Using differential step functioning analysis and Rasch modelling to investigate inter-board comparability of examination standards in GCSE

November 2016

Ofqual/16/6123

Authors

This report was written by Qingping He, Ian Stockford (now at AQA), and Michelle Meadows from Ofqual's Strategy Risk and Research directorate.

Acknowledgements

The authors gratefully acknowledge the support of the awarding organizations in providing the data analysed in this report and the constructive comments made by Beth Black, Tom Bramley and Tom Benton on an early draft of the report.

Contents

Summary

By treating each examination as a polytomous item, and a grade that a student achieved in the exam as a score on the item, the partial credit model (PCM) has been used to analyse data from 16 GCSE examinations administered in 2015 by the four exam boards that provide general qualifications in England. By further treating students taking the exams that tested the same subject areas but were provided by different boards as different subgroups, differential step functioning (DSF) analysis was used to investigate the comparability of standards at specific grades in the examinations between the exam boards. The grade mean ability and average mean GCSE score were also used to investigate the between-board comparability. Main findings include:

- For most of the grades across the examinations, the magnitude of the DSF effect with respect to exam boards for the majority of the subjects studied is small. The size of grade difficulty for individual exam boards relative to the overall difficulty in the unit of grade was found to be less than one fifth of a grade for most of the grades. The effect of DSF varies between subjects and between grades within the same subject, with higher grades shown to be generally more comparable than the lower grades in terms of difficulty or standards between the exam boards.
- The relative grade difficulties derived using Rasch difficulty, grade mean ability and grade average mean GCSE score were broadly consistent.
- Changes in grade outcomes after aligning standards between the exam boards based on Rasch grade difficulty, grade mean ability and grade average mean GCSE score were moderately or highly correlated with those estimated using the existing inter-board statistical screening procedure with mean GCSE score.
- The use of Rasch ability as a performance measure with the existing inter-board screening procedure produced results which were closely similar to those obtained using mean GCSE score.

1. Introduction

1.1 Inter-board comparability of examination standards in GCSE

The provision of GCSE examinations that test the same subject areas by different exam boards, coupled with lack of pretesting or equating between the examinations, raises the issue of comparability of standards in these examinations (inter-board comparability). Inter-board comparability may be defined as the extent to which students in a specific subject, awarded with the same grades by different exam boards, have similar levels of attainment in the required knowledge and skills within a specified domain of content. Both the regulator and the exam boards regularly undertake inter-board comparability studies using both judgemental and statistical methods (see Newton et al., 2007; Ofqual 2009, 2012; Taylor, 2013; Lockyer and Newton, 2015). While judgemental methods involve the use of experts to compare the performance of students on an exam with an explicit standard or the performance of students from other exam boards, statistical methods are based on examining the relationships between the results from the exams concerned and a common performance measure such as the average score over a suite of GCSE examinations taken in the same year or the prior attainment for different exam boards. A recent review of the technical literature on comparability studies discusses the advantages,

disadvantages and the various issues associated with both methods (see Lockyer and Newton, 2015).

One of the statistical approaches routinely used for monitoring and maintaining inter-board comparability in GCSE is the post-award inter-board statistical screening. This involves, for a specific subject, establishing a relationship between the overall grade distribution of students from all exam boards and a performance measure which is assumed to represent a construct similar to the construct measured by the examination being investigated empirically first. This relationship is then examined for the grade distribution of students from individual exam boards. Significant departure from this all-boards relationship for individual boards would suggest inconsistency in standards between the boards which will be taken into consideration in awarding next year. Presently, the average of the numerical GCSE grades in all the subjects taken by the students in the same year is used as the performance measure (the concurrent GCSE performance measure) for inter-board comparability scrutiny. To establish the relationship between the all-boards grade distribution for the subject and the performance measure, the performance measure is normally divided into 10 bands each of which contains similar number of students. The proportions of students from all boards in each band that were awarded individual GCSE grades are then calculated. These proportions form an all-boards matrix (ρ_{ik}), where $i = 1,2,\ldots,10$ and $k = A^*, A, \ldots, U$. For a particular board, the proportions of students in the 10 GCSE bands are calculated which are denoted as (α_i). Application of the all-boards matrix (ρ_{ik}) to the individual board (α_i) will produce a "predicted" or "expected" grade outcome distribution for the board (A_k):

$$A_k = 100 \times \sum_{i=1}^{10} \alpha_i \rho_{ik} \qquad (1)$$

Significant difference between the expected grade distribution (A_k) and the actual observed grade distribution ($A_{k,0}$) for individual boards is assumed to represent a difference in standards between the board and the other exam boards (see Taylor, 2013).

1.2 Measurement invariance and differential step functioning

Measurement invariance (MI) is a measure of the extent to which the relationship between the properties of the measures from a test with respect to the underlying construct or latent trait being measured by the test holds for all subgroups of the population. High level of MI implies that the test measures the same construct in the same way for all subgroups (Reise et al., 1993; Millsap, 2011). Meaningful and fair comparison of test performances between test-takers from different subgroups requires a high level of measurement invariance across the subgroups (Milfont and Fischer 2010; Millsap, 2011). One of the approaches used to study measurement invariance is the application of item response theory (IRT) and Rasch modelling (see, for example, Dai et al., 2011; Millsap, 2011; Zhang et al., 2011; He et al., 2014).

In the IRT approach to measurement invariance, the ability measures of persons from all groups are placed onto the same scale, the level of measurement invariance is investigated by examining the degree of similarity (invariance) of the item response functions (IRFs), the item characteristic curves (ICCs) or values of the item

parameters between the different subgroups. Similar item parameter values would indicate a high level of measurement invariance (i.e. the IRFs or ICCs are the same across different subgroups). Items with parameters significantly different between subgroups would indicate differential item functioning (DIF) between the groups (Clauser and Mazor, 1998; Oshima and Morris, 2008). That is, test-takers from different subgroups with the same level of ability or trait will have different probabilities of succeeding with the same item. DIF items are a potential source of test bias. Recently, there has been research to examine measurement invariance at individual score (category) levels within polytomous items using IRT or Rasch modelling frameworks (see Penfield et al., 2009). If measurement invariance is violated at score levels, there is differential step functioning (DFS) (Penfield et al., 2009; Miller et al., 2010; Gattamorta and Penfield, 2012; El-Komboz et al., 2014; Akour et al., 2015). The net aggregated effect of DSF is DIF.

1.3 Aims of study

The aims of this research are:

- To gain further understanding of the issues with inter-board comparability
- To explore the potential of using differential step functioning (DSF) analysis with Rasch modelling to investigate the comparability of examination standards in GCSEs between exam boards as a partial validation of the current post-award inter-board statistical screening approach

2. Data collection and analysis

2.1 Data collection

Candidate level data for 16 GCSE examinations administered in 2015 by the four exam boards, (which will be labelled Board A, Board B, Board C and Board D hereafter), that provide GCSE and A level qualifications in England were collected for this study. Table 1 lists the total number of candidates taking the 16 GCSE subjects studied. These included information about candidates' subject level grades and Uniform Mark Scale (UMS) marks and gender. Where there were more than one specification in a subject, the specification that the candidate took was also indicated. Grade comparability was assumed for multiple-specification subjects and specification outcomes were combined to form subject outcomes. Rasch analysis was conducted on the subject level data. In order for the results to be more accurate and reliable, candidates taking fewer than two subjects were excluded from the analysis, which resulted in the sample size of the data included in the analysis to be considerably smaller than the original sample size. This has to be kept in mind when interpreting the findings from the study.

Table 1 Number of candidates from the four Exam Boards studied.

Subject name	Sample size
Additional Science	315685
Applications of Mathematics	12431
Biology	129732
Chemistry	128723
English (including English Language)	428291
English Literature	404949
French	148289
Further Additional Science	22917
Geography	213827
German	51497
History	234704
Mathematics	588216
Methods in Mathematics	11428
Physics	129829
Science	315685
Spanish	12431

2.2 Differential step functioning analysis with Rasch modelling and inter-board comparability

The Rasch family of models, including the partial credit model, have been developed for analysing data from tests composed of individual items that measure a single ability in common to establish measurement scales. The partial credit model for a polytomously scored item can be expressed as:

$$\ln \frac{P_k}{P_{k-1}} = \theta - \delta_k \qquad (2)$$

where:

P_k = the probability of a person with ability θ scoring k on the item

P_{k-1} = the probability of a person with ability θ scoring k-1 on the item

δ_k = the step threshold of category (step) k.

Measurement invariance at category score levels requires that this relationship holds for members from all subgroups in the population. If the step parameter for an item is different for different sub-groups, there is differential step functioning (DSF) at category k in this item. Techniques used to investigate differential step functioning include:

- Calibrate items for different subgroups separately and compare item (step) parameter values between subgroups
- Calibrate items using persons from all subgroups and compare average abilities from different subgroups at specific score levels
- Re-estimate item parameters for individual subgroups using their ability distributions estimated with the population and compare item step parameter values between subgroups.

Given the nature of the data analysed here, the last two approaches were used in this study.

Although the Rasch models are primarily used to analyse data from psychological and educational tests (see Rasch 1960; Masters, 1982; Wright and Masters, 1982), the PCM model has recently been used to study the comparability of standards in examinations across different subjects in England and elsewhere (see Coe, 2008; Coe et al., 2008; Bramley, 2011; He and Stockford, 2015; Opposs, 2015). In such investigations, each examination is generally viewed as a polytomous item in a test, and the grades or performance levels assigned to individual examinees in an exam are treated as scores on an item which represent ordered response categories. All exams contained in the analysis form a test. It is assumed that these examinations together define a shared construct which is closely related to the constructs being measured by the individual examinations and that difference in difficulty reflects difference in standards between the exams. It is however to be noted that the use of Rasch models and indeed other statistical methods to investigate inter-subject comparability generally involves the unidimensionality assumption made about the underlying trait or attribute shared by the examinees. This has received sustained criticism, with the main argument being that examinations such as GCSEs and A levels are graded based on standards that are subject specific and that the shared knowledge and skills assessed by the different examinations are insignificant, leading to any such inter-subject comparison being of limited meaning (see Bramley, 2011; Lockyer and Newton, 2015). However, proponents of statistical approaches argue that as long as there is a theoretical basis for the analysis and the interpretation of the results is justified, statistical comparisons would still be appropriate and meaningful.

To facilitate the analysis in the present study, the GCSE letter grades were converted into numerical values representing ordered category scores: U→0, G→1, F→2, E→3, D→4, C→5, B→6, A→7, and A*→8. The maximum score on every item is therefore 8. The Rasch analysis software WINSTEPS which implements the PCM was used to conduct the analysis (Linacre, 2015). To examine inter-board comparability of examination standards using the PCM model, students taking the exams that test the same subject areas but provided by different exam boards were treated as different subgroups. All items (examinations) and subgroups were analysed together first. Comparability of standards in an examination between exam boards at a specific grade can be investigated by comparing the values of the step parameters for different subgroups which were re-estimated for individual subgroup using the PCM by anchoring their ability estimates at the values from the original analysis (R code was developed specifically for this purpose). The existence of significant DSF at specific grades between the exam boards would indicate inconsistency in standards at these grades. However, since the step parameter δ_k in the PCM model cannot be interpreted as the difficulty of step k or the corresponding score category, an alternative definition of step difficulty based on the item characteristic curve (ICC) has been proposed (see Wu and Adams, 2007; Linacre, 2015). ICC shows the relationship between the expected score $E(\theta)$ on the item from a person with ability θ and is defined as:

$$E(\theta) = \sum_{k=0}^{m} kP_k \qquad (3)$$

In Equation (3), *m* is the maximum available mark on the item. 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 $k-0.5$:

$$d_k = \theta\big|_{E(\theta)=k-0.5} \quad (4)$$

This definition is similar to the definition of the item difficulty for dichotomous items and ensures that the step difficulty increases with step monotonically and was used in this study. The grade average Rasch abilities were also used to examine the between board comparability of the 16 subjects.

It is noted that since not every student took all the subjects included in the study (see Table 1), the analysis involved missing data. Using simulations, Bramley (2016) demonstrated that the existence of non-random missing data could produce biased estimates of subject difficulty. However, his comparison was based on analysis of a dataset re-constructed from a complete dataset by non-randomly omitting some of the data points with known values. For the data analysed here, we will not know how the students would perform on the subjects not taken if they had studied those subjects and taken the examinations. The analysis presented here would be similar to the concurrent calibration of items using data collected from a common item nonequivalent groups (CINE) test equating design (see Kolen and Brennan, 2014).

2.3 Use of mean GCSE score to investigate inter-board comparability

In addition to the use of DSF analysis with Rasch modelling, grade average mean GCSE scores were also used to investigate inter-board comparability.

2.4 The effect of aligning inter-board standards on grade outcomes

The relative grade difficulties between the exam boards estimated using Rasch grade difficulties, grade average abilities and mean GCSE scores for a subject were used to estimate the shifts in grade boundary scores that were required for achieving inter-board comparability, and changes in grade outcomes were generated by comparing the new grade distribution with the original grade distribution. These changes were then compared with those predicted using the current inter-board statistical screening procedure using mean GCSE score to investigate the consistency between the different methods. The Rasch ability measure was also used with the existing inter-board screening procedure and the results were compared with those produced using mean GCSE score.

3. Results and discussion

3.1 Model assumptions and model fit

The application of a Rasch model to analyse test data assumes that the unidimensionality requirement of the model is met and that the test data fit the model. Unidimensionality requires that items in a test measure a single construct or underlying latent variable. When the examination 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. The appropriateness of the definition of the latent variable for the Rasch model by the test 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). Table A1 shows the inter-subject correlations which vary from 0.46 between Spanish and Science to 0.93 between Mathematics and Methods in Mathematics, with the majority of the subjects moderately or highly correlated. Analysis of variances suggested that the total variances in the data accounted for by the Rasch model is about 79% and principal components analysis (PCA) of residuals indicated that the ratio of the first contrast to the second contrast in the residuals in eigenvalue terms is about 1.54, suggesting that the dataset could be essentially treated as unidimensional. Model fit for items can be investigated at individual score category and the overall item levels. Frequently used Rasch item fit statistics include some of the residual based fit statistics such as unweighted mean squares fit statistics (outfit MNSQ) and weighted mean squares fit statistics (infit MNSQ) (see Wright and Masters, 1982; Linacre, 2015). Views on the acceptable values for infit and outfit MNSQs vary, depending on the purpose of the analysis. Linacre (2002) suggested that when model fit statistics are above 2.0, the measurement system could be distorted. This value of 2.0 was used to judge whether an item fits the Rasch model sufficiently well in the present study. An inspection of the fit statistics from an initial analysis suggested that the original grade U did not fit the PCM well and was therefore treated as missing. Grade G was then taken to be the lowest score category, and this resulted in only four of the 128 categories with infit slightly over 2.0 (grade G from English Literature and German, and G and A* from Spanish). To account for the effect of misfit of data to the model on the standard errors of item step measures, the model based standard errors were enlarged by a factor calculated as the square root of the infit MNSQ (when larger than 1.0) when calculating the level of significance of the DSF effects (Linacre, 2015). At item level, all subjects had infit less than 1.71. Table 2 list some of the model fit statistics and other statistics at both the overall item and individual category levels.

