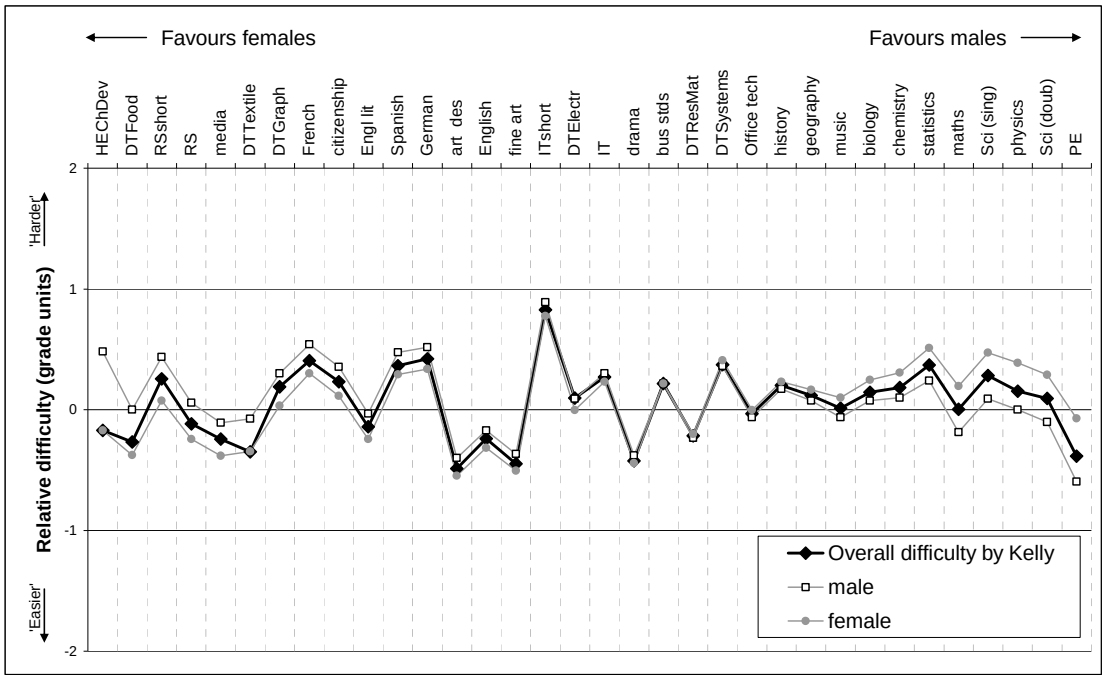


*Figure 21: GCSE subject difficulties, by gender, using Kelly*




For GCSE examinations, there appears to be a more variation in difficulties for the two sexes than at A-level. From the Rasch analysis, three subjects have 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. Behind these, a further four subjects have differences greater than 0.3 of a grade. Overall, the correlation between male and female estimates of difficulty for the 34 subjects is 0.77.

Kelly's method presents a similar picture, with much the same ordering of subjects and scale of differences as from the Rasch model. However, the overall range between the hardest and easiest subjects is smaller for Kelly's method than for Rasch, and the correlation between male and female difficulties is correspondingly lower at 0.66. The mean absolute difference between subject difficulties for the two genders (0.22) is 17% of the overall difference across subjects.

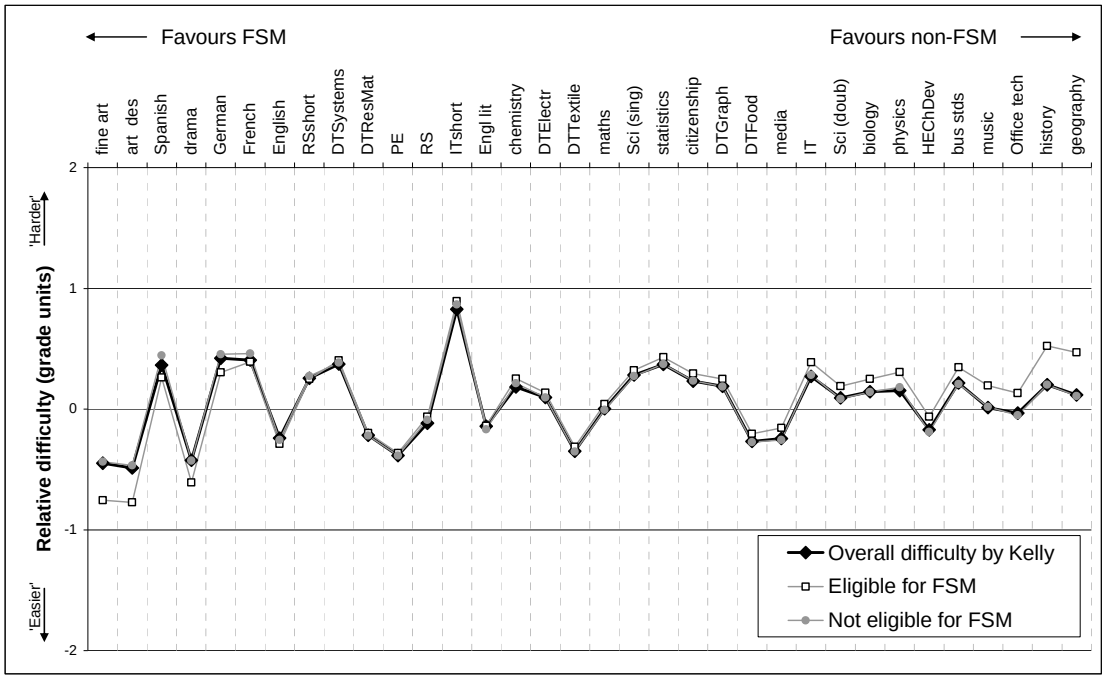
The interpretation of these gender differences is discussed in Section 7.3, below.

### *7.2.2. Differential difficulty by Free School Meals*

A second subgroup split was analysed for the GCSE data, dividing candidates according to their eligibility for Free School Meals (FSM). FSM status is widely used as an indicator of poverty in a child's home background, since only those with household incomes below a certain threshold are eligible. Information on FSM status comes from the Pupil Level Annual School Census (PLASC), which is completed

only by maintained schools. Hence this analysis omits all independent school candidates from both of the groups compared.

*Figure 22: GCSE subject difficulties, by FSM status, using Kelly*




Results of the comparison between difficulty estimates for those eligible for FSM and those not eligible are shown in Figure 22. Of those candidates for whom free School Meals status data were available, about 13% were eligible, with 87% not eligible. Given that they are the overwhelming majority, therefore, it is perhaps not surprising that the relative difficulties for the 'not eligible' group follow the overall difficulties very closely. For most subjects, difficulties for the minority 'eligible' group are also very close to the overall difficulties. In just four subjects, two at each end of the scale, is the difference in difficulty more than 0.2 of a grade. These are Fine Art and Art and Design, which FSM candidates appear to find relatively easy, and History and Geography, which FSM candidates find relatively hard. The overall correlation between difficulties for the two subgroups is 0.93. The mean absolute difference between subject difficulties for the two subgroups (0.11) is 8% of the overall difference across subjects.

The interpretation of these differences has an extra complication, which is that in the case of FSM status, the two groups differ substantially in their overall levels of attainment at GCSE. In fact, their means differ by over a grade: the average GCSE grade achieved by candidates eligible for FSM is just below the mid-point of D and E, while for those not eligible it is just above the mid-point of C and D. This means that when using a method such as Kelly's, that calculates a single difficulty for each subject across all grades, we cannot take account of the fact that the rank order of

difficulty of subjects may depend which grade we are considering, as was shown clearly in the Rasch analysis (see Figure 12, p91). Hence any discrepancy in the relative difficulties for these two groups may not necessarily undermine the claim to unidimensionality that might often be inferred from such differences, but may simply be a consequence of variation in the true sizes of the gaps between grades in different subjects and the inadequacy of assuming that 'difficulty' applies to a whole subject and not to specific grades.

