Inter-Subject Comparability of Exam Standards in GCSE and A Level

ISC Working Paper 3

December 2015

Ofqual/15/5798

Contents

Suggested citation:

Ofqual (2015c) *Inter-Subject Comparability of Exam Standards in GCSE and A Level: ISC Working Paper 3*. Coventry, the Office of Qualifications and Examinations Regulation.

This report was written by Qingping He (Research Fellow) and Ian Stockford (Executive Director for General Qualifications).

Abstract

Results from Rasch analysis of GCSE and A level data over a period of four years suggest that the standards of exams in different subjects are not consistent in terms of the levels of the latent trait, specified in the Rasch model, required to achieve the same grades. Variability in statistical standards between subjects exists at both individual grade level and the overall subject level. Results from linear and multinomial logistic regression analyses based on prior attainment and concurrent performance measures also show substantial between-subject variability in difficulty, in terms of the statistical model that has been specified. Findings from this study are generally consistent with those from previous studies carried out by other researchers, working with similar statistical models. It has been demonstrated that the alignment of statistical standards between subjects based on the Rasch model would likely result in a substantial change in the performance standards of the exams for some subjects, evidenced here by significant changes in grade outcomes.

1. Introduction

1.1 The issue of inter-subject comparability

Inter-subject comparability of standards in GCSEs and A levels has been a matter of debate for a long time (see Newton et al., 2007; Coe et al., 2008; Newton, 2012; Ofqual, 2015b). This subject has been studied extensively, involving the use of both judgemental and statistical approaches to conceptualize and quantify inter-subject comparability. Results from analyses using a variety of statistical models suggested that there has been a consistent pattern in the relative 'difficulty' of exams in GCSE and A level subjects, with some subjects shown to be consistently 'harder' or 'easier' than others. However, there remains disagreement between researchers regarding the interpretation of the characteristics of the examinees on which the comparisons were made, the various assumptions involved in the different statistical models, the interpretation of the differences between subjects, and the implications of the results obtained. There has also been debate about whether, and how, such inconsistencies in statistical standards between different subjects should be addressed. Our review of the technical literature on inter-subject comparability (see Ofqual, 2015b) concluded that:

- There has been no consensus over how inter-subject comparability should be conceptualised and defined in the context of GCSEs and A levels, and whether specific subjects can be justifiably said to be severely or leniently graded.
- There has been no consensus on the most valid method that should be used to measure inter-subject comparability and whether, and how, the issues of inter-subject comparability should be addressed.

1.2 Implications of aligning standards between subjects based on statistical analysis

A range of statistical approaches has been used to investigate inter-subject comparability of exam standards in a variety of contexts (see Coe et al., 2008; Ofqual, 2015b). For example, Coe (2008) and Coe et al. (2008) used the unidimensional Rasch model to link students taking different exams from a range of GCSE and A level subjects to estimate the relative 'difficulty' of the exams at individual grade level. In the Rasch model, a person is characterised by his or her ability and an item (an exam in a subject) by a set of item parameters. A mathematical function is used to describe the probability that a person will have a specific score on a particular item given his or her ability and the characteristics of the item. Based on the concept of a 'linking construct' proposed by Newton (2005, 2010), Coe referred this latent ability to the 'general academic ability' that is shared by all students taking the different subjects and measured by the different exams. Newton (2012) attributed Coe's definition of comparability in this context to his 'cause' definition of comparability. Coe (2008) further suggested that subject standards may need to be statistically aligned when grades from different subjects are used for specific purposes (also see Coe et al., 2008), particularly when they are used interchangeably or as equivalent currencies in situations such as admissions to certain university courses and the use of exam results as part of school accountability measures. Coe also indicated that in some countries such as Australia and Scotland, incomparability of statistical standards between subjects was taken into account for specific uses (also see Lamprianou, 2009). Less attention has, however, been paid to the implications of aligning statistical standards between subjects based on inter-subject comparisons on the performance standards of the exams that are related to subject-specific grade criteria in the context of GCSEs and A levels.

The performance standards at different levels are currently articulated through grade descriptions for individual subjects, which represent a source of evidence used during awarding. If the statistical standards based on inter-subject comparability studies were to be aligned for different subjects that were graded based on subject-specific grade requirements, the consequence would be a mismatch between the grade criteria and the performance standards of the exams (for example, subjects either too 'easy' or too 'hard' based on inter-subject statistical comparisons) and a change in the distribution of grades. This study investigated such potential impacts.

1.3 Aims of the study

The study presented in this report aims to achieve the following objectives, to:

- gain an improved understanding of the issues of inter-subject comparability in GCSEs and A levels;

- gain an understanding of the impact of aligning statistical standards between subjects, based on Rasch analysis, on exam performance standards for individual subjects;
- generate new evidence regarding the impact on performance standards of statistically aligning subjects based on a Rasch analysis of subject difficulty.

2. Data collection and analysis

2.1 Data collection

For this study, student-level data, which included some basic background information and GCSE and A level grades in individual subjects from 2010 to 2013, were collected from the Department for Education's National Pupil Database. These grades were used to perform Rasch and other analyses as described below. For GCSEs, subjects with fewer than 5,000 entries were excluded from the analysis in order to obtain accurate estimates of model parameters. For A levels, subjects with fewer than 1,000 entries were excluded from the analysis. Furthermore, students taking fewer than three GCSE subjects or fewer than two A level subjects were excluded in order for the results to be more accurate and reliable. Tables A1 and A2 in appendix 1 show the subjects initially contained in the Rasch analysis, along with the corresponding number of students.

Based on results from an initial Rasch analysis, a selection of GCSE and A level subjects ranging from 'easy' to 'hard', as defined by their Rasch 'difficulty' measures, were identified for further investigation. To examine the impact on performance standards of aligning statistical standards between the subjects, student-level data for these selected subjects for the 2013 exam series were requested from the exam boards. These data included: subject-level grades, subject-level uniform scale marks (UMS) and unit-level raw marks.

2.2 Data analysis

The Partial Credit Rasch Model and Rasch analysis of subject difficulty

The Rasch family of models has been developed for analysing data from tests composed of individual items to establish measurement scales and improve test development. In Rasch modelling, the underlying ability or latent trait of the person to be measured by the test and the characteristics of the items in the test are specified, and a mathematical function is used to describe the probability that the person will have a specific score on a particular item given his or her ability and the characteristics of the item. In the present study, the unidimensional Partial Credit Rasch Model (PCM) was used to analyse the GCSE and A level exam data collected. Unidimensionality requires that items in a test measure a single ability in common.

There are two types of items: dichotomous items for which there are only two score categories (two possible scores such as 'right' or 'wrong') and polytomous items for which there can be more than two possible scores or ordered score categories. For a polytomous item, each score category acts as a step to a higher score category. Possible scores on the item from a test-taker can vary from 0 to the maximum available score of the item. The Rasch model was originally developed to analyse

tests composed of dichotomous items (see Rasch, 1960/1980) and has been extended subsequently for analysing polytomous items. These extended Rasch models include Andrich's Rating Scale Model, Masters' PCM, and other models (see Andrich, 1978; Masters, 1982; Wright and Masters, 1982; Muraki, 1992). The PCM states that, for a polytomous item with a maximum available score of m (the number of score categories minus one), the probability $P(\theta,x)$ of a person with ability θ scoring *x* on the item can be expressed as:

$$P(\theta,x)=\begin{cases}\dfrac{\exp\sum_{k=1}^{x}(\theta-\delta_k)}{1+\sum_{l=1}^{m}\exp[\sum_{k=1}^{l}(\theta-\delta_k)]} & \text{for } x=1,2,\ldots,m\\[2ex] \dfrac{1}{1+\sum_{l=1}^{m}\exp[\sum_{k=1}^{l}(\theta-\delta_k)]} & \text{for } x=0\end{cases} \qquad (1)$$

where δ_k is the location of the k^{th} step on the latent trait continuum and is referred to as the item step parameter associated with a score category (also frequently referred to as step difficulty or threshold). Model parameters for both items and persons can be estimated using methods such as the joint maximum likelihood estimation and the conditional maximum likelihood estimation. $P(\theta,x)$ is also frequently referred to as the category response function or the item category probability curve (CPC). The step parameters δ_k represent the location of the score category on the ability continuum beyond which the probability of achieving a score of k is higher than that of achieving a score of $k-1$. As can be seen from equation 1, $P(\theta,x)$ is a monotonic increasing function of the difference $\theta-\delta_k$. The PCM reduces to the Rasch model for dichotomous items when the maximum score on the item equals one, or $m=1$. It should be noted that the PCM is not a sequential steps model. That is, there are no clearly defined sequential steps that must be followed in answering the item. Furthermore, δ_k cannot be interpreted as the difficulty of scoring a score of k on the item.

The appropriateness of the use of the unidimensional PCM or the validity of the results from the analysis is dependent on the data being unidimensional and fitting the model, and a unidimensional representation having meaning. When test data meet the unidimensionality requirement of the Rasch model and fit the model, estimation of person ability measures and item difficulty measures can be sample independent – invariance of model parameters. That is person ability estimates using different subsets of items in the test will be similar and item difficulty estimates using data from different groups of persons will also be close.

Another important concept in Rasch modelling is the item characteristic curve (ICC). The ICC shows the relationship between the expected score on the item from a person with ability θ, and it is defined as:

$$E(\theta) = \sum_{x=0}^{m} xP(\theta, x) \qquad (2)$$

The ICC has important applications in developing Rasch scales.

Since the step parameter δ_k cannot be interpreted as the difficulty of the step k or the corresponding score category, an alternative definition of step difficulty based on the ICC has been proposed (see Wu and Adams, 2007). For this definition, the difficulty of a score in category k of the item d_k (the step difficulty) is the ability at which the expected score on the ICC is $k - 0.5$:

$$d_k = \theta\big|_{E(\theta)=k-0.5} \qquad (3)$$

This definition is similar to the definition of the item difficulty for dichotomous items and has been adopted in the present study.

Although each score category in a polytomous item is modelled by the Rasch model individually, Wu and Adams (2007), and others, also suggested that it was possible to use the average of the step parameters to characterise the overall difficulty D of the item:

$$D = \frac{1}{m}\sum_{k=1}^{m} \delta_k \qquad (4)$$

Mathematically, D is the ability at which the probability curve of the first category crosses the probability curve of the last category.

Rasch models are primarily used to analyse data from psychological and educational tests, but also to analyse measurement instruments in other areas where the construct or latent trait to be measured by the instrument is clearly defined. Because the instrument represents an operationalisation of the theoretical construct to be measured, the meaning of the latent trait in the Rasch model is, therefore, clear.

Recently, the PCM and item response theory (IRT) models have been used for investigating the comparability of standards in different subject exams (see Coe, 2008; Coe et al., 2008; Korobko et al., 2008; Bramley, 2011). In such investigations, each exam is generally viewed as a polytomous item in a test, and the grades or performance levels assigned to individual examinees for an exam are treated as scores on an item, which represent ordered response categories. All exams contained in the analysis form a test. When the exam data are unidimensional and represented by an underlying latent trait that is shared by the examinees and fits the

Rasch model, results from the Rasch analysis can be appropriately interpreted. However, it has to be noted that, when the Rasch model is used to analyse such data, the latent trait is operationally defined by the set of exams included in the analysis. This makes it difficult to interpret clearly the latent trait implied. It is likely that such a trait would be dominated by the underlying constructs of the subjects that are highly correlated. As mentioned earlier, in interpreting the results from Rasch analysis of the GCSE exam data, Coe (2008) interpreted such a trait as the 'general academic ability' of the individual students.

When analysing the GCSE and A level data for individual exam series using the Rasch model, the comparability of the performance standards of the exams for similar qualifications from different exam boards is assumed. Furthermore, for the same qualification for a specific year, students' grades from different exam boards are combined to produce the grade distribution for the whole cohort. To facilitate the analysis, the ordered letter grades are converted into numerical values, representing ordered category scores: U→0, G→1, F→2, E→3, D→4, C→5, B→6, A→7, A* →8 for GCSEs, and U→0, E→1, D→2, C→3, B→4, A→5, A*→6 for A levels. The maximum score on an item is 8 for a GCSE exam and 6 for an A level exam. The Rasch analysis software WINSTEPS®, which implements the PCM, was used to conduct the analysis. To estimate student abilities and item step parameters using WINSTEPS®, for each exam series, students' numerical grades were arranged into a two-dimensional matrix, with rows representing persons and columns items (subjects) and the cells the numerical grades. The number of columns is the same as the number of subjects included in the analysis, and the number of rows is the same as the number of students. Since students normally take 8 to 11 subjects for GCSE and 3 to 5 subjects for A level, a large proportion of the cells in the matrices for both GCSE and A level had missing values (particularly for A level exams). However, the existence of missing data does not pose any problems for using the Rasch model to analyse the data, as the model functions at the individual item and person levels. The model parameters can be estimated for all persons and items as long as there is sufficient overlap between them in the score matrix. The ability to deal with incomplete data is one of the advantages of using the Rasch model for studying inter-subject comparability.

As indicated earlier, in the PCM, the probability that a person scores a specific category is a function of the difference of person ability and the item step parameter value, and this introduces indeterminacy into the values of model parameters when establishing the Rasch scale. That is, the origins of person abilities and item step parameters cannot be determined independently. Possible ways to deal with this include setting the average of all item step parameters in the test to zero or setting the average of person abilities to zero to determine the origins of the estimates of both the item step parameters and person ability parameters. For the present study, the average of the ability measures for all students included in the analysis was set to

zero for individual datasets. In this case, the step parameters of the subjects from different exam series may be compared directly if it can be assumed that the ability distribution of the students from different exam series is the same.

Regression analysis of subject difficulty based on prior attainment and concurrent performance measures

In addition to Rasch analysis, we conducted analysis using linear regression and multinomial logistic regression involving the use of prior attainment and concurrent performance measures for the 2013 GCSE and A level data in order to investigate the consistency of the relative subject difficulties estimated using different methods. For GCSEs, the prior attainment used was the average of the normalised scores for Key Stage 2 tests in English, mathematics and science taken by the student five years ago. The concurrent performance measure was the average of the numerical GCSE grades in all the subjects taken by the student in the same year (with a minimum of three subjects). For A levels, the prior attainment used was the average of the numerical GCSE grades in all the subjects taken by the student two years previously (again with a minimum of three GCSE subjects). The concurrent performance measure in this case was the average of the numerical A level grades in all the A level subjects taken by the student in the same year (with a minimum of two subjects).

For simple linear regression, the numerical outcome grade (y) for a subject is linearly related to the prior attainment or concurrent performance measure x:

$$y = a + bx \qquad (5)$$

In equation 5, b is the regression coefficient and a is the intercept. If the values of the parameters a and b are the same for different subjects, the outcome grade distribution for all the subjects will be similar when the distribution of their prior attainment or concurrent performance measure is the same. Variation in a and b will indicate different outcome distributions for different subjects with a similar distribution of prior attainment or concurrent performance. Such variation could be interpreted as reflecting the difference in difficulty between the subjects. For the outcome grade y_k at a particular grade *k*, the difficulty of the subject may be defined as the corresponding value of the prior attainment or concurrent performance measure x:

$$\delta_{k,lr} = x_k = \frac{y_k - a}{b} \qquad (6)$$

For multinomial logistic regression, the relationship between each outcome grade for a subject and the prior attainment or concurrent performance is modelled separately. In the present analysis, for both GCSEs and A levels, the lowest grade was taken as the reference category (category 0). For a specific subject, the logarithm of the ratio of the probability of a student being classified into a specific category (grade) k, P_k,

to the probability of being classified into the reference category P_0, given his or her prior attainment or concurrent performance measure x, is expressed as a linear function of x:

$$\ln \frac{P_k}{P_0} = \alpha_k + \beta_k x \qquad (7)$$

where α_k is the regression coefficient and β_k is the intercept. Variation in α_k and β_k would suggest differences in grade distributions between subjects with similar prior attainment or concurrent performance distribution. From equation 7, for two adjacent categories k and $k-1$, the logarithm of the odds ratio can be expressed as:

$$\ln \frac{P_k}{P_{k-1}} = (\alpha_k - \alpha_{k-1}) + (\beta_k - \beta_{k-1})x \qquad (8)$$

This is similar to the partial credit model, which can be expressed as $\ln P_k / P_{k-1} = \theta - \delta_k$ where θ is the person ability and δ_k is the category threshold. If x in equation 8 is treated as a proxy of the current ability of the student, the difficulty of category k can be defined as the value of x when the probability of being classified into category k is the same as the probability of being classified into category $k-1$ (for example, $P_k = P_{k-1}$):

$$\delta_{k,mr} = -\frac{\alpha_k - \alpha_{k-1}}{\beta_k - \beta_{k-1}} \qquad (9)$$

Variation in α_k and β_k between subjects can, therefore, be interpreted as variation in subject difficulty.

Analysis of the impact on performance standards of aligning statistical standards

The Rasch analysis outlined above was combined with operationally available mark data to investigate the impact of aligning standards on this basis. This analysis provides insight into the likely impact on performance standards of any such adjustment by considering the required changes in grade boundaries. In our study, this adjustment was performed at both subject level and the constituent unit/component level. The calculation of these adjustments and their impact on grade outcomes are presented in section 5.

3. Relative difficulty of GCSE and A level exams based on Rasch analysis

3.1 Unidimensionality and model fit

In the Rasch model, because the probability of succeeding at a score category on an item is specified as a function of the difference between person ability and item difficulty, the person parameters and item parameters are placed on the same measurement scale. The extent to which items (subjects) in a test (the collection of subjects) meet the unidimensionality requirement of the model needs to be investigated. Violation of model assumptions can invalidate the interpretation of results. The application of Rasch and IRT models to analyse test data also assumes that the model reflects the functioning of the test items correctly. An evaluation of how well the data fit the model chosen to represent the data is essential in Rasch modelling to ensure the usability of results. Embretson and Reise (2000) and Reckase (2009) outlined procedures for assessing the fit of Rasch and IRT models by test data.

The unidimensionality assumption of Rasch and IRT models can be investigated using factor analysis of row scores or the residuals of person scores (see Yen, 1993; Smith, 2002; Reeve and Fayers, 2005; Reckase, 2009; Linacre, 2013; He et al., 2014). The residual of a person score on an item is defined as the difference between the person's observed score on the item and his or her Rasch model predicted score. If a distinctive factor or factors cannot be identified for the residuals, the unidimensionality of the test data may be assumed. The degree to which the test data fits the model can be evaluated using model fit statistics for both items and persons. Model fit for items can be investigated at individual score category, overall item and whole test levels. Frequently used Rasch item fit statistics include some of the residual-based fit statistics such as unweighted mean squares fit (outfit) statistics and weighted mean squares fit (infit) statistics (see Wright and Masters, 1982; Wu and Adams, 2007; Linacre, 2013). Both infit and outfit statistics have an expected value of 1. The extent to which the values of infit or outfit statistics depart from 1 reflects the level of dissimilarity between the shapes of the observed ICC or CPC and the theoretical ICC or CPC. Items and persons with infit statistics in the range of 0.70 to 1.30 are normally regarded as fitting the Rasch model well (Keeves and Alagumalai, 1999; Linacre, 2002). However, some researchers set the range of acceptable values for infit and outfit statistics even wider, from 0.60 to 1.40 (Tan and Yates, 2007; Wong, McGrath and King, 2011). Linacre (2002) suggested that when model fit statistics were above 2.0, the measurement system would be distorted. This value of 2.0 has been used to judge whether an item or person fits the Rasch model sufficiently well in the present study.

Initial analysis of the datasets using the Rasch model suggested that, for the GCSE data, the category fit statistics for category 0 (grade U) for a substantial number of

subjects were over 2.0. This was also observed in the work by Coe (2008). Some subjects with high misfit values had disordered category thresholds (that is the difference in the values of the step parameters between G and U was negative). For some of these subjects, the bottom two categories (grades U and G) were also disordered (that is the observed level of ability at grade U was higher than that at grade G). Both disordered thresholds and disordered categories reflect the violation of the measurement construct and the inappropriate functioning of the items as would be expected by the measurement model. For the A level data, results from initial analysis indicated that most of the subjects (except Chinese) had fit statistics within 2.0 at both subject level and category level. To resolve the problems of disordered categories and disordered thresholds and a large misfit of bottom categories for the GCSE exams, the bottom grade U was treated as missing and excluded from the analysis. Grade G was treated as the new bottom grade with a numerical score of 0. Inspection of person fit statistics suggested that about 7 per cent of GCSE students and 10 per cent of A level students had fit statistics over 2.0. In order to obtain more accurate estimates of item parameters (which are the focus of this study), students with fit statistics over 2.0 were excluded from the analysis. Therefore, only a sample of the students from the national cohort was used to calibrate the items (subjects) for each dataset. Results from reanalysis of the new datasets suggested that, for both GCSEs and A levels, the mean squares fit statistics for all the subjects were now below 2.0, both at subject level and individual category level, except A level Chinese for which some fit statistics were still over 2.0. A level Chinese was then excluded from the final analysis. Table A3 in appendix 1 shows the GCSE and A level subjects included in the final analysis.

Figures 1 and 2 below show the distribution of the category infit statistics for the GCSE and A level subjects from the 2010 to 2013 exam series that were included in the final analysis. The error bars show one standard deviation at individual categories. For the GCSE subjects, the average of the infit statistics was above 1.0 and decreased slightly from the bottom category (grade G) to the top category (grade A*). For the A level subjects, the average infit statistics were less than 1.0 for the lower categories (grades U and E), close to 1.0 for the middle categories (grades D, C and B), and slightly over 1.0 for the top categories (grades A and A*). These infit statistics suggest that, at category level, the datasets fit the Rasch model reasonably well.

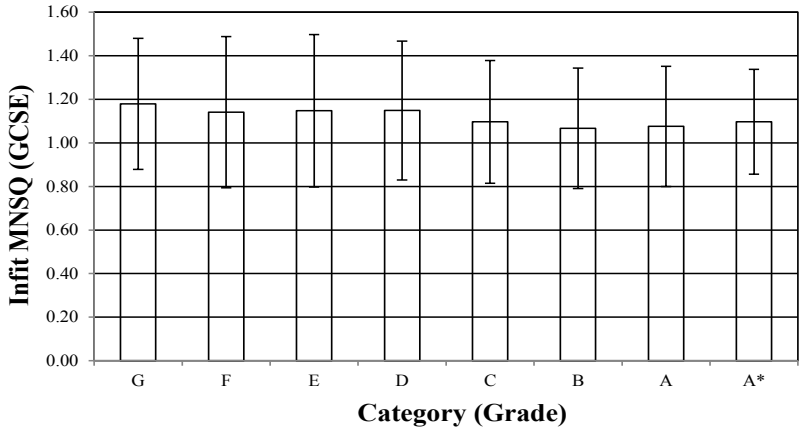

Figure 1: Category infit statistics for GCSE subjects from the 2010–13 exam series.

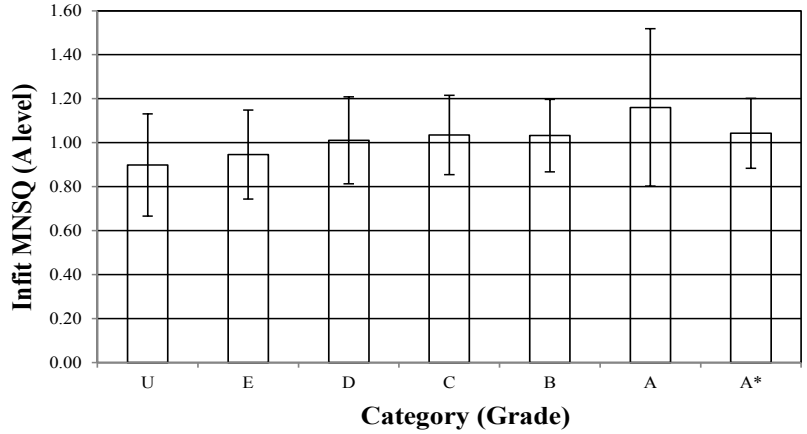

Figure 2: Category infit statistics for A level subjects from the 2010–13 exam series.

When the infit or outfit statistic is less than 1.0 for a category, its observed CPC will be sharper than the model predicted CPC (over-fit, or more discriminative). If the fit statistics are greater than 1.0, the observed CPC will be flatter than the model predicted CPC (under-fit or less discriminative). The CPC (also see equation 1) shows how the probability of scoring a particular category varies with ability. Over-fitting suggests that there is less variability in the observed data than the Rasch model predicted, whilst under-fitting suggests greater variability in the data than the model predicted. Further examination of the fit statistics indicated that, for both GCSEs and A levels, mathematics and the science subjects generally over-fitted the Rasch model (with values of the mean squares statistics less than 1.0), whilst subjects such as music, drama, critical thinking and general studies under-fitted the model. Over-fitting subjects are more discriminative than under-fitting subjects in differentiating students in terms of ability.

Figure 3 depicts the model predicted CPCs for three GCSE and three A level subjects from the 2013 exam series: GCSE biology, with mostly over-fitting

categories (infit statistics varying from 0.52 to 0.74); GCSE art, with mostly under-fitting categories (infit statistics varying from 1.49 to 1.73); GCSE mathematics, with most of the categories fitting the Rasch model very well (infit statistics varying from 0.90 to 1.03); A level further mathematics, with mostly over-fitting categories (infit statistics varying from 0.50 to 0.71); A level general studies, with mostly under-fitting categories (infit statistics varying from 1.15 to 1.68); and A level history, with most of the categories fitting the Rasch model well (infit statistics varying from 0.85 to 1.04). The observed empirical distributions are superimposed on the model predicted CPCs.