Table 2 Rasch model fit statistics and other statistics.

Inter-subject correlation	Range	0.46-0.93
	Mean	0.67
	Standard deviation	0.09
Item infit	Range	0.66-1.71
	Mean	0.98
	Standard deviation	0.31
Step infit	Range	0.62-2.58
	Mean	1.07
	Standard deviation	0.39
Variance explained by the Rasch model (%)		79.3%
Person separation index		3.41
Person reliability		0.92
Average item point – measure correlation		0.87

Figure 1 illustrates the distribution of the Rasch model predicted category probability with ability at individual categories (smooth curves) and that of the observed proportions of persons scoring specific categories (dots) for History, Additional Science and Spanish. The model predicted curves are normally referred to as category probability curves (CPCs) or category characteristic curves (CCCs). As can be seen from the figure, a person with a higher ability will have a larger chance of obtaining a higher score on an item than a person with a lower ability. When the data

fits the model well, the observed and predicted values will be close. The departure of the observed values from the predicted values will reflect the extent to which the data does not fit the model. When the observed distributions are sharper than the predicted curves, there is less variability in the data than the model predicted and the data over-fits the model. When the observed distributions are flatter than the predicted curves, there is more variability in the data than the model predicted and the data under-fits model. Over-fitting items are more effective in differentiating persons with different abilities than under-fitting items. Additional Science over-fits the Rasch model for most of the categories or grades while Spanish under-fits the model. History fitted model very well at all score categories.

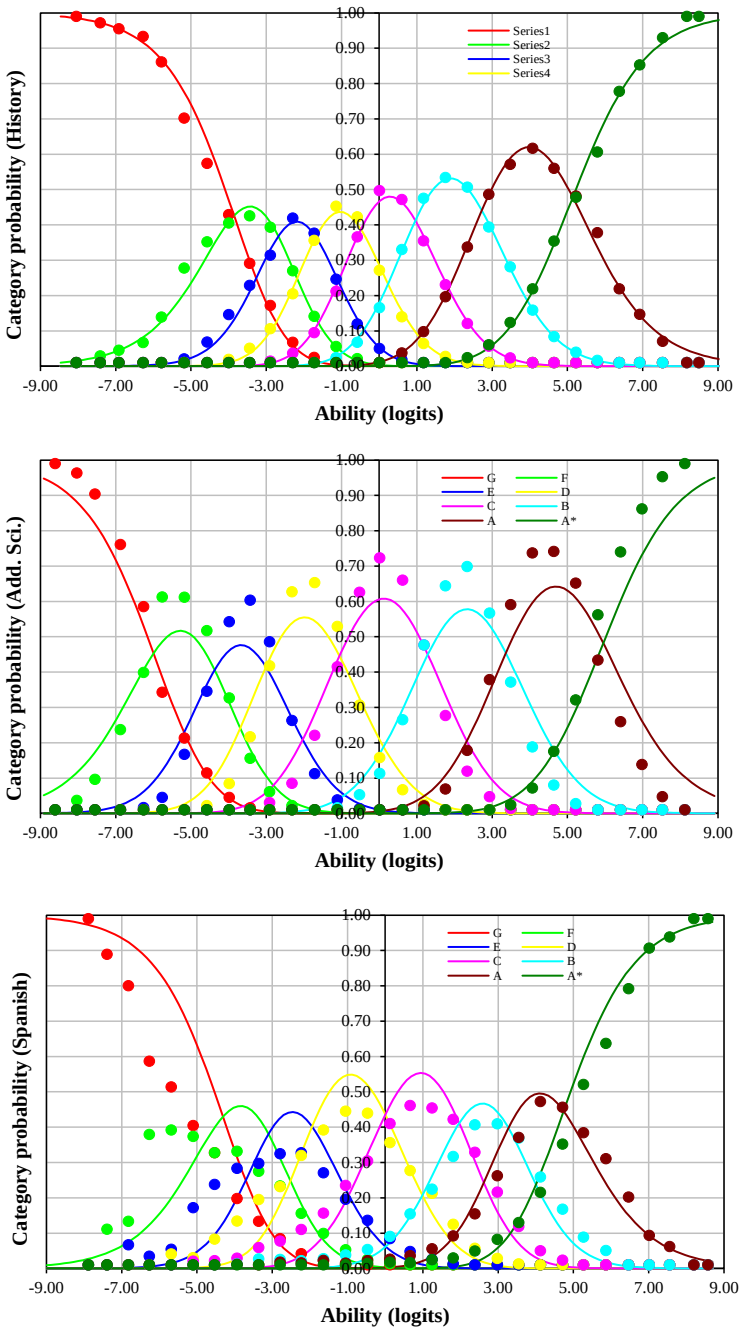

Figure 1 Category probability curves for GCSE History (top), Additional Science (middle) and Spanish (bottom). The smooth curves are model predicted CCCs while the dots are observed values.

3.2 Subject relative difficulty

The top graph in Figure 2 compare the distribution of category step thresholds of the items (subjects) and the distribution of person ability for all persons included in the analysis, with the subject ordered based on their overall difficulty which is defined as the average of the category thresholds. It has to be noted that these values were derived based on the data included in this analysis. It is also worth noting that the

ability distribution in Figure 1 is for the overall sample. Since not every candidate took all the 16 subjects, the ability distribution of the students can vary considerably between the subjects (see discussion below).

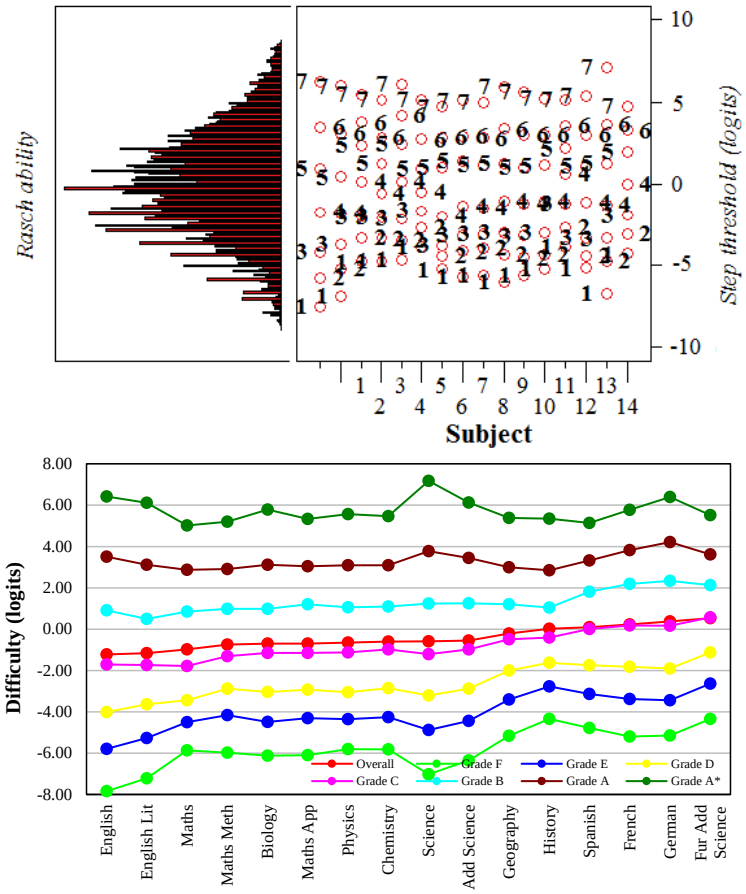

Figure 2 Distributions of person abilities and step thresholds at individual grades for the 16 GCSE exams based on all students (top) and those of grade difficulties.

The bottom graph in Figure 2 shows the distribution of the step (grade) difficulties as defined by Equation (4) (also see Table A2a in Appendix A for actual values). Again the subjects are ordered according to their overall difficulty. 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. It is noted that the gap in difficulty between two adjacent grades varies with the grade, with larger gaps in the higher score categories than lower score categories. For example, difference between the mean difficulty at grade A* across the subjects and the mean difficulty at grade A is about 2.43 logits, while that between grade E and F is about 1.74 logits.

The variability in grade difficulty at higher grades between the subjects is smaller than that at lower grades. For example, at grade A, the threshold varies from 2.85 logits for History to 4.22 logits for German, while that at grade E varies from -5.79 logits for English to -2.64 logits for Further Additional Science. These differences reflects the nonlinearity of grades and variability in difficulty between subjects.

Figures 3 shows how the model expected scores and observed scores on individual items vary with ability (the item characteristic curves – ICCs). The curves on the left of the graph are for subjects which are easy and those on right difficult in terms of the level of ability specified by the Rasch model that is required to achieve the same expected score (grade) in different subjects. The ICCs also spread a slightly wider range of ability for the lower grades than for the upper grades, suggesting that there is a degree of differentiated relative difficulty. This is also consistent with the patterns of grade difficulty for the subjects shown in the bottom graph in Figure 2. If two ICCs do not cross, then the leftmost subject is easier than the rightmost across the full ability range. If two ICCs do cross, then the order of difficulty changes direction at the intersection point. The empirical curves are superimposed on the theoretical ICCs.

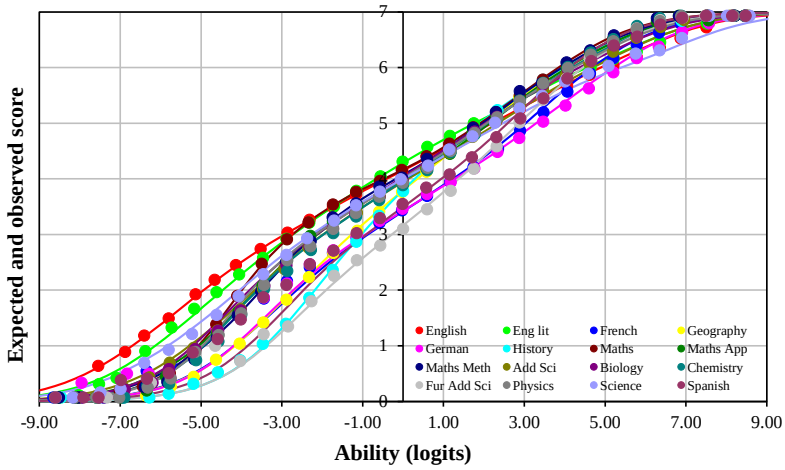

Figure 3 Comparison of the distributions of model expected ICCs and observed ICCs for the 16 GCSE subjects studied.

Figure 4 shows the average ability of students taking the different subjects overall and from the four exam boards based on the data analysed. The average ability of students taking the separate science subjects (Physics, Chemistry and Biology), Further Additional Science and Modern Foreign Languages was considerably higher than that of students taking the other subjects. The average ability of students taking Science is the lowest among the subjects. As schools are free to choose exam boards for their students, there can also be substantial variability in the ability of students between the different exam boards for the same subject. For Biology, Physics and Chemistry, the average abilities of students from Board B were slightly lower than those of students from the other three boards. In contrast, the average abilities of students from Board B who took English (including English Language), English Literature and Modern Foreign Languages were higher than those of students from the other boards.

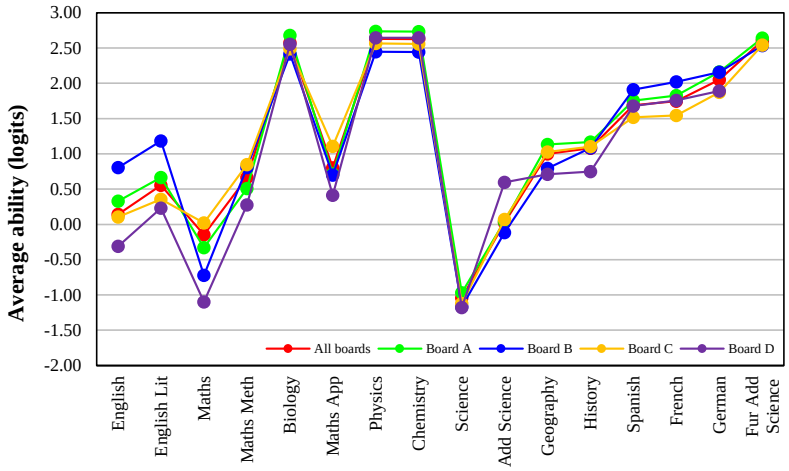

Figure 4 Distributions of average abilities of students taking different exams from different boards.

3.3 Relative between-board grade difficulties based on differential step functioning analysis

Figure 5 shows how the average ability of students taking the different exams at individual grades vary among the subjects and between the exam boards. Difference in average ability at a specific grade to an extent indicates difference in exam difficulty or standard between the boards. For the same subject, the order of average ability between the exam boards can be different at different grades, reflecting differentiated relative grade difficulty. The size of differences in average ability also vary between grades and between subjects. For example, for English, the average ability of students from Board B was higher than that of students from the other boards across the grades, while for Additional Science, its average ability was slightly lower than that of students from the other boards. For French, the average ability of student from Board B was higher than that of students from other boards at grades A*, A and B, but lower at grade F. Similar to the grade difficulty distributions shown in Figure 2, the gap in grade average ability between two adjacent grades for the higher grades are larger than that for the lower grades. The variability in average ability between the subjects at the lower grades is considerably larger than that at the higher grades.

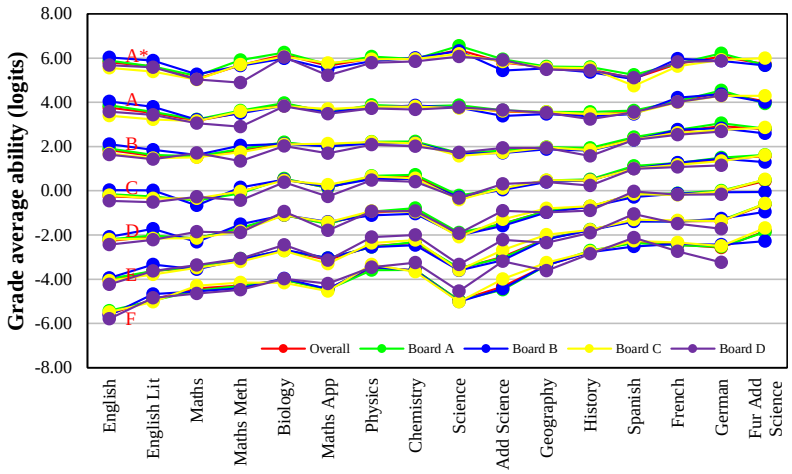

Figure 5 Distribution of average abilities at different grades for different boards and different exams.