### *7.2.3. Differential difficulty by school sector*

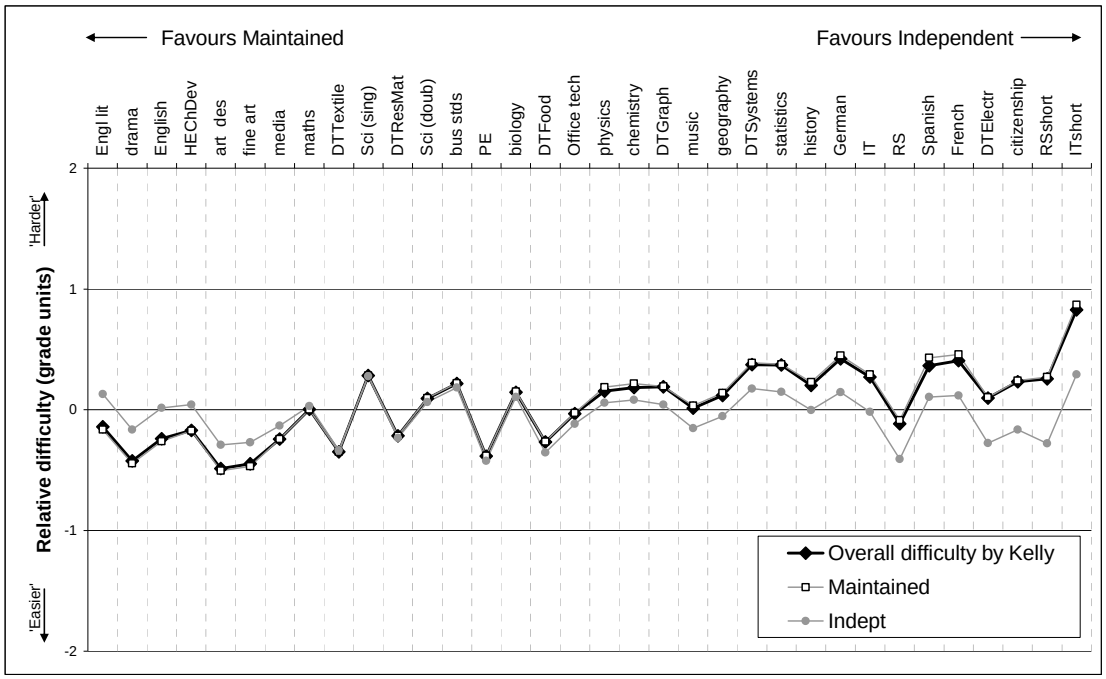
Our final analysis of subgroup differences splits the GCSE sample into candidates in the Independent and Maintained sectors. Results are shown in Figure 23. The correlation between Independent and Maintained sector difficulties is 0.69. The mean absolute difference between subject difficulties for the two genders (0.21) is 16% of the overall difference across subjects.

This split is even more unequal that the previous one, with just 7% of candidates coming from the Independent sector and 93% from Maintained schools. Moreover, the differences in overall achievement between the two groups are even bigger, with the average GCSE grade achieved by candidates in Independent schools between an A and a B, almost two grades higher than the average for those in Maintained schools.

Figure 23 shows that difficulties for Maintained school candidates are almost identical to the overall difficulties, but that for Independent school candidates they do differ somewhat. However, whether this can be taken as evidence against a single unidimensional construct, or simply a reflection of the fact that relative difficulties of subjects at the top grades are somewhat different from those that include the full range of grades (see Figure 12, p91), is open to question.

We note also that the range of relative difficulties for Independent school pupils is significantly less than for those in Maintained schools: the hardest subjects overall are relatively easy for Independent candidates, while the easier subjects are relatively hard for them. However, the most likely explanation for this seems to be again that the variation in GCSE subject difficulties is very much less at the top end than at the lower grades (see Figure 12, p91).

*Figure 23: GCSE subject difficulties, by school sector, using Kelly*




## *7.3. Conclusions*

One of the main arguments against the whole idea of statistical comparisons of subject difficulty has been the issue of subject invariance. We mentioned this issue briefly in Section 2.4.1 and will return to a fuller discussion of it in Section 9.1.4 (p116). However, if the relative difficulties of different subjects differ for different subgroups then it is problematic to talk about the relative difficulty of a subject; instead we may have to talk about the relative difficulty for some particular group.

The empirical evidence presented in this chapter suggests that there may sometimes be a problem with subgroup variation, though this is not always the case. For A-level examinations, there does not appear to be too much of a problem. Relative difficulties for males and females are remarkably close, and it is not obvious what other subgroup divisions might generate bigger differences. Hence, for all practical purposes it does not seem necessary to worry about subgroup invariance at A-level.

For GCSE examinations, the differences in difficulty for the two sexes seem to be large enough to challenge the notion of a single unidimensional construct underlying all these different subjects, and to undermine the notion of 'difficulty' as applying to the subject as a whole. How much of a problem this is depends on how these statistical differences in achievement are interpreted, an issue that we discuss later in Section 9.2 (p117).

The position in relation to other subgroup splits is less clear. We can compare the relative difficulty estimates for groups of differing socioeconomic status (FSM) or school type, but both these comparisons compound substantial differences in overall

attainment with membership of those groups. Given that relative subject difficulties are quite dependent on the particular grade attained, it is hard to say whether these subgroup differences constitute real subgroup invariance or just reflect unequal gaps between grades both within and across subjects.

# 8. ARE STEM SUBJECTS MORE DIFFICULT?

In this chapter we summarise the evidence presented in the previous chapters on the question of whether the STEM subjects are more difficult than other subjects. We have already shown that the different statistical methods of analysing subject difficulties broadly agree with each other (Chapter 5), that subject difficulties (at least at A-level) are highly stable over time (Chapter 6) and that although difficulties vary somewhat by gender at GCSE, there is little variation at A-level or for other subgroup divisions (Chapter 7). Hence the concept of 'difficulty' seems robust enough for us to ask whether STEM subjects are more difficult that other subjects.

In Section 3.3.4 (p68) we defined the STEM subjects as those that are widely required as a basis for further study in science, technology, engineering and mathematics. In the context of A-level subjects we took this to include biology, chemistry, physics and mathematics. A-level further maths was not included in any of the existing analyses we summarised there, but given its value as a foundation for study in many STEM subjects it seems appropriate to add it to our list. However, we note that further maths A-level is not strictly a requirement for many degree courses.

*Figure 24: Difficulty of STEM and non-STEM subjects at A-level*

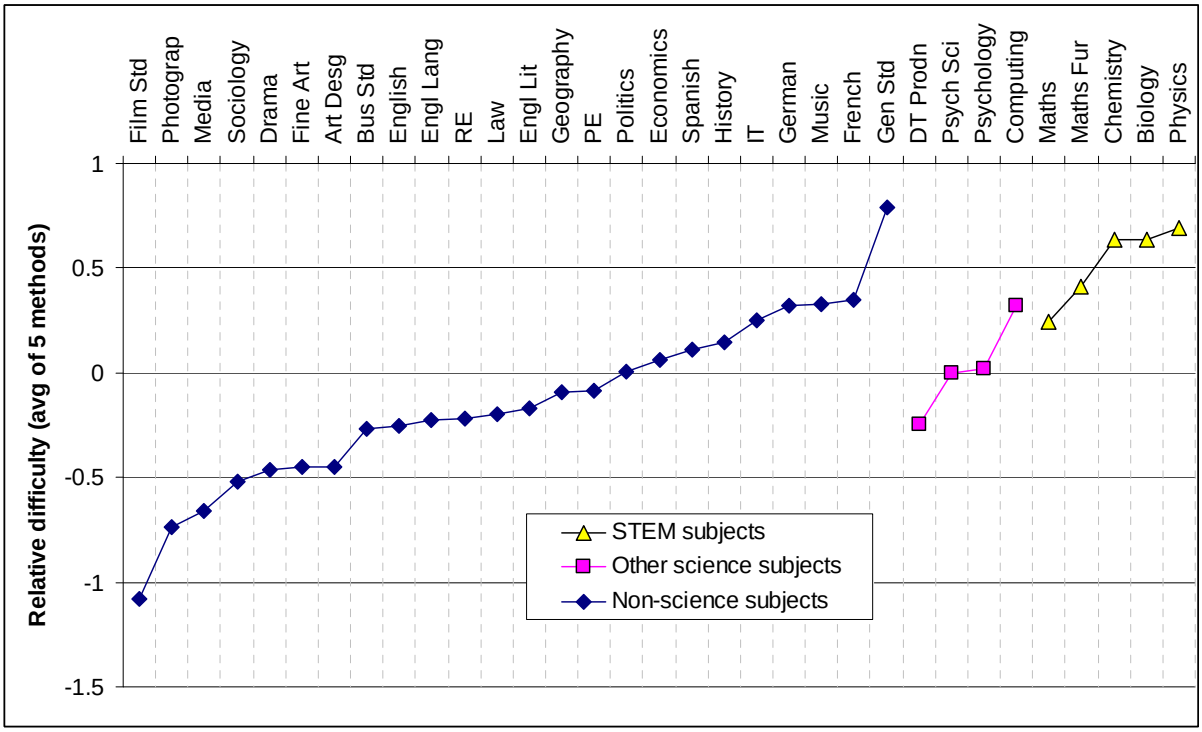


Figure 24 shows the average estimate of relative difficulty of A-level subjects from the five methods used in Section 5.2 (pp80-89), with subjects separated according to whether they are 'STEM', 'Other science', or 'Non-science'. The STEM subjects are not just more difficult on average than the non-sciences, they are actually 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 it is perhaps less clear which subjects should count as 'STEM' since very few subjects are specifically required for further study. In England and Wales, the separate sciences, biology, chemistry and physics, are not offered in many state schools and are not a requirement for any A-level or university course. However, as with further maths at A-level, they are widely seen as providing a good foundation for further study in science, so we have included them in our list of STEM subjects. We have also included double science GCSE, the science examination with the largest entry, and maths, both of which are compulsory subjects in many schools.