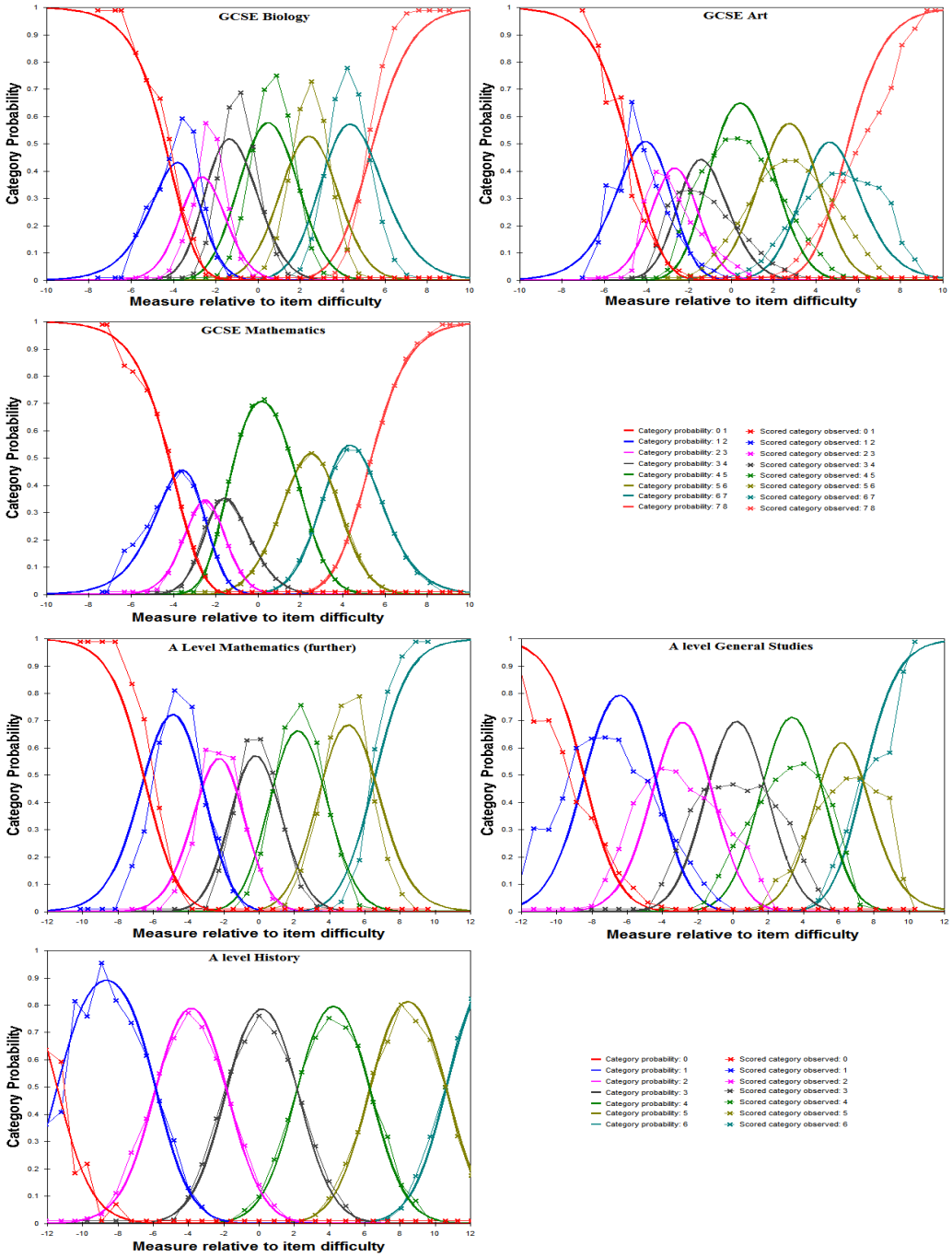

Figure 3: Model predicted and observed CPCs for three GCSE and three A level subjects from the 2013 exam series: GCSE biology and A level further mathematics (over-fitting); GCSE art and A level general studies (under-fitting); and GCSE mathematics and A level history (good model fit).

Figures 4 and 5 show the item infit statistics at the overall item level for the GCSEs and A levels from the 2013 exam series. These statistics vary from about 0.53 to 1.63 for GCSEs and from 0.69 to 1.58 for A levels, suggesting that, at the overall subject level, the data also fits the Rasch model reasonably well. The fit statistics at

the overall item level are generally consistent with category fit statistics. That is, if a subject is over-fitting/under-fitting the Rasch model at the category level, then the item will also generally over-fit/under-fit the model. GCSEs in physical education, music, short course IT, drama, and art have infit statistics considerably higher than those for other subjects, and they fit the Rasch model less well. For A levels, the values of the infit statistics for general studies and critical thinking are substantially higher than those for other subjects, suggesting that the traits assessed by these two subjects may differ considerably from those inferred from the Rasch model for all the subjects included in the analysis.

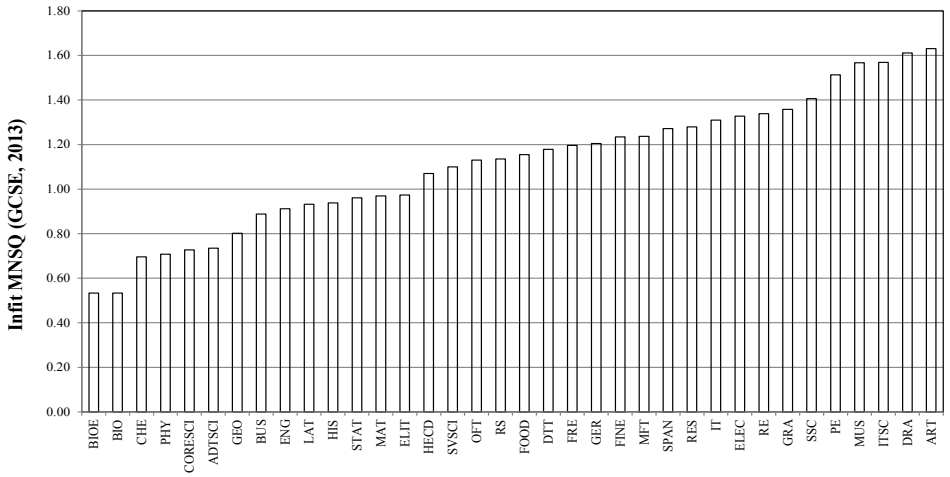

Figure 4: Item (subject) infit statistics for GCSE subjects from the 2013 exam series.

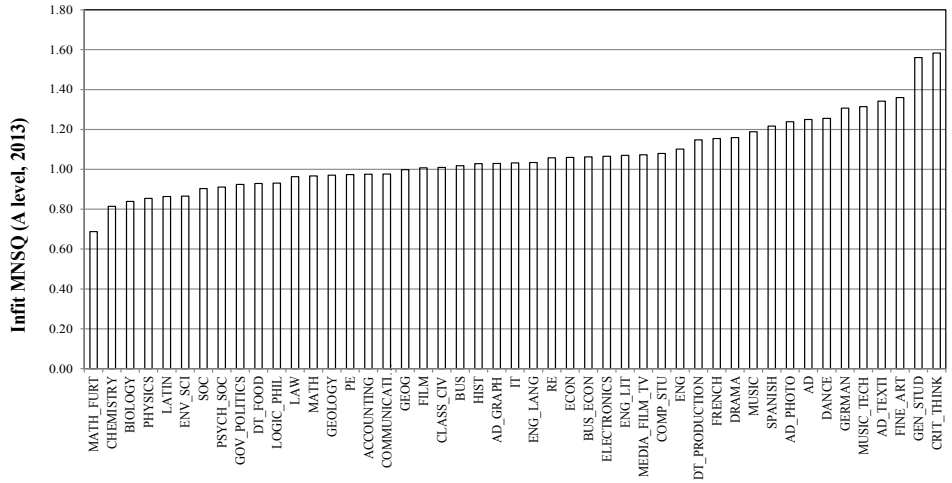

Figure 5: Item (subject) infit statistics for A level subjects from the 2013 exam series.

Analysis of variances suggested that the total variances of the datasets which could be accounted for by the Rasch model were about 78 per cent for the GCSEs and 83

per cent for the A levels (see table 1). Principal components analysis of residuals indicated that the ratio of the first contrast to the second contrast in the residuals in eigenvalue terms was about 1.4 for GCSEs and 1.1 for A levels, suggesting that these contrasts were of relatively equal importance in explaining the variance unexplained by the Rasch model, and, therefore, it may be assumed that no meaningful second dimension could be constructed for the original numerical grades. Therefore, these datasets could be essentially treated as unidimensional (see Linacre, 2013; Pae, 2012). Some of the other statistics showing how well the Rasch model functioned in establishing the measurement scale for the various datasets are also listed in table 1. These include the person separation index, defined as the ratio of the standard deviation of the person measures and its average standard error of estimation, person reliability, which is related to the separation index and is defined as one minus the square of the ratio of the person average measurement error, and the standard deviation of the person measures (similar to the definition of reliability in classical test theory (CTT)), and the average item point-measure correlation between the observations on an item and the corresponding person measures. With Rasch modelling, the reliability can be estimated even where there are missing data. This is not possible in CTT. These indicators suggested that the Rasch model functioned reasonably well overall.

Table 1: Variances explained by Rasch measures, person separation index and reliability, and average item point-measure correlation for the datasets analysed.

	GCSE subjects				**A level subjects**			
	2010	**2011**	**2012**	**2013**	**2010**	**2011**	**2012**	**2013**
Variance explained by Rasch measures (%)	78.6	78.5	78.1	77.9	83.3	83.4	83.7	83.7
Person separation index	4.79	4.62	4.55	4.64	3.33	3.34	3.39	3.37
Person reliability	0.96	0.96	0.95	0.96	0.92	0.92	0.92	0.92
Average item point-measure correlation	0.85	0.85	0.84	0.85	0.90	0.90	0.90	0.90

3.2 Relative difficulty of GCSE exams

The graph on the left in figure 6 shows the model predicted and observed ICCs for the 36 GCSE subjects from the 2013 exam series. The curves on the left of the graph are for subjects that are generally ‘easy’, and those on the right are for ‘difficult’

subjects in terms of the level of ability specified by the Rasch model which is required to achieve the same expected score (grade) in different subjects.

The spread of the ICCs also shows a slightly wider range of ability for the lower grades than for the higher grades, suggesting there is a degree of differentiated relative difficulty. If two ICCs do not cross, then the leftmost subject is easier than the rightmost across the full ability range. Most of the ICCS do not cross one another. If two ICCs do cross, then the order of difficulty changes direction at the intersection point. For example, on the right graph in figure 6, the ICC of GCSE in graphic products (design technology) crosses the ICC of GCSE in physical education at the ability of about 2.20 logits (with an expected score of about 5.3). The empirical curves are superimposed on the theoretical ICCs. On the left side of the intersection between the two ICCs, GCSE in physical education is easier than GCSE in graphic products (that is students with similar abilities will have a higher expected score on physical education than on graphic products). In contrast, on the right side of the intersection, GCSE physical education is harder than GCSE graphic products.

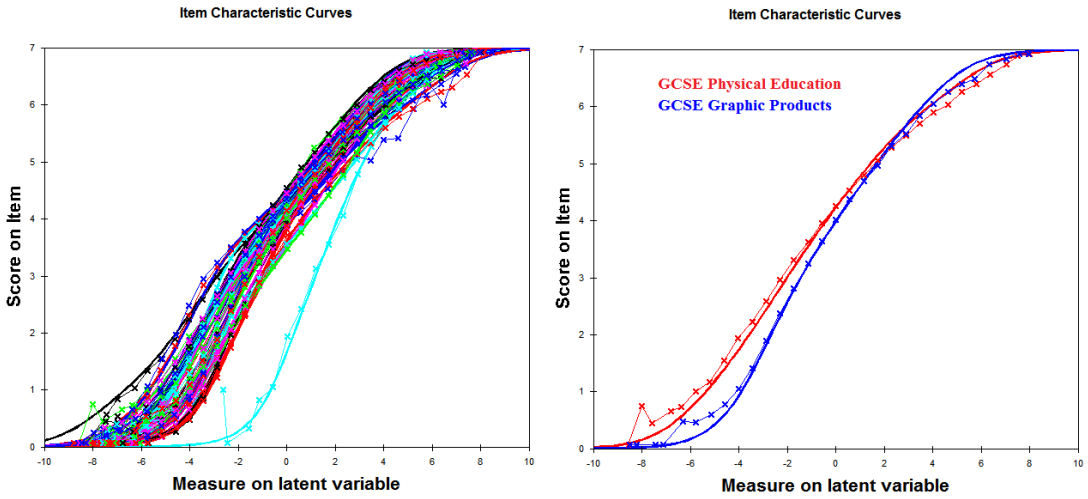

Figure 6: Comparison of the distributions of model predicted ICCs and observed ICCs for the 36 GCSE subjects from the 2013 exam series (left), and between the ICC of GCSE physical education (red line) and the ICC of GCSE graphic products (blue line) (right).

Figure 7 compares the model predicted and observed ICCs of GCSE art and GCSE French. As can be seen from the graph, French is considerably harder than art across the full range of ability. For example, a student with an ability of 1.0 logits will score 4 (grade C) in French but 5 (grade B) in art. The ICCs of the two subjects are also not parallel, suggesting that the difference in difficulty between the two subjects varies with ability.

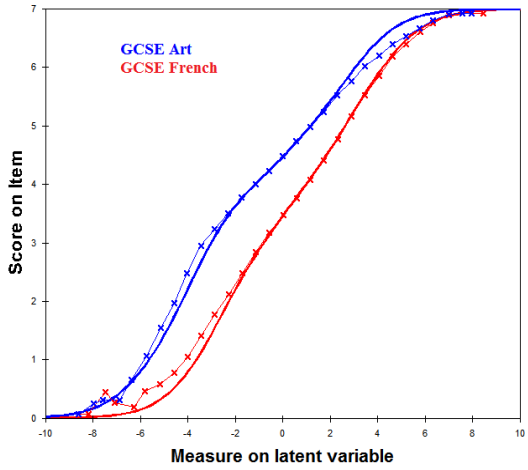

Figure 7: The model predicted and observed ICCs of GCSE art and GCSE French.

Individual grade difficulty and difficulty gap between grades

Since each score category (grade) in a subject is characterised by its own parameter and modelled using the Rasch model, it is possible to compare individual grades between the subjects. As indicated earlier, the step difficulty defined using equation 3 is used as a measure of difficulty rather than the step parameter associated with a score category.

Figure 8 compares the difficulty of individual grades between the 36 GCSE subjects from the 2013 exam series, with the subjects ordered by the mean of the category step parameter estimates, which is used as a measure of the overall difficulty of the item (subject). The subjects on the left of the graph are generally easier at all individual grades, based on their Rasch difficulty measures, whilst those on the right are harder. Art, fine art and English are among the easiest subjects, whilst Spanish, German, French and short course IT are among the hardest. Mathematics is easier than biology, chemistry and physics, which are of average difficulty. Latin is considerably harder than most of the other subjects in almost the entire ability range.

Although the distributions of grade difficulties are generally consistent with the distribution of the overall subject difficulty, there is considerable variability in difficulty between the subjects at individual grades. For example, the overall difficulty of geography is 0.80 logits, about 1.26 logits higher than the overall difficulty of English. The grade difficulty of English at A* is 5.43 logits, 0.63 logits higher than the grade difficulty of 4.80 for geography. That is, at A*, English is harder than geography in terms of the level of the Rasch ability required to achieve this grade. However, at grade E, the grade difficulty of English is -5.51 logits, whilst that of geography is -3.04. Therefore, under this definition, English is considerably easier than geography

at grade E. At the top grades (A*, A and B), the core science is harder than the separate sciences (biology, chemistry, physics), but at the lower grades it is easier. Again, Latin is considerably harder at all grades than the other subjects. These findings of the difficulty or easiness of the various subjects relative to one another are broadly similar to those reported by Coe (2008) and Coe et al. (2008).

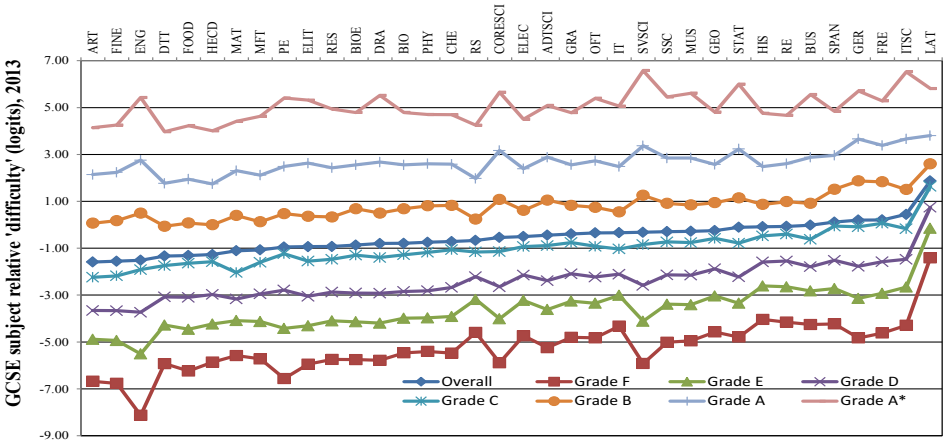

Figure 8: Comparison of the overall subject difficulty and the difficulty at individual grades for the 36 GCSE subjects from the 2013 exam series.

It is evident from figure 8 that the difficulty gaps between two adjacent grades for individual subjects vary across the range of the grades. For example, for English, the differences in difficulty between adjacent grades are 2.67 (A* to A), 2.26 (A to B), 2.41 (B to C), 1.82 (C to D), 1.78 (D to E), and 2.62 logits (E to F), respectively. These differences would suggest that different relative progress or, more specifically, a different amount of the latent trait specified by the Rasch model, would be required for progressing from one grade to another at different grades. Figure 9 illustrates the difficulty gaps between two adjacent grades for the 2013 GCSE subjects. As with the grade difficulty distributions, the difficulty gaps vary between subjects and between grades. The gap between grade D and grade E is the smallest, whilst that between A* and A is the largest.

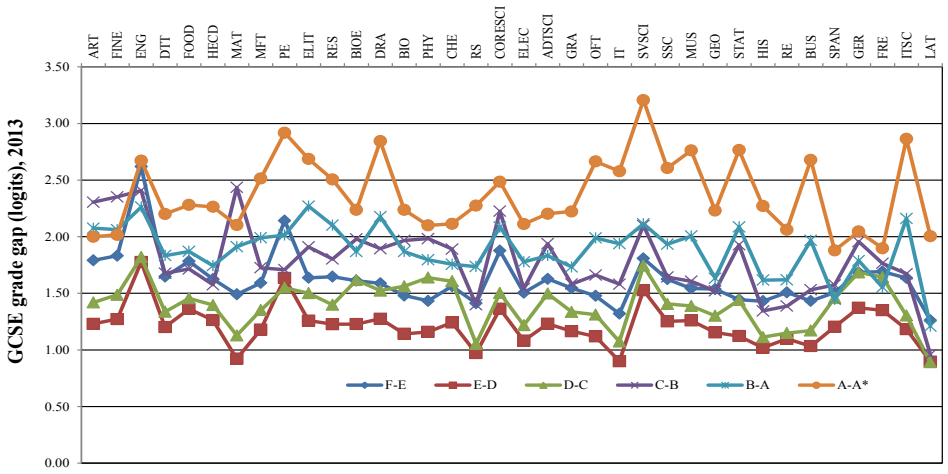

Figure 9: Differences in difficulty between two adjacent grades for the 36 GCSE subjects from the 2013 exam series.

Variation of subject and grade difficulties over time

The overall difficulties of the GCSE subjects from the 2010 to 2013 exam series are shown in figure 10 (the values are also listed in table A4 in appendix 1). In figure 10, the subjects are ordered by the overall subject difficulty for the 2010 series. It should be noted that, since the data for each exam series were analysed using the Rasch model separately, the difficulty measures are not necessarily on the same scale as a result of the use of a sub-sample rather than the whole GCSE population in the calibration. Since the ability distribution of the students used for calibration may vary between the exam series, a systematic shift in the difficulty parameter over time does not necessarily suggest a change in actual difficulty over time. Instead, the focus here is on the patterns of the distributions of average difficulty between subjects that are generally consistent over the four year period. This is not unexpected because of the strong use of statistical evidence, albeit a different methodology to that presented here, to maintain standards over time in these subjects. Subjects like art, fine art, physical education and child development (home economics) are generally among the easier subjects, whilst statistics, modern foreign languages and Latin are among the harder subjects.

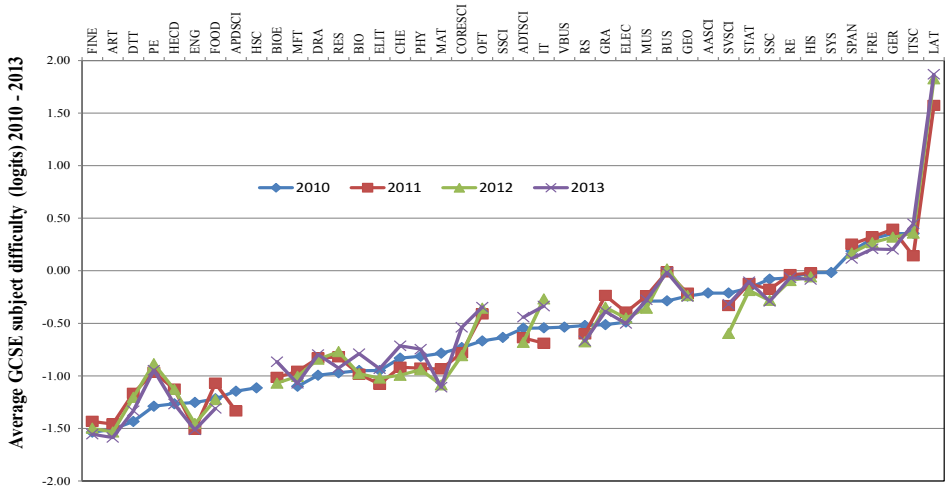

Figure 10: Variation of subject level difficulty of GCSE subjects from 2010 to 2013.

Figure 11 shows how the grade difficulty of GCSE subjects at grades A and C (which are judgemental grades) and E (which is a non-judgemental grade) varies from 2010 to 2013 (the subjects are ordered based on the overall subject difficulty for 2010. See table A5 in appendix 1 for the values.) In general, the grade difficulties remain relatively stable over the four year period of study.

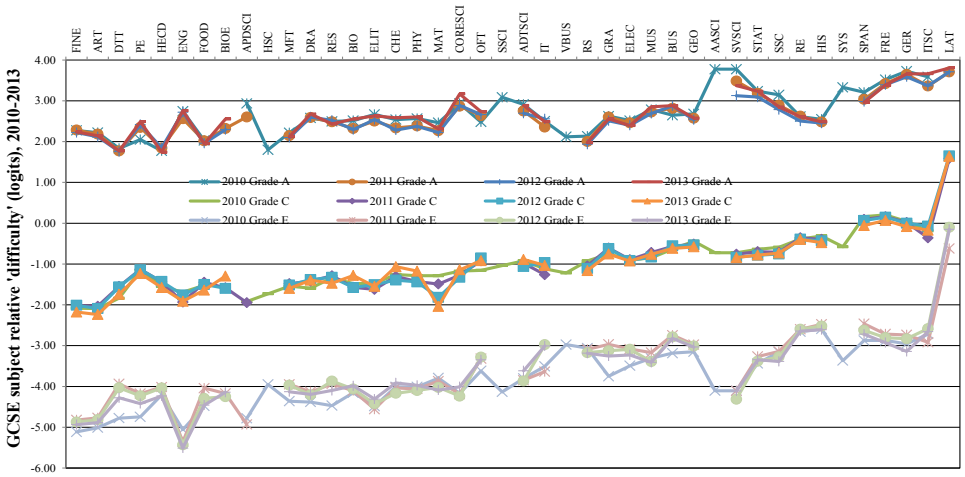

Figure 11: Variation of relative grade difficulties of GCSE exams at grades A, C and E from 2010 to 2013**.**

3.3 Relative difficulty of A level exams

The graph on the left in figure 12 shows the model predicted and observed ICCs for the 47 A level subjects from the 2013 exam series. Similar to the GCSEs, the ICCs spread more widely for the lower grades than for the higher grades. Again, the

comparison discussed here is based on the latent ability specified in the Rasch model.

The graph on the right of figure 12 compares the model predicted and observed ICCs of graphics (art and design) and physics. The ICCs are relatively parallel, and physics is considerably harder than graphics across the full ability range. Students with a score of 5 (grade A) in graphics will possess the same level of ability as those who achieve a score of 3 (grade C) in physics. Similarly, the level of ability for achieving a grade E in physics is the same as that for achieving a grade C in graphics. It has to be noted that a direct comparison between the A level ICCs depicted in figure 12 and the ICCs for the GCSEs shown in figure 6 is inappropriate as the graphs are on different measurement scales. Furthermore, the meaning of the trait for GCSEs may be different from that for A levels.