The top graph in Figure 6 shows the distributions of grade difficulties across the subjects for individual exam boards re-estimated by fixing their ability estimates at the values from the original analysis using all students (also see Tables A2b to A2e). For most of the grades across the subjects, the relative positions of grade difficulty for individual boards are generally consistent with those for the average grade ability shown in Figure 5, which is demonstrated in the bottom graph in Figure 6. However, the magnitude of variability in grade difficulty between the subjects and boards is generally slightly larger than that in grade average ability. This reflects the fact that estimation of the Rasch grade difficulty involves all students that took the subject (and the other subjects) while the average grade ability is calculated based on the abilities of the students who achieved a particular grade in the subject. The average Rasch grade abilities would therefore likely exhibit less variability between the subjects and between the exam boards than the Rasch grade difficulties. The relative difference in grade difficulty between the harder subjects and the easy subjects at the lower grades is smaller than the relative difference in grade average ability. Significant differences in grade difficulty between exam boards at a specific grade for a specific examination suggests differential step functioning (DSF) which may be interpreted as the differences in standards between the exam boards at that grade.

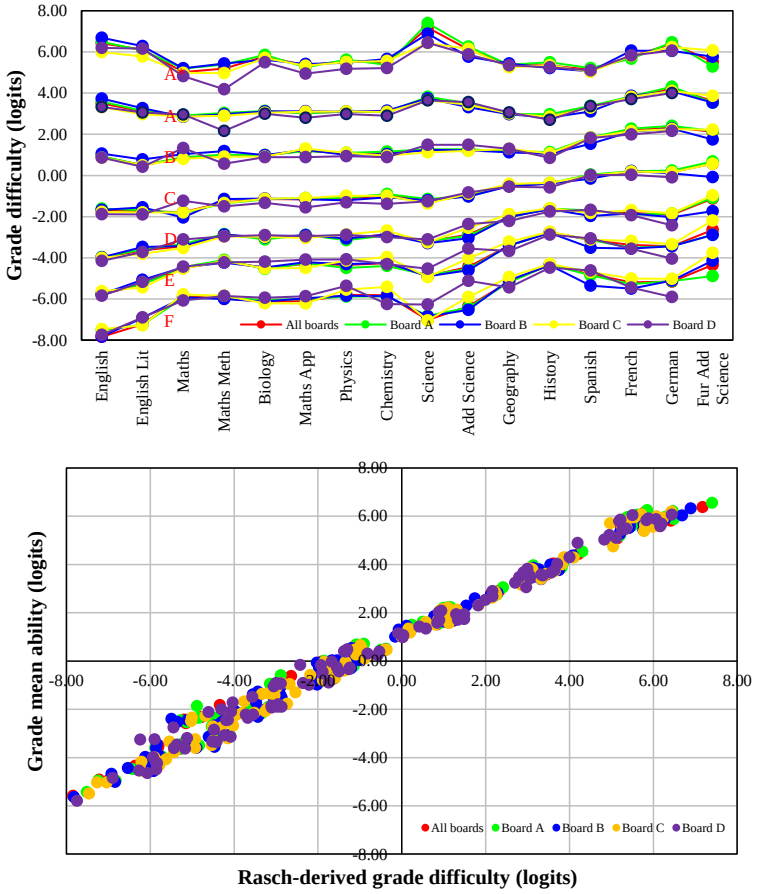

Figure 6 Distribution of Rasch grade difficulties (top) and their relationship with grade average ability (bottom).

To see how DSF affects the performance of students from different boards on individual subjects in more detail, Figure 7 shows the distributions of abilities of students from different boards taking two exams, English which is a relatively easy subject and Additional Science which is a more difficult subject (top graph), and the corresponding ICCs (bottom graph). The standard deviations of abilities of students taking English are considerably larger than those of students taking Additional Science. Further, the variability in the ability distributions between the exam boards for English is also substantially larger than that for Additional Science. Students from Board B taking English had generally higher abilities than those from the other three boards. However, the opposite is true for Additional Science. The average ability of students taking English was slightly higher than that of the students taking Additional Science.

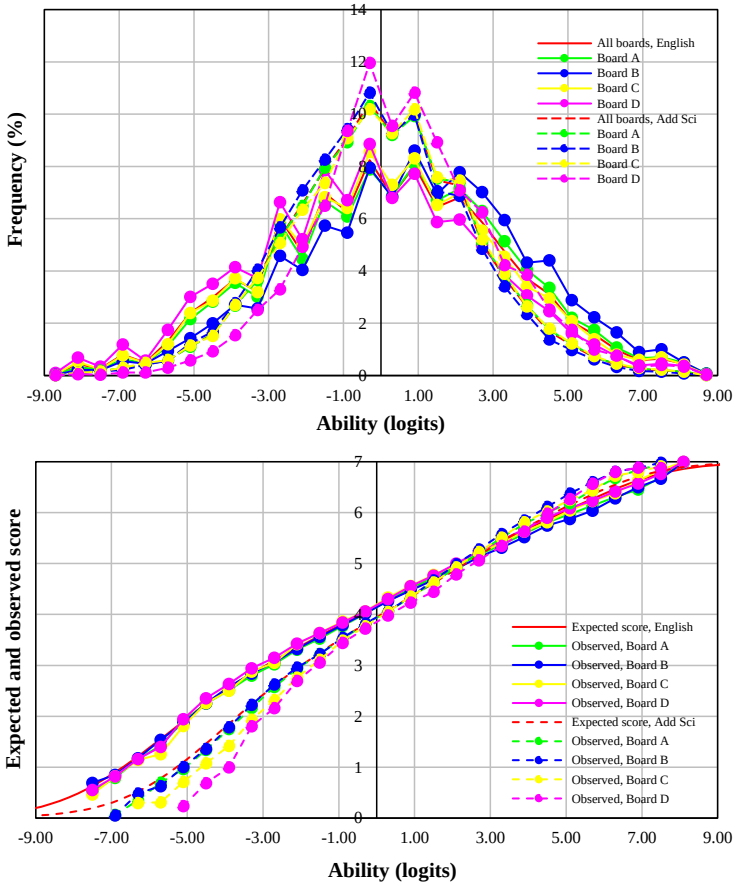

Figure 7 Distributions of ability of students from the four exam boards taking English and Additional Science (top) and the item characteristic curves of the two subjects (bottom).

Below the ability of 3.0 (corresponding to grades below B), the Additional Science is slightly more difficult than English. However, above 5.0, it is easier than English. It is evident from Figure 7 that for the lower grades, the variability in expected score between the exam boards is very small for English, while that for Additional Science is substantial. For the middle and higher grades, difference in expected score between the exam boards is generally less than 0.3 grades for both subjects. To see how students from different boards performed on the two subjects at individual grades, Figure 8 depicts the category probability curves for individual boards. In Figure 8, the smoothed curves are the Rasch model CCCs for students from all boards, while the symbols show how the observed proportions of students in each grade from individual boards vary with ability. It is to be noted that the extent to which the observed values depart from the model predictions is not a measure of DSF but a measure of the fit of the data to the Rasch model. While English fitted the Rasch model for the majority of the grades reasonably well, Additional Science generally over-fitted the model at all grades. The positions of the CCCs also reflect the difficulty of the corresponding grades. For the same grade, the subject with CCC on the right will be more difficult than the subject on the left. For grade C to F, comparisons of the CCCs between the two subjects suggested that these grades are harder for

Additional Science than for English. For the other grades, the difficulty for both subjects is similar.

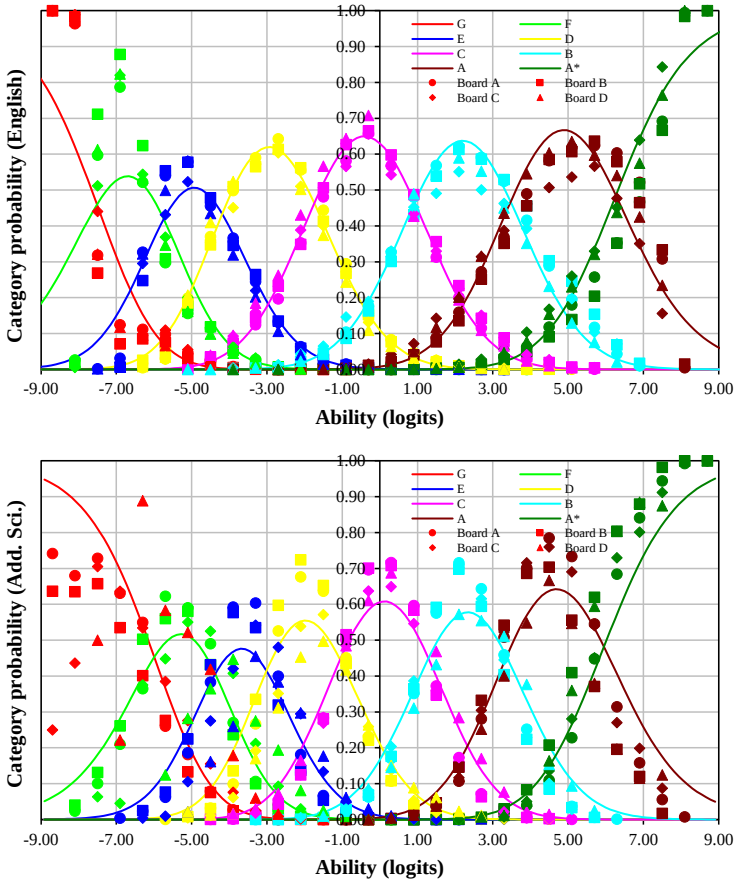

Figure 8 Model predicted category characteristic curves and observed proportions of students in each grade for English (top) and Additional Science (bottom).

The larger the variability in the observed proportions in each grade between the exam boards, the larger the effect of differential step functioning or inconsistency in grade standards between the exam boards. The patterns shown in Figure 8 are generally consistent with the patterns shown in Figure 7. For the lower grades, there is less variability in the observed proportions for English than for Additional Science. For the higher grades, there is similar level of variability between the boards for both subjects. For a specific grade in a subject, if the observed proportions for a particular exam board are to the left of the proportions for the other boards, the exam from the board is easier at this grade than the exams from the other boards. If the observed proportions are to the right of those of the other boards, it is more difficult. For example, for English at grades A* and A, the observed proportions for Board B are slightly to the right of the corresponding proportions for the other three boards while those of Board C are slightly to the left. Therefore, at grades A* and A, the exam from Board B is harder than the exam from Board C. For Additional Science, the observed proportions at grades A* and A for Board A are slightly to the right of those for the other boards, but the observed proportions for Board B are slightly to the left. This exam is harder for Board A at these two grades than the exam from Board B.

The relative difficulty $d_{k,R}$ at a specific grade for a subject from a specific exam board can be defined as the difference between its grade difficulty d_k and the overall difficulty $d_{k,ALL}$ estimated based on students from all exam boards (i.e. the corresponding grade difficulty value from the bottom graph shown in Figure 2):

$$d_{k,R} = d_k - d_{k,ALL} \tag{5}$$

If $d_{k,R}$ is negative, the exam from the board at this specific grade is easier than the overall difficulty for all boards; if, on the other hand, it is positive, its exam is more difficult at this grade. The relative difficulty is a measure of the DSF effect. The significance level of DSF can be tested using a t-test, involving the use of the relative difficulty and the standard errors of the two step difficulty measures. Figure 9 shows the relative grade difficulties for the exam boards across the 16 examinations. For most of the grades across the examinations, particularly the higher grades, the relative grade difficulty is small (with the absolute value less than 0.3 logits), with the majority shown to be significant (see Tables A3a to A3d in Appendix A). Significant difference in relative difficulty would indicate significant difference in standards between the exam boards. There are a few lower grades (grade F or E) that have moderate to larger DSF values (from 0.3 logits to 0.6 logits and above). Higher grades are generally more comparable between the exam boards than the lower grades for most of the subjects. It is to be noted that for a subject, larger-entry boards are likely to contribute more than small-entry boards to the overall difficulty at individual grades. The majority of the exams from Board A were close to the all-boards difficulty at most grades. However, for Further Additional Science, Board A was relatively easier than those from the other boards, particularly at A*. For Board B, Science, Additional Science and Further Additional Science were generally easier than those from the other boards, but French and Methods in Mathematics were harder. For Board C, its exams in Chemistry, Additional Science and Further Additional Science were harder than those from other boards, but English Literature and Science were easier. The relative grade difficulties for Board D show larger variability than those for the other boards, with exams in German, Methods in Mathematics and Chemistry being easier than those from the other boards.

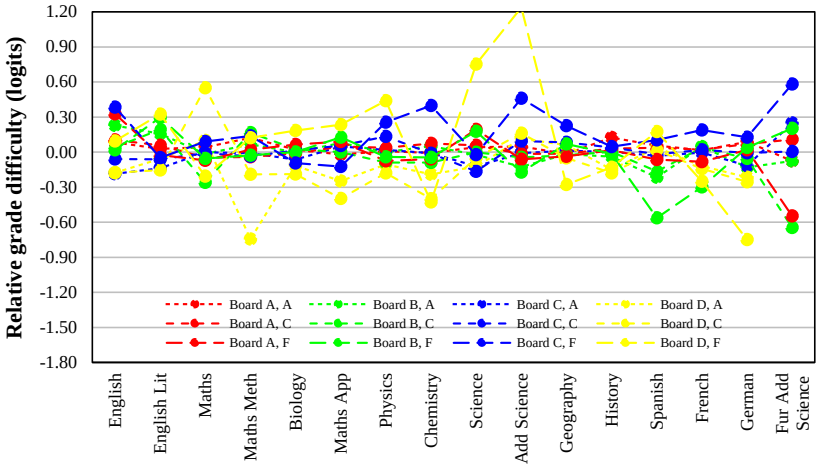

Figure 9 Distribution of relative grade difficulty in logits based on DSF analysis.

To make the comparison at specific grades between the boards more intuitive, the relative grade difficulties expressed in logits shown in Figure 9 were converted into units of grade. To achieve this, the average grade gap across the subjects Δ (logits) was used:

$$\Delta = \frac{1}{N_G N_S} \sum_{i=1}^{N_s} (d_{i,A} - d_{i,E}) \quad (6)$$

where

N_G = number of grade gaps between A and E

N_S = number of subjects included in the analysis

$d_{i,E}$ = the difficulty of grade E

$d_{i,A}$ = the difficulty of A

It is to be noted that since the gap between grades varies between subjects and between grades within the same subject, the average grade gap can be defined differently for different subjects and for different grades. For example, for grades A* to C, the average grade gap may be defined using Equation (6) by setting the upper grade as A* and the lower grade as C; for grade E to F it may be defined by setting grade C as the upper grade and F as the lower grade. Average grade gap for a subject can also be defined based on the grade gaps within the subject only. Dividing the relative difficulty $d_{k,R}$ by the average grade gap in logits gives the relative grade difficulty in the unit of grade:

$$d_{k,RG} = \frac{d_{k,R}}{\Delta} \quad (7)$$

Equation (7) can be used to compare the relative grade difficulty between the boards further and to estimate the amount of adjustment in boundary scores that would be needed when aligning inter-board standards for different subjects. Figure 10 depicts the relative grade difficulty between the exam boards in the unit of grade (also see Table A4a to A4d in Appendix A). The patterns in the distributions of relative grade difficulties between the exam boards across the subjects shown in Figure 10 are similar to those shown in Figure 9. The difference in relative grade difficulty between the most difficulty board and the easiest board for most of the grades were less than 0.3 grade. There are a few subjects, including Science and the separate sciences, the difference between the hardest board and the easiest board were over 0.4 grade.