*Figure 25: Difficulty of STEM and non-STEM subjects at GCSE*

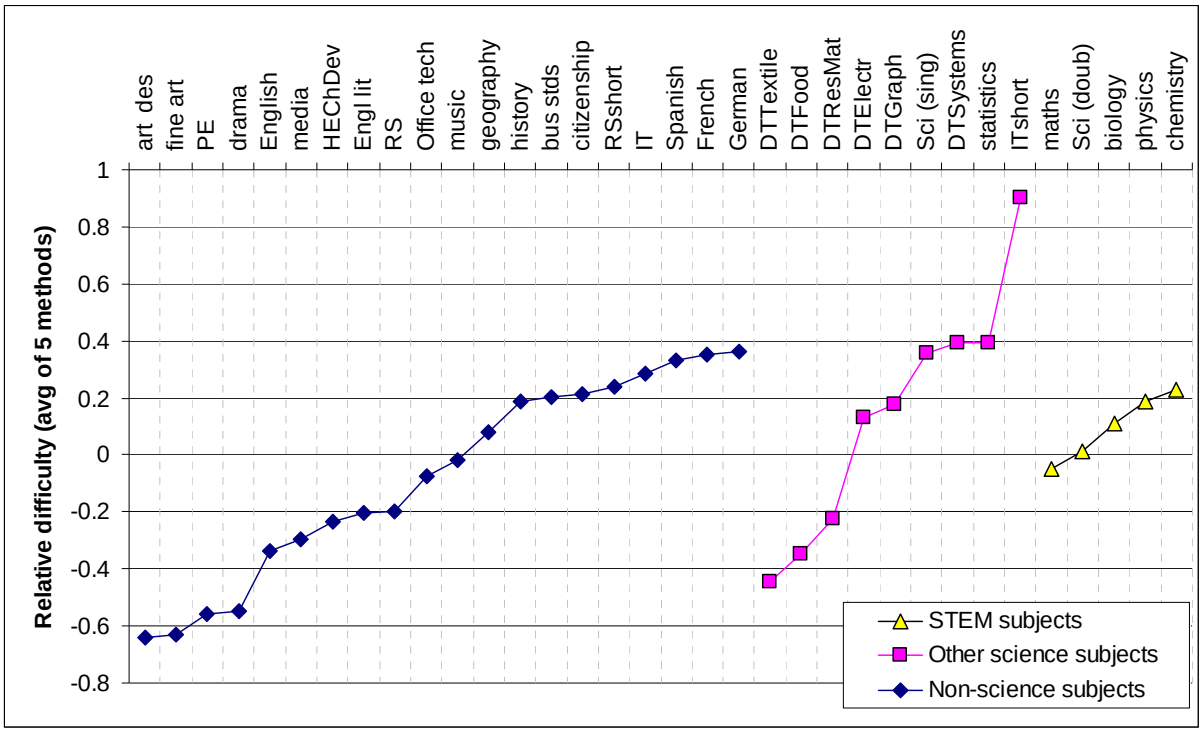



Figure 25 shows the average estimate of relative difficulty of GCSE subjects from the five methods used in Section 5.3 (pp89-97), with subjects separated according to whether they are 'STEM', 'Other science', or 'Non-science'. This time the STEM subjects are a little more difficult on average than the non-sciences, though the difference is a lot less clear than at A-level. The 'other science' group is the hardest on average, though again the difference is not large. In interpreting this comparison for

GCSE subjects, we note that the variation in difficulty at GCSE was found to be less than at A-level, and that the problem of subgroup variation was greater. Hence the interpretation of subject difficulties at GCSE may be more problematic than at A-level.

Overall, we can summarise the comparison between STEM and non-STEM subjects by saying that at A-level it is clear that the STEM subjects are harder. At GCSE the concept of subject difficulty may be more problematic, and the difference between STEM and non-STEM less marked, but there is still a tendency for STEM subjects to be the ones in which students are likely to get lower grades.

# *PART IV*
# *CONCLUSIONS*

# 9. CONCEPTUAL ISSUES

It is clear from the statistical evidence presented in the previous chapters that there are significant differences in the grades achieved by apparently similar students in different subjects. This much is not controversial. Disagreement concerns the question of how – if at all – such differences can be interpreted, how much they matter and what, if anything, should be done about them.

We begin this chapter with a summary of some of the criticisms that have been made of the use of statistical methods to investigate comparability across examinations in different subjects. Most of these critiques have indeed focused on the methods, but our position is that one should criticise a method only in relation to the claims and interpretations that are made for it. The same method may be valid in one application interpreted one way, and invalid in another. For this reason, one cannot evaluate these criticisms in the abstract, but only in relation to particular interpretations of their results, and a number of different interpretations may be possible. We therefore next present an outline of the different possibilities for interpreting such statistical differences.

## 9.1. Criticisms of statistical methods

This section draws heavily on Coe (2007).

A number of writers have discussed issues arising from the use of these methods, including Christie and Forrest (1981), Newbould (1982), Forrest and Vickerman (1982), Alton and Pearson (1995), Pollitt (1996), Cresswell (1996), Goldstein and Cresswell (1996), Fitz-Gibbon and Vincent (1997), Newton (1997; 2005) and Jones (2003). The main criticisms of statistical comparisons may be listed under six headings, and are summarised briefly below.

### *9.1.1. Factors other than difficulty*

A number of writers point out that examination performance is affected by many factors apart from difficulty, so, unless we are prepared to assume (or can show) that these factors are equal or unimportant, we cannot judge difficulty simply by comparing outcomes. Just because common candidates typically get lower grades in one subject than in others it does not necessarily follow that it is more difficult. A number of other factors, including 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, etc., etc., etc., could all affect performance, without making that subject more

difficult (Alton and Pearson, 1995; Goldstein and Cresswell, 1996; Newton, 1997).

### *9.1.2. Multidimensionality / Incommensurability*

In order to compare standards in different subjects we have to have some basis for comparing them. This amounts to saying that different subjects must all measure the same thing, or at least have some significant trait in common. In other words, that subjects being compared must be unidimensional – which of course they are not. It is meaningless to say, for example, that 'art is easier than physics'; they are just different. Goldstein and Cresswell (1996) give an example of a comparison between a spelling test and a degree in English; in theory one could make the two equal in difficulty, at least in a statistical sense, though it would be absurd to say that they were equivalent.

Another subtle variation on this argument is provided by Pollitt (1996). He gives an example of a set of examination results in different subjects where five candidates of equal 'general ability' choose different subjects in which they have different specific aptitudes. When these different aptitudes in different subjects are allowed to interact with different reasons for choosing them, Pollitt shows that the illusion of differential difficulty is created. Some subjects (English and economics in his example) are chosen by those who are best in those subjects, while others (maths) are chosen by those who are relatively weak in that subject. The result is that candidates typically do worse in maths than in their other subjects and hence maths appears 'harder'.

### *9.1.3. Unrepresentativeness*

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. This point is made by, for example, Goldstein and Cresswell, (1996).

Newton (1997) discusses the question of exactly who it is any group of candidates on whom a statistical comparison is made should be representative of. In other words, to what population do we want any claims of comparability to apply? He argues that this should be the whole population of students in the cohort, whether or not they choose actually to take a particular subject. In this case, unless we have truly representative (e.g. random) samples taking every subject, any claims about comparability are very problematic.

### *9.1.4. Subgroup differences*

If we analyse subject difficulties for different subgroups (e.g. males and females) we get quite different results. For example, for males, history may appear 'harder' than maths, while for females maths is the 'harder' (Pollitt, 1996). We might also find that for candidates who take mathematics with it,

physics is really no harder than any other subject, whereas for those who take it without mathematics, physics appears substantially more difficult (Rutter, 1994). Hence a judgement about whether one subject is 'harder' than another depends very much on who happened to take those subjects. And if the characteristics of the entry change, so would the supposed difficulties. (Alton and Pearson, 1995; Pollitt, 1996; Newton, 1997; Sparkes, 2000). The existence of different relative difficulties for different subgroups is also a challenge to the assumption of unidimensionality.

#### *9.1.5. Disagreement among statistical methods*

Some critiques have claimed that different methods of estimating relative difficulties would give different corrections, and there is no clear consensus about which method is best (eg Alton and Pearson, 1995).

#### *9.1.6. Problems of forcing equality*

Adjusting the difficulties of different subjects to make them all equivalent would cause problems during any changeover period for users, the public and professional bodies. Some of the currently 'harder' subjects would need to have absurdly high pass rates, while subjects currently graded 'leniently' would have to be failed by most candidates. This situation would be satisfactory for neither group. Requiring grade boundaries to be modified in this way would change the nature of the examining process and could delay the publication of results. (Alton and Pearson, 1995; Goldstein and Cresswell, 1996).

### *9.2. Interpreting statistical differences*

Given that there are statistical differences in the grades achieved in different subjects, how can these differences be interpreted? A number of possibilities are listed.

#### *9.2.1. No interpretation: Statistical differences are meaningless*

This is the null position which holds that just because we can calculate a number it does not follow that it means anything. This view takes seriously the criticisms listed above (Section 9.1) and concludes that no valid meaning can be attached to differences in the grades achieved in different subjects. Even if we could satisfactorily define what we meant by the relative 'difficulty' of two subjects that are so different as to be incomparable, we could not either in practice or in principle rule out all of the possible reasons for the differences as alternatives to the difficulty of the examinations. The existence of so many anomalies in the statistical evidence (such as subgroups for whom the difficulty is reversed) illustrate the absurdity of trying to interpret statistical differences in any defensible way. The obvious

interpretation – that one subject is 'harder' than another – may be superficially attractive but is fatally over-simplistic and should be resisted.