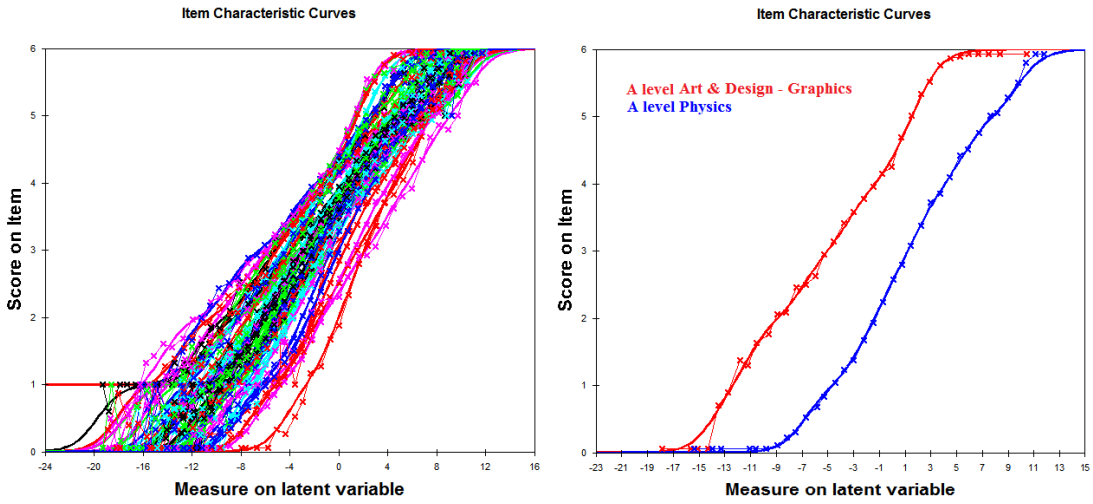

Figure 12: Distributions of model predicted ICCs and observed ICCs for the 47 A level subjects from the 2013 exam series (left), and between the ICC of graphics (red line) and the ICC of physics (blue line) (right).

Individual grade difficulty and difficulty gap between grades

Figure 13 compares the overall difficulty and the difficulty of individual grades between the 47 A level subjects from the 2013 exam series, with the subjects ordered by the overall subject difficulty. Communication studies, film studies, graphics, photography (art and design), media studies, and textiles (art and design) are among the easiest subjects, whilst science, technology, engineering and mathematics (STEM) subjects, modern foreign languages, general studies and critical thinking are among the hardest subjects. Further mathematics and Latin are the hardest subjects. Similar to the GCSEs, the grade difficulties exhibit greater variability between the subjects than the subject level overall difficulties. Again, these

findings are broadly similar to those reported by Coe et al. (2008). As indicated above, figure 13 cannot be directly compared with figure 8.

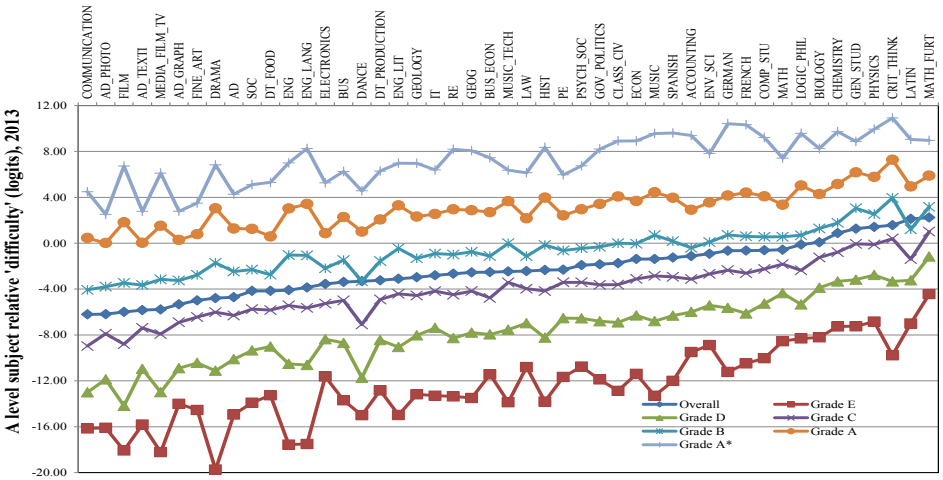

Figure 13: Comparison of the overall subject difficulty and the difficulty at individual grades for the 47 A level subjects from the 2013 exam series.

Figure 14 shows the distribution of grade gaps in logits for the A level subjects from the 2013 exam series. The gap between A* and A is among the smallest for the easier subjects but similar to the gap between D and E for the harder subjects. This is different from the GCSEs, where the gap between A* and A is the largest. For the mid-difficulty subjects, the gaps between B and C and C and D are the smallest. The gap between D and E is the largest for almost all subjects. The grade gaps also become smaller for harder subjects. For example, further mathematics, which is one of the hardest subjects, has an average grade gap of 3.0 logits, whilst film studies, one of the easiest subjects, has an average grade gap of 4.0 logits.

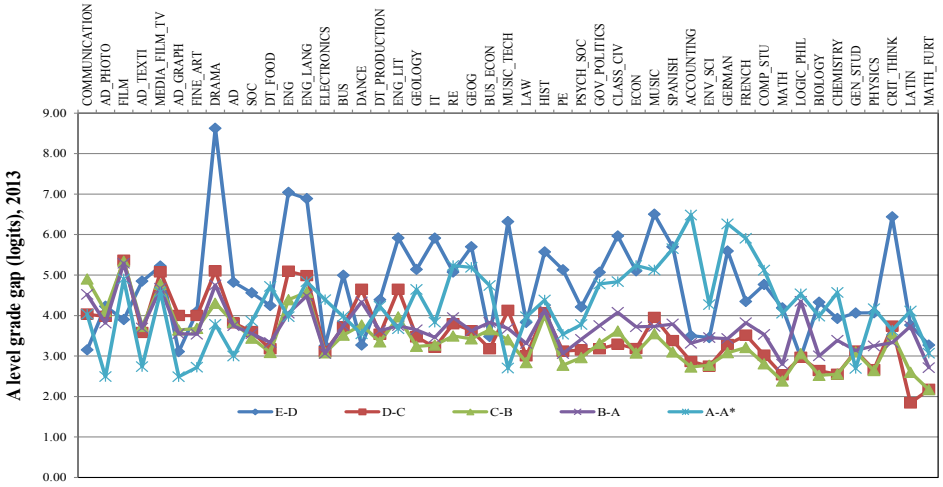

Figure 14: Differences in difficulty between two adjacent grades for the 47 A level subjects from the 2013 exam series.

Variation of subject and grade difficulties over time

Figure 15 shows the distribution of the overall difficulty for the A level subjects from 2010 to 2013, with the subjects ordered by the average subject difficulty for the 2010 subjects (see table A6 in appendix 1 for the values). Again, as with GCSEs, a systematic shift in difficulty over time for the subjects does not imply a change in difficulty of the subjects over time. As can be seen from figure 15, the relative difficulties between the subjects are relatively stable from 2010 to 2013. Again, this is likely to be a reflection of the use of comparable outcomes in awarding A levels. Subjects such as film studies, media studies, and graphics, among others, are consistently easy, whilst subjects like mathematics, the sciences, and modern foreign languages are consistently hard.

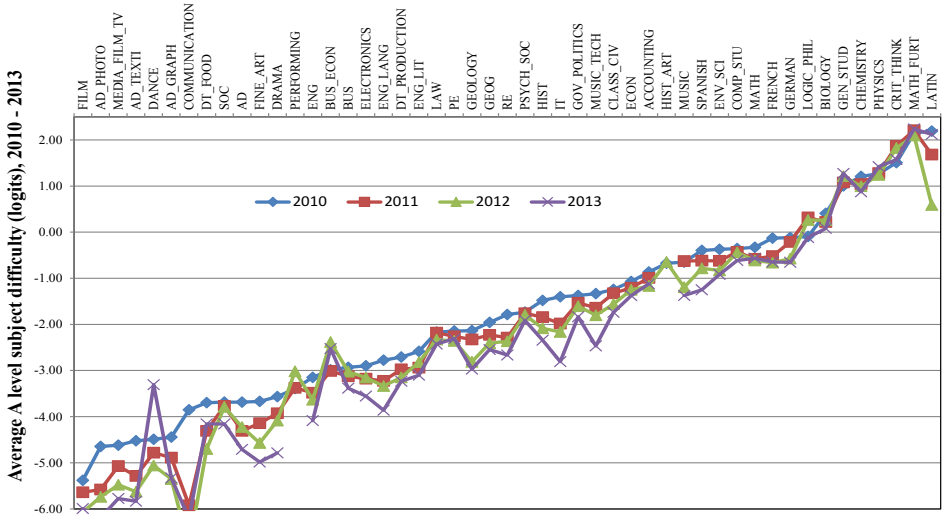

Figure 15: Variation of subject level difficulty of A level subjects from 2010 to 2013.

Figure 16 shows how the grade difficulty of A level subjects at grades A (a judgemental grade) and D (a non-judgemental grade) varies from 2010 to 2013 (the subjects are ordered based on the overall subject difficulty for 2010. Values of the grade difficulties are also listed in table A7 in appendix 1.) In general, the grade difficulties remain relatively stable over the four year period of study.

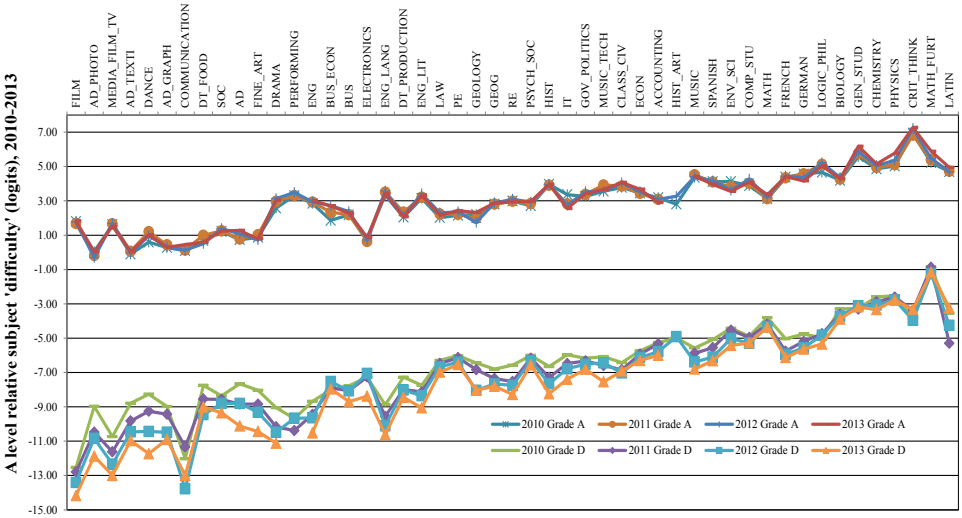

Figure 16: Variation of relative grade difficulties of A level exams at grades A and D from 2010 to 2013.

4. Relative subject difficulty based on regression analysis using prior attainment and concurrent performance measures

This section presents results from the regression analyses using prior attainment and concurrent performance measures for the GCSE and A level datasets from the 2013 exam series.

4.1 Results from the simple linear regression analysis

GCSE subjects

Figure 17 shows segments of the regression lines of GCSE grades with respect to the mean normalised Key Stage 2 test score and the mean numerical GCSE grade for the 36 subjects from the 2013 exam series. For each of the subjects, the line is centred on the mean of the numerical outcome grades and extends one standard deviation both sides of the mean. Values of coefficient of determination (R^2) vary from 0.21 for Latin to 0.59 for mathematics, when regressing on the mean normalised Key Stage 2 score. For regression on the mean numerical GCSE grade, these values vary from 0.55 for art to 0.82 for biology, for which the value is considerably higher. This is expected as the Key Stage 2 tests were taken five years previously and will have less predictive power than the mean GCSE grades, which will be a better measure of the current ability of the student. The regression lines based on the Key Stage 2 score spread wider vertically and are also not as close to parallel as the regression lines based on the mean GCSE grade.

For each set of the regression lines, considerable variability exists in the values of the slope and intercept parameters. The relative positions of the regression lines reflect the relative 'difficulty' of the subjects. Subjects with lower intercept values may be interpreted as 'harder' subjects, whilst those with higher values are 'easier' subjects. This is because for similar average Key Stage 2 scores or mean numerical GCSE grades, subjects with higher intercept values will generally have better grade outcomes than subjects with lower values. However, since the slope of the regression lines varies between the subjects, the relative subject difficulty varies at different grades. When two lines cross, the order of the relative difficulty changes direction at the intersection point. Furthermore, the relative positions for some of the regression lines based on concurrent GCSE performance measure are considerably different from those based on the mean normalised Key Stage 2 score. For example, GCSE biology would be 'easier' than GCSE English if the comparison was based on the mean normalised Key Stage 2 score. However, it would be 'harder' than English if the comparison was based on the mean GCSE grade. Latin is considerably 'easier' than most of the subjects based on the mean normalised Key Stage 2 score, but it is the 'hardest' subject based on the mean GCSE grade. A large proportion of the

subjects are judged to be either hard or easy consistently by both the mean normalised Key Stage 2 score and the mean GCSE grade.

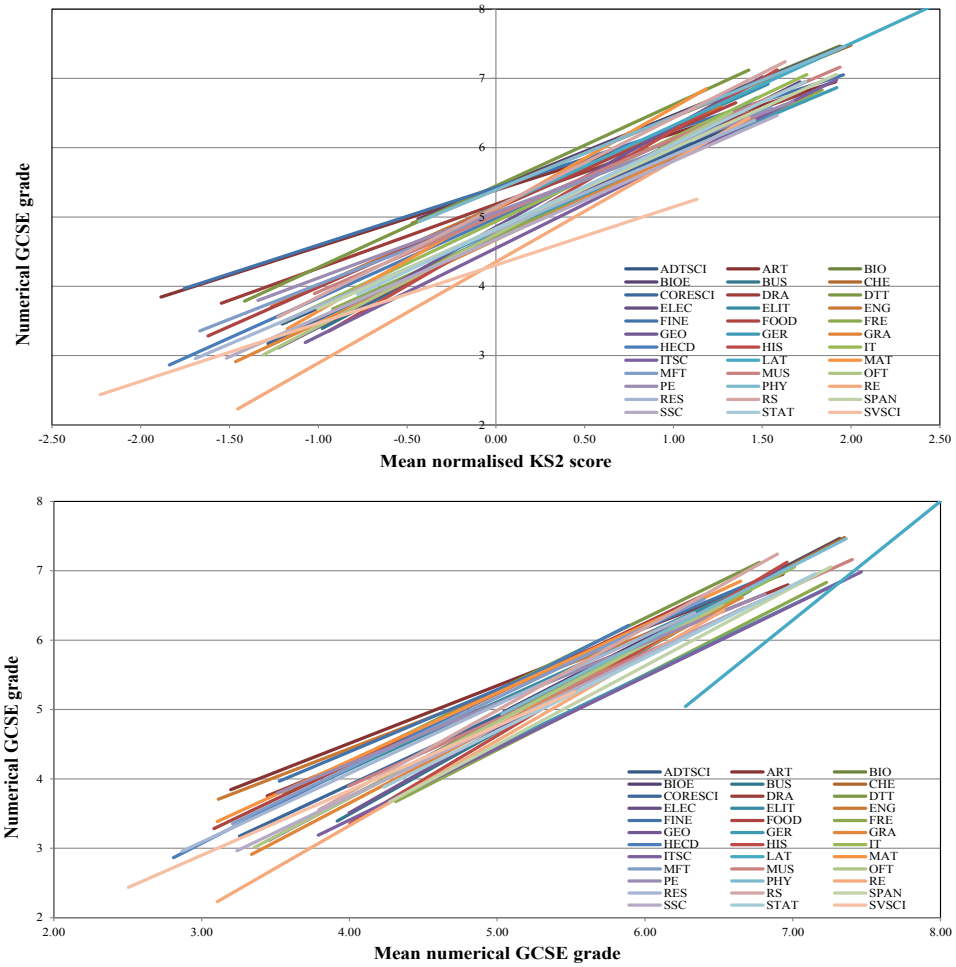

Figure 17: Regression lines of GCSE grades on the mean normalised Key Stage 2 score (top) and mean numerical GCSE grade (bottom) for the 36 GCSE subjects from the 2013 exam series**.**

As indicated earlier, it is possible to define the relative subject difficulty at individual grades for each subject based on the linear regression approach. If a horizontal line at each numerical GCSE grade is drawn on the *y*-axis, it will cross all the regression lines. The difficulty of a subject at that grade can be defined as the corresponding mean normalised Key Stage 2 score or mean numerical GCSE grade on the *x*-axis. The relative difficulty of the subject at that grade can be defined as the difference between its difficulty and the mean of the difficulties of all the subjects. An overall difficulty of the subject may also be defined as the mean of the difficulties at individual grades. In the case of using Key Stage 2 scores as a basis for comparison, the unit of the difficulty is one standard deviation of the mean normalised Key Stage 2 scores (which is close to 1.0). In the case of using the mean GCSE grade, the unit will be one grade. Figure 18 shows the overall relative difficulty of the 36 GCSE

subjects from the 2013 exam series. Subjects such as fine art, English, biology and religious studies are 'easier' subjects when judged by both mean normalised Key Stage 2 scores and mean GCSE grades, whilst subjects like short course IT, graphic products, Spanish, German and statistics are 'harder' subjects. There are also subjects such as Latin, mathematics, chemistry and others that are judged to be easier by one performance measure but harder by the other.

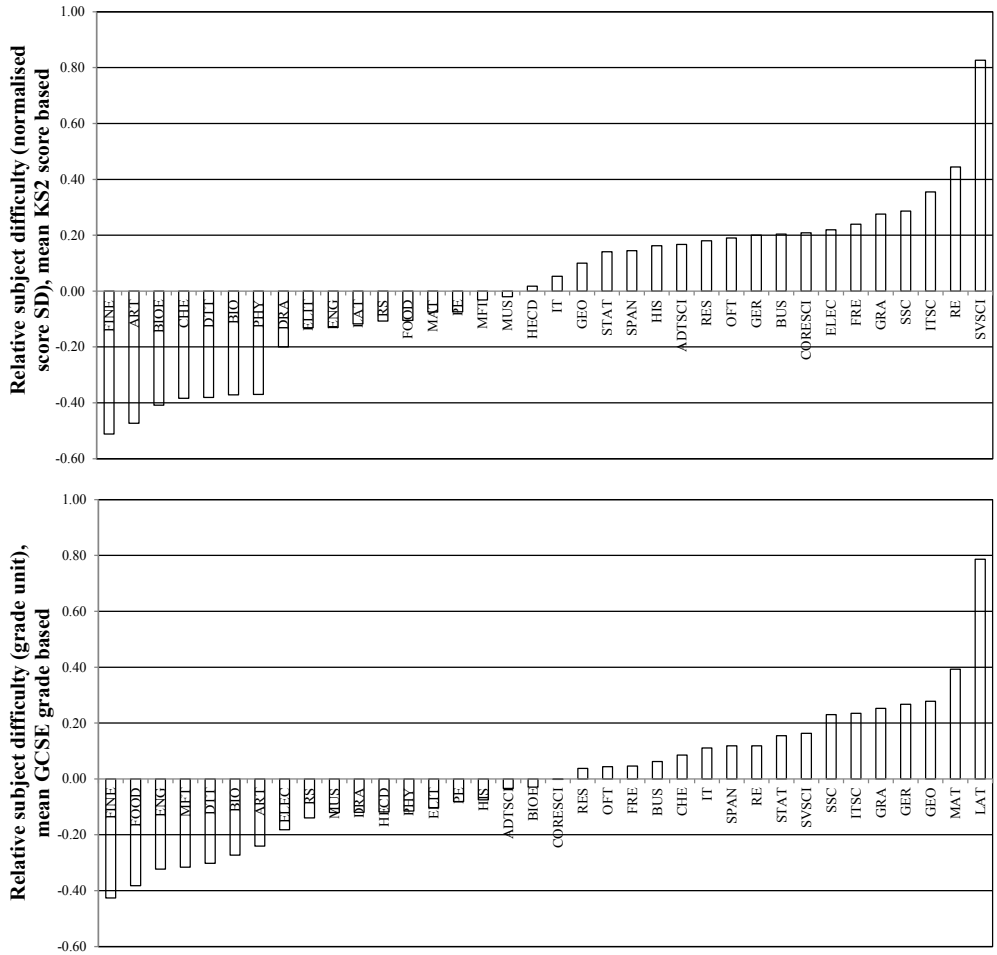

Figure 18: The overall relative difficulty of the GCSE subjects from the 2013 exam series based on the mean normalised Key Stage 2 score (top) and mean numerical GCSE grade (bottom).

Figure 19 compares the difficulties at individual grades and the overall difficulty for the 36 GCSE subjects, derived using the two performance measures with the difficulties derived using the Rasch model discussed in section 3 (grade G was excluded). These difficulties are generally positively correlated for all grades, although a number of subjects, including Latin and English, were judged considerably differently by the different methods. Furthermore, the consistency in difficulty between the subjects is higher for the mid-grades than for the top (A*) or bottom (F) grades. This may suggest a higher level of inconsistency in setting the performance

standards at A* and F for the subjects. The difficulties derived using the mean GCSE grade correlate better with the Rasch difficulties than the difficulties derived using the mean normalised Key Stage 2 test score, indicating that the order of difficulty of the subjects derived using the mean GCSE grade is more consistent with that derived using the Rasch model than the difficulty order derived using the mean normalised Key Stage 2 score.

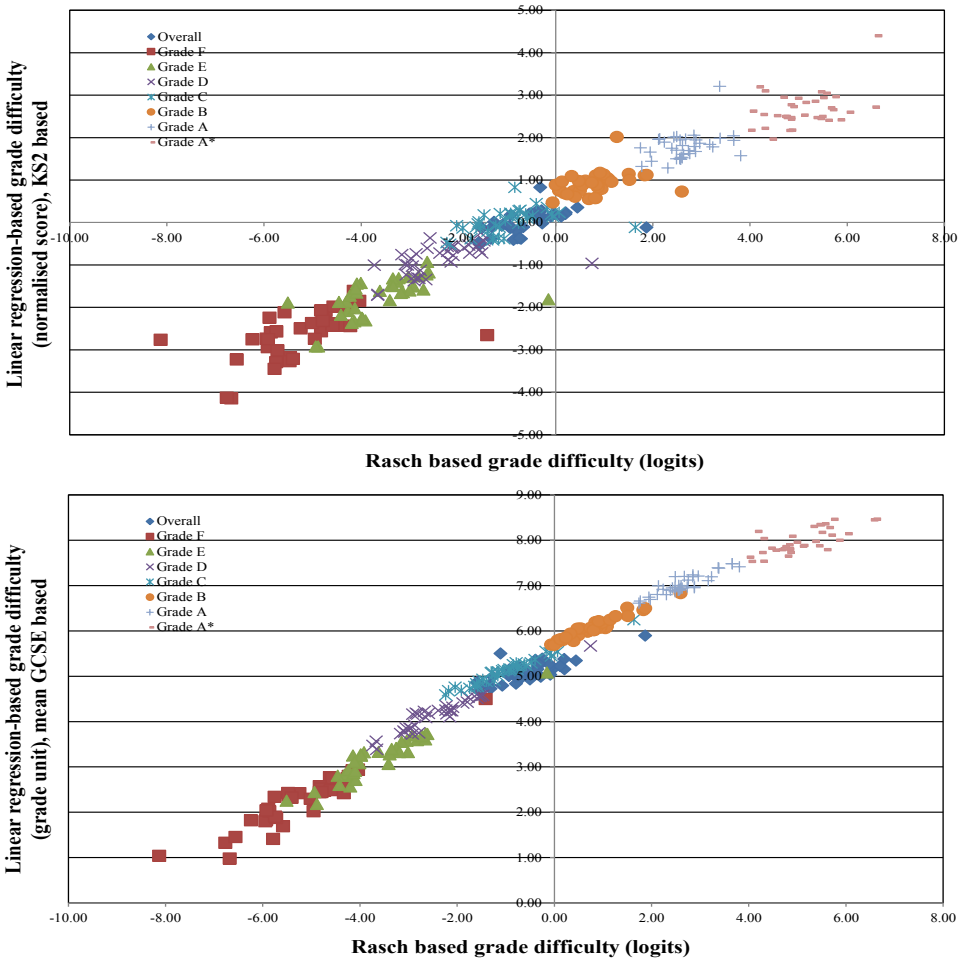

Figure 19: Comparison of subject difficulty at individual grades and the overall difficulty derived using the mean normalised Key Stage 2 test scores (top) and the mean GCSE grades (bottom) with the Rasch-model-derived difficulties for the GCSE subjects from the 2013 exam series.

A level subjects

Figure 20 shows the regression lines of the numerical A level outcome grades on the mean numerical GCSE grade and the mean numerical A level grade for the 47 A level subjects from the 2013 exam series. As in the case of the GCSEs, each of the regression lines is centred on the mean of the A level grades of the subject and extends one standard deviation both sides of the mean. Values of R^2 vary from 0.21 for photography to 0.52 for English literature, when regressing on the mean GCSE

grade. For regression on the average numerical A level grade, values of R^2 vary from 0.53 for critical thinking to 0.84 for physics. The regression lines with respect to the mean GCSE grade spread considerably wider and are not as close to parallel as those based on the mean A level grade. There is considerable variability in the values of the slope and intercept parameters between the regression lines. Similar to the case of the GCSE subjects discussed above, subjects with lower intercept values may be interpreted as 'harder' than subjects with higher intercept values. For example, when comparison between the subjects is based on the mean GCSE grade, biology, chemistry, physics, Latin, and sports/physical education studies are harder subjects, whilst photography, communications, film, dance and Spanish are easier subjects. When comparison is based on the mean A level grade, business studies and economics, graphics, psychology, further mathematics, accounting and English literature are harder subjects, whereas general studies, German, photography and history are easier subjects.