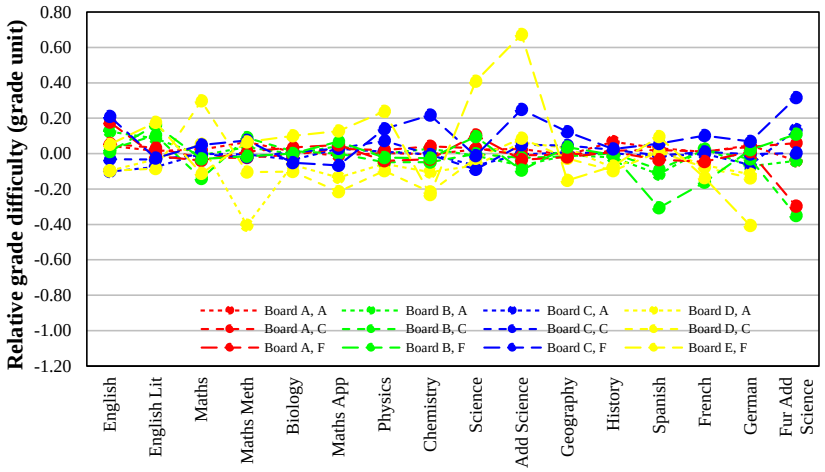

Figure 10 Distribution of relative grade difficulties in the unit of grade based on Rasch grade difficulty.

From the average grade Rasch abilities for the overall sample and the individual boards at a specific grade shown in Figure 5, the relative grade difficulty in the unit of grade for each board can also be defined using the approach described above, and the top graph in Figure 11 shows the distribution of the relative grade difficulties thus derived. The pattern of the relative grade difficulties based on grade mean ability is broadly similar to that derived using the Rasch grade difficulty. However there are substantial differences between the two difficulty measures for some of the subjects in terms of magnitude and direction, reflecting the difference in how grade difficulty was conceptualised. The size of the relative grade difficulties based on grade mean ability is generally smaller than that derived using Rasch grade difficulty.

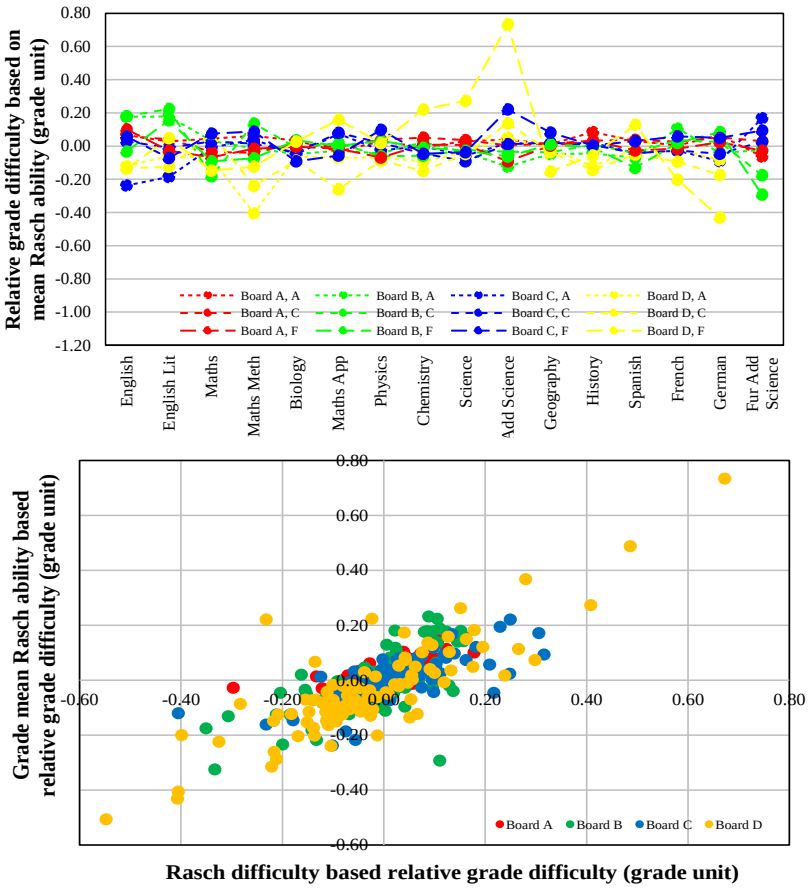

Figure 11 Distribution of relative grade difficulties in the unit of grade based on average grade Rasch ability (top) and their relationship with those derived using Rasch grade difficulty (bottom).

3.4 Relative grade difficulty based on mean GCSE score

This section explores the use of the concurrent GCSE performance measure to investigate inter-board comparability. The top graph in Figure 12 depicts the distribution of the average GCSE scores for the students included in the analysis. The mean GCSE score for a student was calculated as the average of numeric GCSE grades on the subjects (from the 16 subjects) that were taken by the student. The distribution is slightly negatively skewed and different from the ability distribution shown in the top graph in Figure 2. The ability distribution is relatively symmetric as a result of the non-linear transformation of the raw scores to ability measures through the Rasch model. The standard deviations of mean GCSE scores and Rasch abilities for all students included in the analysis are 1.54 and 3.08 (logits) respectively. The bottom graph in Figure 12 shows the relationship between Rasch ability and mean GCSE score for students from all the four exam boards. Mean GCSE score and ability are highly correlated (with a correlation of 0.98). This is expected as although the Rasch model represents a nonlinear regression of expected score (which is related to the mean GCSE score) on the latent trait there is strong linearity in the middle ability range. However, the relationship is not monotonic. That is, for similar values of ability, there can be a range of corresponding mean GCSE scores,

particularly in the lower to middle ability range. This reflects the fact that different students took different sets of subjects and different subjects have different difficulties. Further, there is large variability in difficulty between the subjects at lower to middle grades (see Figure 3).

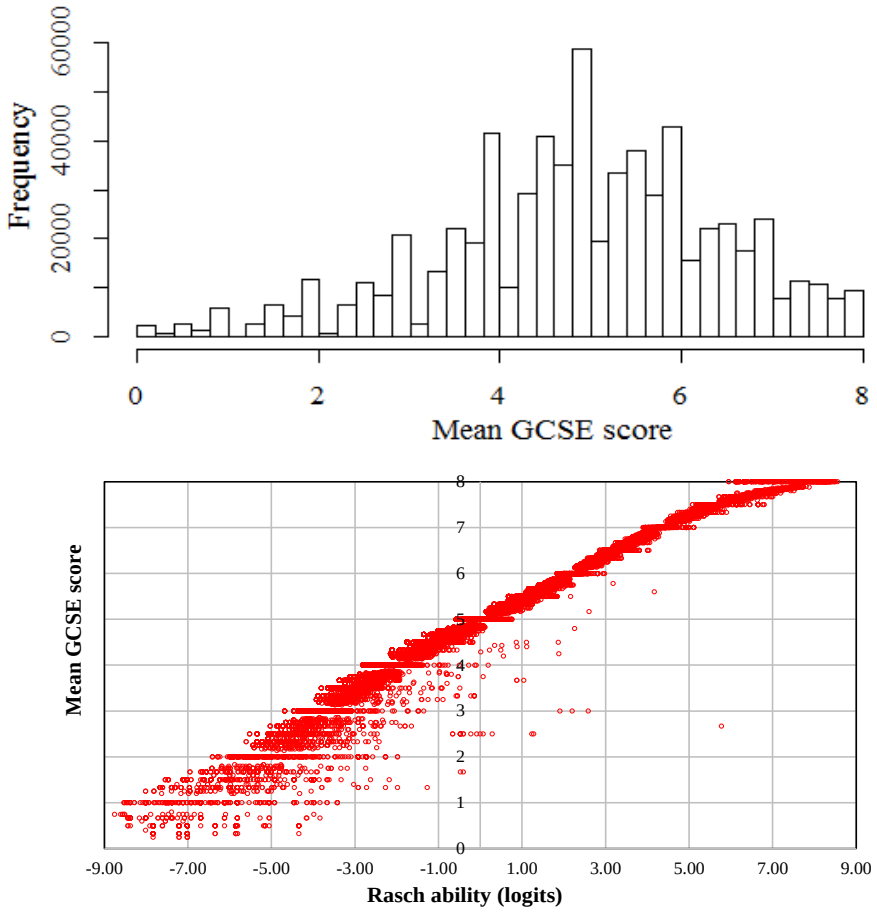

Figure 12 Distribution of mean GCSE score for all students (top) and the relationship between mean GCSE score and Rasch ability (bottom).

The top graph in Figure 13 shows the distributions of the average mean GCSE scores for students from the four exam boards at individual grades across the 16 subjects. Here the order of the subjects is the same as that in Figure 6. At the same grade in a subject, the higher the average mean GCSE score for a board, the harder its examination at this grade compared with the examinations from the other boards. There are similarities and differences between the distributions shown in Figure 13 and those shown in Figures 5 and 6. For the higher grades, variability in the average mean GCSE score between the subjects is relatively small, while that at the lower grades is large. For example, at grade A, the average mean GCSE score varies from 6.35 for Board D in English to 6.97 for Board A in German. At grade E, the average mean GCSE score varies from 2.82 in English for Board D to 4.55 in Further Additional Science for Board A. Variability in average mean GCSE score between boards is also general smaller at higher grades than that at lower grades. The bottom graph in Figure 13 shows the relationship between grade average mean GCSE score and Rasch grade difficulty and grade mean ability. Grade mean GCSE scores are

strongly correlated with the grade mean ability. This is expected as mean GCSE score and ability were calculated in the same way and are highly correlated.

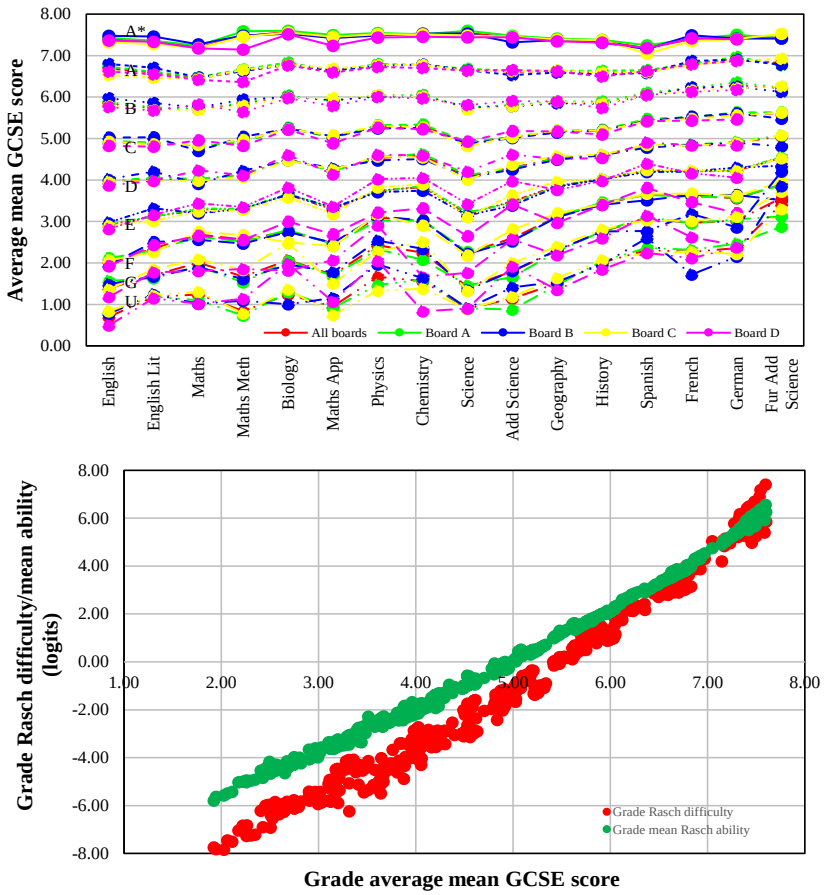

Figure 13 Distributions of average mean GCSE scores for students from different exam boards (top) and the relationships between grade average GCSE score, Rasch grade difficulty and grade mean ability (bottom).

Figure 14 shows how the average scores of students in each subject varies with their mean GCSE score. The curves are similar to the observed ICCs from the Rasch analysis shown in Figure 3. Since mean GCSE score is assumed to be a performance measure related to the constructs measured by the individual exams, the relative positions of the observed mean GCSE score derived ICCs reflect the relative difficulty of the subjects. Subjects on the left may be viewed as easy subjects while those on the right hard subjects. There are similarities and differences between the mean GCSE score derived ICCs and the Rasch derived ICCs in terms of the shape of the curves. Variability in the average score between the subjects is smaller for students with higher mean GCSE scores than those with lower mean GCSE scores. While the mean GCSE score derived ICCs are relatively linear at A and A*, the Rasch derived ICCs curved to the right at these two grades, reflecting the non-linear nature of the Rasch model.

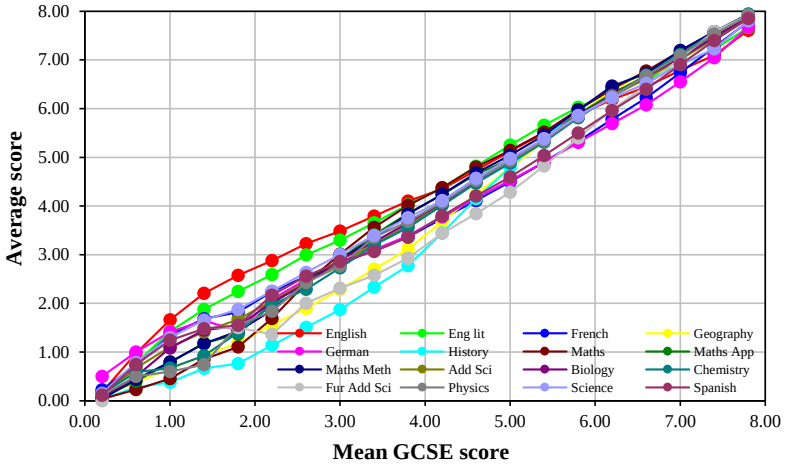

Figure 14 Distributions of subject average score with mean GCSE score for the 16 subjects.

Figure 15 shows how the average score of students from different exam boards varies with their mean GCSE score for English and Additional Science. As can be seen from the graph, for English, the variability in average score between the four boards is small in the lower grades. However, for Additional Science, the average scores of students from Board D were considerably lower than those of students from the other three boards, suggesting that its exam was substantially harder than those from the other boards at these grades. At the higher grades, variability in the average score between the boards is relatively small and the grades were more comparable.