### *9.2.2. 'Learning gains' interpretation: Statistical differences may indicate the relationship between grades and learning gains in different subjects, provided other factors are taken into account*

This interpretation acknowledges some of the difficulties of inferring difficulty from statistical differences, but takes the view that in practice those difficulties can be satisfactorily overcome. If we have a sophisticated model of learning, which specifies which factors affect the learning gains made by students with particular characteristics in particular contexts, together with adequate measures of those factors, we may interpret any statistical differences between the outcomes that would be expected for particular candidates and the grades they were actually awarded in a specific subject as evidence of the difficulty of that subject.

This interpretation deals with the criticism that other factors may account for the differences by trying to measure and adjust for those factors directly. The issues of multidimensionality, the representativeness of any samples and subgroup differences are effectively dealt with by the model of learning. Although different kinds of learning may take place, the amount of expected learning gain can be compared, so multidimensionality need not be a problem. If actual candidates are different from potential ones on key factors in the model, then by measuring those differences we can adjust for their effects within the model, so representativeness may not be a problem. Differential gains for different subgroups can either be directly modelled, thus removing the anomalies, or viewed as a bias in the outcome measure; either way, they do not present a particular threat to the notion of comparability. We may still be left with the problems caused by trying to force examinations to be equally difficult, but these are really practical and political problems rather than technical ones.

Clearly a key issue for the 'equal learning gains' view of comparability is to decide what factors should be included in the model. A number of the critics of statistical methods point out that 'other factors' may account for differences in performance, and some give examples of such factors, including teaching quality, motivation, different reasons for choosing different subjects, etc. Nowhere have we found any attempt to list these factors comprehensively, however. Such a list may be important if we are to evaluate the validity of the 'learning gains' interpretation of statistical differences, and we present a first attempt to specify a list of factors in Table 31. The list consists of all possible reasons that may influence examination achievement in different subjects.

*Table 31: Possible reasons for systematic differences in the grades achieved in two examinations (e.g. higher grades in X than Y)*

| Type | | Reason for higher grades in X than Y |
|---|---|---|
| Examining and assessment process | 1 | Lack of comparability in the grade awarding process means that different grades are awarded to work of equal standard in different examinations. |
| | 2 | Differences in the assessment structure (e.g. modular vs terminal, or coursework vs timed examination) make the same 'standard' of work easier to demonstrate through route X than route Y. |
| | 3 | Differences in the levels of specification detail of the two examinations (e.g. syllabus X more detailed/explicit than Y) means students are generally better prepared for the specific demands of X than Y. The standard reached in each might nevertheless be equivalent. |
| | 4 | Students doing badly in X are more likely to drop out, hence less likely to be entered for the examination, than those doing badly in Y. For example, if X is modular or relies on continuous assessment, low attaining students will know they have no chance of succeeding so will not enter. |
| Teaching and learning | 5 | Systematically better teaching on X leads to more and better learning than on Y |
| | 6 | More curriculum time devoted to X than Y |
| | 7 | X is inherently more interesting/relevant/motivating/needed than Y |
| Differential choice/selection | 8 | Students choosing (or allowed to do/forced to do) X are generally more motivated/able/advantaged than those who choose/are allowed/forced to do Y. These factors are general in the sense that they may also be expected to affect any other examinations taken by that student. |
| | 9 | Students taking X tend to have specific interest or ability in that subject (e.g. native speakers, musicians). These factors are specific to that particular examination. |
| | 10 | Students taking Y are more likely to do it from need rather than choice, or to choose it even though it is not their best subject |
| | 11 | Schools that offer X are generally better resourced, have better teaching |

The thinking behind this is as follows. If we can specify all the possible reasons why candidates might systematically perform better in one examination than another, including differential difficulty, we may be able to test empirically how strongly each factor is related to achievement grades. Then by eliminating all the 'other reasons' we will be left with differential difficulty as the only explanation. Of course, before we can do this we must decide which of these reasons constitute 'differential difficulty' and which are 'other reasons'. For example, Reason 2, that different assessment routes may make the same 'standard' easier to demonstrate, could probably be put in either category, depending what we mean by 'difficulty'.

### *9.2.3. 'Chances of success' interpretation: Statistical differences may indicate the relative chances of success in different subjects*

This interpretation appears to take a slightly different meaning of the word 'difficulty', emphasising that although the word 'difficult' can mean 'challenging' or 'making great demands', it can also mean 'rarely achieved'. For example, if we find that students typically achieve lower grades in science subjects than in other subjects, the former may be seen as more 'difficult' in the sense that the same grades are typically less likely to be achieved, other things being equal.

This interpretation is similar to the previous one, but differs in that factors that are judged to be external to the individual student, such as the quality of teaching available, are part of what is understood by the 'difficulty' of a particular subject. From the point of view of a potential candidate, there is no reason to distinguish between a subject in which they are likely to do badly because it is generally badly taught and one in which they will do badly because the examination is graded more severely; both would be judged as 'difficult'.

In this interpretation, 'statistical' comparability across different examinations would be indicated by a condition of *equipoise* in which a typical candidate would have no reason to choose to take one subject rather than another on grounds of expecting better grades. The discovery of any statistical differences in the grades achieved by candidates in different subjects would be interpreted as indicating that this condition of *equipoise* has not been met.

Examples of this interpretation of statistical differences can be found in the literature, though they are seldom unequivocal. The statement by Nuttall *et al*. (1974),

> … we can see no logical reason why, if a large *groups* of candidates representative of the population took, for example,

> both English and mathematics, their average grades should not be the same. (p12)

seems to adopt this perspective. A similar view can also be found in Fitz-Gibbon and Vincent (1994),

> The term 'difficult' cannot be taken as meaning *necessarily* or *intrinsically* difficult. Rather, subjects are said to be either 'difficult' or 'severely graded' if the grades awarded are generally lower than might have been reasonably expected on the basis of adequate statistics. (p *i*)

One feature of this interpretation is that it seems to have less difficulty with the problem of subgroup differences. On average, candidates who take both subjects may have a better chance of success in English than they do in mathematics. However, if we know that a candidate is male, for example, our estimate of their relative chances may change. From this viewpoint there is no particular reason why an estimate of chances of success should not depend on the characteristics of the candidate as well as on the subjects taken.

While it might be considered a desirable characteristic of different examinations that their relative difficulties should be the same for, say, males and females, it is not a prerequisite for an understanding of 'chances of success' comparability between them. Of course, if it turns out that relative difficulties are indeed quite different for different subgroups, then we can no longer talk about comparability of subjects *per se* but only of subjects in relation to particular subgroups.

The 'chances of success' conception of comparability is particularly broad-minded on the question of which subjects can be compared. If comparability is based on the concept of chances of success then there seems to be no reason why any groups of subjects cannot be compared. This conception makes no requirement for different subjects to be related in any way, only that a particular level of achievement should be equally rare in each.

### *9.2.4. 'Linking construct' interpretation: Statistical differences may indicate differences in the relationship between grades achieved and some underlying construct such as 'general ability'*

This interpretation starts from the position that examination grades in a range of subjects can be seen as indicators of a common construct such as 'general ability', 'aptitude' or 'capacity for academic learning'. If we accept that different examinations can be interpreted in that way, then the statistical differences in the grades achieved may be seen as evidence of differences in the conversion rate between a particular grade and the ability it indicates.

For example, the analysis presented in Section 6.2 (p99) on A-level difficulties over a thirteen-year period shows that on average, students who take psychology A-level achieve 0.9 grades higher in that subject than comparable students who take chemistry. This is a sizeable difference, especially given its consistency over such a long period. One of the main uses of A-level grades is as an entrance qualification to higher education. So, how should an admissions tutor in, say, history or economics treat otherwise equivalent candidates with grades in these subjects?

It seems reasonable to assume that an admissions tutor in an unrelated subject is not particularly interested in the specific skills accredited by these examinations, but interprets them as an indicator of a student's generalisable capacity for learning in other academic contexts. In this case, it might be true that psychology has been better taught than chemistry; candidates in the former might genuinely deserve their better grades. But if so, the higher grades may not reflect a greater capacity for future learning in a different context. Equally, students who chose psychology may have had a special talent or passion for it and so again deserve their higher grades, but again these qualities may not transfer to their future learning in a different context. We can carry on trying to think of reasons why students might have done so much better in one subject than the other, but in all cases the interpretation seems to be the same. Whatever the reasons for better performance, if we want to get a fair indication of a candidate's suitability for a non-overlapping course, those grades in psychology should be reduced by 0.9 to make them comparable with chemistry.

If the logic of this argument is accepted then there is a compelling case for interpreting the statistical differences in grades achieved as indicating something about the relative value of grades achieved in different subjects. Chemistry is 'harder' than psychology in the sense that the same level of general academic skills and abilities is indicated by lower grades in chemistry than in psychology.