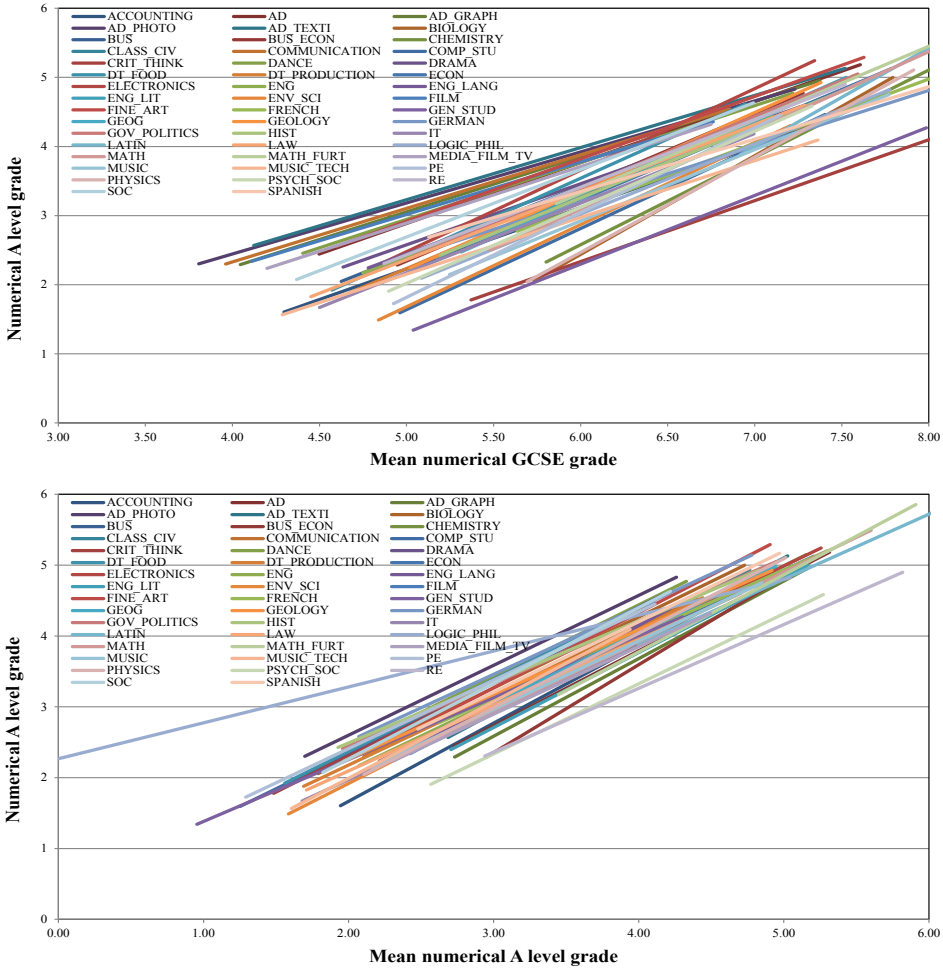

Figure 20: Regression lines of A level grades on the mean GCSE grade (top) and mean A level grade (bottom) for the 47 A level subjects from the 2013 exam series**.**

The relative positions of the regression lines based on the mean GCSE grade are substantially different from those based on the mean A level grade, suggesting that

the order of the subject difficulty is affected by the measures on which the comparison between the subjects is based. Whilst a large number of the subjects are judged to be either easy or hard by both the mean GCSE grade and the mean A level grade, a considerable number of the subjects are also judged to be easy by one measure but hard by the other measure, or vice versa. For example, photography, communications, dance, and drama are easy subjects when judged by both the mean GCSE grade and the mean A level grade, and chemistry, physics, mathematics, and English literature are hard subjects. Musical technology, film, and graphics are hard subjects if judged by the mean A level grade, but easy subjects if judged by the mean GCSE grade. The relative difficulty at each grade, and overall for the A level subjects, can be defined using the procedure described above. Figure 21 shows the overall relative subject difficulty of the 47 A level subjects.

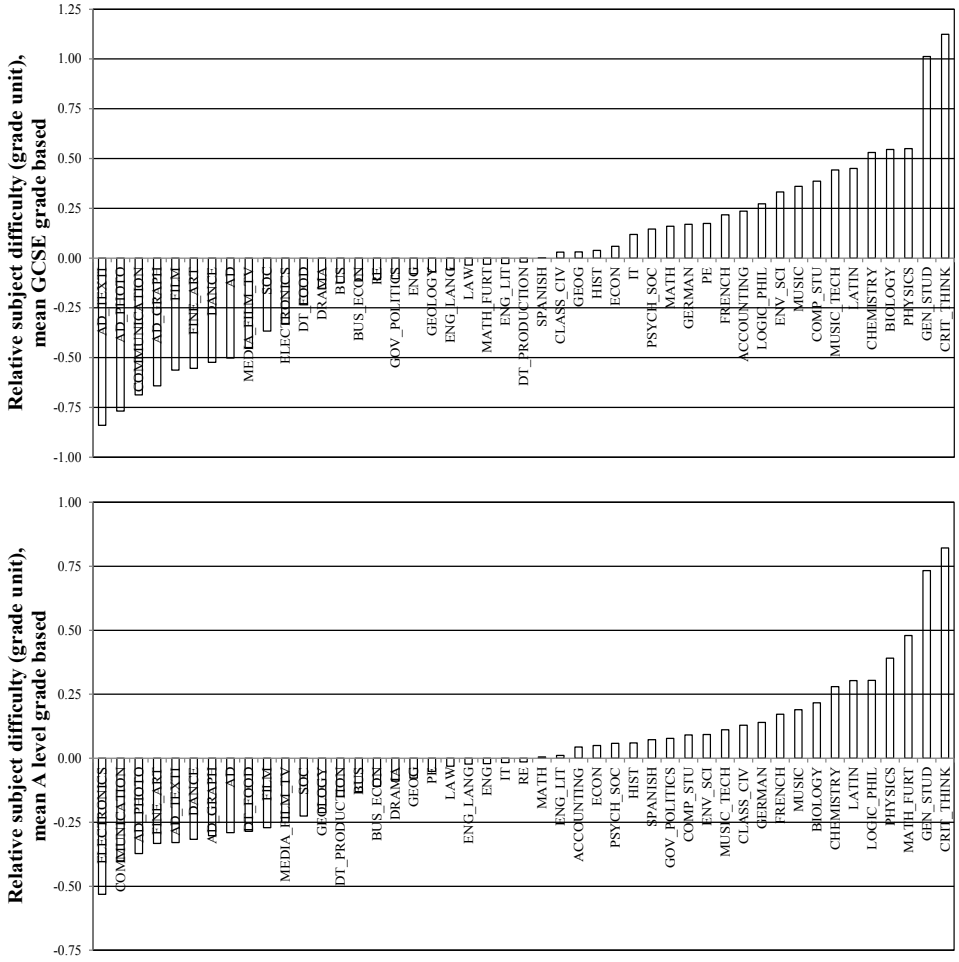

Figure 21: The overall relative difficulty of the A level subjects from the 2013 exam series based on the mean numerical GCSE grade (top) and the mean numerical A level grade (bottom).

Figure 22 compares the overall subject difficulty and the difficulties at individual grades, derived using the mean GCSE grade and the mean A level grade with the

difficulties derived using the Rasch model for the 47 A level subjects. These difficulties are generally positively correlated. Again, the consistency in difficulty between the subjects is higher for the mid-grades than for the top (A*) and bottom (E) grades, particularly when comparison is based on the mean GCSE grade. The difficulties derived using the mean A level grade correlate better with the Rasch model difficulties than the difficulties derived using the mean GCSE grade.

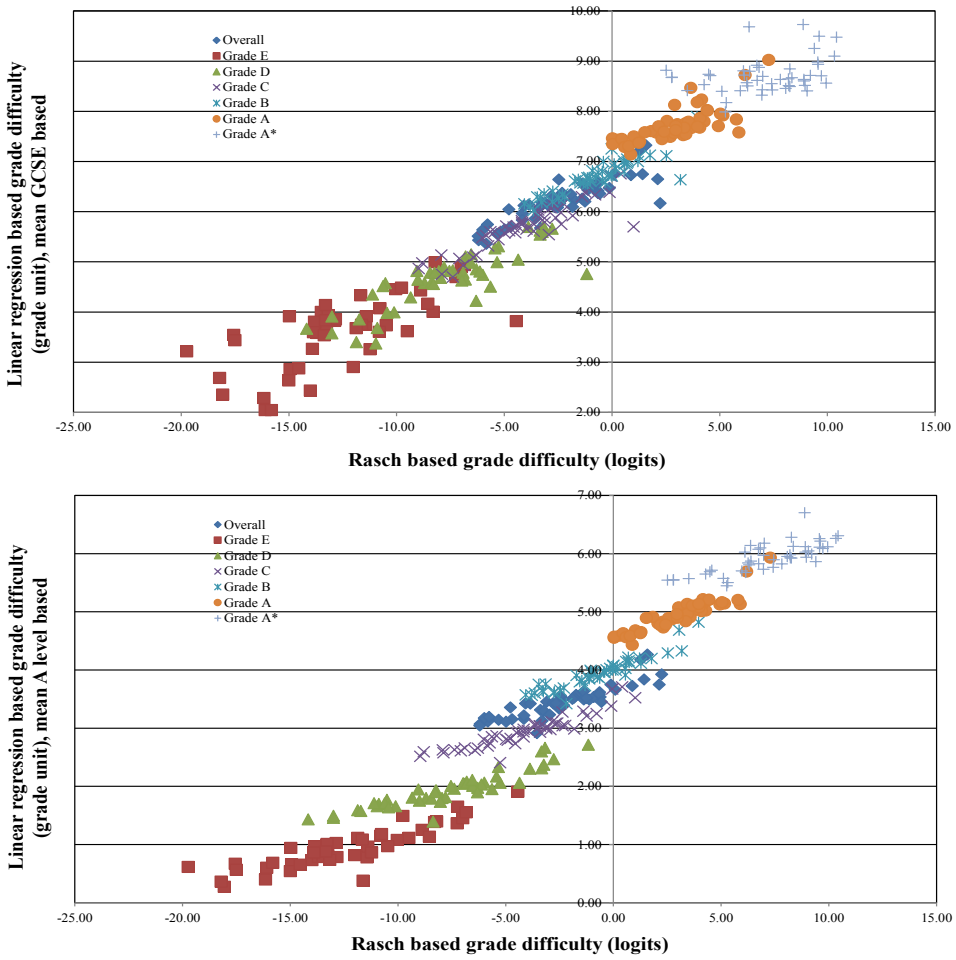

Figure 22: Comparison of subject difficulty at individual grades and the overall difficulty derived using the mean GCSE grade (top) and the mean A level grade (bottom) with the Rasch-model-derived difficulties for the A level subjects from the 2013 exam series.

4.2 Results from multinomial logistic regression analysis

GCSE subjects

With multinomial logistic regression, the relationship between a category relative to the reference category and the prior attainment or concurrent performance measure is modelled separately. Once the model parameters are estimated, the probability of a student with a fixed measure of prior attainment or concurrent performance being classified into each GCSE grade category can be calculated. In order to compare the

results from multinomial logistic regression analysis with those from the Rasch analysis, grade U was excluded from the analysis, and grade G was used as the reference category. As an example, figure 23 depicts the category probability distributions against the mean normalised Key Stage 2 score and the mean numerical GCSE grade for GCSE art and French from the 2013 exam series. The probability of being classified into a specific category varies with the mean normalised Key Stage 2 score and the mean GCSE grade for both subjects.

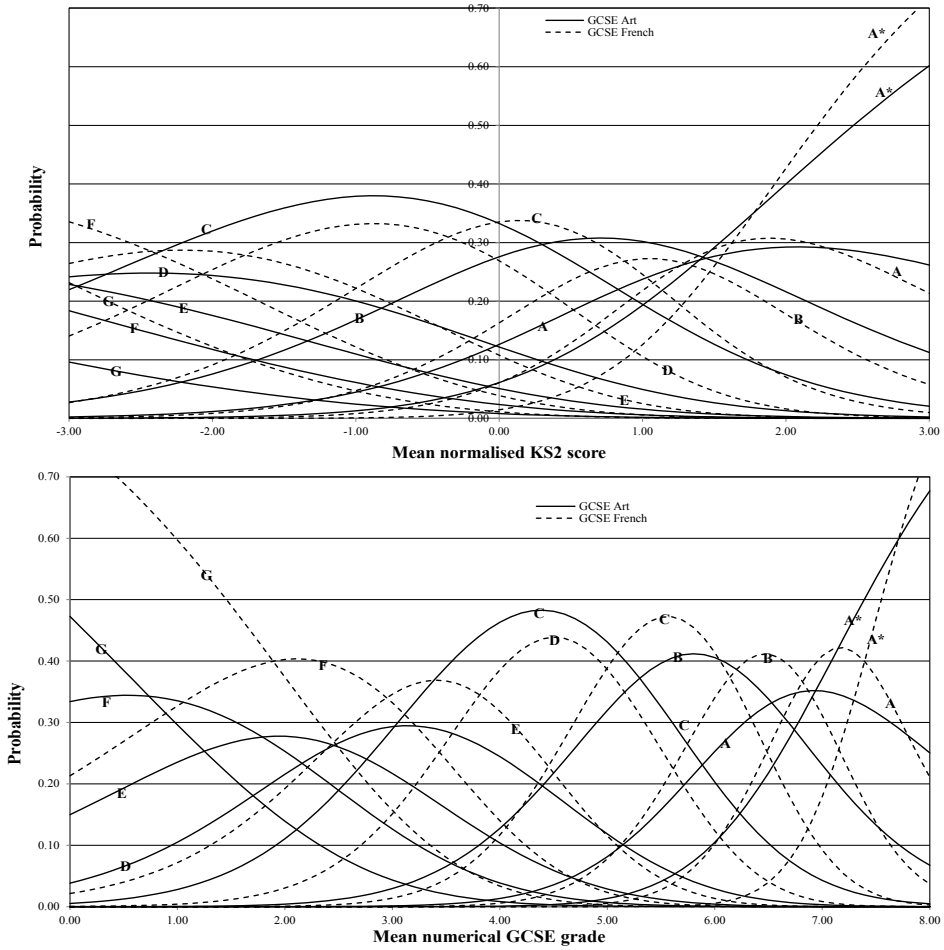

Figure 23: Category probability distributions against mean normalised Key Stage 2 test score (top) and mean numerical GCSE grade (bottom) for GCSE art and French from the 2013 exam series.

If the category probability curves for two subjects are compared, the subject with the curves on the left will have better GCSE outcomes than the other subject, if the subjects have a similar prior attainment or concurrent performance distribution. That is, the subject on the left will be 'easier' than the subject on the right when the prior attainment or concurrent performance measure is used as a basis for comparison. For example, for both graphs in figure 23, the probability curve for grade C of GCSE art is to the left of that of GCSE French, and French can be said to be 'harder' than

art at this grade. With a mean normalised Key Stage 2 score of -1.0, the probability of the student being awarded a grade C in art is 0.38, whilst the probability of being awarded a grade C in French is 0.22. Similarly, with a mean GCSE score of 4.50, the probability of being awarded a grade C in art is 0.48, whereas the probability of being awarded a grade C in French is 0.24.

Figures 24 and 25 compare the model predicted category probability distributions and the observed proportions of students being classified into the corresponding categories against the mean Key Stage 2 score and the mean GCSE grade for art and French, respectively. The model appears to fit the mid-grades better than the top or bottom grade. It is noted that when comparison is based on the mean normalised Key Stage 2 score, at grade F (category 1), the regression coefficient is negative for biological science, English, English literature, food, and additional science, which suggests that, for these subjects, with the increase in mean Key Stage 2 score the probability of being awarded a grade F decreases, whilst the probability of being awarded a grade G (category 0 or the reference category) increases. This may be an indication of a high level of unreliability associated with low Key Stage 2 test scores and inconsistency in setting the performance standards of the bottom grades.

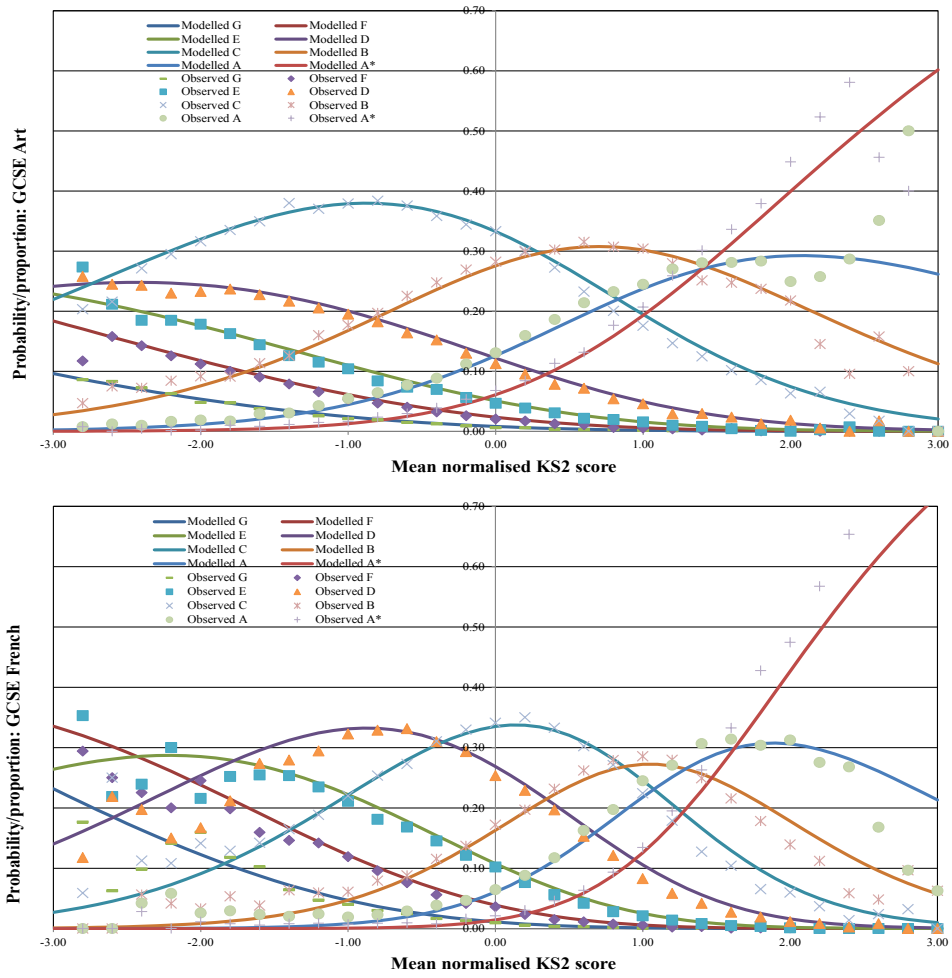

Figure 24: Model predicted category probability distributions and observed proportions of students being classified into the corresponding categories against mean normalised Key Stage 2 test score for GCSE art (top) and GCSE French (bottom) from the 2013 exam series.

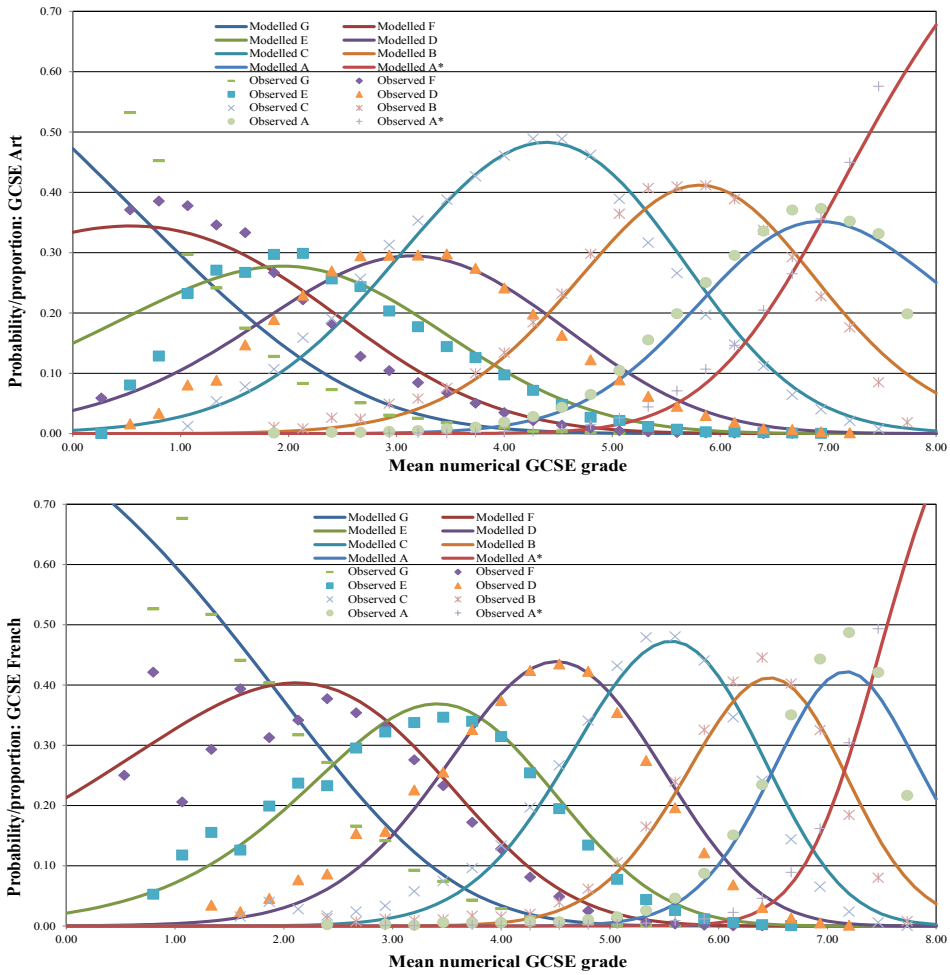

Figure 25: Model predicted category probability distributions and observed proportions of students being classified into the corresponding categories against mean numerical GCSE grade for GCSE art (top) and GCSE French (bottom) from the 2013 exam series.

As demonstrated earlier, mathematically, the relationship between category probabilities and the prior attainment or concurrent performance measure represented using logistic regression is similar to the relationship between category probabilities, category difficulty and person ability represented using the Rasch model. A 'difficulty' parameter that is linked to the regression coefficient and the intercept for each grade for a subject can be defined through equation 9. For the 36 GCSE subjects, figure 26 compares the grade difficulties and the overall difficulty derived using logistic regression on the mean normalised Key Stage 2 test score and the grade difficulties derived using the Rasch model. The difficulties are generally positively correlated. However, the strength of the correlation is only moderate, which may reflect the difference in the nature of the two performance measures. Whilst the mean normalised Key Stage 2 test score represents an ability measure from five

years ago, the Rasch ability measure represents the current ability, which is estimated based on the GCSE subjects that the student has taken in the same year.

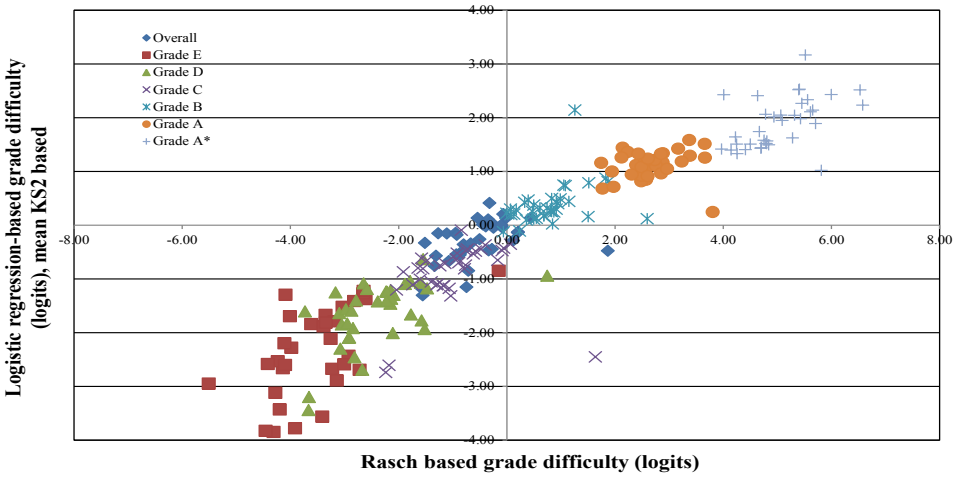

Figure 26: Comparison of the overall difficulty and the difficulty at individual grades based on logistic regression analysis on the mean normalised Key Stage 2 score and the Rasch-model-derived difficulties for the 36 GCSE subjects from the 2013 exam series.

The top graph in figure 27 shows the distribution of the overall difficulty and the difficulty at individual grades derived using the mean GCSE grade for the 36 GCSE subjects, with the subjects ordered by the overall difficulty. The bottom graph in figure 27 compares the difficulties derived using logistic regression with the difficulties derived using the Rasch model. The logistic-regression-derived difficulties and the Rasch-model-derived difficulties are highly correlated, indicating a high level of consistency in classifying the students by the two performance measures. Art, fine art, food, and English are judged to be easy subjects by both approaches, whilst French, German, Latin and statistics are hard subjects. Again, the high correlation between the difficulties derived using logistic regression and Rasch modelling indicates that the mean GCSE grade represents a performance measure similar to the Rasch person ability measure, which in turn is estimated based on all the GCSE subjects taken by the student.