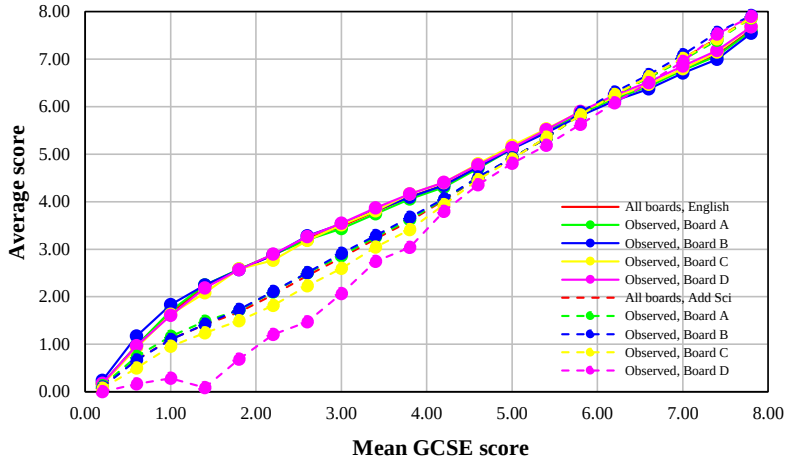

Figure 15 Distributions of subject average score with mean GCSE score for students from different exam board for English and Additional Science.

To see how students with similar mean GCSE scores performed on the exam at individual grades in a subject, the mean GCSE scores for students from all boards were grouped into 10 bands with similar number of students in each band for analysis. Figure 16 shows the distributions of students from different exam boards in each of the mean GCSE score bands achieving individual grades in English and

Additional Science. The solid lines represent the observed values for students from all four exam boards, and the symbols are the values for individual boards. These are similar to the observed category characteristic curves (CCCs) derived from the Rasch analysis discussed above. It is to be noted that the mean GCSE score bands used for English were different from those used for Additional Science in terms of the actual ranges of the absolute mean GCSE scores covered. The bands for each subject were determined by the distributions of the mean GCSE scores of the students from all four boards who took the subject. These mean GCSE score derived CCCs are similar to the Rasch derived CCCs. With an increase in mean GCSE score, the probability of achieving higher grades increases. Variability between the exam boards in the proportions of students in each mean GCSE score band achieving different grades can be seen to reflect inconsistency in standards between the boards.

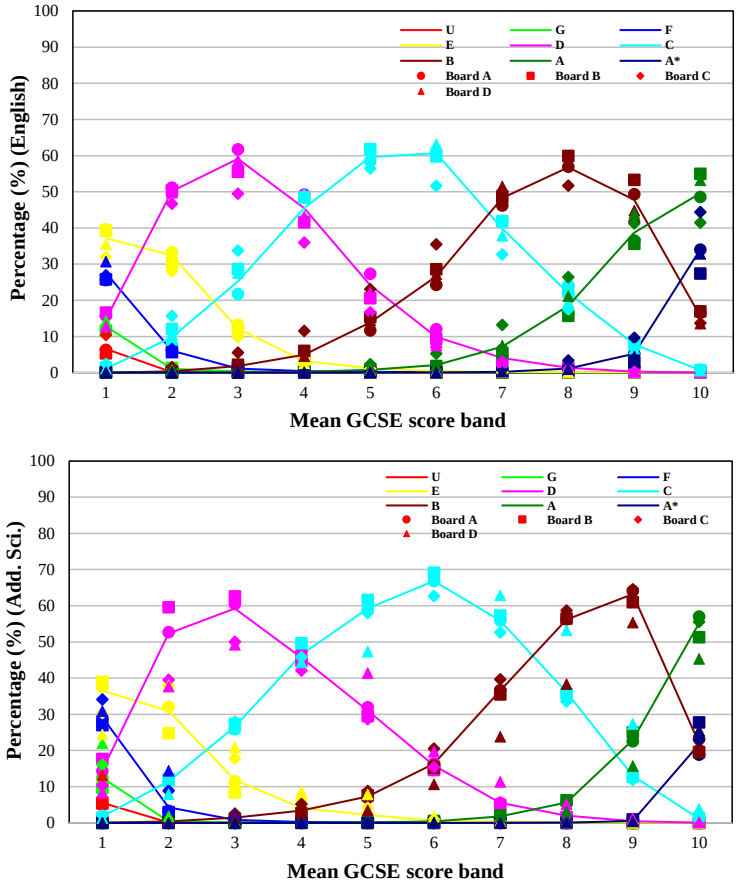

Figure 16 Distributions of proportions of students in each mean GCSE score band achieving different grades in English (top) and Additional Science (bottom).

If the average mean GCSE scores at individual grades shown in Figure 13 are assumed to reflect the difficulties of the grades, these can be expressed in the unit of grade using the procedure described above for the Rasch analysis. This is shown in Figure 17. The patterns shown in Figure 17 are broadly similar to those shown in Figures 10 and 11. However, the magnitude of the relative grade difficulties based on mean GCSE score is generally larger than those based on the Rasch grade difficulty and average grade ability. It is clear from Figure 17, the exams in Science and

Additional Science from Board D were harder than those from the other boards at most grades, but Methods in Mathematics, French and German were easier. For Board B, English Literature was slightly harder than the other three boards, but Spanish and Further Additional Science were easier. Biology was slightly easier for Board C, but Additional Science was more difficult.

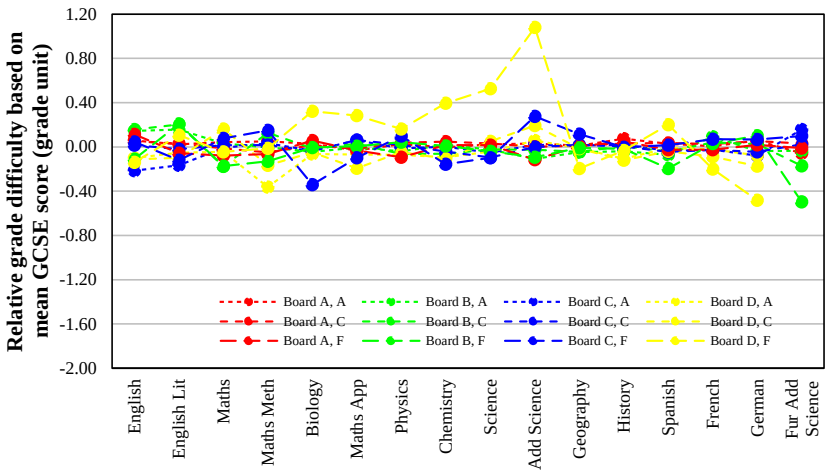

Figure 17 Distribution of relative grade difficulty in the unit of grade derived using mean GCSE score.

The top graph in Figure 18 shows the relationship between the relative grade difficulties derived based on Rasch grade difficulty and those derived using grade average ability and mean GCSE score. Differences in the different difficulty measures again reflect the differences in how grade difficulty was defined. If grade difficulties are used as measures of standards, the different definitions will result in slightly different levels of the underlying attainment at the same grade. The relationship between the relative grade difficulties derived using grade average mean GCSE score and mean ability is very strong (see the bottom graph in Figure 18).

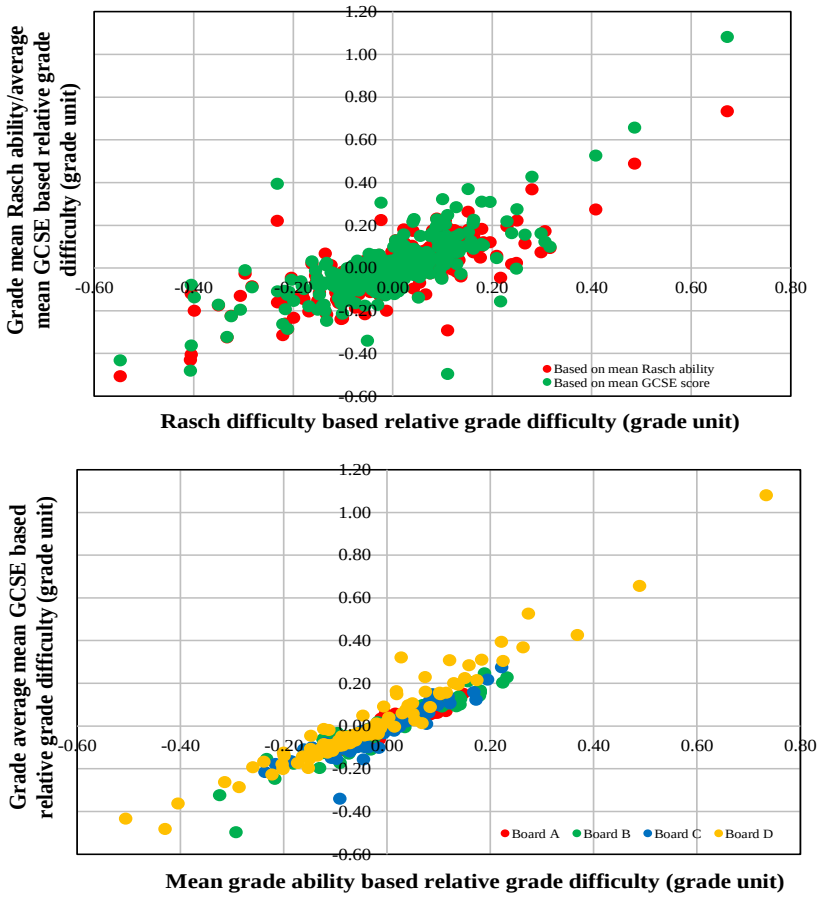

Figure 18 The relationships between the relative grade difficulties derived using Rasch grade difficulty, grade mean Rasch ability and average mean GCSE score.

3.5 Comparison of the effects of aligning inter-board standards on grade outcomes

This section looks at the effects of aligning standards between exam boards.

3.5.1 Aligning inter-board standards by changing grade boundary scores

For GCSEs, grade boundaries can be viewed as the operationalization of performance standards, and aligning standards between exam boards for a specific subject would necessarily involve adjusting the boundary marks for some exam boards. 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 standards with the standards for all boards will be:

$$b'_k = b_k - wd_{k,RG} \tag{7}$$

Most of the exams studied here are unitized and the UMS mark scale is used. When UMS is used for a qualification, the grade interval at subject level is 10% of the maximum available uniform mark. Application of the new grade boundaries to the UMS mark distribution for a subject will produce a new grade distribution which will

be different from the original grade distribution. Change in grade distribution will depend on the UMS mark distribution and the magnitude of adjustment in grade boundary scores.

Figure 19 shows the changes in percentages of candidates being awarded individual grades across the 16 subjects and the four exam boards after aligning the standards to the overall standards defined by students from all four boards based on the Rasch grade difficulty, mean grade Rasch ability and average grade mean GCSE score. In the case of using Rasch grade difficulty, to make the results more comparable with those from the inter-board screening method, the average grade gap for a subject used in Equation (6) was estimated using the overall grade gaps within the subject when generating the new grade boundaries. Further, the average gap for grade A* to C was based on the gap between A* and C, and the average grade gap for D to F was based on the gap between C and F. Positive values indicate an increase in percentage of students classified into the grade after alignment of standards between the exam boards, while negative values indicate a decrease. At A, the change varies from -1.4% in Further Additional Science for Board A to 2.9% in Physics for Board D. At grade E, this varies from -2.0% in Further Additional Science for Board C and 1.4% in Further Additional Science for Board A. Mathematics was not included as both linear and unitised specifications were offered by all the four boards which made it difficult to estimate the changes in boundary scores that were required. The patterns of estimated changes in grade outcomes based on grade mean Rasch ability and average mean GCSE score are broadly similar to the patterns produced based on Rasch grade difficulty (see the bottom graph in Figure 19). However, considerable differences in grade outcomes between the different approaches also exist. For example, large changes in grade outcomes for Spanish for Board B were predicted based on Rasch grade difficulties, while substantial changes were predicted for Board C based on grade mean ability and average mean GCSE score. As indicated earlier, the differences reflect the difference in the definition of grade difficulty by the different approaches. As with the relative grade difficulties, the changes derived using grade average mean GCSE score are strongly correlated with the changes derived using grade mean Rasch ability. The majority of the predicted changes from the different methods are within the tolerances estimated for the predictions of subject outcomes using mean GCSE score with similar sample size by Benton and Sutch (2014).

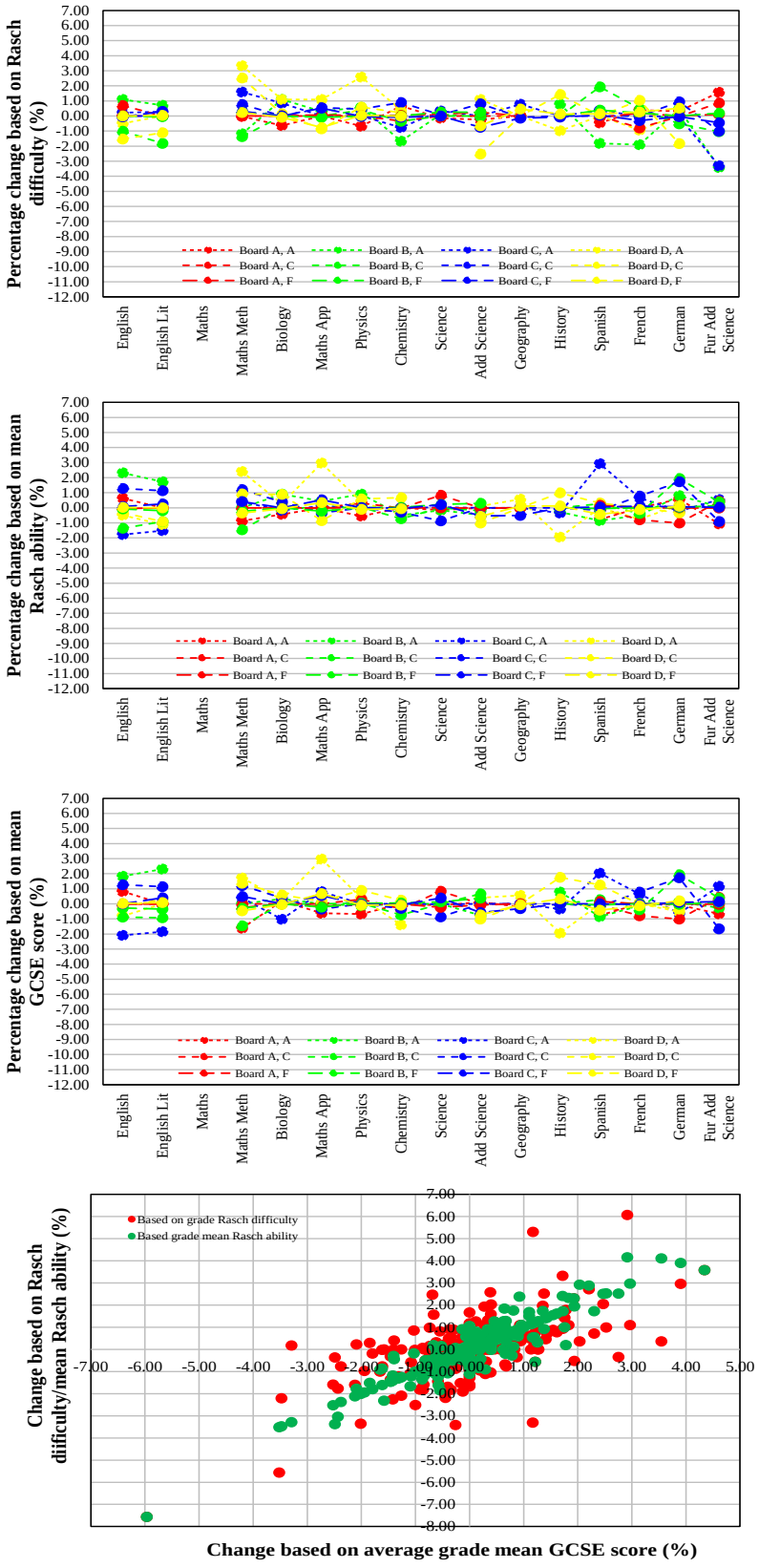

Figure 19 Changes in predicted percentage of candidates being classified into individual grades after aligning standards between the exam boards based on Rasch grade difficulty, grade mean Rasch ability and grade mean GCSE score (top three) and their relationships (bottom).