Defining the comparability between examinations in terms of a particular linking construct has been called 'construct comparability' by Coe (2007b). It is generally demonstrated by a combination of judgement applied to observed examination outputs, and statistical modelling, as, for example in equating two parallel forms of an examination, or applying a latent trait model to the results of different examinations. For this version of comparability, the 'standard' of a particular examination performance depends on the level of the linking construct that it signifies. One examination is 'harder' than another if it indicates a higher level of the linking construct.

An example of this kind of comparability can be found in Fitz-Gibbon and Vincent (1997) who talk about the 'common currency' of A-level grades. By this they mean that for some purposes, such as when admissions tutors in UK universities make decisions about which applicants to accept, grades in different subjects may be treated as interchangeable.

> What our analyses suggest is that the 'common currency', i.e. that which can be seen as the information contained in any grade about general aptitudes, can be better operationalised by recognising differences between the subjects in 'difficulty'. (p293-4)

Here the linking construct is 'general aptitudes', though other linking constructs could be imagined. In the context of an admissions tutor using grades in a subject other than their own to infer a candidate's suitability for entry, we might speculate that the construct of interest would be that student's generalisable capacity for learning in another academic context, a construct that is probably reasonably well summarised by the term 'general aptitudes'.

If it is accepted that all the subjects being compared measure (at least to some extent) 'general aptitudes', 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 aptitudes'. So in saying that, for example, physics is 'harder' than biology we mean that a particular grade in physics indicates a higher level of 'general aptitudes' than would the same grade in biology.

Another context mentioned by Fitz-Gibbon and Vincent (1997) is the use of examination grades in school performance tables. Here the grades, perhaps after adjustment for the effects of prior attainment or other factors, might be taken as an indication of the effectiveness of the teaching received, so we have an alternative linking construct from the same analysis and with it the possibility of an entirely different interpretation of the differences in difficulty that were found. Indeed, the same study was cited above as exemplifying a 'chances of success' conception of comparability, so it is clear that the same method can support more than one interpretation.

The problem of subgroup differences is not entirely solved by adopting the 'construct' conception of comparability. Some level of unidimensionality is required by the assumption of a linking construct and significant variation in relative difficulties for different subgroups would undermine this. There are a number of possible ways to get around this problem, though none of them is really completely satisfactory.

One approach would be to limit any comparison to groups of subjects in which there were no substantial subgroup differences. For example, we might say that maths and English cannot really be compared in relation to a common construct because they are not sufficiently unidimensional. We cannot infer levels of 'general aptitudes' from performance in these two subjects since the aptitudes required by each appear to be too specific. On the other hand, despite the differences in the comparison of maths and English for different sexes, grades in these two subjects are highly correlated[15], so in

[15] For example, analysis of the national GCSE 2004 dataset shows a correlation of r=0.77

this particular case we might well conclude that they are sufficiently unidimensional for our purposes.

Another approach would be to limit comparisons to particular subgroups, so we could compare maths and English separately for males and females, and accept that 'comparability' will be gender-specific. However, this requires us to invoke gender-specific linking constructs such as 'male general aptitudes' and 'female general aptitudes' – a rather strange idea.

The question of which subjects can be compared is of course related to this issue. If our interpretation of differences between subjects draws on the notion of a linking construct such as 'general aptitudes' to provide a basis for comparability, then comparisons must be limited to subject examinations that broadly measure that trait. Achievements in any subjects being compared would therefore have to correlate reasonably well with each other, so we might adopt a largely empirical criterion to decide whether subjects are comparable or not.

Note that this last criterion is likely to contrast with the idea of 'cognate' subjects since it would be quite possible for examinations even with the same subject title to correlate very poorly. An example of this might be found at GCSE where under the heading of 'science' we would find biology, chemistry and physics, along with combined science, but also vocational science. Variations in the modes of assessment used could make more difference to the correlations among syllabuses than their nominal content.

### *9.3. Criticisms of judgement methods*

The rationale for the use of judgements by experts to set standards has generally been defended as a form of criterion referencing. Here the standard is defined in terms of a set of criteria that must be demonstrated for it to be awarded. Experts are required to judge whether or not a particular piece of work does indeed meet the criteria.

It often seems to be the case that people's default assumption about what we mean by the 'standard' of an examination is that it relates to some external definition of what a candidate has to do in order to achieve that standard, in other words, that it is criterion-referenced. This is implied, for example, in Tomlinson's (2002) review of A-level standards and is explicitly assumed in the development of National Vocational Qualifications (NVQs) (Wolf, 2002). It is clear from a range of literature (see Baird, 2007, for a summary), however, that such an apparently simple approach is in practice extremely problematic. We now summarise six specific problems with the use of judgment methods based on criterion-referencing.

### *9.3.1. 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. In fact this problem arises even for criteria within a subject, and even in subjects where it might be thought that skills could be relatively easily identified and described.

For example, it is common to find that a sum such as '11+3' is easier than '3+11'. Hence even an apparently well-defined criterion such as 'add two numbers below 20' can be operationalised in different ways, with different facilities, leading to different judgements about whether it has been achieved, depending on the precise items chosen. Clearly, if criteria are required to be applicable to different subjects, a similar problem will arise. Criteria always have to be interpreted and operationalised; they do not define the standard alone. Operationalisations that are valid and reliable and apply across subjects may be extremely difficult to achieve.

### *9.3.2. Crediting responses to different levels of demand*

Early attempts to judge the standard of work where there might be a choice of tiers of entry leading to the same grade found that found that when the same candidates are given questions with different levels of demand, their responses to the easier questions tend to be judged more favourably (Good and Cresswell, 1988). Given a choice between a good answer to an easy question (eg one that is structured) and the inevitably less good answer that the same candidate might have given to a harder question, even experienced examiners tend to give more credit to the former.

Acknowledgement of this problem has led some to change the justification of the standard from pure 'criterion referencing' to what Baird et al (2000) have called 'weak criterion referencing'. In the latter justification, the standard is still defined in terms of criteria, but the judgement about whether it has been met takes into account the level of demand of the questions asked. In practice, asking examiners to try to take this into account in their judgements has been found to lead to significant fluctuations in pass rates (Baird, 2007), so these judgements are often moderated by statistical data. Hence, this is no longer strictly a judgement method, but a pragmatic compromise between judgement and statistical methods that implicitly acknowledge the limitations of the former.

### *9.3.3. Crediting different types of performance*

This difficulty is illustrated in the comparisons in the QCA (2008) study between different subjects. For example, judges found that at GCSE geography was dominated by 'short-answer questions, which focused on very specific items of knowledge' (p4). History, on the other hand, was 'more open-ended, with essay-style questions requiring considerable intellectual

and communication skills to structure a logical response, with a greater emphasis on literary demands and quality of written communication' (p21). Given such differences it is hard to know exactly how a common conception of 'difficulty' can be applied to both. The claim that the judges were apparently able to quantify the levels of difficulty on a common scale seems unconvincing in the absence of detail about precisely how this was done or about the psychometric properties of the resulting scale.

### *9.3.4. Even 'judgement' methods are underpinned by statistical comparisons*

The claim that has been made in applying expert judgements to compare the standards in different subjects, assessed by different types of task under different conditions, is that those judges can assess the demands of the task and the quality of the response on a common scale of 'difficulty'. Even if it could be done, it is hard to know how one might define the construct that is common to both subjects and different assessment modes, some threshold of which defines the standard that equates them. An interesting feature of the QCA (2008) study is that it specified that the judges comparing two subjects should have experience of teaching both. Although this seems like a sensible requirement, in practice it could mean that the construct they used was essentially a statistical one, based on their experience of what a comparable student would be likely to achieve in each subject and assessment mode.

If a judgement about 'difficulty' is based on the experience of seeing the kinds of tasks at which students typically succeed or fail, then it is conceptually no different from the kinds of statistical methods we have used in this study. If the verdict that one subject is more difficult than another boils down to a judgement that one would expect typical students to be less likely to succeed at it, then it is essentially statistically based. In that case we might as well take the need for making a subjective judgement, and the limitations of the sample of personal experience on which this would be based, out of this process and simply calculate the chances of success by more rigorous statistical methods using representative samples.

If one were to argue that judgements about difficulty were not underpinned by some such statistical justification, it would be necessary to show that a subject expert who had no experience of seeing candidates attempt the kinds of tasks being compared could nevertheless abstract from those tasks their intrinsic difficulty in relation to some criteria. No comparability study has ever suggested that such a person exists, let alone tried to use them for this purpose. Indeed, the idea seems absurd.

### *9.3.5. Interpretation and context.*

A further difficulty is that the context in which an examination is taken can make a lot of difference to its difficulty. An example might arise in a comparison between a specification with a single terminal examination, and one with a series of modular examinations contributing to the overall grade.

The content of the two syllabuses may be the same; even the examination questions may be the same. Yet by breaking it up into smaller units, perhaps even allowing examinations to be retaken, the modular specification effectively makes the same 'standard' easier to reach.