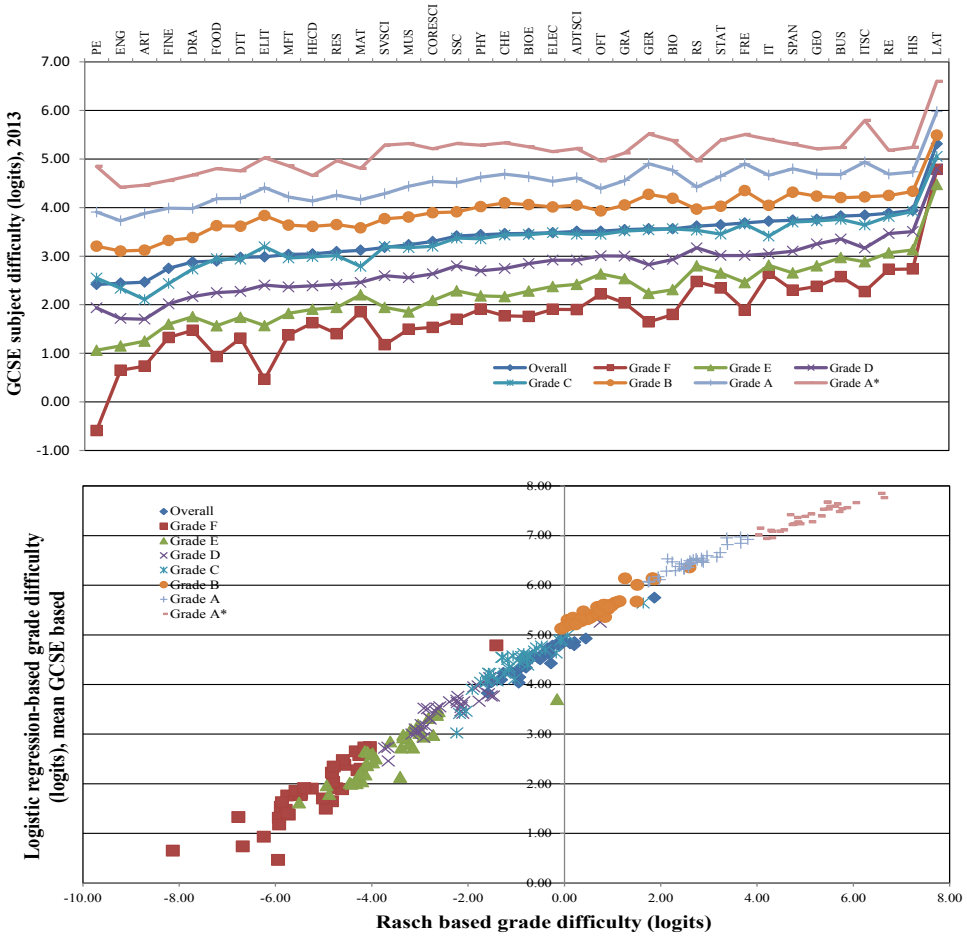

Figure 27: Comparison of the overall difficulty and the difficulty at individual grades for the 36 GCSE subjects from the 2013 exam series, based on logistic regression on the mean numerical GCSE grade (top), and the relationship between logistic-regression-derived subject difficulty and Rasch-model-derived difficulty (bottom).

A level subjects

For the 47 A level subjects, grade U was used as a reference category. Similar to the case of the GCSEs, the relative positions of the category probability curves between the subjects indicate their relative difficulties in terms of the level of prior attainment or concurrent performance required to achieve the same grade. Figure 28 depicts the category probability distributions against the mean numerical GCSE grade and the mean numerical A level grade for A levels in graphics and physics from the 2013 exam series. The probability of being classified into a specific category varies with the mean GCSE grade and the mean A level grade for both subjects. Physics is harder than graphics at all grades when judged by both the mean GCSE grade and the mean A level grade, as its category probability curves are to the right of those for graphics. Figures 29 and 30 compare the distributions of the observed proportions of students being awarded the individual grades, and the corresponding model

predicted category probabilities for the two subjects. Again, the model fits the data better in the mid-grades than at the top or bottom grade.

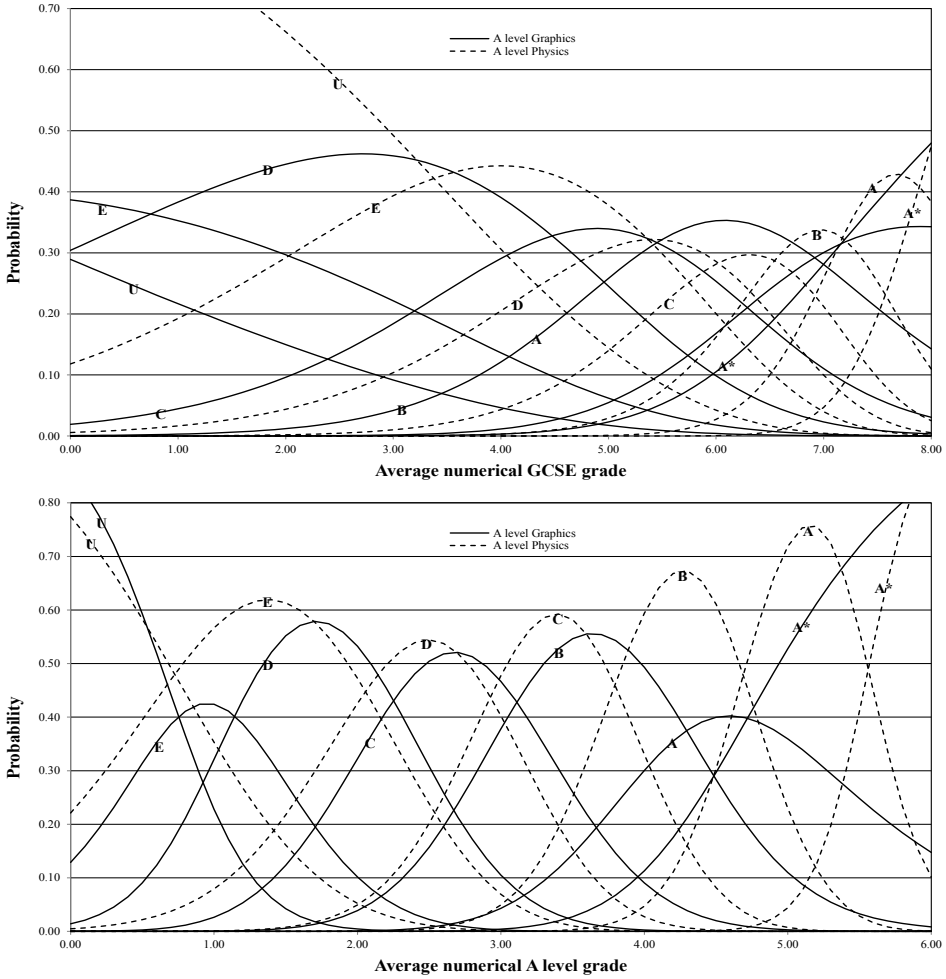

Figure 28: Category probability distributions against the mean numerical GCSE grade (top) and the mean numerical A level grade (bottom) for A level graphics and A level physics from the 2013 exam series.

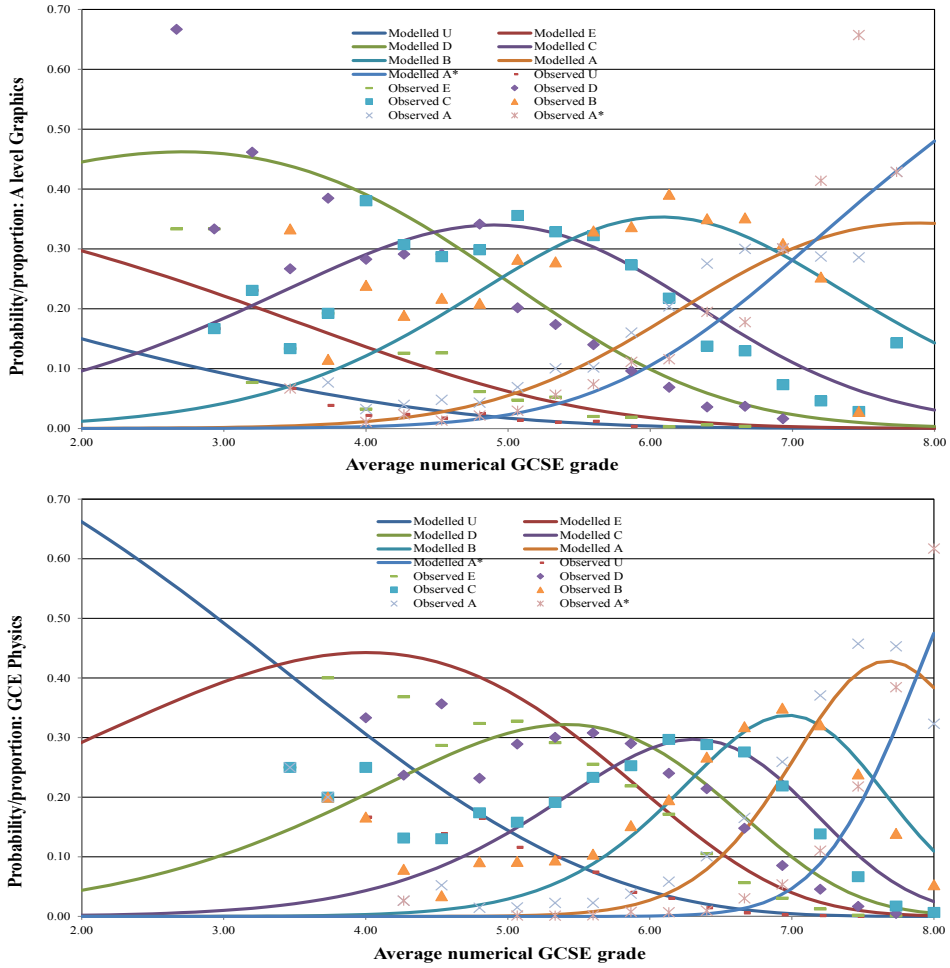

Figure 29: Model predicted category probability distributions and observed proportions of students being classified into the corresponding categories against the mean numerical GCSE grade for A level graphics (top) and A level physics (bottom) from the 2013 exam series.

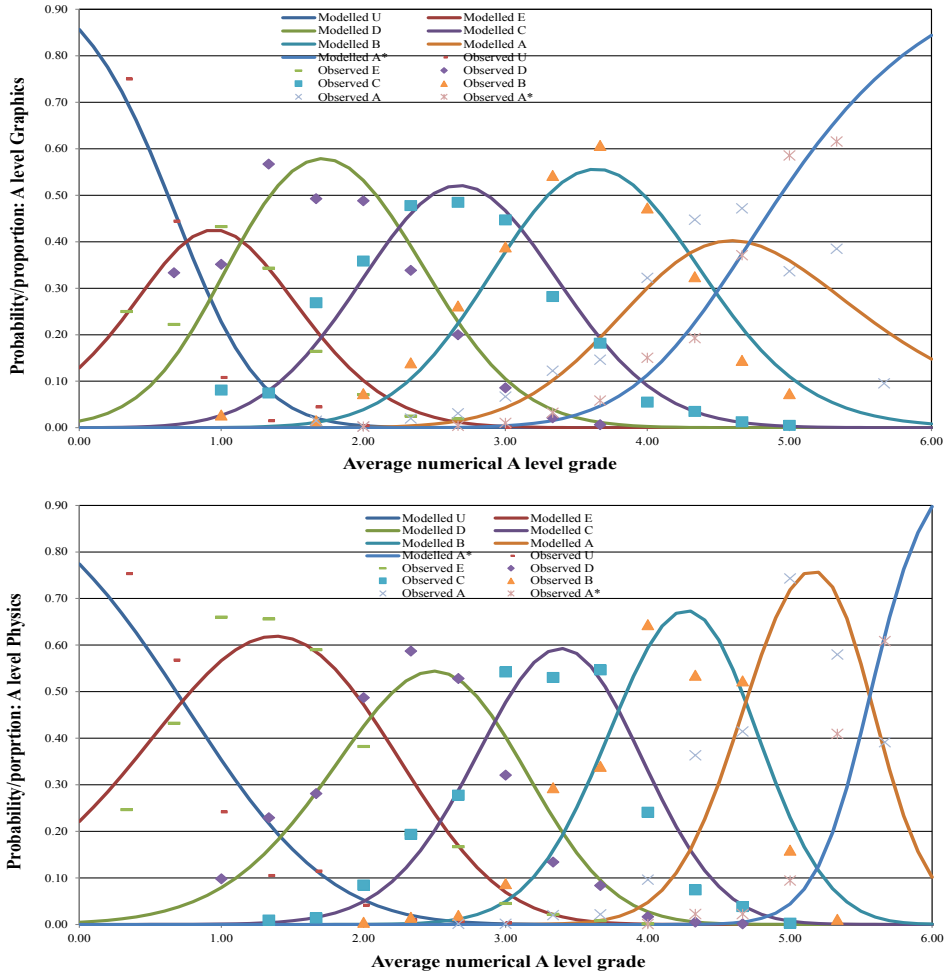

Figure 30: Model predicted category probability distributions and observed proportions of students being classified into the corresponding categories against the mean numerical A level grades for A level graphics (top) and A level physics (bottom) from the 2013 exam series.

For the 47 A level subjects, figure 31 compares the grade difficulties and the overall difficulty derived using logistic regression on the mean GCSE grade and the difficulties derived using the Rasch model. As with the GCSEs, the difficulties are generally positively correlated, but the strength of the correlation is weak for the bottom and top categories. The low level of consistency between the difficulties may again reflect the difference in the nature of the mean GCSE grade and the Rasch ability measure. The GCSEs were taken two years previously, but the Rasch ability measure is based on all the A level subjects being taken by the students presently. Furthermore, there would be a substantial number of common GCSE subjects that were taken by the students, but the total number of A level subjects taken by a student normally ranges from three to five, and the number of common subjects taken by the students would be much smaller.

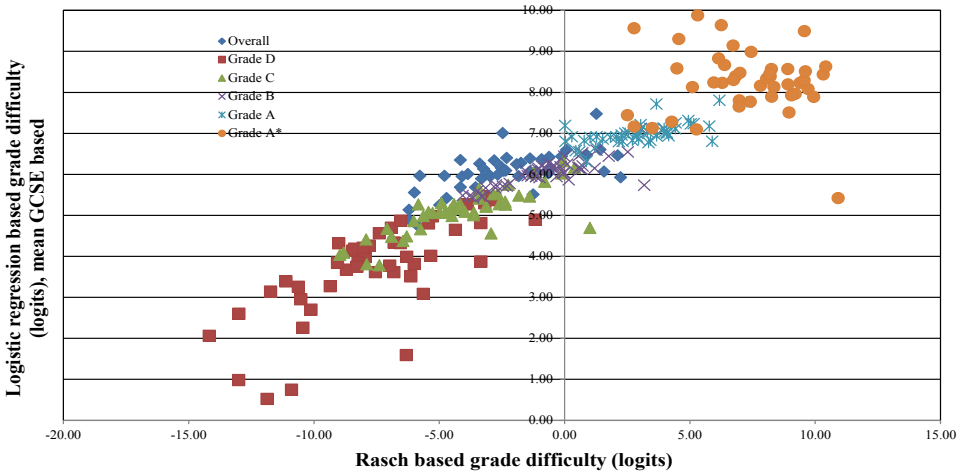

Figure 31: Comparison of the overall difficulty and the difficulty at individual grades based on logistic regression on the mean GCSE grade and the Rasch-model-derived difficulties for the 47 A level subjects from the 2013 exam series.

The top graph in figure 32 shows the distribution of the overall difficulty and the difficulty at individual grades derived using the mean A level grade for the 47 A level subjects, with the subjects ordered by the overall difficulty. The bottom graph in figure 32 compares the difficulties derived using the logistic regression approach with those derived using the Rasch model, and the two are highly correlated. Communications, photography, graphics, food and sociology are judged to be easy subjects by both approaches, whilst biology, physics, chemistry, Latin and critical thinking are hard subjects. The high level of correlation indicates that the mean A level grade is a performance measure similar to the Rasch model estimated person ability measure.

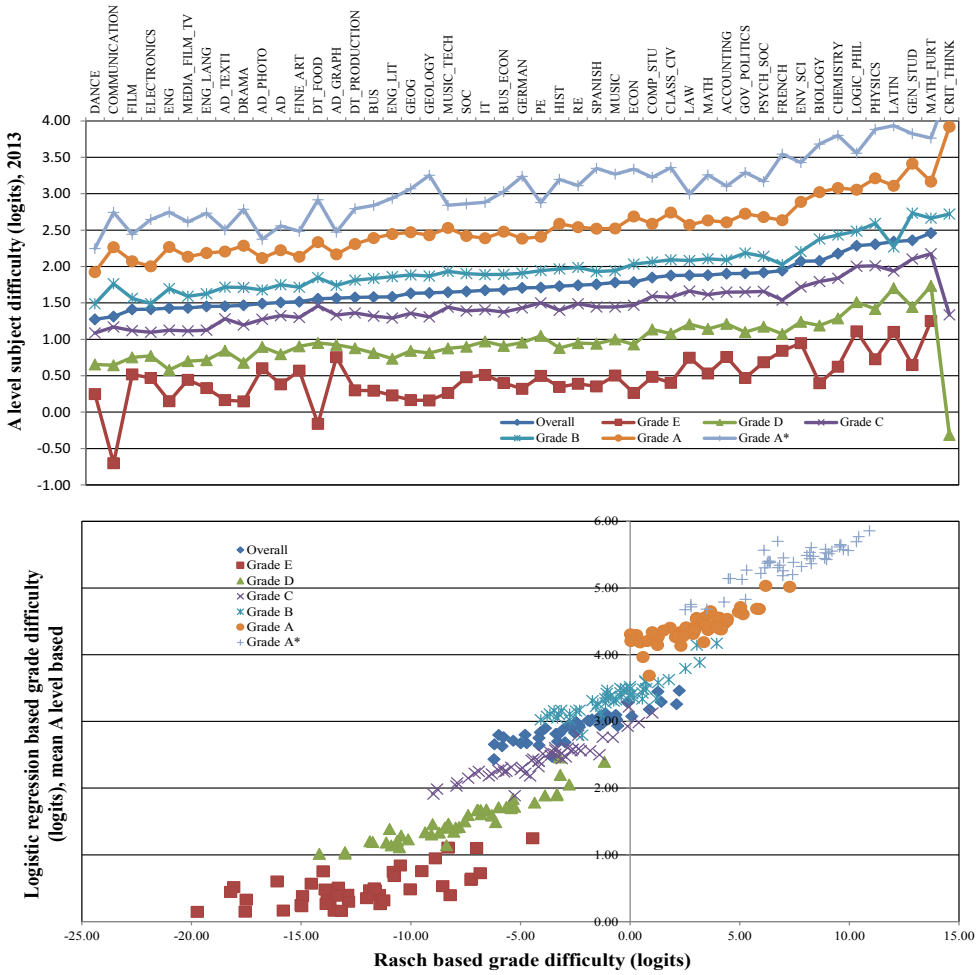

Figure 32: Comparison of the overall difficulty and the difficulty at individual grades for the 47 A level subjects from the 2013 exam series, based on logistic regression on the mean A level grade (top), and the relationship between logistic-regression-derived difficulty and Rasch-model-derived difficulty (bottom).

5. Impact of aligning statistical standards on performance standards

This section explores the impact of aligning statistical standards based on results from the Rasch analysis presented in section 4. There are two stages of this analysis, which will consider the impact on grade outcomes and performance standards. The first stage involves analysis of the subject level mark distributions[1] and adjustment of the subject level grade boundaries to represent the alignment of subjects. A limitation of this approach is that, whilst the impact on performance standards can be approximated, adjustment of subject level grade boundaries is only a proxy for what would be required operationally – the adjustment of unit/component level grade boundaries. The adjustment of unit/component level grade boundaries is the second stage of the analysis. This stage will provide stronger evidence regarding the impact on performance standards of aligning statistical standards.

5.1 Impact on performance standards

Grade gap and difference in difficulty between grades

The Rasch analysis above makes it possible for grade difficulty to be compared both within a subject and between two subjects, based on the underlying latent trait assumed to be measured by the exams. Furthermore, the ability and difficulty scale from the Rasch analysis is linear (that is the same difference between two points on the Rasch ability continuum has the same meaning). As discussed above, the gap between two adjacent grades is not a constant but varies between grades for the same subject, suggesting that the numerical grades are not on a linear scale in terms of representing the underlying latent trait (see figures 6 and 12). Figures 6 and 12 indicate that the grades in the middle range are approximately linear. The grade gap in logits also varies between subjects. Although the difference in difficulty between two adjacent grades is not a constant, an average grade gap Δ (logits) in units of difficulty can be defined across all grades and subjects as:

$$\Delta = \frac{1}{N_G N_S} \sum_{i=1}^{N_S} (d_{i,A} - d_{i,F/E}) \qquad (10)$$

where:

N_G = number of grade gaps (five for GCSEs, between A and F; four for A levels, between A and E);

N_S = number of subjects;

[1] Subject level mark distributions are expressed in uniform marks for modular specifications.

$d_{i,F/E}$ = the difficulty of grade F for GCSEs and E for A levels;

$d_{i,A}$ = the difficulty of grade A for both GCSEs and A levels.

Relative grade difficulty between subjects by grade units

As indicated above, the measurement scale established using the Rasch modelling approach is linear, whilst the ordered numerical grade scale is not. Here, we have assumed that the numerical grade scale is linear but have used the Rasch grade difficulty measure to compare the relative difficulty between subjects at individual grades. At a specific grade *k* for a specific subject, the difference between the grade difficulty of this subject and the mean difficulty of all subjects at this grade is the relative difficulty $d_{k,R}$ of this grade (relative to the mean of all subjects at the same grade):

$$d_{k,R} = d_k - \frac{1}{N_S}\sum_{i=1}^{N_S} d_{ik} \qquad (11)$$

If $d_{k,R}$is negative, the subject concerned at this specific grade is easier in relation to subjects of average difficulty at this grade. If, on the other hand, it is positive, the subject is harder at this grade. Dividing $d_{k,R}$ by the average grade gap in logits gives the relative grade difficulty in the units of grade:

$$d_{k,RG} = \frac{d_{k,R}}{\Delta} \qquad (12)$$

Equation 12 will be used to compare the relative difficulty between the subjects further and to estimate the amount of adjustment in boundary scores that would be needed when aligning statistical standards between the different subjects.

Impact of aligning statistical standards on exam performance standards

As $d_{k,RG}$ is already in the units of grade, it represents the proportion of average grade width in scaled score units (for example, UMS marks) to be adjusted. If it is negative, the subject is relatively too easy, and the corresponding boundary score should increase. If it is positive, the subject is relatively too hard, and the corresponding boundary score should decrease.

GCSEs and A levels adopt a standards-based results reporting system to support their defined purposes, and the grades awarded to students should be interpreted as the levels of attainment in individual subjects (see the later discussion). There are well-established grade descriptions for both GCSEs and A levels for the judgemental grades, which represent a source of evidence used during awarding. In addition, there are also expectations from users of grades regarding the level of performance that grades represent. A shift in grade boundary scores will likely imply a different

performance standard from those established boundary scores (officially or unofficially), which would impact on the interpretation of grades in GCSEs and A levels.

Figure 33 shows the distribution of the relative grade difficulties in units of grade for both the GCSE and A level subjects from the 2013 exam series (see tables A8 and A9 in appendix 1 for actual values). The subjects are arranged according to their overall difficulty. As can be seen, the relative grade difficulties vary within a subject. Although, for most subjects, the signs of the relative grade difficulties are consistent across the grades (either negative or positive), there are a few subjects for which the relative grade difficulties have both positive and negative values. For example, for the GCSE core science, the relative grade difficulty is positive for grades A* to B, but negative for grades C to F. Similarly, for A level English language, the relative grade difficulty is positive for grades A* and A, but negative for grades B to E.

At a specific grade, for the GCSE subjects except Latin, the hardest subjects are about half a grade harder than the subjects of mean difficulty, and the easiest subjects are about half a grade easier. The hardest subjects are, therefore, about one grade harder than the easiest subjects. GCSE Latin is over two grades harder than the easiest subjects at lower grades. For the A level subjects except further mathematics, the hardest subjects (the STEM subjects and modern foreign languages) are nearly two grades harder than the easiest subjects.