3.5.2 Comparison with inter-board screening based on mean GCSE score

For the inter-board statistical screening procedure using mean GCSE score, the change at a specific grade k in a subject for an exam board can be calculated as the difference between the expected percentage (A_k) and the original observed percentage ($A_{k,0}$) at the grade (see Equation 1). Figure 20 shows changes in percentage of students classified into individual grades using the inter-board screening procedure (see Tables A5a to A5d in Appendix A). At grade A, the change varied from -2.1% in Application of Mathematics for Board B to 3.7% in Further Additional Science for Board C. At grade E, the change varied from -3.2% in Mathematics for Board D to 1.4% in English for Board C.

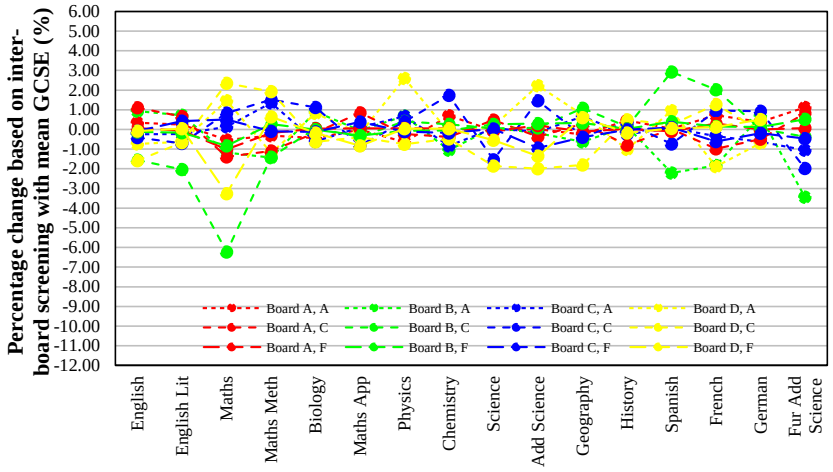

Figure 20 Changes in predicted percentage of candidates classified into individual grades after aligning standards between the exam boards, based on inter-board screening using mean GCSE score.

Figure 21 compares changes in grade outcomes predicted using the inter-board screening approach with mean GCSE score with those predicted using Rasch grade difficulty, mean grade ability and grade average mean GCSE score discussed above. Although the changes are general positively correlated, substantial differences exist for some subjects between the inter-board screening approach and the other methods. These differences to a large extent reflect the difference in the operationalisation of standards alignment. The inter-board screening approach retains the original grade boundary scores and therefore the original performance standards, but the students were redistributed within the grades. In contrast, the other three methods involve changing grade boundary scores to align performance standards. The requirement for boundary scores to be integer would also have introduced additional variability in changes in grade outcomes.

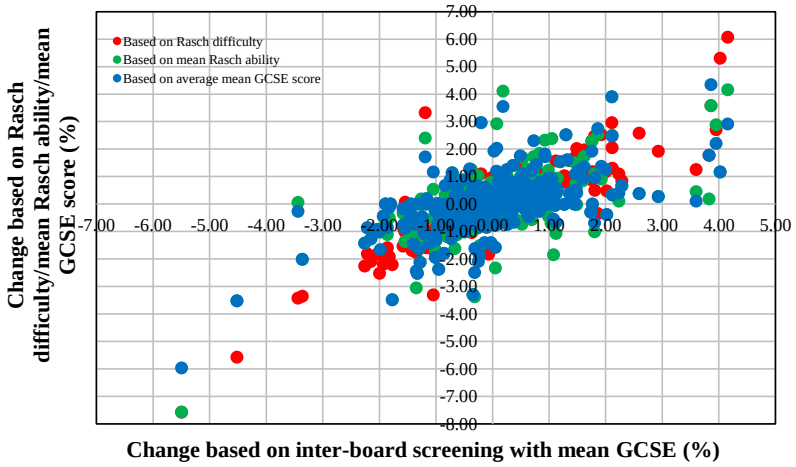

Figure 21 Relationship between predicted percentage changes in grade outcomes estimated using inter-board screening with mean GCSE score and those using Rasch grade difficulty, grade mean ability and average mean GCSE score.

As indicated earlier, the grade difficulties vary between the subjects. However, the mean GCSE score for a student, which was calculated as the mean of the numerical grades achieved on all the subjects that the student took, does not consider the differences in difficulty between the subjects. Since the Rasch ability estimated for a student takes into account the variability in grade difficulty between the subjects, it would be a more appropriate performance measure than the mean GCSE score when used for comparing students. For example, a mean GCSE score obtained from a set of more difficult subjects may represent a higher level of the underlying attainment than the same mean GCSE score obtained from a set or easier subjects. There can also be situations where, for a specific subject, the total number of other subjects taken by the students was small and varies between the exam boards and the subject itself makes a large contribution to the overall mean GCSE score. The contribution of the subject itself to the overall mean GCSE score can therefore vary between the subjects, as demonstrated by Table A6 in Appendix A which shows the correlations between the 16 subjects and the mean GCSE score and the mean GCSE score calculated without the subject itself.

The Rasch ability measure may also be used as a performance measure for inter-board screening and Figure 22 compares changes in grade outcomes using the Rasch ability measure with the existing inter-board screening procedure described above and those using mean GCSE score (see Tables A5a to A5d and A7a to A7d). These changes are highly consistent at most grades across the subjects, suggesting that taking into account the difference in subject difficulty at individual grades using the Rasch model makes little or no difference in predicted changes compared with the changes estimated using mean GCSE score. These findings are consistent with the findings by Benton (2015) who compared the predicted A level subject outcomes using mean GCSE score and more complex grade and subject based models and found that the two methods produced similar results. However, in the present study, for Board D, the expected changes for several grades derived using Rasch ability are considerably different from those derived using mean GCSE score, and this is shown in Figure 23.

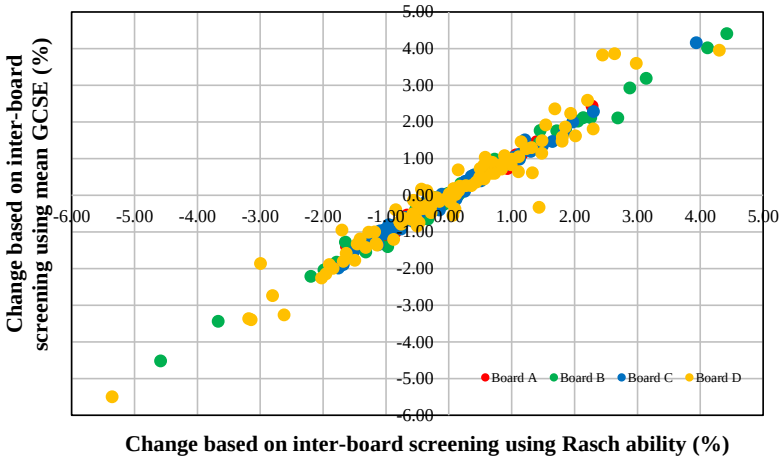

Figure 22 Relationship between predicted changes in grade outcomes estimated using inter-board screening with mean GCSE score and Rasch ability.

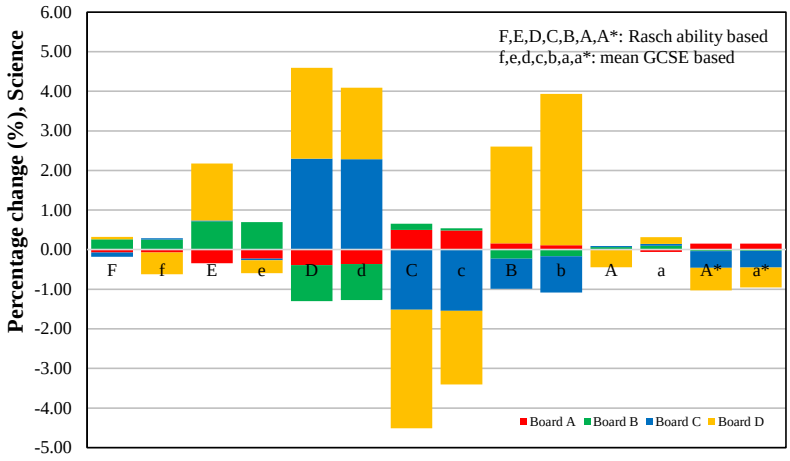

Figure 23 Comparison of changes at individual grades predicted based on Rasch ability with the existing inter-board screening procedure (capital letters) and those predicted based on mean GCSE score (lower case letters) for Science.

The differences in expected changes produced using the two measures with the existing screening procedure to an extent reflects the difference in the distributions of the two measures between the grades and their relationships. Figure 24 shows the relationship between mean GCSE score and Rasch ability for students from the four exam boards who took GCSE Science. As with the relationship for students from all four boards shown in Figure 12, there is substantial variability in mean GCSE scores in the low to middle ability range. Further, the mean GCSE scores for students from Board D appear to be slightly higher than the scores for students from the other boards with similar abilities.

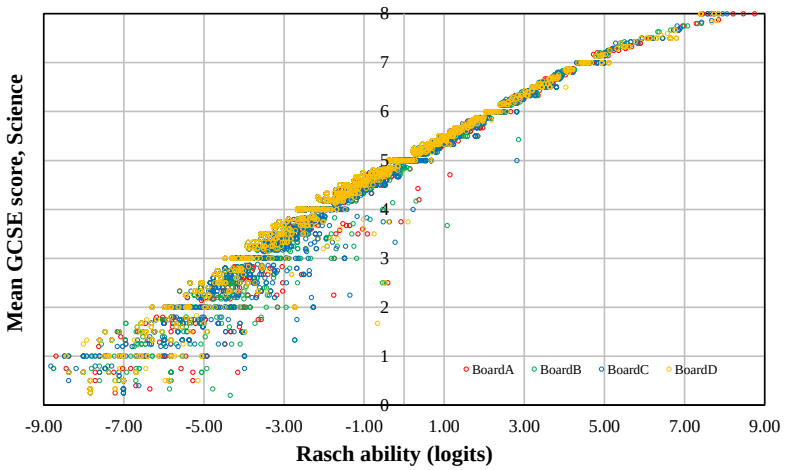

Figure 24 Relationship between mean GCSE score and Rasch ability for students from the four exam boards who took the Science subject.

The relationship between mean GCSE score and Rasch ability was examined further using a "value added" or residual-based approach proposed by He and Tymms (2014). If the mean GCSE scores and the Rasch abilities are standardised to have a mean of 0 and standard deviation of 1, then the two measures are directly comparable. The difference or residual, v_i, between the two standardised measures, $x_{i,GCSE}$ and $x_{i,Ability}$, for student i is $v_i = x_{i,GCSE} - x_{i,Ability}$ which may be termed "value added", would be a measure of the relative "progress" made by the student in relation to the average "progress" made by all students. When the difference is positive, the student's mean GCSE score is higher than the mean of all students with similar level of ability. If the difference if negative, his/her mean GCSE score is lower than the average. For an exam board, a mean "value added" or residual measure can be calculated. This average residual measure may be used as a measure for identifying any systematic difference in the relationship between the mean GCSE score and ability among the exam boards. As the average residual measure for all boards is zero, an exam board with a positive (negative) average residual will have higher (lower) average mean GCSE score than the mean of all boards with similar level of ability. The top graph in Figure 25 depicts the average residual measures for students taking Science from the four exam boards at individual grades. As can be seen, from grades B to F, the average residual measure for students from Board D are significantly higher than those for students from the other three boards. This indicates that for similar abilities, students from Board D generally had higher mean GCSE scores than those from the other boards, particularly from grades B to F. For Board C, the average residual measures at grade E and F are slightly lower than those from the other boards. The bottom graph in Figure 25 shows the total number of subjects taken by the students from the four boards who also took Science. The average number of subjects taken by students from Board D was about one less than students from the other boards. When the total number of subjects taken was small, the subject itself would make a substantial contribution to the overall mean GCSE score. Science was a relatively easy subject which may partly explain the higher residuals estimated for students from Board D. Figure 25 suggests that the use of

mean GCSE score as a performance measure with the inter-board screening procedure would not sufficiently reflect the different difficulties for different subjects.