In a pure criterion-referenced approach one might ignore the different contexts and say that if the same criteria are demonstrated, the standard is the same. However, the 'weak criterion-referenced' approach, as advocated by Baird (2007) requires that this context be taken into account. In practice, examiners will not have all the relevant information about this context. They may not know how many times a module has been retaken; they will not know the nature of the support that has been given to any given candidate; nor will they know the conditions under which the examination has been taken. How would one judge how much difference it makes to the 'difficulty' of the examination to have it broken into modules, without making use of any statistical data on performance?

### *9.3.6. Aggregating judgements.*

A final problem with criterion-referenced definitions of standards is that in practice the award of a particular grade is likely to depend on performance in a range of tasks with different criteria. Examinations in the UK and elsewhere typically allow compensation, so that a strong performance in one area can make up for a weaker performance in another. Candidates who achieve the same grade may have very different profiles of achievement on the different elements of the examination. This means that there is no single criterion that represents a particular grade. Indeed, the awarding process must specify some method for aggregating the individual criterion-referenced judgements. The resulting grade no longer corresponds to a single criterion and cannot be interpreted and justified as doing so.

Having outlined these six criticisms of using criterion-based judgements to compare standards, we must finally consider an alternative definition.

### *9.3.7. The 'conferred power' definition of comparability*

One response to the difficulties of criterion-referencing, as outlined above, has been to say that the standard cannot be defined explicitly but arises purely from the judgements of experts who are chosen for this purpose. This definition of 'standards' is called the 'conferred power' definition by Baird (2007), or the 'sociological definition' by Baird et al (2000). With this definition one cannot question the 'standard' since the expert decree is effectively a 'speech act' (Searle, 1969). If we were to ask the question 'How do you know that the standards of these two subjects are comparable?' the answer would be 'Because the experts say so.' Although this offers an escape to the problem of comparability that may be philosophically defensible, it seems both scientifically and politically very unsatisfactory. In fact it is not an alternative definition of comparability (as Baird, 2007, claims) at all, but an

acknowledgment of our inability to define it satisfactorily. It also seems unlikely to be convincing to those (e.g. Woodhead or Dunford, cited by Baird, 2007) who have forcefully criticised what they see as the current lack of comparability across subjects

### *9.4. Discussion of conceptual issues*

There seems to be little disagreement that analysis of data from examinations such as GCSE and A-level shows that candidates who are at least in some ways similar tend to achieve quite different grades in different subjects. The statistical methods that have been used to reveal these differences have been subjected to forceful criticism, though in the context of comparing examinations in different subjects, it is not clear that judgement methods offer a viable alternative. Hence the choice seems to be between seeking some valid interpretation of the statistical differences and having to concede that the whole problem of inter-subject comparability cannot be solved.

In relation to the criticisms that have been made of the statistical approaches, we have argued that the validity of a particular method depends on how its results are interpreted and used. Hence we need to be clear exactly how these statistical differences might be interpreted if we are to evaluate the criticisms. Although some would claim that statistical differences are not meaningful (see Section 9.2.1), we have put forward three specific interpretations that we believe are defensible and valid.

The first (9.2.2) is the 'learning gains' interpretation, that statistical differences may indicate the relationship between grades and learning gains in different subjects, provided other factors are taken into account. Precisely what the relevant 'other factors' are, however, may be less clear, and more research is needed before this interpretation can be argued convincingly.

The second (9.2.3) is the 'chances of success' interpretation, that statistical differences may indicate the relative chances of success in different subjects. We have argued that at least a crude form of this interpretation ('a candidate's relative chances of success, on the assumption they are typical of the current entry') does not depend on knowing anything about any 'other factors' that might account for statistical differences, or on some of the assumptions that have been claimed to be a requirement for statistical analyses of comparability (eg unidimensionality).

The third (9.2.4) is the 'linking construct' interpretation, that statistical differences may indicate differences in the relationship between grades achieved and some underlying construct such as 'general ability'. This interpretation probably does require examinations to be unidimensional and

may also require subgroup invariance, but certainly does not depend on knowledge of any 'other factors' that might account for differences in performance.

# 10. POLICY IMPLICATIONS

The issue of comparability of examinations in different subjects is clearly of interest to both public and policy makers. In this Chapter, we attempt to outline a set of options for addressing the policy question.

## 10.1. Policy options

We believe there are three possible policy responses to the issue of subject difficulty. One could make grades statistically comparable, scale grades for specific purposes or simply leave them alone. We present and discuss each option in turn.

### *10.1.1. Leave things alone*

The first choice is to do nothing. The evidence of differential difficulties of different subjects is not robust enough to be a basis for change. As Alton and Pearson (1996) conclude, 'It is hard to resist the conclusion that any attempt to tinker with standards would lead to more unjustifiable outcomes rather than fewer.' It seems that in England at least, the priority for comparability has been to ensure that standards are maintained over time, and any changes to the relative grade standards for different subjects would certainly compromise this.

Some have suggested that although it is widely believed that some subjects are genuinely harder than others, this does not cause too many problems in practice because selection requirements for universities generally specify particular subjects as well as grades (eg Wolf, 2003). In fact, as anyone who types a search term such as 'minimum UCAS points' into a search engine will see, there appear to be hundreds of UK university courses, to say nothing of an equivalent number of employers – including many prestigious companies – for which different subjects are treated as interchangeable qualifications for entry. This argument therefore seems unconvincing to us.

Superficially more convincing is the argument that although there may be substantial differences in the difficulty of different subjects, for the subjects that may present a genuine choice for most students, the difference is much smaller. The number of A-level students who are likely to be faced with a real decision about whether to take, for example, further maths or film studies as their third subject is probably quite small. Given that the more able candidates tend to take the harder subjects anyway, it may be that taking account of subject difficulties would make little difference to the rank order of candidates.

In fact, if we compare the rank order of candidates when grades are treated as equivalent irrespective of subject (as, for example in the UCAS tariff) with the order given by an analysis that takes into account the difficulty of each grade in each subject (as the Rasch analysis of A-level grades reported in Section 5.2), we do find a reasonably high correlation: 0.95. However, it is interesting that even with a high correlation such as this, the choice of one rank order over another could make quite a difference to some selection decisions. For example, if a university wanted to select the top 10% of A-level candidates and used unadjusted UCAS-type tariffs, about one in six of the students selected would not have been chosen, had they adjusted for subject and grade difficulties. If they wanted to select the top 5%, almost half (48%) of those applicants who would be offered places on the basis of raw grades would not have been, had subject difficulties been taken into account.[16]

Furthermore, if examination grades are treated as equivalent, then the current system provides quite strong and potentially undesirable incentives to take some subjects rather than others. For example, we know that, other things being equal, candidates typically achieve 0.9 of a grade higher in A-level psychology than they do in chemistry.[17] For the careful researcher, words like 'other things being equal' and 'typically' may ring alarms that indicate such a result must be treated with a good deal of caution. For the potential candidate, who knows that their entry to university or a good job will depend on their grades in a way that is unlikely to take full account of the scale of this difference, such academic scepticism may seem a luxury likely to be appreciated only by someone whose future would not depend on it.

Hence if we want to be able to use examination grades for purposes such as selection for higher education or employment or in league tables, where grades are treated as interchangeable, we may need to address this question of the incentive to take 'easier' subjects. However, we do not know to what extent candidates' (or schools') choices of examination subjects are actually influenced by their grading severity or, perhaps more importantly, whether changing subject difficulties would change their behaviour.

### *10.1.2. Make grades statistically comparable*

The second policy option would be to level the standards of grades in different subjects to make them statistically comparable. If we accept that the kinds of statistical differences between subjects identified in this report are meaningful in any of the ways outlined in Section 9.2, then this statistical equating seems like an obvious, and perhaps even necessary, thing to do.

[16] This example may be a little unfair since almost 10% of A-level candidates gain straight A grades, so there is no way any university can actually select the top 5% on the basis of raw grades.

[17] See p121

Advantages of such a change are, firstly, that comparability would be transparent and fair in a way that could be much clearer and more explicit than the current system. Stakeholders would no longer have to trust the benign expertise of the awarding bodies and regulators to ensure comparability; the mechanisms for defining and ensuring it could be explicit and objective. Secondly, there would no longer be any overall incentive for either students or teachers to favour one subject over another on grounds of anticipated grade. Thirdly, grades in different subjects would be demonstrably equivalent for use in selection for university entrance or employment. This would contrast with the current system where it is widely believed that some are worth more than others, but making a fair and informed decision about which ones are worth more and how much is far from easy. Fourthly, it could be argued that levelling would actually enhance the status of the currently 'easy' subjects, putting them on a par with even the most 'difficult' and removing their 'Mickey Mouse' status in the eyes of some.

Despite these advantages, there are also some disadvantages. Many of the critics of statistical methods of analysing comparability have assumed that such levelling must be the end result, and a number of problems with this course of action have been identified.

The first is that in order to make subjects comparable we would first have to decide on a particular conception of comparability; different perspectives on what 'comparability' means would lead to different statistical models and hence different 'corrections'. In this context, deciding on a conception of comparability amounts to identifying a particular linking construct in terms of which different subjects can be compared and grades can be interpreted. In other words, we must decide what we think examination grades, considered collectively, should represent. Newton (2005) has identified a choice between privileging one particular view of comparability, described as the *diktat* model, and balancing the needs of more than one view, described as the *contest* model. He argues that because the latter lacks a coherent interpretation of what comparability would mean, the former is to be preferred. Politically, however, a compromise that supports most of the desired interpretations of examination grades moderately well might be preferable to having to choose one and allow others to become untenable.