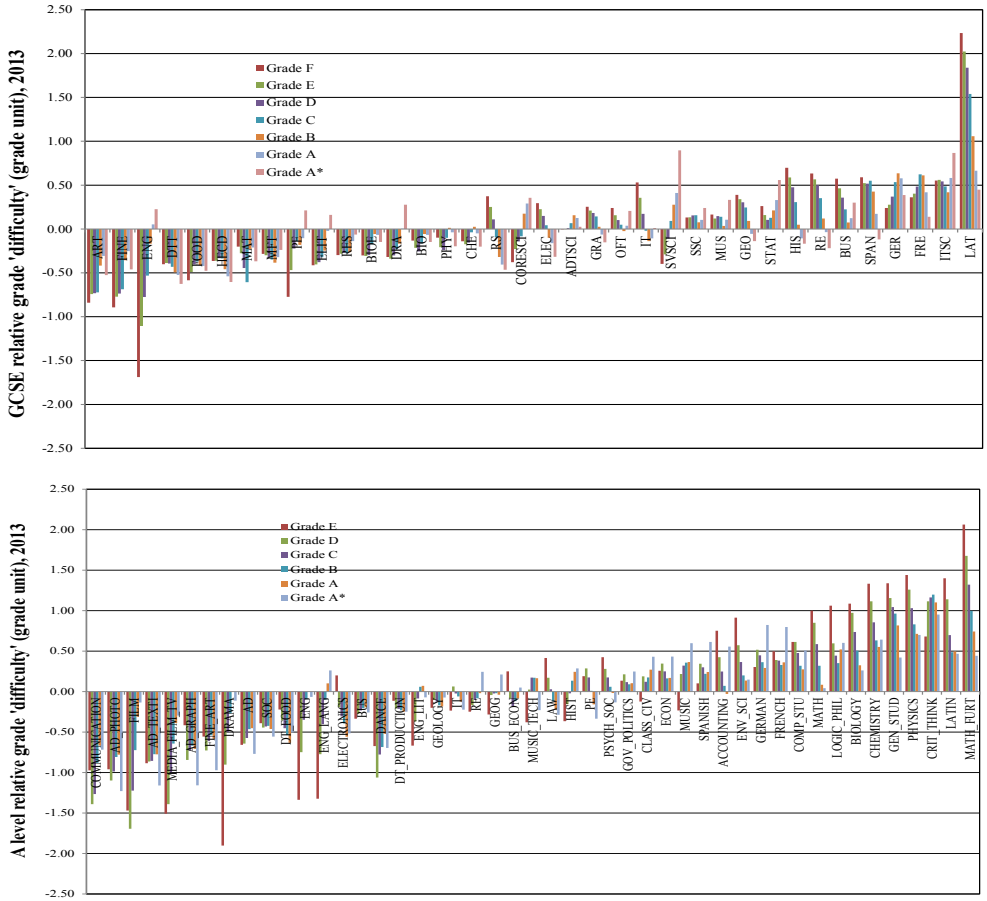

Figure 33: Relative grade difficulty (in units of grade) of GCSE subjects (top) and A level subjects (bottom) for individual grades from the 2013 exam series.

Similar to the definition of relative grade difficulty in grade units for individual grades, relative subject difficulty in units of grade for subjects can be defined as the difference in difficulty between the subject and the mean difficulty of all subjects divided by the grade gap in logits. Figure 34 depicts the relative difficulty of the subjects in grade units for the GCSE and A level subjects from the 2013 exam series. On average, the hardest subjects are about one grade harder than the easiest subject for the GCSEs, and two grades harder for the A level subjects.

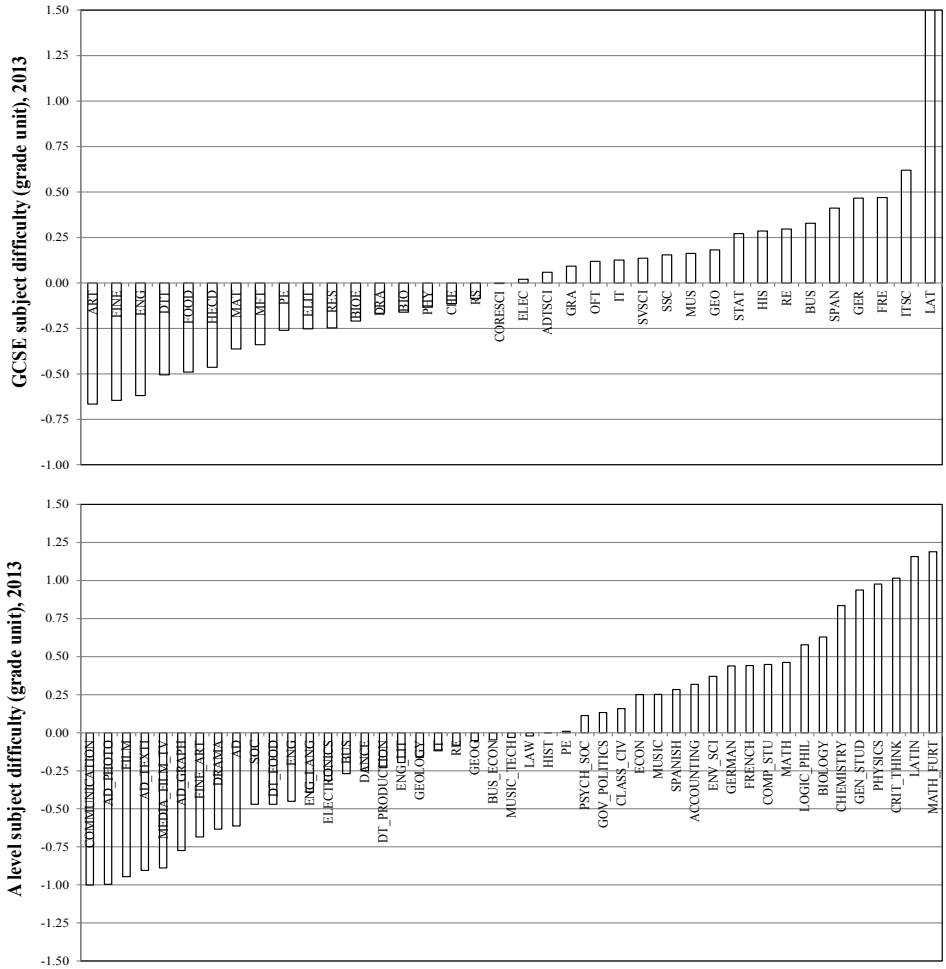

Figure 34: Relative overall subject difficulty (in units of grade) of GCSE subjects (top) and A level subjects (bottom) from the 2013 exam series.

For GCSEs and A levels, grade boundaries could be viewed as the operationalization of performance standards, and aligning statistical standards between subjects would necessarily involve changing the boundary marks for certain subjects. Assuming that the original subject level grade boundary score and grade interval (grade width) at grade *k* are b_k and w respectively for a subject, the new grade boundary b'_k after the alignment of statistical standards with other subjects based on results from the Rasch analysis or other approaches will be:

$$b'_k = b_k - wd_{k,RG} \qquad (13)$$

Equation 13 can be used to investigate the impact of aligning statistical standards on grade distributions (see the following sections).

5.2 Impact on subject grade distributions

Equation 13 presented above has been applied to the GCSE and A level subjects selected to investigate the impact of such alignment on grade distribution that was

originally based on subject-specific performance standards. When UMS is used for a qualification, the grade interval at subject level is, by definition, 10 per cent of the maximum available uniform marks. When raw scores are used (for linearly assessed specifications), the grade interval is the mean of the subject-level grade bandwidths. For A level subjects, the effect on A* was not examined. This is because of the non-linear rules for determining those students that achieve an A* in specifications using UMS, meaning that adjustment of the A* subject-level grade boundary is, conceptually, not straightforward. Instead, the impact on the combination of A* and A was examined.

Impact on grade distributions for GCSE subjects

Table 2 below shows the original grade distributions for three GCSE subjects from the 2013 exam series: English (including GCSE English language), chemistry and German. English has a relatively low value of difficulty parameter, whilst German has a high value of difficulty. Chemistry is of medium difficulty, but with a difficulty parameter slightly below the median. Percentage changes in students at individual grades and the changes in the cumulative percentages of students as well as the shift in grade boundary scores, represented using a percentage of the maximum available UMS marks after alignment with the Rasch statistical standards for the average of all subjects, are also listed in the table. For English, if the statistical standards were to be aligned, the cumulative percentage of students receiving grade A* would increase by about 1.7 per cent. The cumulative percentage of students at grade A (that is those receiving an A or A*) would go up by about 1 per cent. The cumulative percentage of students at grade C (that is those receiving a grade C or above) would drop by about 18 per cent. The UMS boundary score at grade C would need to increase by over 5 per cent of the maximum available marks. At grade F, the cumulative percentage of students would drop by about 11 per cent. For chemistry, there would be a drop of over 5 per cent in students receiving grade A*. The cumulative percentage of students at grade A would decrease by slightly over 1 per cent. For German, the alignment would result in an increase of over 5 per cent of students receiving grade A*. The cumulative percentages of students at grades A and C would increase by about 13.6 per cent and 11 per cent, respectively. At grade C, the UMS boundary mark would need to be reduced by about 5.8 per cent of the maximum available UMS marks.

Table 2: Changes in percentages of students receiving individual grades and cumulative percentages of students for the selected GCSE subjects from the 2013 exam series after alignment of statistical standards.

Subject	Number of can.	Grade distribution (%) and grade boundary change (% of max UMS marks)								
			A*	A	B	C	D	E	F	G+U
English and English language	672,005	Original (Ind.)	3.30	11.04	20.46	29.20	21.33	9.10	3.68	1.89
		New (Ind.)	5.02	10.33	15.18	15.37	22.85	13.63	5.04	12.58
		Change (Ind.)	1.72	-0.71	-5.28	-13.84	1.52	4.53	1.37	10.69
		Original (Cum.)	3.30	14.34	34.80	64.00	85.34	94.44	98.11	100.00
		New (Cum.)	5.02	15.35	30.53	45.89	68.75	82.38	87.42	100.00
		Change (Cum.)	1.72	1.00	-4.28	-18.11	-16.59	-12.06	-10.69	0.00
		Boundary shift	-2.44	-0.56	1.83	5.77	8.41	11.97	18.29	
Chemistry	158,386	Original (Ind.)	16.56	25.45	26.88	20.98	7.83	1.68	0.41	0.21
		New (Ind.)	11.26	29.34	28.94	19.56	7.94	2.15	0.55	0.26
		Change (Ind.)	-5.29	3.90	2.05	-1.43	0.11	0.47	0.14	0.05
		Original (Cum.)	16.56	42.00	68.88	89.87	97.70	99.38	99.79	100.00
		New (Cum.)	11.26	40.61	69.54	89.10	97.04	99.19	99.74	100.00
		Change (Cum.)	-5.29	-1.40	0.66	-0.77	-0.65	-0.19	-0.04	0.00
		Boundary shift	2.19	0.54	-0.27	0.40	1.70	1.88	1.51	
German	60,724	Original (Ind.)	9.15	14.80	23.20	27.75	16.29	5.89	2.11	0.82
		New (Ind.)	14.51	23.03	29.31	19.10	8.42	3.62	1.49	0.57
		Change (Ind.)	5.37	8.23	6.11	-8.65	-7.87	-2.27	-0.62	-0.25
		Original (Cum.)	9.15	23.94	47.14	74.89	91.18	97.07	99.18	100.00
		New (Cum.)	14.51	37.54	66.84	85.95	94.36	97.98	99.47	100.00
		Change (Cum.)	5.37	13.60	19.71	11.05	3.18	0.91	0.29	0.00
		Boundary shift	-4.19	-6.27	-6.87	-5.79	-4.00	-3.00	-2.53	

Impact on grade distributions for A level subjects

Table 3 below shows the original grade distributions for four A level subjects from the 2013 exam series: English language, further mathematics, physics and German. Percentage changes in students at individual grades and the changes in the cumulative percentages of students as well as the shift in grade boundary scores, represented using a percentage of the maximum available UMS marks after alignment with the Rasch statistical standards for the average of all subjects, are also listed in the table. For English language, if the statistical standards were to be aligned, the cumulative percentage of students receiving grade A would go up by about 2 per cent, whilst that of students at grade C would drop by about 13 per cent. For further mathematics and German, such alignment would result in over 11 per cent and 9 per cent increase in the cumulative percentage of students at grade A, respectively. At grade C, the cumulative percentage of students would increase by over 6 per cent for further mathematics and 5 per cent for German. For physics, the cumulative percentage of students at grade A would go up by nearly 17 per cent, which is broadly similar to the findings reported by Alton and Pearson (1996), who attempted to produce a grade distribution of A level physics that was close to the average grade distribution of a set of A level subjects. At grade C, the cumulative percentage of students would increase by over 14 per cent, and the UMS boundary mark would decrease by over 10 per cent of the maximum available UMS marks.

Table 3: Changes in percentages of students receiving individual grades and cumulative percentages of students for the selected A level subjects from the 2013 exam series after alignment of statistical standards.

Subject	Number of can.	Grade distribution (%) and grade boundary change (% of max UMS marks)						
			A*+A	B	C	D	E	U
English language	24,600	Original (Ind.)	12.76	29.76	35.76	17.77	3.52	0.42
		New (Ind.)	14.76	23.86	26.13	19.35	8.65	7.25
		Change (Ind.)	2.00	-5.90	-9.63	1.58	5.13	6.83
		Original (Cum.)	12.76	42.52	78.28	96.05	99.57	99.99
		New (Cum.)	14.76	38.62	64.75	84.10	92.75	100.00
		Change (Cum.)	2.00	-3.90	-13.53	-11.95	-6.82	0.00
		Boundary shift	-1.02 (A)	1.01	4.02	7.74	13.24	
Further mathematics	13,642	Original (Ind.)	56.27	21.61	11.73	5.92	3.00	1.46
		New (Ind.)	67.45	22.16	6.47	3.03	0.78	0.11
		Change (Ind.)	11.18	0.55	-5.26	-2.90	-2.22	-1.35
		Original (Cum.)	56.27	77.88	89.61	95.54	98.54	100.00
		New (Cum.)	67.45	89.61	96.08	99.11	99.88	99.99
		Change (Cum.)	11.18	11.73	6.47	3.57	1.35	0.00
		Boundary shift	-7.42 (A)	-9.98	-13.21	-16.77	-20.64	
Physics	35,781	Original (Ind.)	30.94	23.29	19.41	14.28	9.00	3.07
		New (Ind.)	47.80	22.94	17.57	9.91	1.64	0.14
		Change (Ind.)	16.86	-0.35	-1.85	-4.37	-7.36	-2.93
		Original (Cum.)	30.94	54.24	73.65	87.93	96.92	100.00
		New (Cum.)	47.80	70.75	88.31	98.22	99.86	100.00
		Change (Cum.)	16.86	16.51	14.67	10.29	2.93	0.00
		Boundary shift	-7.13 (A)	-8.30	-10.30	-12.60	-14.41	
German	4,032	Original (Ind.)	41.69	26.22	17.96	9.65	3.82	0.67
		New (Ind.)	50.72	24.98	15.33	6.85	1.74	0.40

		Change (Ind.)	9.03	-1.24	-2.63	-2.80	-2.08	-0.27
		Original (Cum.)	41.69	67.91	85.86	95.51	99.33	100.00
		New (Cum.)	50.72	75.69	91.02	97.87	99.60	100.00
		Change (Cum.)	9.03	7.79	5.16	2.36	0.27	0.00
		Boundary shift	-2.93 (A)	-3.63	-4.49	-5.18	-3.03	

5.3 Impact on unit grade distributions

A GCSE or A level qualification generally contains a number of components or units. This section discusses the impact of aligning standards at subject level on grade distributions at component or unit level for a small selection of GCSE and A level subjects. These subjects include: GCSE English and English language, GCSE German, A level English language and A level physics, for which the impact of aligning statistical standards on grade distributions at subject level was analysed in section 5.2 (see tables 2 and 3). All these qualifications are unitised. As different exam boards may use a different number of units with different raw and uniform marks for the same qualification, the analysis presented here focussed on the qualifications provided by particular exam boards. Since the qualification-level UMS marks represent a linear combination of unit-level UMS marks, the subject-level percentage changes in UMS grade boundaries, which were required for aligning subject standards, were also assumed to represent the levels of impact on unit level standards and applied equally to the unit UMS grade boundaries for all units in a qualification.

GCSE English and GCSE English language

GCSE English and GCSE English language have three units, with two units shared between the two qualifications (Unit 1 – E1 and Unit 2 – E2). Unit 1 is tiered, with a higher tier (E1H) and a foundation tier (E1F). Both qualifications also have a unique unit - Unit 3 (E3 for English and L3 for English language).

Table 4 below shows the number of students sitting the individual units of the GCSE English and English language qualifications, and the changes in unit grade distributions after aligning subject statistical standards. The pattern of changes in grade distributions at unit level is broadly similar to that at subject level. However, the magnitude of changes in percentages of students and the cumulative percentages of students at individual grades vary between the units, which to a certain extent reflect their differences in the distribution of UMS marks. For a small number of grades, the changes are slightly larger than those at subject level. However, for the majority of the grades, the changes at unit level are slightly smaller than the changes at subject level.

Table 4: Changes in grade distributions at unit level for GCSE English and English language from a particular exam board for the 2013 exam series after alignment of statistical standards at subject level.

Unit	Number of can.	Grade distribution (%)								
			A*	A	B	C	D	E	F	G+U
E1F	184,187	Original (Ind.)				17.23	47.77	22.77	7.23	4.99
		New (Ind.)				2.20	23.18	32.08	13.92	28.62
		Change (Ind.)				-15.03	-24.60	9.30	6.68	23.64
		Original (Cum.)				17.23	65.01	87.78	95.01	100.00
		New (Cum.)				2.20	25.38	57.46	71.38	100.00
		Change (Cum.)				-15.03	-39.62	-30.32	-23.64	0.00
E1H	249,030	Original (Ind.)	7.47	14.77	25.27	25.57	15.39			6.77
		New (Ind.)	9.47	12.78	19.73	16.38	18.66			16.41
		Change (Ind.)	2.00	-1.99	-5.55	-9.19	3.27			9.65
		Original (Cum.)	7.47	22.24	47.51	73.08	88.47			100.00
		New (Cum.)	9.47	22.25	41.97	58.35	77.01			100.00
		Change (Cum.)	2.00	0.00	-5.54	-14.73	-11.47			0.00
E2	400,546	Original (Ind.)	11.93	15.32	24.95	22.99	14.61	5.67	2.50	2.04
		New (Ind.)	11.93	15.32	18.56	18.47	14.99	8.69	3.40	8.64
		Change (Ind.)	0.00	0.00	-6.39	-4.51	0.38	3.01	0.91	6.60
		Original (Cum.)	11.93	27.25	52.20	75.18	89.79	95.46	97.96	100.00
		New (Cum.)	11.93	27.25	45.81	64.28	79.27	87.95	91.36	100.00
		Change (Cum.)	0.00	0.00	-6.39	-10.90	-10.52	-7.51	-6.60	0.00
L3	114,296	Original (Ind.)	1.77	3.86	10.63	20.76	30.75	19.30	8.52	4.40
		New (Ind.)	2.49	3.14	8.40	9.87	18.91	20.68	8.20	28.31
		Change (Ind.)	0.72	-0.72	-2.23	-10.89	-11.84	1.38	-0.32	23.90
		Original (Cum.)	1.77	5.63	16.26	37.02	67.77	87.08	95.60	100.00

		New (Cum.)	2.49	5.63	14.03	23.90	42.81	63.49	71.69	100.00
		Change (Cum.)	0.72	0.00	-2.23	-13.12	-24.96	-23.58	-23.90	0.00
E3	292,167	Original (Ind.)	6.52	14.53	27.61	28.14	15.17	4.93	1.83	1.25
		New (Ind.)	9.58	14.28	18.09	19.91	19.49	8.73	2.73	7.19
		Change (Ind.)	3.07	-0.25	-9.52	-8.23	4.31	3.79	0.90	5.94
		Original (Cum.)	6.52	21.05	48.66	76.80	91.98	96.91	98.75	100.00
		New (Cum.)	9.58	23.86	41.96	61.86	81.35	90.08	92.81	100.00
		Change (Cum.)	3.07	2.81	-6.71	-14.94	-10.63	-6.83	-5.94	0.00

GCSE German

GCSE German has four units. Units 1 and 2 are tiered, with higher tiers (G1H and G2H) and foundation tiers (G1F and G2F). The other two units are not tiered (G3 and G4). Table 5 below shows the changes in grade distributions for individual units after aligning subject statistical standards. Compared with other GCSE subjects, German is relatively hard. At grades A* and A, the magnitude of changes in percentages and cumulative percentages of students at individual grades is generally larger than that at subject level.. At grade C, the changes in cumulative percentages at individual grade level are generally smaller than those at subject level, but at grade F the changes are slightly larger.

Table 5: Changes in grade distributions at unit level for GCSE German from a particular exam board for the 2013 exam series after alignment of statistical standards at subject level.

Unit	Number of can.	Grade distribution (%)								
			A*	A	B	C	D	E	F	G+U
G1F	11,684	Original (Ind.)				46.64	28.73	15.85	6.62	2.15
		New (Ind.)				56.66	24.69	12.86	4.55	1.24
		Change (Ind.)				10.01	-4.04	-3.00	-2.07	-0.91
		Original (Cum.)				46.64	75.38	91.23	97.85	100.00
		New (Cum.)				56.66	81.35	94.21	98.76	100.00
		Change (Cum.)				10.01	5.97	2.98	0.91	0.00
G1H	22,078	Original (Ind.)	26.87	18.57	21.22	18.10	10.49			1.67
		New (Ind.)	32.80	26.41	20.07	9.91	7.93			0.91
		Change (Ind.)	5.93	7.84	-1.15	-8.19	-2.55			-0.76
		Original (Cum.)	26.87	45.44	66.66	84.76	95.25			100.00
		New (Cum.)	32.80	59.21	79.28	89.19	97.12			100.00
		Change (Cum.)	5.93	13.77	12.62	4.43	1.88			0.00
G2F	10,892	Original (Ind.)				31.57	30.83	19.27	11.36	6.97
		New (Ind.)				46.50	22.40	17.47	8.53	5.10
		Change (Ind.)				14.93	-8.43	-1.80	-2.83	-1.87
		Original (Cum.)				31.57	62.40	81.67	93.03	100.00
		New (Cum.)				46.50	68.90	86.38	94.90	100.00
		Change (Cum.)				14.93	6.50	4.70	1.87	0.00
G2H	22,965	Original (Ind.)	14.48	16.79	30.25	26.60	10.63			0.23
		New (Ind.)	20.42	29.97	30.43	14.56	4.21			0.14
		Change (Ind.)	5.94	13.18	0.18	-12.04	-6.42			-0.09
		Original (Cum.)	14.48	31.27	61.52	88.12	98.75			100.00

		New (Cum.)	20.42	50.39	80.81	95.37	99.58			100.00
		Change (Cum.)	5.94	19.12	19.29	7.25	0.83			0.00
G3	32,263	Original (Ind.)	16.83	16.20	24.89	23.51	10.26	4.73	2.20	1.38
		New (Ind.)	22.73	27.69	24.17	12.68	7.08	3.03	1.44	1.19
		Change (Ind.)	5.90	11.50	-0.72	-10.83	-3.18	-1.71	-0.77	-0.19
		Original (Cum.)	16.83	33.03	57.92	81.43	91.69	96.42	98.62	100.00
		New (Cum.)	22.73	50.42	74.59	87.27	94.35	97.38	98.81	100.00
		Change (Cum.)	5.90	17.39	16.67	5.84	2.66	0.95	0.19	0.00
G4	32,302	Original (Ind.)	13.26	10.73	27.55	26.15	13.43	6.02	1.94	0.93
		New (Ind.)	18.47	23.04	26.85	17.98	8.01	3.69	1.21	0.75
		Change (Ind.)	5.20	12.30	-0.70	-8.17	-5.41	-2.32	-0.73	-0.17
		Original (Cum.)	13.26	24.00	51.54	77.69	91.12	97.14	99.07	100.00
		New (Cum.)	18.47	41.50	68.35	86.33	94.35	98.04	99.25	100.00
		Change (Cum.)	5.20	17.51	16.81	8.64	3.23	0.90	0.17	0.00

A level English language (Specification B)

A level English language, provided by the exam board studied here, has two specifications: Specification A and Specification B. Specification B, which has a larger entry, was used in this analysis. This specification has four assessment units (ELB1, ELB2, ELB3 and ELB4). Table 6 below shows the changes in grade distributions for individual units after aligning subject statistical standards. English language is a relatively easy subject compared with other A level subjects. At grade A, the changes in percentages of students and the cumulative percentages at individual grades are slightly larger than the changes at subject level. At other grades, some of the changes are slightly larger than those at subject level but others are smaller.

Table 6: Changes in grade distributions at unit level for A level English language (Specification B) from a particular exam board for the 2013 exam series after alignment of statistical standards at subject level.

Unit	Number of can.	Grade distribution (%)						
			A	B	C	D	E	U
ELB1	16538	Original (Ind.)	26.46	30.07	26.49	13.33	3.23	0.42
		New (Ind.)	29.16	24.04	18.62	14.97	6.05	7.17
		Change (Ind.)	2.70	-6.03	-7.87	1.64	2.81	6.75
		Original (Cum.)	26.46	56.53	83.02	96.35	99.58	100.00
		New (Cum.)	29.16	53.19	71.81	86.78	92.83	100.00
		Change (Cum.)	2.70	-3.34	-11.21	-9.57	-6.75	0.00
ELB2	16,533	Original (Ind.)	25.83	34.11	25.47	11.21	2.93	0.45
		New (Ind.)	28.41	27.38	21.60	11.19	5.24	6.18
		Change (Ind.)	2.58	-6.73	-3.87	-0.02	2.31	5.73
		Original (Cum.)	25.83	59.94	85.41	96.62	99.55	100.00
		New (Cum.)	28.41	55.79	77.38	88.57	93.82	100.00
		Change (Cum.)	2.58	-4.16	-8.03	-8.04	-5.73	0.00
ELB3	16,556	Original (Ind.)	8.05	17.74	30.22	26.59	13.53	3.87
		New (Ind.)	9.04	14.39	17.95	20.02	13.20	25.40
		Change (Ind.)	0.99	-3.35	-12.27	-6.58	-0.33	21.53
		Original (Cum.)	8.05	25.79	56.01	82.60	96.13	100.00
		New (Cum.)	9.04	23.43	41.38	61.40	74.60	100.00
		Change (Cum.)	0.99	-2.36	-14.63	-21.21	-21.53	0.00
ELB4	16,532	Original (Ind.)	21.76	29.20	26.91	14.92	5.60	1.62
		New (Ind.)	24.33	23.22	19.53	15.88	4.63	12.41
		Change (Ind.)	2.58	-5.98	-7.38	0.96	-0.97	10.80
		Original (Cum.)	21.76	50.96	77.87	92.78	98.38	100.00

		New (Cum.)	24.33	47.55	67.08	82.96	87.59	100.00
		Change (Cum.)	2.58	-3.41	-10.79	-9.82	-10.80	0.00

A level physics (Specification A)

The A level physics qualification studied here also has two specifications, Specification A and Specification B. Specification A has a larger entry and was analysed here. This specification has six assessment units, with Units 3 and 6 having two options and Unit 5 having four options (see table 7). Physics is a relatively hard subject compared with other A level subjects. Table 7 below shows the changes in grade distributions for individual units after aligning subject statistical standards. The pattern of changes in grade distributions for the units is broadly similar to that of changes at subject level. The changes in percentages of students and the cumulative percentages at individual grades also vary between the units, with some of the grades having changes that are larger than the change at subject level, whilst others have smaller changes.