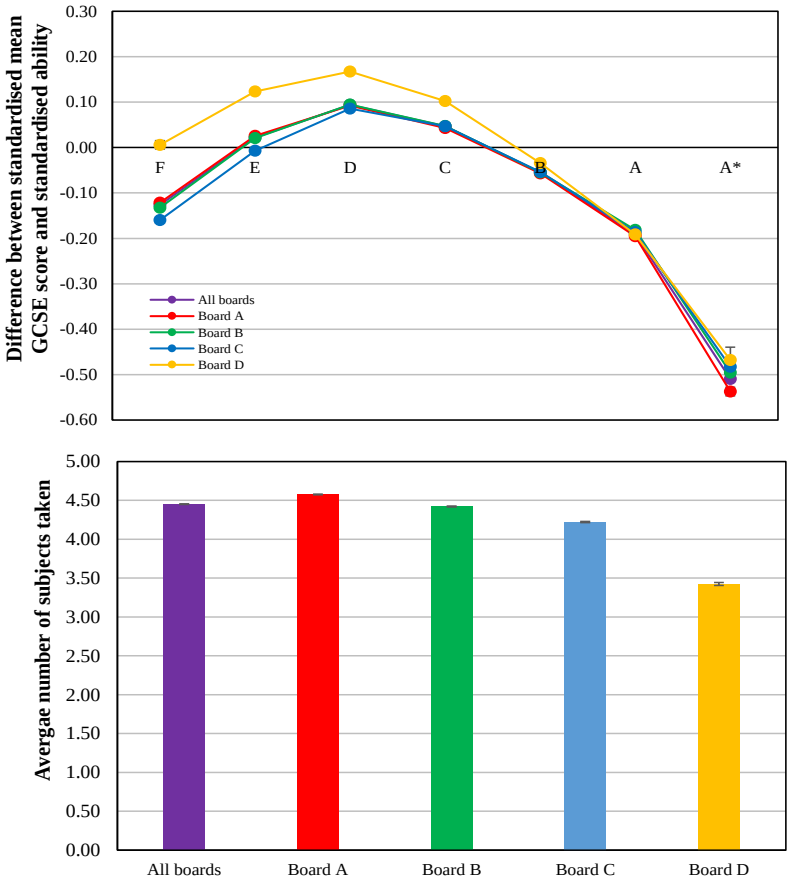

Figure 25 Difference between standardised mean GCSE score and ability at individual grades for students taking GCSE Science (top) and the average total number of subjects taken (from the 16 subjects included in the analysis) (bottom).

4. Concluding remarks

Results from DSF analysis of the 16 GCSE subjects using the partial credit Rasch model suggested that the data fitted the model reasonably well and the subjects shared a common construct measured by the Rasch ability. The grade (step) difficulty at a specific grade in an exam derived using the Rasch model was used as a measure of the performance standard at the grade. The existence of significant differential step functioning (DFS) with respect to exam boards at a specific grade would therefore suggest difference in standard between the exam boards. For most of the grades across the 16 subjects studied, the effect of DSF is small or moderate. Comparability of standards in the exams between the exam boards for the higher grades was found to be higher than that for the lower grades. This may partly reflect the difficulty in setting grade boundaries for the lower grades where evidence of performance on the exam is likely to be poor.

The relative grade difficulties derived using the Rasch grade difficulty, grade mean ability and average mean GCSE score were broadly consistent. However,

considerable differences exist between the methods, reflecting the difference in how grade difficulty (standard) was conceptualised in the different approaches.

It was found that changes in grade distributions after aligning standards between the exam boards based on Rasch grade difficulty, grade mean ability and grade average mean GCSE score were broadly consistent with those estimated using the existing inter-board statistical screening method that uses the mean GCSE score as a performance measure to compare grade distributions between the exam boards. This may suggest that both mean GCSE score and Rasch ability may be used as measures of the constructs assessed by the individual examinations. However, there were also considerable differences in the results for some subjects between the different methods, reflecting the differences in the definition of grade difficulty and the operationalization of standards alignment.

The use of Rasch ability as a performance measure with the existing inter-board screening procedure produced results which were closely similar to those produced using mean GCSE score. Since the ability measure takes into account the difference in difficulty between the subjects, its use with the existing inter-board screening procedure is likely to produce fairer results than the use of mean GCSE score, particularly for subjects where the total number of other subjects taken by the students was small and varies between exam boards and the subjects are either too hard or too easy.

It is worth noting that while the current inter-board statistical screening method employs an empirical approach when establishing the relationship between the subject grade distribution and the concurrent GCSE performance distribution, the Rasch analysis approach is based on a measurement model which links the examinations with a common construct explicitly. The use of the partial credit Rasch model and differential step functioning to investigate comparability of examination standards in GCSEs between exam boards represents a new approach. As demonstrated in this study, this approach could be used to validate some of the existing methods used for maintaining inter-board comparability.

References

Akour, M. Sabah, S. and Hammouri, H. (2015) Net and Global Differential Item Functioning in PISA Polytomously Scored Science Items: Application of the Differential Step Functioning Framework. *Journal of Psychoeducational Assessment* 33, 166–176.

Benton, T. and Sutch, T. (2014). Analysis of use of Key Stage 2 data in GCSE predictions. Ofqual, Coventry, UK. Available online at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/429074/2014-06-16-analysis-of-use-of-key-stage-2-data-in-gcse-predictions.pdf

Benton, T. (2015) Can we do better than using 'mean GCSE grade' to predict future outcomes? An evaluation of Generalised Boosting Models, *Oxford Review of Education* 41, 587-607.

Bramley, T. (2011) Subject difficulty – the analogy with question difficulty. Research Matters: A Cambridge Assessment Publication, Special issue 2: Comparability, 27-33.

Bramley, T. (2016). *The effect of subject choice on the apparent relative difficulty of different subjects*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Clauser, B., and Mazor, K. (1998) An NCME instructional module on using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice* 17, 31-44.

Coe, R. (2008) Comparability of GCSE examinations in different subjects: An application of the Rasch model. *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, Report for SCORE (Science Community Supporting Education)*, CEM Centre, Durham University. Available online at: http://www.score-education.org/media/3194/relativedifficulty.pdf

Dai, G., Han, K., Hu, H., and Colarelli, S. (2011) Cultural differences and measurement invariance of selection tools: A case of examining Chinese NEO PI-R conscientiousness scale. *Journal of Chinese Human Resource Management* 1, 95 - 114.

El-Komboz, B., Zeileis, A. and Strobl, C. (2014) Detecting Differential Item and Step Functioning with Rating Scale and Partial Credit Trees. Technical Report Number 152, 2014 Department of Statistics, University of Munich. Available online at: http://epub.ub.uni-muenchen.de/17984/1/TR152_pctrees.pdf

Gattamorta, K. and Penfield, R. (2012) A Comparison of Adjacent Categories and Cumulative Differential Step Functioning Effect Estimators. *Applied Measurement in Education* 25, 142-161.

Gattamorta, K., Penfield, R. and Myers, N. (2012) Modelling Item-Level and Step-Level Invariance Effects in Polytomous Items Using the Partial Credit Model. *International Journal of Testing* 12, 252-272.

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, 211-239.

He, Q. and Stockford, I. (1015) Inter-Subject Comparability of Exam Standards in GCSE and A Level. Ofqual: Coventry, UK. Available online at:

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/486936/3-inter-subject-comparability-of-exam-standards-in-gcse-and-a-level.pdf

He, Q. and Tymms, P. (2014). The principal axis approach to value added calculation. *Educational Research and Evaluation* 20, 25-43.

Kolen, M. and Brennan, R. (2014) *Test equating, scaling, and linking. Methods and practices (Third Edition).* New York, NY. Springer.

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

Linacre, J. (2015) Winsteps® Rasch measurement computer program User's Guide. Beaverton, Oregon: Winsteps.com.

Lockyer, C. and Newton, P. (2015) Inter-subject comparability: a review of the technical literature. Ofqual: Coventry.

Masters, G. (1982) A Rasch model for partial credit scoring. *Psychometrika* 47, 149-74.

Milfont, T. and Fischer, R. (2010) Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological research* 3, 111-121.

Miller, T., Saad, C. and Childs, R. (2010). Detecting Differential Item Functioning and Differential Step Functioning Due to Differences that Should Matter. *Practical Assessment, Research & Evaluation* 15(10). Available online: http://pareonline.net/getvn.asp?v=15&n=10.

Millsap, R. (2011) *Statistical approaches to measurement invariance*. New York, USA: Routledge.

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.

Office of Qualifications and Examinations Regulation (Ofqual) (2012) Review of standards in GCE A level Critical Thinking. Ofqual: Coventry, UK. Available online at: http://dera.ioe.ac.uk/14200/1/2012-04-27-review-of-standards-in-gce-a-level-critical-thinking.pdf

Office of Qualifications and Examinations Regulation (Ofqual) (2011) Review of Standards in Art and Design (GCSE 1999/2009, GCE 1999/2099). Available online at: http://www.rewardinglearning.org.uk/docs/regulation/standards_reports/review_of_standards_in_art_and_design_gce_gcse.pdf

Opposs, D. (2015) Inter-subject comparability: an international review. Ofqual: Coventry, UK. Available online at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/486939/4-inter-subject-comparability-an-international-review.pdf

Oshima, T. and Morris, S. (2008) Raju's Differential Functioning of Items and Tests (DFIT). *Items – Instructional Topics in Educational Measurement* 27(3). National Council on Measurement in Education.

Penfield, R., Alvarez, K., and Lee, O. (2009). Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education 22*, 61–78.

Penfield, R., Gattamorta, K., and Childs, R. (2009) An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement: Issues and Practice* 28, 38–49.

Rasch, G. (1960) *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Paedagogiske Institute.

Reckase, M. (2009) *Multidimensional item response theory*. New York, USA: Springer-Verlag.

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

Reise, S., K. Widaman, and R. Pugh. 1993. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin* 114, 552-566.

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: 205-231.

Taylor, M. (2013). GCSE (and level 1/2 project and functional skills) statistical screening. Assessment and Qualifications Alliance (AQA).

Wright, B. and Masters, G. (1982) *Rating scale analysis: Rasch measurement*. Chicago, USA: MESA Press.

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

Yen, W. (1993) Scaling performance assessment: Strategies for managing local item dependence. *Journal of Educational Measurement* 20, 187-213.

Zhang, B., Fokkema, M., Cuijpers, P., Smits, N. and Beekman, A. (2011) Measurement invariance of the Center for Epidemiological Studies Depression Scale (CES-D) among Chinese and Dutch elderly. *Medical Research Methodology* 20, 11:74

Appendix A Additional tables

Table A1 Inter-subject correlations for the 16 subjects studied.

	English	English Lit	French	Geog.	German	His.	Maths	Maths (App)	Maths (Meth)	Add Sci	Bio.	Chem.	Fur Add Sci	Phy.	Sci.	Spanish
English	1 (428291)	0.79 (285710)	0.62 (102223)	0.73 (143311)	0.61 (37149)	0.73 (159939)	0.70 (292998)	0.68 (7487)	0.69 (6003)	0.65 (205882)	0.61 (93391)	0.57 (93823)	0.52 (15811)	0.55 (94844)	0.69 (126974)	0.57 (56284)
English Lit		1 (404943)	0.59 (105888)	0.73 (144860)	0.59 (38718)	0.74 (161067)	0.66 (353986)	0.64 (7830)	0.68 (6389)	0.64 (206106)	0.60 (96172)	0.57 (97369)	0.52 (16467)	0.55 (98509)	0.67 (133399)	0.55 (57881)
French			1 (148289)	0.64 (63561)	0.76 (4516)	0.63 (72002)	0.59 (131234)	0.62 (3323)	0.65 (2641)	0.58 (76519)	0.62 (46188)	0.62 (46556)	0.58 (8849)	0.60 (47014)	0.52 (38466)	0.73 (6749)
Geography				1 (213827)	0.62 (22707)	0.81 (55895)	0.73 (189899)	0.76 (4868)	0.77 (4054)	0.75 (112712)	0.72 (57299)	0.69 (57701)	0.67 (9946)	0.69 (58361)	0.75 (62104)	0.61 (34294)
German					1 (51497)	0.62 (25905)	0.59 (45387)	0.60 (1450)	0.63 (1308)	0.58 (24573)	0.60 (19633)	0.61 (19531)	0.59 (2658)	0.58 (19808)	0.54 (11728)	0.69 (1283)
History						1 (234704)	0.69 (209746)	0.69 (5032)	0.73 (4140)	0.70 (125470)	0.69 (62043)	0.66 (62391)	0.63 (11305)	0.64 (63178)	0.70 (68821))	0.61 (38971)
Maths							1 (588216)	0.87 (263)	0.72 (312)	0.75 (293101)	0.70 (110541)	0.73 (110771)	0.69 (21094)	0.76 (112711)	0.75 (203988)	0.54 (74256)
Maths (App)								1 (12413)	0.93 (8179)	0.77 (5945)	0.69 (3791)	0.71 (3769)	0.73 (367)	0.74 (3638)	0.79 (3229)	0.57 (1681)
Maths (Meth)									1 (11428)	0.77 (4650)	0.71 (3057)	0.75 (3057)	0.70 (205)	0.76 (2953)	0.78 (4016)	0.62 (1414)
Add Sci										1 (315685)	0.77 (148)	0.63 (112)	0.83 (17760)	0.63 (123)	0.85 (147032)	0.54 (43123)
Biology											1 (129732)	0.84 (120816)	0.73 (17)	0.82 (121474)	0.68 (88)	0.61 (25067)
Chemistry												1 (128723)	0.69 (13)	0.85 (122978)	0.52 (43)	0.61 (25382)
Fur Add Sci													1 (22917)	0.65 (19)	0.76 (3579)	0.52 (4696)
Physics														1 (129829)	0.71 (74)	0.59 (25745)
Science															1 (239465)	0.46 (22803)
Spanish																1 (84040)

Table A2a Grade difficulties for the 16 GCSE subjects derived using the partial credit Rasch model (PCM) based on students from all four boards

	Difficulty (logits)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	-7.84	-5.79	-4.02	-1.71	0.91	3.51	6.42
English Lit	-7.22	-5.27	-3.64	-1.74	0.50	3.11	6.11
French	-5.20	-3.38	-1.82	0.18	2.19	3.83	5.78
Geography	-5.16	-3.40	-2.01	-0.49	1.21	3.00	5.38
German	-5.15	-3.44	-1.91	0.17	2.33	4.22	6.40
History	-4.35	-2.77	-1.62	-0.40	1.05	2.85	5.34
Maths	-5.87	-4.50	-3.44	-1.79	0.85	2.87	5.02
Maths App	-6.09	-4.30	-2.93	-1.15	1.20	3.05	5.34
Maths Meth	-5.98	-4.16	-2.88	-1.31	0.98	2.91	5.20
Add Science	-6.36	-4.44	-2.87	-0.98	1.25	3.45	6.13
Biology	-6.12	-4.49	-3.04	-1.14	0.98	3.12	5.79
Chemistry	-5.82	-4.26	-2.86	-0.98	1.10	3.09	5.47
Fur Add Science	-4.34	-2.64	-1.12	0.57	2.13	3.62	5.52
Physics	-5.80	-4.35	-3.06	-1.12	1.06	3.10	5.57
Science	-7.02	-4.88	-3.21	-1.21	1.24	3.77	7.18
Spanish	-4.79	-3.13	-1.75	0.01	1.81	3.33	5.14

Table A2b Grade difficulties for the 16 GCSE subjects for Board A derived using the PCM.