A second problem is that rigid adherence to a statistical rationale for comparability without any role for judgement would be likely to lead to some anomalies. Examples that might be cited to illustrate the kinds of problems that could arise include the case of Urdu, which often comes out of statistical analyses as one of the easiest subjects (see p62 for an example from Standard Grade). In the UK context, a likely explanation for this seems to be that a large proportion of candidates in this subject are native speakers of the language, so the fact that they do better in this subject than in their others does not necessarily mean it is 'easier', at least not in the sense that the level of linguistic skills and knowledge that must be demonstrated for a particular grade is any less. Whether it would be desirable to lower the grades awarded

in Urdu to bring them into line with what similar candidates achieve in their other subjects depends on what we think those grades should represent. If we think they are primarily certifying a level of attainment of specific skills and knowledge, then the answer is probably no. On the other hand, if we think they primarily signify more general abilities – as for example when they are used for selection – then the answer could be yes. Unfortunately, we cannot really have it both ways.

We might also judge the desirability of statistically equating the standard by its likely consequences. Imagine a scenario in which Urdu grades were dramatically reduced to bring them into statistical equivalence and as a result the proportion of native speakers entering that subject fell. As the entry changed each year, the grade awarded to an examination performance of identical merit would steadily rise. At any point, a candidate who was not typical of those taking this subject, such as a non-native speaker when the majority were native speakers, might reasonably feel that the level of 'difficulty' of the subject, although ostensibly 'fair' for the entry as a whole, was not fair for them.

A third difficulty with statistical equating of grades is that we would have to decide whether to bring all subjects up to the level of the hardest, bring all down to the level of the lowest, move all to the level of the average, or adopt some other benchmark. None of these options is free from problems. At one extreme the change could be accused of 'dumbing down' standards to the lowest common denominator; at the other of creating an elitist examination in which most candidates' experience would be of failure and low grades.

Related to this is a fourth difficulty that the grade profiles for some subjects would become significantly skewed if they were to retain the same calibre of entrants after statistical equating. This is because the wide range of difficulties of different subjects is reflected in their respective ability profiles. For example, if the difficulty of A-level film studies were raised to the level of further maths, fewer than one in five candidates would actually pass. Equally, if further maths were made as 'easy' as film studies, virtually all current entrants would achieve either an A or B, including many of those who currently fail.[18] In neither scenario would it be easy to create an examination with appropriate levels of challenge.

A fifth difficulty is that establishing statistical comparability across subjects would destroy comparability over time. This would be problematic for many stakeholders in the assessment system who would need to distinguish between, and avoid directly comparing, qualifications taken before and after the change.

[18] These estimates are based on the Rasch analysis of A-level grades, presented in Figure 7, p81.

A sixth and final objection that has been made is that requiring this kind of statistical analysis before grades could be awarded would substantially delay the awarding process (Alton and Pearson, 1996). However, we find it hard to believe that this could not be done within the timescales available if it were felt to be necessary.

### *10.1.3. 'Scaling' at the point of use*

The third policy option is to leave grading standards as they are, but to apply a fair conversion rate whenever grades in different subjects are to be treated as equivalent. To do this, a 'scaling' process, such as AMS or the Rasch model (see Sections 2.1.2 and 2.1.3) would be used to calculate the true equivalence among grades. For example, instead of the current UCAS tariff, in which the same number of points are awarded to each grade, regardless of subject, we would have a variable tariff that allocated different points depending on the difficulty of that particular grade.

This system would allow the examining process to award grades much as it currently does, using a combination of statistical evidence and judgement to reflect the level of performance demonstrated by candidates. This would enable examinations, as currently in the UK, to continue to certify attainment in specific domains, to attempt to maintain comparability of subject standards over time and maintain a spread of grades and appropriate levels of challenge despite widely differing ability profiles of candidates in different subjects.

However, if examination grades are used for a specific purpose such as selection to particular higher education courses, an equivalence among them could be calculated for this particular interpretation of what is signified by those grades. For example, if A-level grades are taken as an indication of a potential candidate's general capacity for further academic study, differential points can be assigned to each grade in each subject according to the true level of this construct to which that grade best corresponds.

The advantage of this policy option is that it retains all the advantages of leaving things alone, whilst removing any incentives for students or schools to take easier subjects. In fact in a number of parts of the world, including Australia, Fiji and Cyprus, this kind of 'scaling' is exactly what is done.

One disadvantage that has been pointed out is that the statistical methods used to scale different examinations are inevitably complex and may not be perceived as transparent and fair by all interested parties (Lamprianou, 2007). A second potential disadvantage is that certain examinations might be shown by the scaling process not to be comparable at all. For example, in the Rasch analysis of A-level difficulties presented earlier (Section 5.2.2), we found that General Studies did not fit the model well, implying that General Studies is measuring something different from the other subjects. This really amounts to saying that for a particular purpose, such as academic selection, grades achieved in these subjects are not relevant. Although this may be a position

taken informally by selectors in various contexts, stating it explicitly could be politically controversial.

### *10.2. Encouraging take-up of sciences*

This report has been mainly concerned with the question of whether STEM subjects are more difficult than other subjects. However, the background to that question is a concern that such relative difficulty is part of the reason for the decline in take-up of many STEM subjects and, by implication, that addressing the problem of difficulty would help to redress the decline.

It seems clear from the evidence we have presented that the sciences are both objectively harder and widely perceived to be so. It is also clear that current structures, such as school league tables and UCAS tariffs, that treat all subjects as equivalent therefore create incentives to take easier subjects. From a moral perspective, it is clear that this is unfair. What is not clear, however, is whether such unfairness actually changes people's behaviour. A student who wants to take sciences may well do so in spite of incentives to do otherwise; a school that believes in offering a balanced programme of subjects to meet its students' educational needs will continue to support the more difficult subjects, despite any impact on its league table position. Does difficulty really make a difference to people's choices? More importantly, would addressing the problem of difficulty actually encourage more people to take the STEM subjects?

These are hard questions and the answers are by no means obvious. It is quite possible that if we were to eliminate the discrepancy in the difficulties of examinations in different subjects, we might see no effect on the take-up of sciences. Science subjects were more difficult than others even when they were more popular, so even if difficulty is a factor in take-up, it cannot be the only factor. Indeed it is possible that the decline would even be accelerated. Perhaps the status that science subjects have depends in part on their difficulty; take that away and you remove part of their attractiveness for some. In short, it is very hard to predict the consequences of different ways of attempting to address the problem of subject difficulties, or indeed of any other attempt to encourage more people to study sciences.

We therefore end this report with a plea for a scientific approach from those who are in a position to influence and to change policy in relation to the science curriculum, its assessment and the structures and context within which these things operate. The world of education policy is littered with changes that were made for the best of reasons, with the best of intentions, but with consequences that were ineffective or even counter-productive. It is so hard to predict the results of policy changes that our good intentions, even if supported by a strong theoretical rationale, are not enough. Instead, we must adopt the approach of an engineer: try some likely strategies and evaluate

them properly to see whether any of them does indeed produce the desired effect. Failure to do this may still leave us with the satisfaction of feeling that we have done something, but is unlikely to solve the problem. The urge to act is always strong, but the need to do the right thing, and the fear of making it worse by failing to do so, should be stronger. Only by adopting a cautious, scientific approach can we be sure that we will not ultimately do more harm than good.

# 11. REFERENCES


Adams, R (2007) 'Cross-Moderation Methods', 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.

Alis (2004) 'A-Level Subject Difficulties'. The Advanced Level Information System, Curriculum, Evaluation and Management Centre, University of Durham. Available at http://www.cemcentre.org/renderpage.asp?linkid=11625001

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

AQA (2006). *AQA Grade Boundaries (June 2006, GCSE*) [online]. Available from: http://www.aqa.org.uk/over/stat_pdf/AQA-GCSE-BOUND-JUN06.PDF [Accessed 7 June 2007].

Backhouse, J.K. (1972) 'Reliability of GCE examinations: a theorietical and empirical approach', in *British examinations : techniques of analysis* by Desmond L. Nuttall and Alan S. Willmott. Slough: National Foundation for Educational Research.

Backhouse, J.K. (1978) *Comparability of grading standards in science subjects and GCE A-level*. Schools Council Examinations Bulletin 39. London: Evans/Methuen

Baker, E., McGaw, B. and Sutherland, S. (2002), Maintaining GCE A level standards: The findings of an Independent Panel of Experts. London: Qualifications and Curriculum Authority.

Baird, J., Cresswell, M. and Newton, P. (2000). 'Would the real gold standard please stand up?'. Research Papers in Education, 15(2), pp. 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.

Barmby, P. & Defty, N. (2006). 'Secondary school pupils' perceptions of physics'. *Research in Science & Technological Education*, 24(2), pp. 199-215.

Barnes, G., McInerney, D. M. & Marsh, H. W. (2005). 'Exploring sex differences in science enrolment intentions: An application of the general model of academic choice'. *The Australian Educational Researcher*, 32(2), pp. 1-23.