Table 7: Changes in grade distributions at unit level for A level physics (Specification A) from a particular exam board for the 2013 exam series after alignment of statistical standards at subject level.

Unit	Number of can.	Grade distribution (%)						
			A	B	C	D	E	U
PA1	14,308	Original (Ind.)	48.69	26.13	15.10	6.69	2.54	0.85
		New (Ind.)	66.03	21.18	9.40	2.61	0.50	0.28
		Change (Ind.)	17.33	-4.95	-5.70	-4.07	-2.04	-0.57
		Original (Cum.)	48.69	74.83	89.92	96.61	99.15	100.00
		New (Cum.)	66.03	87.21	96.61	99.22	99.72	100.00
		Change (Cum.)	17.33	12.38	6.69	2.61	0.57	0.00
PA2	14,306	Original (Ind.)	54.65	23.75	12.15	6.26	2.26	0.94
		New (Ind.)	69.91	18.38	8.51	2.41	0.50	0.29
		Change (Ind.)	15.26	-5.36	-3.64	-3.84	-1.76	-0.65
		Original (Cum.)	54.65	78.39	90.54	96.80	99.06	100.00
		New (Cum.)	69.91	88.29	96.80	99.21	99.71	100.00
		Change (Cum.)	15.26	9.90	6.26	2.41	0.65	0.00
PA3-1	9,599	Original (Ind.)	36.87	29.92	18.57	8.94	3.82	1.88
		New (Ind.)	57.61	22.65	14.04	4.27	1.22	0.21
		Change (Ind.)	20.74	-7.27	-4.53	-4.67	-2.60	-1.67
		Original (Cum.)	36.87	66.79	85.36	94.30	98.12	100.00
		New (Cum.)	57.61	80.26	94.30	98.57	99.79	100.00
		Change (Cum.)	20.74	13.47	8.94	4.27	1.67	0.00
PA3-2	5,234	Original (Ind.)	42.34	28.62	14.54	8.16	4.18	2.16
		New (Ind.)	63.24	17.86	12.55	4.72	1.26	0.36
		Change (Ind.)	20.90	-10.76	-1.99	-3.44	-2.92	-1.80
		Original (Cum.)	42.34	70.96	85.50	93.66	97.84	100.00
		New (Cum.)	63.24	81.10	93.66	98.38	99.64	100.00

		Change (Cum.)	20.90	10.15	8.16	4.72	1.80	0.00
PA4	14349	Original (Ind.)	34.47	17.99	16.03	11.21	8.81	11.50
		New (Ind.)	45.81	19.79	14.10	9.70	6.31	4.30
		Change (Ind.)	11.34	1.80	-1.93	-1.51	-2.50	-7.20
		Original (Cum.)	34.47	52.46	68.49	79.69	88.50	100.00
		New (Cum.)	45.81	65.59	79.69	89.39	95.70	100.00
		Change (Cum.)	11.34	13.14	11.21	9.70	7.20	0.00
PA5-1	6,518	Original (Ind.)	29.40	14.50	14.25	12.49	8.61	20.76
		New (Ind.)	37.76	17.54	15.34	10.29	9.99	9.08
		Change (Ind.)	8.36	3.04	1.09	-2.19	1.38	-11.68
		Original (Cum.)	29.40	43.89	58.15	70.64	79.24	100.00
		New (Cum.)	37.76	55.29	70.64	80.93	90.92	100.00
		Change (Cum.)	8.36	11.40	12.49	10.29	11.68	0.00
PA5-2	1,027	Original (Ind.)	23.97	12.50	14.27	11.94	9.05	28.26
		New (Ind.)	31.44	16.23	15.02	11.10	12.13	14.09
		Change (Ind.)	7.46	3.73	0.75	-0.84	3.08	-14.18
		Original (Cum.)	23.97	36.47	50.75	62.69	71.74	100.00
		New (Cum.)	31.44	47.67	62.69	73.79	85.91	100.00
		Change (Cum.)	7.46	11.19	11.94	11.10	14.18	0.00
PA5-3	2,245	Original (Ind.)	27.48	15.99	15.59	12.83	9.58	18.53
		New (Ind.)	37.02	19.11	15.77	10.78	9.53	7.80
		Change (Ind.)	9.53	3.12	0.18	-2.05	-0.04	-10.73
		Original (Cum.)	27.48	43.47	59.06	71.89	81.47	100.00
		New (Cum.)	37.02	56.12	71.89	82.67	92.20	100.00
		Change (Cum.)	9.53	12.65	12.83	10.78	10.73	0.00
PA5-4	4,606	Original (Ind.)	28.44	13.70	14.89	11.59	8.45	22.93
		New (Ind.)	36.58	17.93	14.11	10.51	9.18	11.68
		Change (Ind.)	8.14	4.23	-0.78	-1.09	0.74	-11.25

		Original (Cum.)	28.44	42.14	57.03	68.63	77.07	100.00
		New (Cum.)	36.58	54.52	68.63	79.14	88.32	100.00
		Change (Cum.)	8.14	12.38	11.59	10.51	11.25	0.00
PA6-1	9,138	Original (Ind.)	30.32	15.21	15.87	11.40	10.01	17.18
		New (Ind.)	38.26	15.64	18.91	10.01	15.20	1.98
		Change (Ind.)	7.93	0.43	3.04	-1.39	5.19	-15.20
		Original (Cum.)	30.32	45.54	61.40	72.81	82.82	100.00
		New (Cum.)	38.26	53.90	72.81	82.82	98.02	100.00
		Change (Cum.)	7.93	8.36	11.40	10.01	15.20	0.00
PA6-2	5,314	Original (Ind.)	35.72	15.66	13.79	12.97	7.06	14.81
		New (Ind.)	46.37	14.11	17.65	7.06	11.22	3.59
		Change (Ind.)	10.65	-1.54	3.86	-5.91	4.16	-11.22
		Original (Cum.)	35.72	51.37	65.17	78.13	85.19	100.00
		New (Cum.)	46.37	60.48	78.13	85.19	96.41	100.00
		Change (Cum.)	10.65	9.11	12.97	7.06	11.22	0.00

6. Conclusions

Results from Rasch analysis of GCSE and A level data from over a period of four years suggest that the standards of exams from different subjects are not consistent in terms of the levels of the latent trait specified in the Rasch model that is required to achieve the same grades. There is considerable variability in statistical standards between subjects at both individual grade level and the overall subject level. Results from linear and multinomial logistic regression analyses based on prior attainment and concurrent performance also show substantial inter-subject variability in difficulty, in terms of the statistical model that has been specified. Although the difficulties derived using prior attainment are positively correlated with the difficulties derived using the Rasch model, the strength of the correlation is moderate for the mid-grades and weak for the bottom or top grade. The difficulties derived using the concurrent performance measure are highly correlated with the Rasch-model-derived difficulties. Findings from this study are broadly consistent with those from studies reported by other researchers.

It has been demonstrated that the alignment of statistical standards between subjects based on comparisons using the Rasch model would result in a substantial change in grade distributions and a likely change in performance standards that are based on subject-specific grade criteria.

References

Alton, A. and Pearson, S. (1996) *Statistical Approaches to Inter-Subject Comparability*. Report for the Joint Forum of the GCSE and GCE. Unpublished.

.

Andrich, D. (1978) *A binomial latent trait model for the study of Likert-style attitude questionnaires*. British Journal of Mathematical and Statistical Psychology 31, pp. 84–98.

Bramley, T. (2011) *Subject difficulty – the analogy with question difficulty*. Cambridge, Cambridge Assessment, Research Matters: Special Issue 2: Comparability, pp. 27–33.

Coe, R. (2008) *Comparability of GCSE examinations in different subjects: an application of the Rasch model*. Oxford, Oxford Review of Education, 34, pp. 609–636.

Coe, R., Searle, J., Barmby, P., Jones, K. and Higgins, S. (2008) *Relative difficulty of examinations in different subjects*. Durham, Centre for Evaluation and Monitoring, Durham University.

Elliot, G. (2013) *A guide to comparability terminology and methods*. Cambridge, Cambridge Assessment.

Embretson, S. and Reise, S. (2000) *Item Response Theory for Psychologists*. New Jersey, USA: Lawrence Erlbaum Associates.

He, Q., Anwyll, S., Glanville, M. and Opposs, D. (2014) *An investigation of measurement invariance of Key Stage 2 National Curriculum science sampling test in England*. Research Papers in Education 29, pp. 211–239.

Keeves, J., and Alagumalai, S. (1999) Item banking. In *Advances in measurement in educational research and assessment* ed. G. Masters and J. Keeves, 23-42. The Netherlands: Elsevier Science.

Korobko, O., Glas, C., Bosker, R. and Luyten, J. (2008) Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement*, 45 (2), pp. 139–157.

Lamprianou, I. (2009) Comparability of examination standards between subjects: an international perspective. *Oxford Review of Education*, 35 (2), pp. 205–226.

Linacre, J. (2002) *What do Infit and Outfit, Mean-square and Standardized mean?* Rasch Measurement Transactions 16, p. 878.

Linacre, J. (2013) *Winsteps® Rasch measurement computer program User's Guide*. Beaverton, Oregon, Winsteps.com.

Masters, G. (1982) *A Rasch Model for Partial Credit Scoring*. Psychometrika 47, pp. 149–74.

Muraki, E. (1992) *A Generalized Partial Credit Model: Application of an EM Algorithm*. Applied Psychological Measurement 16, pp. 159–176.

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

Newton, P.E. (2010) Contrasting conceptions of comparability. *Research Papers in Education*, 25 (3), pp. 285–292.

Newton, P.E. (2012) Making sense of decades of debate on inter-subject comparability in England. *Assessment in Education: Principles, Policy & Practice*, 19 (2), pp. 251–273.

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

Ofqual (2014) *GCSE (9 to 1) Qualification Level Conditions and Requirements*. Coventry, Ofqual. Available at: www.gov.uk/government/uploads/system/uploads/attachment_data/file/371219/2014-04-09-gcse-qualification-level-conditions-and-requirements-april.pdf (accessed 29/11/2015)

Ofqual (2015) *GCE Qualification Level Conditions and Requirements*. Coventry, Ofqual. Available at: www.gov.uk/government/uploads/system/uploads/attachment_data/file/423720/gce-qualification-level-conditions-and-requirements.pdf (accessed 29/11/2015)

Ofqual (2015b) *Inter-Subject Comparability: A Review of the Technical Literature: ISC Working Paper 2*. Coventry, the Office of Qualifications and Examinations Regulation.

.

Pae, H. (2012) *A psychometric measurement model for adult English language learners: Pearson Test of English Academic*. Educational Research and Evaluation 18, pp. 211–229.

Rasch, G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark Paedagogiske Institute (Expanded edition (1980). Chicago, University of Chicago Press.)

Reckase, M. (2009) *Multidimensional Item Response Theory*. New York, Springer-Verlag.

Reeve, B. and Fayers, P. (2005) *Applying item response theory modelling for evaluating questionnaire item and scale properties.* In Fayers, P. and Hays, R. (Eds.) *Assessing quality of life in clinical trials: Methods and practice*. New York, Oxford University Press.

Smith, E. (2002) *Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals*. Journal of Applied Measurement 3, pp. 205–231.

Tan, J., and Yates, S. (2007) A Rasch analysis of the Academic Self-Concept Questionnaire. *International Education Journal* 8, 470-484.

Wong, H., C. McGrath, and King, N. (2011) Rasch validation of the early childhood oral health impact scale. *Community Dent Oral Epidemiology* 39, 449–457.

Wright, B. and Masters, G. (1982) *Rating Scale Analysis: Rasch measurement*. Chicago, MESA Press.

Wu, M. and Adams, R. (2007) *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne, Educational Measurement Solutions.

Yen, W. (1993) *Scaling Performance Assessments: Strategies for Managing Local Item Dependence*. Journal of Educational Measurement 20, pp. 187–213.

Appendix 1: Additional tables

Table A1: GCSE subjects from the 2010–13 exam series initially included in the Rasch analysis.

GCSE subject		Sample size			
Code	Full name	2010	2011	2012	2013
AASCI	Vocational GCSE additional applied science	38,366			
ADTSCI	Additional science	288,205	251,328	244,391	248,167
APDSCI	Applied science	8,383	6,426		
ART	Art and design	161,720	154,103	150,200	154,923
BIO	Biology	116,186	139,310	157,091	167,067
BIOE	Biological science	68,715	139,127	156,679	166,552
BUS	Business studies	113,891	61,332	58,856	64,941
CHE	Chemistry		137,436	155,116	165,464
CORESCI	Core science	403,379	347,039	308,458	317,732
DRA	Drama	80,563	74,595	70,263	69,639
DTT	Design and technology: textiles technology	35,700	33,531	31,895	27,690
ELEC	Design and technology: electronic products	10,929	9,822	9,115	8,478
ELIT	English literature	469,944	454,536	443,888	440,644
ENG	English	594,033	581,785	576,286	576,144
FINE	Fine art	45,419	50,578	48,647	50,493
FOOD	Design and technology: food technology	61,900	53,862	49,420	43,310
FRE	French	160,728	144,306	141,640	166,260
GEO	Geography	169,316	164,638	168,707	204,316
GER	German	65,825	58,866	56,232	62,303
GRA	Design and technology: Graphic products	51,134	43,592	40,624	36,135
HECD	Home economics: child development	20,450	17,605	16,462	16,193
HIS	History	198,491	197,297	202,823	240,460
HSC	Health and social care	8,928			
IT	Information technology	43,731	36,678	40,041	59,304
ITSC	Short GCSE information technology	39,393	25,591	13,557	10,653
LAT	Latin		8,214	8,361	8,986
MAT	Mathematics	582,614	585,413	578,605	591,449

MFT	Media, film and television Studies	58,502	51,715	49,389	48,610
MUS	Music	45,515	43,392	41,064	41,548
OFT	Office technology	27,544	18,608	13,290	12,106
PE	Physical education	122,512	105,190	97,749	101,529
PHY	Physics	112,723	137,179	154,880	165,132
RE	Short GCSE religious studies	223,282	194,793	173,019	154,696
RES	Design and technology: resistant materials technology	67,005	58,124	53,974	52,154
RS	Religious studies	176,504	195,336	210,863	229,126
SPAN	Spanish	58,184	59,623	66,302	86,753
SSC	Short GCSE social science citizenship	90,965	72,692	63,594	49,183
SSCI	Science	189,408			
STAT	Statistics	70,092	61,392	52,110	44,183
SVSCI	Additional applied science	38,366	28,867	17,475	11,712
SYS	Design and technology: Systems and control	5,623			
VBUS	Vocational GCSE applied business	5,611			

Table A2: A level subjects from the 2010–13 exam series initially included in the Rasch analysis.

A level subject		Sample size			
Code	Full name	2010	2011	2012	2013
ACCOUNTING	Accounting/finance	3,485	3,178	3,004	2,815
AD	Art and design	7,360	6,677	6,245	6,102
AD_GRAPH	Art and design (graphics)	3,632	3,675	3,718	3,600
AD_PHOTO	Art and design (photography)	9,413	9,924	10,932	10,494
AD_TEXTI	Art and design (textiles)	3,255	3,415	3,313	2,968
BIOLOGY	Biology	51,536	53,393	54,618	55,698
BUS	Business studies	27,056	25,060	24,111	23,105
BUS_ECON	Business studies and economics	1830	1888	1862	1925
CHEMISTRY	Chemistry	39,881	42,691	44,117	46,515
CHINESE	Chinese	1,093	1,528	1,761	1,570
CLASS_CIV	Classical civilisation	3,413	3,264	3,579	3,492
COMMUNICATION	Communication studies	1,777	1,712	1,799	1,471
COMP_STU	Computer studies/computing	3,492	3,286	3,232	3,165
CRIT_THINK	Critical thinking	1,787	1,484	968	843
DANCE	Dance	1,849	1,910	1,695	1,615
DRAMA	Drama and theatre studies	14,193	13,518	12,665	11,452
DT_FOOD	Design/tech and food technology	1,215	1,080	1,006	927
DT_PRODUCTION	Design/tech and production design	10,208	10,065	9,092	8,113
ECON	Economics	19,858		20,733	22,526
ELECTRONICS	Electronics	1,002	963	926	912
ENG	English	14,980	14,694	13,882	13,615
ENG_LANG	English language	21,482	21,393	22,024	22,160
ENG_LIT	English literature	43,193	42,544	42,411	42,230
ENV_SCI	Science: environmental	1,340	1,321	1,239	995
FILM	Film studies	5,815	5,971	5,782	5,175
FINE_ART	Fine art	13,563	13,248	12,918	11,998
FRENCH	French	11,636	10,839	10,303	9,276
GEN_STUD	General studies	43,979	39,671	34,942	30,736
GEOG	Geography	27,819	26,673	27,421	27,836
GEOLOGY	Geology	1,598	1,565	1,751	1,860
GERMAN	German	4,493	4,046	3,745	3,358
GOV_POLITICS	Government and politics	12,322	12,930	13,295	13,363
HIST	History	43,782	43,919	44,200	44,779

HIST_ART	History of art	1,003	8,139	996	
IT	Information technology	8,616	20,268	7,548	6,924
LATIN	Latin	1,254	1,254	1,319	1,219
LAW	Law	13,421	12,364	11,560	10,956
LOGIC_PHIL	Logic/philosophy	3,095	2,939	2,985	2,680
MATH	Mathematics	65,287	69,712	72,078	73,860
MATH_FURT	Further mathematics	10,797	11,361	12,270	12,808
MEDIA_FILM_TV	Media/film/television studies	21,763	21,549	20,226	18,007
MUSIC	Music	5,462	5,280	5,061	4,665
MUSIC_TECH	Music technology	2,813	2,834	2,570	2,386
PE	Sport/physical education studies	17,631	16,042	14,096	12,041
PERFORMING	Performing studies	1,320	1,207	1,003	
PHYSICS	Physics	27,487	28,830	30,425	31,364
PSYCH_SOC	Psychology	50,056	49,871	50,400	49,941
RE	Religious studies	17,123	17,532	17,847	18,030
SOC	Sociology	24,636	25,000	25,558	25,188
SPANISH	Spanish	6,047	5,896	5,700	5,955

Table A3: GCSE and A level subjects from the 2010–13 exam series included in the final Rasch analysis.

GCSE subject	Sample size				A level subject	Sample size			
	2010	2011	2012	2013		2010	2011	2012	2013
Vocational GCSE additional applied science	35231				Accounting/finance	3,059	2,760	2,568	2,410
Additional science	26,6347	232,073	225,394	228,540	Art and design	5,954	5,340	5,020	4,942
Applied science	7,423	5,775			Art and design (graphics)	3,083	3,059	3,093	3,042
Art and design	146,942	138,295	134,326	138,485	Art and design (photography)	7,798	8,250	9,091	8,702
Biology	109,002	131,309	147,835	157,883	Art and design (textiles)	2,723	2,815	2,769	2,445
Biological science		131,205	147,579	157,579	Biology	46,689	48,181	49,574	50,260
Business studies	63,349	56,940	54,757	60,536	Business studies	23,608	21,711	20,937	19,996
Chemistry	107,110	129,715	146,295	156,696	Business studies and economics	1,596	1,662	1,622	1,667
Core science	371,469	318,599	283,234	290,899	Chemistry	36,335	38,806	40,347	42,449
Drama	73,551	67,760	63,668	63,078	Classical civilisation	3,021	2,856	3,147	3,023
Design and technology: textiles technology	33,067	30,967	29,364	25,530	Communication studies	1,576	1,514	1,590	1,295
Design and technology: electronic products	9,595	8,722	8,150	7,630	Computer studies/computing	3,012	2,858	2,763	2,732
English literature	432,849	418,323	410,264	406,056	Critical thinking	1,475	1,195	762	659
English	548,045	536,283	531,414	531,456	Dance	1,538	1,605	1,407	1,361
Fine art	40,935	45,175	43,348	44,632	Drama and theatre studies	12,213	11,581	10,835	9,876
Design and technology: food technology	57,190	49,575	45,811	40,120	Design/tech and food technology	1,069	950	881	821
French	150,125	134,206	131,686	154,677	Design/tech and production design	8,633	8,448	7,646	6,895
Geography	157,898	153,887	158,088	190,628	Economics	17,561	17,875	18,294	19,843
German	61,394	54,781	52,370	58,221	Electronics	877	844	822	804

Design and technology: graphic products	45,291	38,919	36,419	32,269	English	12,974	12,857	12,042	11,794
Home economics: child development	18,971	16,222	15,093	14,828	English language	18,951	18,940	19,351	19,461
History	183,620	182,257	187,151	221,690	English literature	37,832	37,230	36,930	36,657
Health and social care	7,912				Science: environmental	1,192	1,198	1,082	881
Information technology	40,016	33,501	36,397	53,535	Film studies	5,120	5,216	5,086	4,555
Short GCSE information technology	34,865	22,894	12,400	9,655	Fine art	11,082	10,792	10,575	9,757
Latin		7,955	8,077	8,641	French	9,989	9,259	8,775	7,897
Mathematics	535,120	537,953	531,385	542,072	General studies	35,712	31,840	27,889	24,003
Media, film and television studies	52,750	46,727	44,776	44,012	Geography	24,683	23,732	24,381	24,424
Music	41,245	39,294	37,182	37,596	Geology	1,421	1,394	1,559	1,641
Office technology	24,985	16,936	12,174	11,060	German	3,721	3,324	3,005	2,682
Physical education	113,973	97,410	90,712	94,325	Government and politics	10,901	11,362	11,719	11,755
Physics	105,992	129,497	146,061	156,389	History	38,796	38,715	38,903	39,302
Short GCSE religious studies	197,835	170,344	150,675	134,651	History of art	870		839	
Design and technology: resistant materials technology	60,427	52,645	48,845	47,433	Information technology	7,588	7,105	6,600	6,010
Religious studies	162,638	179,722	193,879	210,376	Latin	1,173	1,164	1,205	1,128
Spanish	53,158	54,270	60,331	79,104	Law	11,748	10,895	10,133	9,607
Short GCSE social science citizenship	80,942	64,193	56,438	44,410	Logic/philosophy	2,695	2,557	2,573	2,327
Science	175,211				Mathematics	57,893	61,521	63,705	65,347

Statistics	64,064	56,138	48,104	40,559	Further mathematics	9,273	9,685	10,529	11,002
Additional applied science	35,231	26,456	16,000	10,615	Media/film/television studies	18,905	18,869	17,716	15,789
Design and technology: systems and control	5,022				Music	4,576	4,329	4,238	3,878
Vocational GCSE applied business	4,811				Music technology	2,294	2,279	2,040	1,911
					Sport/physical education studies	15,673	14,221	12,548	10,648
					Performing studies	1,124	1,035	876	
					Physics	24,513	25,714	27,195	27,970
					Psychology	44,748	44,407	44,796	44,236
					Religious studies	14,977	15,278	15,503	15,705
					Sociology	21,806	22,185	22,728	22,320
					Spanish	5,047	4,900	4,705	4,939

Table A4: Average subject difficulty of GCSE subjects from 2010–13.