Bramley, T. (2005). 'Accessibility, easiness and standards'. *Educational Research*, 47 (2), pp. 251-261.

Cheng, Y., Payne, J. and Witherspoon, S. (1995). *Science and mathematics in full-time education after 16: England and Wales Youth Cohort Study*. Sheffield: Department for Education and Employment.

Christie, T and Forrest, G M (1981). Defining public examination standards. Schools Council Publications. Macmillan Education.

Coe, R. (2007) 'Common Examinee Methods' 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 (2008) (in press) 'Comparability of GCSE examinations in different subjects: an application of the Rasch model' *Oxford Review of Education,* 34, 5, (October 2008) DOI: 10.1080/03054980801970312

Coe, R, Tymms, P and Fitz-Gibbon, C (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.

Crawley, F. E. & Black, C. B. (1992). 'Causal modelling of secondary science students' intentions to enroll in physics'. *Journal of Research in Science Teaching*, 29(6), pp. 585-599.

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.

DeBoer, G. E. (1984) 'A study of gender effects in the science and mathematics course taking behaviour of a group of students who graduated from college in the late 1970s'. *Journal of Research in Science Teaching*, 21(1), pp. 95-103.

DeBoer, G. E. (1987). 'Predicting continued participation in college chemistry for men and women'. *Journal of Research in Science Teaching*, 24(6), pp. 527-538.

DfES (Department for Education and Skills) (2007) *Languages Review*. HMSO, ref: 00212-2007DOM-EN

Duckworth, D. & Entwistle, N. J. (1974). 'The swing from science: A perspective from hindsight'. *Educational Research*, 17(1), pp. 48-53.

Fearnley, A.J. (1998) Update on an investigation of methods of analysis of subject pairs by grade. (Unpublished NEAB research paper) [ref in Jones, 2003]

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

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

Forrest GM & Vickerman C (1982) Standards in GCE: subject pairs comparisons, 1972-1980. Occasional Publication 39. Manchester, Joint Matriculation Board.

Forrest, G.M. and Smith, G.A. (1972). Standards in subjects at the Ordinary level of the GCE, June 1971. OP 34. Manchester. Joint Matriculation Board.

Fowles, D.E. (1996) The translation of GCE and GCSE grades into numerical values. (Unpublished NEAB research paper) [ref in Jones, 2003]

Garratt, L. (1986) 'Gender differences in relation to science choice at A-level'. *Educational Review*, 38(1), pp. 67-77.

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

House of Lords (2006) *Science Teaching in Schools*. House of Lords Science and Technology Select Committee, 10th Report of Session 2005-6. London, The Stationery Office.

Hughes S, Pollitt A & Ahmed A (1998) The development of a tool for gauging the demands of GCSE and A-level exam questions. Paper presented at BERA, Queen's University Belfast, August 1998.

Johnson, S. & Bell, J. F. (1987). 'Gender differences in science: option choices'. *School Science Review*, Dec. 87, pp. 268-276.

Jones BE (2003) Subject Pairs over time: a review of the evidence and the issues. *Unpublished AQA research paper RC/220.*

Jones BE (2004) Inter-subject standards: an investigation into the level of agreement between qualitative and quantitative evidence for four apparently discrepant subjects. *Unpublished AQA research paper RC/255.*

Jones BE (1997) Comparing Examination Standards: is a purely statistical approach adequate? Assessment in Education, vol 4, No 2, 249-263.

Kelly, A (1975). The relative standards of subject examinations. Research Intelligence, 1, pp 34-38

Kelly, A (1976b). The comparability of examining standards in Scottish Certificate of Education Ordinary and Higher Grade examinations. Scottish Certificate of Education Examining Board. Dalkeith.

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

Kessels, U., Rau, M. and Hannover, B. (2006). 'What goes well with physics? Measuring and altering the image of science'. *British Journal of Educational Psychology*, 76, pp. 761-780.

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.

Massey, A J (1981). Comparing standards between AL English and other subjects. Test Development and Research Unit, RR 05. Oxford and Cambridge Schools Examination Board.

McGaw, B., Gipps, C. and Godber, R. (2004) Examination standards: Report of the independent committee to QCA. London: QCA.

Murphy, R. (2007) 'Common Test Methods' 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.

Murphy, P. & Whitelegg, E. (2006). Girls in the physics classroom: A review of research on the participation of girls in physics. London: Institute of Physics.

Newbould C A and Schmidt, C C (1983). Comparison of grades in physics with grades in other subjects. Test Development and Research Unit, RR 07. Oxford and Cambridge Schools Examination Board.

Newbould, C A (1982). Subject preferences, sex differences and comparability of standards. British Educational Research Journal, 8 (2), pp 141-146.

Newton PE (2005) 'Examination standards and the limits of linking'. *Assessment in Education*, 12, 2, 105-123.

Newton PE (1996) Subject-pair analysis and the enigma of between-subject comparability. Unpublished Research Report, RAC/713. Associated Examining Board.

Newton PE (1997) Measuring Comparability of Standards between Subjects: why our statistical techniques do not make the grade. British Educational Research Journal, vol 23, No 4, 433-449.

Nuttall DL, Backhouse JK & Willmott AS (1974) Comparability of Standards between Subjects. Schools Council Examinations Bulletin 29.

Osborn, L. G. (1939) 'Relative Difficulty of High School Subjects'. *The School Review*, 47(2), pp. 95-100.

Osborne, J., Driver, R. and Simon, S. (1998). 'Attitudes to science: Issues and concerns'. *School Science Review*, 79(288), pp. 27-33.

Patrick, H. (1996) 'Comparing Public Examinations Standards over time'. Paper presented at BERA Conference, September 1996.

Partis, M.T. (1997) 'Scaling of Tertiary Entrance Marks in Western Australia. Western Australia Curriculum Council: Osbourne Park, WA. Avaliable at http://www.curriculum.wa.edu.au/files/pdf/114537_1.pdf [accessed 12 Apr 2006].

Pell, A. W. (1977). 'Subject swings at A-level: Attitudes to physics'. *School Science Review*, 58, pp. 763-770.

Pell, A. W. (1985). 'Enjoyment and attainment in secondary school physics'. *British Educational Research Journal*, 11(2), pp. 123-132.

Pollitt A 1996 The "Difficulty" of A-level subjects. Unpublished UCLES research paper.

QCA (2007) Code of Practice, GCSE, GCE, GNVQ and AEA, April 2007. London: Qualifications and Curriculum Authority, Welsh Assembly Government, Council for the Curriculum, Examinations and Assessment. (Ref QCA/07/3082)

QCA (2008) *Inter-subject comparability studies*. London: Qualifications and Curriculum Authority, February 2008 (ref QCA/08/3568).

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.

Scottish Qualifications Authority (SQA). 2005. *The Scottish Standard: a guide to pass mark meetings for National Courses* [online]. Midlothian, Scotland: SQA. Available from: <http://www.sqa.org.uk/files_ccc/TheScottishStandard.pdf> [Accessed 22 May 2007]

Searle, J.R. (1969). *Speech acts. An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Seneta, E. (1987) The University of Sydney Scaling System for the New South Wales Higher School Certificate: A Manual. Department of Mathematical Statistics, University of Syney.

Sharp, C., Hutchinson, D., Davis, C. & Keys, W. (1996). *The take-up of advanced mathematics and science courses – summary report*. National Foundation for Educational Research report for SCAA 19.

Spall, K., Barrett, S., Stanisstreet, M., Dickson, D. & Boyes, E. (2003). 'Undergraduates' Views about Biology and Physics'. *Research in Science and Technological Education*, 21(2), pp. 193-208.

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

TQA (Tasmanian Qualifications Authority) (2000) 'Using Rasch analysis to scale TCE subjects'. Available at http://www.tqa.tas.gov.au/4DCGI/_WWW_doc/003675/RND01/Rasch_Intro.pdf [Accessed 4.9.07]

Tomlinson, M. (2002). Inquiry into A level standards. Final Report. London: Department for Education and Skills.

WACC (Western Australia Curriculum Council) (1998) 'Scaling'. Western Australia Curriculum Council: Osbourne Park, WA. Avaliable at www.curriculum.wa.edu.au/files/pdf/scaling.pdf [accessed 24.4.06]

Wiliam D 1996a Meaning and consequences in standard setting. Assessment in Education 3, 3, 287-307.

Wiliam D (1996b) Standards in examinations: a matter of trust? The Curriculum Journal, 7, (3), 293-306.

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.

Wolf, A. (2002). *Does education matter? Myths about education and economic growth*. London: Penguin.

Wolf, A. (2003) An 'English Baccalaureate': exactly what do we want it to do? Oxford Magazine, 216, pp. 16-17.

Wood R (1976) Your Chemistry equals my French. The Times Educational Supplement, 30 July 1976.

Wright, B.D., & Stone, M.H. (1979). *Best Test Design*. Chicago, IL: MESA Press.

Yellis (2006) 'Relative ratings'. Year 11 Indicator System, Curriculum, Evaluation and Management Centre, University of Durham (www.yellisproject.org)