GCSE subject		Difficulty (logits)			
Code	Full name	2010	2011	2012	2013
AASCI	Vocational GCSE additional applied science	-0.21			
ADTSCI	additional science	-0.55	-0.64	-0.68	-0.44
APDSCI	applied science	-1.15	-1.33		
ART	Art and design	-1.51	-1.46	-1.54	-1.59
BIO	Biology	-0.95	-0.98	-0.98	-0.79
BIOE	Biological science		-1.01	-1.07	-0.87
BUS	Business studies	-0.29	-0.01	0.02	-0.02
CHE	Chemistry	-0.83	-0.92	-0.99	-0.71
CORESCI	Core science	-0.73	-0.78	-0.81	-0.54
DRA	Drama	-0.99	-0.83	-0.84	-0.80
DTT	Design and technology: textiles technology	-1.43	-1.17	-1.21	-1.33
ELEC	Design and technology: electronic products	-0.49	-0.39	-0.45	-0.50
ELIT	English literature	-0.95	-1.08	-1.02	-0.93
ENG	English	-1.25	-1.51	-1.46	-1.51
FINE	Fine art	-1.54	-1.43	-1.50	-1.56
FOOD	Design and technology: food technology	-1.22	-1.07	-1.23	-1.31
FRE	French	0.31	0.32	0.27	0.21
GEO	Geography	-0.24	-0.21	-0.24	-0.25
GER	German	0.35	0.39	0.32	0.20
GRA	Design and technology: graphic products	-0.51	-0.24	-0.35	-0.39
HECD	Home economics: child development	-1.27	-1.13	-1.13	-1.27
HIS	History	-0.02	-0.02	-0.06	-0.08
HSC	Health and social care	-1.11			
IT	Information technology	-0.54	-0.69	-0.27	-0.34
ITSC	Short GCSE information technology	0.35	0.14	0.36	0.45
LAT	Latin		1.57	1.83	1.87
MAT	Mathematics	-0.79	-0.94	-1.09	-1.11
MFT	Media, film and television studies	-1.10	-0.96	-1.00	-1.07
MUS	Music	-0.29	-0.24	-0.36	-0.28
OFT	Office technology	-0.67	-0.41	-0.36	-0.35

PE	Physical education	-1.29	-0.97	-0.88	-0.95
PHY	Physics	-0.81	-0.93	-0.94	-0.75
RE	Short GCSE religious studies	-0.07	-0.04	-0.09	-0.07
RES	Design and technology: resistant materials technology	-0.97	-0.82	-0.77	-0.93
RS	Religious studies	-0.52	-0.60	-0.68	-0.67
SPAN	Spanish	0.19	0.25	0.17	0.12
SSC	Short GCSE social science citizenship	-0.08	-0.18	-0.28	-0.29
SSCI	Science	-0.64			
STAT	Statistics	-0.16	-0.12	-0.19	-0.11
SVSCI	Additional applied science	-0.21	-0.33	-0.60	-0.32
SYS	Design and technology: systems & control	-0.02			
VBUS	Vocational GCSE applied business	-0.54			

Table A5: Variation of grade difficulty in logits for GCSE exams from 2010–13.

GCSE subject		Difficulty (logits)											
Code	Full name	Grade A				Grade C				Grade E			
		2010	2011	2012	2013	2010	2011	2012	2013	2010	2011	2012	2013
AASCI	Vocational GCSE additional applied science	3.77				-0.72				-4.10			
ADTSCI	Additional science	2.91	2.75	2.70	2.89	-0.92	-0.98	-1.06	-0.89	-3.80	-3.85	-3.86	-3.62
APDSCI	Applied science	2.94	2.60			-1.92	-1.95			-4.78	-4.93		
ART	Art and design	2.22	2.18	2.12	2.14	-2.09	-2.03	-2.09	-2.24	-5.01	-4.77	-4.82	-4.89
BIO	Biology	2.51	2.32	2.29	2.55	-1.48	-1.57	-1.58	-1.28	-4.15	-4.10	-4.06	-3.99
BIOE	Biological science		2.32	2.29	2.55		-1.58	-1.60	-1.30		-4.17	-4.25	-4.14
BUS	Business studies	2.65	2.82	2.87	2.88	-0.64	-0.57	-0.55	-0.62	-3.18	-2.74	-2.79	-2.82
CHE	Chemistry	2.54	2.34	2.28	2.58	-1.26	-1.30	-1.39	-1.06	-4.01	-4.03	-4.17	-3.91
CORESCI	Core science	2.93	2.87	2.89	3.17	-1.17	-1.27	-1.33	-1.14	-4.17	-4.19	-4.24	-4.01
DRA	Drama	2.58	2.59	2.60	2.68	-1.59	-1.42	-1.38	-1.40	-4.38	-4.13	-4.21	-4.20
DTT	Design and technology: textiles technology	1.83	1.77	1.76	1.77	-1.84	-1.55	-1.56	-1.74	-4.78	-3.94	-4.03	-4.28
ELEC	Design and technology: electronic products	2.52	2.46	2.40	2.39	-0.90	-0.89	-0.91	-0.93	-3.49	-3.08	-3.09	-3.23
ELIT	English literature	2.67	2.50	2.54	2.63	-1.47	-1.62	-1.50	-1.55	-4.48	-4.56	-4.42	-4.31
ENG	English	2.74	2.56	2.72	2.76	-1.69	-1.94	-1.76	-1.91	-5.06	-5.36	-5.43	-5.51
FINE	Fine art	2.27	2.28	2.22	2.24	-2.08	-2.02	-2.01	-2.18	-5.11	-4.82	-4.87	-4.94
FOOD	Design and technology: food technology	2.03	2.02	1.96	1.95	-1.52	-1.45	-1.50	-1.64	-4.48	-4.04	-4.29	-4.46
FRE	French	3.52	3.42	3.39	3.39	0.21	0.16	0.15	0.07	-2.87	-2.72	-2.80	-2.93
GEO	Geography	2.68	2.58	2.53	2.58	-0.44	-0.51	-0.53	-0.58	-3.15	-2.97	-3.01	-3.04

GER	German	3.73	3.65	3.59	3.66	0.04	0.02	0.00	-0.09	-2.95	-2.74	-2.83	-3.14
GRA	Design and technology: graphic products	2.63	2.60	2.51	2.56	-0.75	-0.61	-0.62	-0.76	-3.75	-2.97	-3.13	-3.26
HECD	Home economics: child development	1.77	1.85	1.85	1.74	-1.55	-1.44	-1.43	-1.58	-4.21	-4.01	-4.04	-4.24
HIS	History	2.55	2.48	2.45	2.49	-0.31	-0.39	-0.43	-0.48	-2.59	-2.48	-2.52	-2.61
HSC	Health and social care	1.80				-1.73				-3.95			
IT	Information technology	2.50	2.36	2.56	2.49	-1.12	-1.26	-0.97	-1.03	-3.51	-3.64	-2.98	-3.01
ITSC	Short GCSE information technology	3.59	3.36	3.37	3.66	-0.19	-0.36	-0.07	-0.17	-2.72	-2.92	-2.58	-2.66
LAT	Latin		3.71	3.71	3.81		1.59	1.65	1.64		-0.62	-0.10	-0.15
MAT	Mathematics	2.46	2.27	2.22	2.31	-1.29	-1.49	-1.82	-2.04	-3.79	-3.87	-4.02	-4.09
MFT	Media, film and television studies	2.22	2.15	2.10	2.12	-1.54	-1.48	-1.51	-1.60	-4.36	-3.95	-3.97	-4.13
MUS	Music	2.78	2.73	2.69	2.85	-0.85	-0.72	-0.83	-0.76	-3.31	-3.17	-3.39	-3.41
OFT	Office technology	2.48	2.64	2.66	2.73	-1.16	-0.92	-0.85	-0.92	-3.62	-3.32	-3.28	-3.35
PE	Physical education	2.05	2.35	2.42	2.49	-1.15	-1.13	-1.14	-1.23	-4.74	-4.18	-4.23	-4.42
PHY	Physics	2.57	2.38	2.37	2.61	-1.29	-1.43	-1.44	-1.17	-4.01	-4.06	-4.10	-3.97
RE	Short GCSE religious studies	2.62	2.62	2.50	2.61	-0.39	-0.35	-0.39	-0.40	-2.66	-2.60	-2.60	-2.65
RES	Design and technology: resistant materials technology	2.49	2.48	2.52	2.43	-1.34	-1.29	-1.32	-1.47	-4.47	-3.94	-3.87	-4.10
RS	Religious studies	2.13	2.00	1.93	1.97	-0.93	-1.05	-1.11	-1.16	-3.06	-3.09	-3.17	-3.19
SPAN	Spanish	3.21	3.03	2.98	2.96	0.16	0.10	0.07	-0.06	-2.87	-2.46	-2.62	-2.72
SSC	Short GCSE social science citizenship	3.15	2.89	2.78	2.85	-0.59	-0.71	-0.75	-0.73	-3.15	-3.14	-3.27	-3.39

SSCI	Science	3.09				-1.04				-4.14			
STAT	Statistics	3.24	3.18	3.09	3.23	-0.64	-0.70	-0.79	-0.78	-3.44	-3.27	-3.38	-3.35
SVSCI	Additional applied science	3.77	3.48	3.13	3.37	-0.72	-0.76	-0.83	-0.84	-4.10	-4.11	-4.31	-4.11
SYS	Design and technology: systems and control	3.33				-0.58				-3.37			
VBUS	Vocational GCSE applied business	2.12				-1.22				-2.97			

Table A6: Average subject difficulty of A level subjects from 2010–13.

A level subject		Difficulty (logits)			
Code	Full name	2010	2011	2012	2013
ACCOUNTING	Accounting/finance	-0.86	-0.99	-1.17	-1.12
AD	Art and design	-3.69	-4.31	-4.22	-4.71
AD_GRAPH	Art and design (graphics)	-4.44	-4.89	-5.36	-5.33
AD_PHOTO	Art and design (photography)	-4.65	-5.58	-5.74	-6.19
AD_TEXTI	Art and design (textiles)	-4.53	-5.29	-5.63	-5.83
BIOLOGY	Biology	0.41	0.22	0.26	0.08
BUS	Business studies	-2.93	-3.13	-3.03	-3.39
BUS_ECON	Business studies and economics	-2.99	-3.01	-2.38	-2.52
CHEMISTRY	Chemistry	1.21	1.04	0.99	0.88
CLASS_CIV	Classical civilisation	-1.24	-1.33	-1.56	-1.74
COMMUNICATION	Communication studies	-3.85	-5.92	-6.83	-6.21
COMP_STU	Computer studies/computing	-0.36	-0.43	-0.44	-0.62
CRIT_THINK	Critical thinking	1.51	1.87	1.83	1.57
DANCE	Dance	-4.49	-4.78	-5.06	-3.31
DRAMA	Drama and theatre studies	-3.57	-3.93	-4.09	-4.79
DT_FOOD	Design/tech and food technology	-3.70	-4.31	-4.70	-4.16
DT_PRODUCTION	Design/tech and production design	-2.71	-2.97	-3.15	-3.23
ECON	Economics	-1.07	-1.21	-1.25	-1.38
ELECTRONICS	Electronics	-2.90	-3.18	-3.15	-3.55
ENG	English	-3.15	-3.48	-3.64	-4.09
ENG_LANG	English language	-2.78	-3.23	-3.34	-3.86
ENG_LIT	English literature	-2.59	-2.94	-2.82	-3.10
ENV_SCI	Science: environmental	-0.37	-0.62	-0.83	-0.92
FILM	Film studies	-5.38	-5.64	-6.07	-5.99
FINE_ART	Fine art	-3.67	-4.14	-4.57	-4.99
FRENCH	French	-0.13	-0.52	-0.66	-0.65
GEN_STUD	General studies	1.00	1.08	1.21	1.27
GEOG	Geography	-1.96	-2.23	-2.41	-2.54
GEOLOGY	Geology	-2.14	-2.33	-2.81	-2.97
GERMAN	German	-0.12	-0.21	-0.58	-0.65
GOV_POLITICS	Government and politics	-1.37	-1.53	-1.60	-1.83
HIST	History	-1.48	-1.84	-2.08	-2.34
HIST_ART	History of art	-0.67		-0.65	
IT	Information technology	-1.40	-1.98	-2.16	-2.81
LATIN	Latin	2.19	1.68	0.59	2.12
LAW	Law	-2.17	-2.18	-2.33	-2.43
LOGIC_PHIL	Logic/philosophy	-0.10	0.32	0.27	-0.12

MATH	Mathematics	-0.33	-0.58	-0.62	-0.57
MATH_FURT	Further mathematics	2.15	2.21	2.10	2.24
MEDIA_FILM_TV	Media/film/television studies	-4.62	-5.07	-5.48	-5.78
MUSIC	Music	-0.65	-0.63	-1.19	-1.38
MUSIC_TECH	Music technology	-1.33	-1.64	-1.80	-2.47
PE	Sport/physical education studies	-2.14	-2.26	-2.36	-2.31
PERFORMING	Performing studies	-3.39	-3.38	-3.02	
PHYSICS	Physics	1.27	1.28	1.24	1.42
PSYCH_SOC	Psychology	-1.74	-1.76	-1.80	-1.91
RE	Religious studies	-1.79	-2.29	-2.37	-2.66
SOC	Sociology	-3.69	-3.77	-3.79	-4.16
SPANISH	Spanish	-0.40	-0.62	-0.79	-1.25

Table A7: Variation of grade difficulty in logits for A level exams from 2010–13.

A level subject		Difficulty (logits)							
Code	Full name	Grade A				Grade D			
		2010	2011	2012	2013	2010	2011	2012	2013
ACCOUNTING	Accounting/finance	3.19	3.08	3.08	2.91	-5.26	-5.33	-5.80	-6.00
AD	Art and design	0.74	0.78	1.08	1.28	-7.66	-8.84	-8.80	-10.11
AD_GRAPH	Art and design (graphics)	0.28	0.44	0.32	0.30	-8.98	-9.43	-10.48	-10.90
AD_PHOTO	Art and design (photography)	0.05	-0.18	-0.31	0.02	-8.98	-10.48	-10.84	-11.87
AD_TEXTI	Art and design (textiles)	-0.10	0.06	-0.05	0.04	-8.79	-9.82	-10.46	-10.97
BIOLOGY	Biology	4.20	4.27	4.36	4.28	-3.30	-3.61	-3.72	-3.88
BUS	Business studies	2.19	2.19	2.37	2.28	-7.77	-8.06	-8.07	-8.70
BUS_ECON	Business studies and economics	1.86	2.33	2.72	2.71	-8.00	-7.90	-7.53	-7.96
CHEMISTRY	Chemistry	4.86	4.95	5.06	5.15	-2.60	-2.89	-3.07	-3.34
CLASS_CIV	Classical civilisation	3.78	3.86	4.05	4.08	-6.43	-6.84	-7.03	-6.91
COMMUNICATION	Communication studies	0.10	0.14	0.12	0.45	-12.02	-11.34	-13.78	-13.00
COMP_STU	Computer studies/computing	3.88	4.02	4.26	4.09	-4.88	-4.96	-5.29	-5.27
CRIT_THINK	Critical thinking	6.91	6.83	7.23	7.28	-3.83	-3.38	-3.96	-3.34
DANCE	Dance	0.60	1.20	0.96	1.02	-8.26	-9.26	-10.44	-11.73
DRAMA	Drama and theatre studies	2.58	2.97	3.13	3.04	-9.05	-10.17	-10.47	-11.12
DT_FOOD	Design/tech and food technology	0.68	0.99	0.51	0.59	-7.75	-8.55	-9.46	-9.01
DT_PRODUCTION	Design/tech and production design	2.05	2.34	2.13	2.07	-7.27	-8.00	-8.02	-8.45

ECON	Economics	3.45	3.44	3.53	3.68	-5.73	-5.93	-6.12	-6.30
ELECTRONICS	Electronics	0.81	0.64	0.72	0.87	-7.29	-7.26	-7.06	-8.38
ENG	English	2.85	2.93	2.95	3.03	-8.68	-9.44	-9.64	-10.52
ENG_LANG	English language	3.33	3.50	3.57	3.42	-8.88	-9.55	-10.11	-10.62
ENG_LIT	English literature	3.14	3.21	3.42	3.30	-7.75	-8.13	-8.35	-9.06
ENV_SCI	Science: environmental	4.13	3.83	3.62	3.55	-4.42	-4.53	-5.02	-5.43
FILM	Film studies	1.80	1.68	1.78	1.83	-12.53	-12.79	-13.41	-14.17
FINE_ART	Fine art	0.87	1.02	0.73	0.78	-8.04	-8.84	-9.31	-10.44
FRENCH	French	4.38	4.34	4.45	4.42	-5.04	-5.76	-5.94	-6.13
GEN_STUD	General studies	5.58	5.76	5.87	6.18	-3.26	-3.31	-3.11	-3.18
GEOG	Geography	2.81	2.81	2.88	2.88	-6.81	-7.35	-7.67	-7.80
GEOLOGY	Geology	2.21	1.97	1.74	2.32	-6.44	-6.84	-8.04	-8.04
GERMAN	German	4.56	4.57	4.35	4.16	-4.76	-5.19	-5.57	-5.64
GOV_POLITICS	Government and politics	3.26	3.38	3.53	3.43	-6.17	-6.33	-6.52	-6.81
HIST	History	3.96	3.91	3.96	3.97	-6.65	-7.29	-7.65	-8.24
HIST_ART	History of art	2.81		3.26		-4.99		-4.91	
IT	Information technology	3.37	2.82	2.74	2.55	-5.97	-6.48	-6.78	-7.39
LATIN	Latin	4.70	4.73	4.69	4.94	-3.56	-5.29	-4.26	-3.23
LAW	Law	2.03	2.24	2.30	2.17	-6.29	-6.48	-6.67	-6.97
LOGIC_PHIL	Logic/philosophy	4.67	5.13	5.21	5.04	-4.98	-4.75	-4.82	-5.34
MATH	Mathematics	3.13	3.13	3.21	3.36	-3.81	-4.19	-4.34	-4.36
MATH_FURT	Further mathematics	5.23	5.38	5.47	5.89	-0.85	-0.87	-1.24	-1.17
MEDIA_FILM_TV	Media/film/television studies	1.61	1.66	1.68	1.53	-10.74	-11.62	-12.34	-13.00
MUSIC	Music	4.37	4.51	4.41	4.44	-5.57	-5.88	-6.39	-6.80
MUSIC_TECH	Music technology	3.56	3.93	3.55	3.67	-6.07	-6.59	-6.44	-7.54

PE	Sport/physical education studies	2.16	2.22	2.36	2.42	-6.02	-6.14	-6.40	-6.53
PERFORMING	Performing studies	3.31	3.31	3.49		-9.69	-10.38	-9.67	
PHYSICS	Physics	5.03	5.12	5.36	5.78	-2.52	-2.61	-2.76	-2.77
PSYCH_SOC	Psychology	2.72	2.84	2.93	2.96	-6.00	-6.16	-6.32	-6.56
RE	Religious studies	2.99	2.98	3.09	2.95	-6.57	-7.53	-7.75	-8.28
SOC	Sociology	1.22	1.25	1.39	1.25	-8.36	-8.57	-8.81	-9.35
SPANISH	Spanish	4.12	4.10	4.17	3.96	-5.08	-5.55	-6.10	-6.31

Table A8: Relative grade difficulty in grade units for GCSE exams from 2013.

GCSE subject	Relative difficulty (grade unit)						
	F	E	D	C	B	A	A*
Art and design	-0.91	-0.80	-0.79	-0.78	-0.45	-0.33	-0.57
Fine art	-0.97	-0.84	-0.80	-0.74	-0.39	-0.27	-0.50
English	-1.83	-1.20	-0.84	-0.58	-0.18	0.06	0.24
Design and technology: textiles technology	-0.44	-0.42	-0.43	-0.47	-0.54	-0.57	-0.68
Design and technology: food technology	-0.63	-0.53	-0.44	-0.40	-0.45	-0.46	-0.52
Home economics: child development	-0.39	-0.39	-0.36	-0.36	-0.50	-0.58	-0.66
Mathematics	-0.22	-0.30	-0.48	-0.66	-0.25	-0.23	-0.40
Media, film and television studies	-0.31	-0.33	-0.35	-0.38	-0.42	-0.35	-0.26
Physical education	-0.84	-0.51	-0.24	-0.15	-0.20	-0.11	0.23
English literature	-0.45	-0.44	-0.41	-0.35	-0.27	-0.02	0.17
Design and technology: resistant materials technology	-0.32	-0.31	-0.30	-0.30	-0.29	-0.15	-0.07
Biological science	-0.33	-0.33	-0.32	-0.19	-0.06	-0.07	-0.16
Drama	-0.35	-0.37	-0.33	-0.25	-0.18	0.00	0.30
Biology	-0.14	-0.23	-0.28	-0.18	-0.06	-0.07	-0.16
Physics	-0.11	-0.23	-0.26	-0.11	0.02	-0.04	-0.21
Chemistry	-0.15	-0.19	-0.17	-0.04	0.03	-0.05	-0.22
Religious studies	0.40	0.27	0.12	-0.10	-0.35	-0.44	-0.50
Core science	-0.41	-0.25	-0.15	-0.09	0.19	0.32	0.39
Design and technology: electronic products	0.32	0.25	0.16	0.05	-0.11	-0.18	-0.34
Additional science	0.00	0.00	0.01	0.07	0.17	0.14	0.03
Design and technology: graphic products	0.27	0.23	0.20	0.15	0.03	-0.07	-0.17
Office technology	0.26	0.17	0.11	0.05	-0.03	0.04	0.22
Information technology	0.57	0.38	0.19	-0.02	-0.15	-0.12	0.01
Additional applied science	-0.43	-0.31	-0.12	0.10	0.30	0.44	0.97
Short GCSE social science citizenship	0.14	0.14	0.17	0.17	0.08	0.11	0.26
Music	0.18	0.13	0.16	0.15	0.04	0.12	0.36
Geography	0.42	0.37	0.33	0.27	0.10	-0.06	-0.15
Statistics	0.28	0.17	0.11	0.14	0.23	0.36	0.60
History	0.76	0.64	0.51	0.33	0.05	-0.12	-0.18
Short GCSE religious studies	0.69	0.61	0.54	0.38	0.13	-0.04	-0.24

Business studies	0.62	0.50	0.39	0.24	0.08	0.13	0.33
Spanish	0.64	0.57	0.56	0.60	0.46	0.19	-0.13
German	0.26	0.30	0.40	0.58	0.69	0.63	0.42
French	0.39	0.44	0.52	0.68	0.66	0.45	0.15
Short GCSE information technology	0.60	0.61	0.59	0.52	0.45	0.63	0.94
Latin	2.42	2.19	1.99	1.67	1.15	0.72	0.49

Table A9: Relative grade difficulty in grade units for A level exams from 2013.

A level subject	Relative difficulty (grade unit)					
	E	D	C	B	A	A*
Communication studies	-0.97	-1.39	-1.27	-0.88	-0.67	-0.72
Art and design (photography)	-0.96	-1.10	-0.99	-0.81	-0.78	-1.23
Film studies	-1.47	-1.69	-1.22	-0.72	-0.31	-0.14
Art and design (textiles)	-0.88	-0.86	-0.86	-0.77	-0.78	-1.16
Media/film/television studies	-1.51	-1.39	-0.99	-0.64	-0.39	-0.30
Art and design (graphics)	-0.42	-0.85	-0.73	-0.67	-0.71	-1.16
Fine art	-0.56	-0.73	-0.61	-0.54	-0.58	-0.97
Drama and theatre studies	-1.90	-0.90	-0.50	-0.27	0.00	-0.11
Art and design	-0.66	-0.64	-0.57	-0.46	-0.45	-0.77
Sociology	-0.39	-0.44	-0.43	-0.42	-0.46	-0.56
Design/tech and food technology	-0.22	-0.36	-0.45	-0.54	-0.63	-0.50
English	-1.34	-0.75	-0.35	-0.09	0.00	-0.07
English language	-1.32	-0.77	-0.40	-0.10	0.10	0.26
Electronics	0.20	-0.19	-0.31	-0.39	-0.56	-0.51
Business studies	-0.34	-0.28	-0.23	-0.21	-0.19	-0.26
Dance	-0.67	-1.06	-0.78	-0.69	-0.52	-0.70
Design/tech and production design	-0.11	-0.21	-0.21	-0.23	-0.25	-0.25
English literature	-0.67	-0.37	-0.08	0.06	0.07	-0.07
Geology	-0.20	-0.10	-0.12	-0.17	-0.18	-0.07
Information technology	-0.23	0.06	-0.02	-0.06	-0.12	-0.22
Religious studies	-0.25	-0.17	-0.10	-0.08	-0.02	0.24
Geography	-0.28	-0.04	-0.02	-0.02	-0.04	0.21
Business studies and economics	0.25	-0.08	-0.18	-0.12	-0.08	0.05
Music technology	-0.38	0.02	0.17	0.17	0.17	-0.23
Law	0.41	0.17	0.03	-0.12	-0.22	-0.29
History	-0.37	-0.16	-0.02	0.13	0.24	0.29
Sport/physical education studies	0.19	0.29	0.17	0.01	-0.16	-0.33
Psychology	0.42	0.28	0.18	0.06	-0.02	-0.13
Government and politics	0.14	0.21	0.12	0.09	0.10	0.25
Classical civilisation	-0.12	0.19	0.12	0.17	0.27	0.43
Economics	0.26	0.35	0.25	0.16	0.17	0.43
Music	-0.23	0.22	0.32	0.36	0.36	0.60
Spanish	0.10	0.34	0.30	0.22	0.24	0.61
Accounting/finance	0.75	0.43	0.25	0.07	-0.03	0.56
Science: environmental	0.91	0.57	0.37	0.20	0.14	0.15
German	0.30	0.52	0.45	0.36	0.29	0.82
French	0.50	0.39	0.38	0.33	0.36	0.80
Computer studies/computing	0.61	0.61	0.48	0.32	0.27	0.51

Mathematics	1.00	0.85	0.59	0.32	0.09	0.04
Logic/philosophy	1.06	0.60	0.44	0.35	0.52	0.60
Biology	1.09	0.97	0.74	0.51	0.32	0.26
Chemistry	1.33	1.12	0.86	0.63	0.55	0.64
General studies	1.34	1.16	1.04	0.96	0.82	0.42
Physics	1.44	1.26	1.03	0.83	0.71	0.70
Critical thinking	0.68	1.11	1.16	1.20	1.10	0.95
Latin	1.40	1.14	0.70	0.49	0.50	0.47
Further mathematics	2.06	1.68	1.32	1.00	0.74	0.44

We wish to make our publications widely accessible. Please contact us at publications@ofqual.gov.uk if you have any specific accessibility requirements.

This publication is available at www.gov.uk/ofqual.

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

2nd Floor
Glendinning House
6 Murray Street
Belfast BT1 6DN

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346