	Difficulty (logits)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	-7.51	-5.71	-3.96	-1.60	0.93	3.59	6.49
English lit	-7.25	-5.29	-3.59	-1.68	0.50	3.14	6.08
French	-5.28	-3.51	-1.90	0.20	2.27	3.84	5.68
Geography	-5.19	-3.42	-2.01	-0.50	1.18	2.98	5.38
German	-5.13	-3.42	-1.85	0.24	2.41	4.31	6.47
History	-4.33	-2.77	-1.61	-0.37	1.13	2.98	5.50
Maths	-5.94	-4.45	-3.33	-1.83	0.89	2.91	5.20
Maths App	-6.00	-4.26	-2.88	-1.10	1.19	3.03	5.25
Maths Meth	-5.96	-4.10	-2.84	-1.35	1.02	3.03	5.39
Add Science	-6.42	-4.54	-2.93	-1.00	1.26	3.50	6.26
Biology	-6.05	-4.53	-3.11	-1.15	0.99	3.14	5.85
Chemistry	-5.87	-4.39	-2.91	-0.90	1.16	3.08	5.38
Fur Add Science	-4.89	-2.88	-1.09	0.68	2.17	3.54	5.30
Physics	-5.88	-4.49	-3.14	-1.08	1.09	3.10	5.63
Science	-6.82	-4.83	-3.18	-1.15	1.28	3.81	7.40
Spanish	-4.85	-3.13	-1.70	0.05	1.85	3.38	5.21

Table A2c Grade difficulties for the 16 GCSE subjects for Board B derived using the PCM.

	Difficulty (logits)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	-7.82	-5.77	-3.97	-1.67	1.06	3.74	6.69
English Lit	-6.92	-5.07	-3.47	-1.54	0.78	3.27	6.29
French	-5.49	-3.55	-1.84	0.22	2.12	3.87	6.07
Geography	-5.09	-3.40	-2.02	-0.52	1.13	2.98	5.44
German	-5.11	-3.43	-2.03	0.11	2.28	4.09	6.06
History	-4.36	-2.79	-1.63	-0.40	1.04	2.82	5.21
Maths	-5.92	-4.47	-3.43	-2.04	1.04	2.85	5.21
Maths app	-5.97	-4.21	-2.90	-1.15	1.25	3.12	5.41
Maths Meth	-6.00	-4.23	-2.87	-1.14	1.18	2.98	5.44
Add Science	-6.53	-4.61	-3.05	-1.03	1.22	3.32	5.76
Biology	-6.12	-4.46	-3.01	-1.12	0.99	3.11	5.63
Chemistry	-5.86	-4.24	-2.88	-1.06	1.04	3.14	5.66
Fur Add Science	-4.14	-2.88	-1.74	-0.07	1.74	3.54	5.77
Physics	-5.84	-4.31	-3.06	-1.21	1.00	3.10	5.48
Science	-6.84	-4.90	-3.29	-1.21	1.24	3.75	6.89
Spanish	-5.35	-3.50	-1.97	-0.16	1.53	3.11	5.06

Table A2d Grade difficulties for the 16 GCSE subjects for Board C derived using the PCM.

	Difficulty (logits)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	-7.46	-5.62	-4.00	-1.77	0.86	3.32	5.99
English Lit	-7.27	-5.43	-3.80	-1.80	0.50	2.98	5.78
French	-5.01	-3.18	-1.69	0.20	2.15	3.82	5.80
Geography	-4.93	-3.22	-1.88	-0.40	1.29	3.01	5.25
German	-5.02	-3.34	-1.84	0.16	2.26	4.10	6.24
History	-4.30	-2.73	-1.59	-0.36	1.07	2.86	5.36
Maths	-5.78	-4.53	-3.52	-1.79	0.80	2.87	4.98
Maths App	-6.22	-4.49	-3.03	-1.10	1.31	3.10	5.33
Maths Meth	-5.84	-4.15	-2.95	-1.34	0.91	2.90	4.97
Add Science	-5.90	-4.02	-2.54	-0.89	1.20	3.44	6.15
Biology	-6.21	-4.51	-2.99	-1.13	0.91	3.05	5.71
Chemistry	-5.42	-3.96	-2.67	-0.99	0.97	3.08	5.56
Fur Add Science	-3.76	-2.18	-0.96	0.58	2.23	3.87	6.08
Physics	-5.54	-4.13	-2.87	-0.99	1.10	3.12	5.52
Science	-7.04	-4.93	-3.26	-1.37	1.12	3.65	6.43
Spanish	-4.68	-3.12	-1.82	-0.01	1.84	3.29	5.04

Table A2e Grade difficulties for the 16 GCSE subjects for Board D derived using the PCM.

	Difficulty (logits)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	-7.75	-5.83	-4.15	-1.89	0.86	3.33	6.20
English Lit	-6.90	-5.17	-3.72	-1.89	0.42	3.06	6.16
French	-5.45	-3.57	-1.93	0.04	1.99	3.71	5.83
Geography	-5.43	-3.67	-2.21	-0.54	1.30	3.06	5.35
German	-5.90	-4.04	-2.43	-0.09	2.16	4.00	6.06
History	-4.48	-2.87	-1.74	-0.58	0.85	2.71	5.26
Maths	-6.07	-4.42	-3.11	-1.24	1.34	2.97	4.83
Maths App	-5.86	-4.08	-2.95	-1.55	0.89	2.80	4.95
Maths Meth	-5.85	-4.22	-2.96	-1.50	0.58	2.16	4.19
Add Science	-5.12	-3.55	-2.36	-0.82	1.49	3.56	5.88
Biology	-5.93	-4.19	-2.90	-1.33	0.89	3.00	5.50
Chemistry	-6.24	-4.30	-3.00	-1.38	0.88	2.90	5.21
Fur Add Science							
Physics	-5.36	-4.07	-2.89	-1.30	0.94	2.99	5.18
Science	-6.27	-4.52	-3.11	-1.25	1.49	3.65	6.44
Spanish	-4.61	-3.05	-1.67	0.03	1.82	3.36	5.11

Table A3a Relative grade difficulties for the 16 GCSE subjects for Board A derived using the PCM.

	Difficulty (logits)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	0.33*	0.09*	0.06*	0.10*	0.02	0.08*	0.07*
English Lit	-0.02	-0.02	0.04*	0.06*	0.00	0.03*	-0.03
French	-0.08	-0.12*	-0.08*	0.02	0.08*	0.02	-0.10*
Geography	-0.04	-0.02	0.00	-0.01	-0.03	-0.01	0.00
German	0.02	0.02	0.06	0.07	0.08*	0.09*	0.07
History	0.02	0.00	0.01	0.03	0.08*	0.12*	0.15*
Maths	-0.07	0.05	0.11*	-0.05*	0.04*	0.04	0.18*
Maths App	0.09	0.04	0.05	0.05	-0.02	-0.02	-0.09
Maths Meth	0.02	0.05	0.04	-0.04	0.03	0.12	0.20
Add Science	-0.06*	-0.09*	-0.06*	-0.02	0.02	0.05*	0.13*
Biology	0.06	-0.05	-0.07	-0.01	0.01	0.02	0.07*
Chemistry	-0.06	-0.13	-0.05	0.08*	0.07*	-0.01	-0.10*
Fur Add Science	-0.55	-0.24	0.04	0.11*	0.04	-0.08	-0.22*
Physics	-0.07	-0.14	-0.09	0.04	0.04	0.01	0.06*
Science	0.19*	0.05*	0.03	0.06*	0.03*	0.04	0.23*
Spanish	-0.06	0.00	0.05	0.04	0.04	0.06	0.07*

*significant at $p<0.05$

Table A3b Relative grade difficulties for the 16 GCSE subjects for Board B derived using the PCM.

	Difficulty (logits)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	0.02	0.02	0.04	0.04	0.15*	0.23*	0.27*
English Lit	0.30*	0.20*	0.16*	0.19*	0.28*	0.16*	0.18*
French	-0.30	-0.17	-0.02	0.04	-0.07	0.05	0.29*
Geography	0.07	0.00	-0.01	-0.03	-0.08*	-0.02	0.06
German	0.04	0.00	-0.12	-0.06	-0.05	-0.13	-0.34*
History	-0.01	-0.02	-0.01	0.00	0.00	-0.04	-0.13*
Maths	-0.06	0.04	0.01	-0.26*	0.19*	-0.02	0.19*
Maths App	0.13	0.09	0.03	0.00	0.04	0.08	0.08
Maths Meth	-0.03	-0.07	0.01	0.17	0.20	0.07	0.24
Add Science	-0.17*	-0.16*	-0.18*	-0.04*	-0.03	-0.13*	-0.37*
Biology	0.00	0.03	0.03	0.02	0.02	0.00	-0.16*
Chemistry	-0.05	0.02	-0.02	-0.09*	-0.05*	0.05*	0.19*
Fur Add Science	0.20	-0.24	-0.61*	-0.64*	-0.39*	-0.08	0.25*
Physics	-0.04	0.04	-0.01	-0.09*	-0.06*	0.00	-0.09*
Science	0.18*	-0.01	-0.08*	0.00	0.00	-0.02	-0.29*
Spanish	-0.56	-0.37*	-0.22*	-0.16*	-0.28*	-0.22*	-0.08

*significant at $p<0.05$

Table A3c Relative grade difficulties for the 16 GCSE subjects for Board C derived using the PCM.

	Difficulty (logits)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	0.38*	0.18*	0.02	-0.06	-0.05	-0.19*	-0.43*
English Lit	-0.05	-0.16*	-0.17*	-0.06*	0.01	-0.14*	-0.33*
French	0.19*	0.21*	0.14*	0.02	-0.04*	0.00	0.02
Geography	0.23	0.18*	0.13*	0.08*	0.08*	0.01	-0.13*
German	0.13	0.10	0.06	-0.01	-0.08	-0.12*	-0.16*
History	0.05	0.04	0.03	0.04	0.02	0.00	0.01
Maths	0.09*	-0.03	-0.08*	-0.01	-0.05*	0.00	-0.03*
Maths App	-0.12	-0.19	-0.11	0.05	0.11	0.05	0.00
Maths Meth	0.14	0.00	-0.07	-0.04	-0.07	-0.01	-0.23
Add Science	0.46*	0.42*	0.34*	0.09*	-0.05*	-0.01	0.02
Biology	-0.09	-0.02	0.05	0.01	-0.07	-0.07	-0.07
Chemistry	0.40	0.30	0.18	-0.01	-0.13*	-0.01	0.08
Fur Add Science	0.58	0.46*	0.17	0.01	0.09	0.25*	0.56*
Physics	0.26	0.22	0.19	0.13*	0.04	0.02	-0.05
Science	-0.02	-0.05	-0.05	-0.16*	-0.12*	-0.12*	-0.75*
Spanish	0.10	0.01	-0.07	-0.02	0.03	-0.03	-0.10*

*significant at $p<0.05$

Table A3d Relative grade difficulties for the 16 GCSE subjects for Board D derived using the PCM.

	Difficulty (logits)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	0.09	-0.04	-0.13*	-0.18*	-0.05*	-0.18*	-0.22*
English Lit	0.32*	0.10*	-0.08*	-0.15*	-0.08*	-0.06*	0.05
French	-0.25	-0.18*	-0.11	-0.14*	-0.20*	-0.12*	0.05
Geography	-0.28*	-0.27*	-0.21*	-0.05	0.09*	0.06	-0.02
German	-0.75	-0.60*	-0.52*	-0.25*	-0.17*	-0.22*	-0.34*
History	-0.13	-0.10	-0.12*	-0.18*	-0.20*	-0.15*	-0.09
Maths	-0.21*	0.08*	0.33*	0.55*	0.49*	0.10*	-0.18*
Maths App	0.23	0.22	-0.02	-0.40*	-0.31*	-0.25	-0.39
Maths Meth	0.12	-0.07	-0.08	-0.19	-0.41*	-0.74*	-1.01*
Add Science	1.24*	0.89*	0.51*	0.16*	0.24*	0.11	-0.25*
Biology	0.18	0.30	0.14	-0.19	-0.09	-0.12	-0.28*
Chemistry	-0.43	-0.04	-0.15	-0.40*	-0.21*	-0.19*	-0.26*
Fur Add Science							
Physics	0.44	0.28	0.17	-0.18	-0.11	-0.11	-0.38*
Science	0.75*	0.36*	0.10	-0.05	0.24*	-0.12	-0.73*
Spanish	0.18	0.08	0.08	0.02	0.01	0.04	-0.03

*significant at $p<0.05$

Table A4a Relative grade difficulties in unit of grade for Board A derived using the PCM.

	Difficulty (unit of grade)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	0.18	0.05	0.03	0.06	0.01	0.04	0.04
English Lit	-0.01	-0.01	0.02	0.03	0.00	0.02	-0.02
French	-0.05	-0.07	-0.04	0.01	0.04	0.01	-0.06
Geography	-0.02	-0.01	0.00	-0.01	-0.02	-0.01	0.00
German	0.01	0.01	0.03	0.04	0.04	0.05	0.04
History	0.01	0.00	0.00	0.02	0.04	0.07	0.08
Maths	-0.04	0.03	0.06	-0.03	0.02	0.02	0.10
Maths App	0.05	0.02	0.02	0.03	-0.01	-0.01	-0.05
Maths Meth	0.01	0.03	0.02	-0.02	0.02	0.07	0.11
Add Science	-0.03	-0.05	-0.03	-0.01	0.01	0.03	0.07
Biology	0.03	-0.03	-0.04	-0.01	0.01	0.01	0.04
Chemistry	-0.03	-0.07	-0.03	0.04	0.04	0.00	-0.05
Fur Add Science	-0.30	-0.13	0.02	0.06	0.02	-0.04	-0.12
Physics	-0.04	-0.08	-0.05	0.02	0.02	0.00	0.03
Science	0.11	0.03	0.02	0.03	0.02	0.02	0.12
Spanish	-0.04	0.00	0.02	0.02	0.02	0.03	0.04

Table A4b Relative grade difficulties in unit of grade for Board B derived using the PCM.

	Difficulty (unit of grade)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	0.01	0.01	0.02	0.02	0.08	0.12	0.15
English Lit	0.16	0.11	0.09	0.11	0.15	0.09	0.10
French	-0.16	-0.09	-0.01	0.02	-0.04	0.03	0.16
Geography	0.04	0.00	-0.01	-0.02	-0.04	-0.01	0.03
German	0.02	0.00	-0.07	-0.03	-0.03	-0.07	-0.18
History	-0.01	-0.01	-0.01	0.00	0.00	-0.02	-0.07
Maths	-0.03	0.02	0.00	-0.14	0.10	-0.01	0.11
Maths App	0.07	0.05	0.02	0.00	0.02	0.04	0.04
Maths Meth	-0.01	-0.04	0.01	0.09	0.11	0.04	0.13
Add Science	-0.09	-0.09	-0.10	-0.02	-0.01	-0.07	-0.20
Biology	0.00	0.01	0.02	0.01	0.01	0.00	-0.09
Chemistry	-0.03	0.01	-0.01	-0.05	-0.03	0.03	0.10
Fur Add Science	0.11	-0.13	-0.33	-0.35	-0.21	-0.04	0.14
Physics	-0.02	0.02	0.00	-0.05	-0.03	0.00	-0.05
Science	0.10	-0.01	-0.05	0.00	0.00	-0.01	-0.16
Spanish	-0.31	-0.20	-0.12	-0.09	-0.15	-0.12	-0.04

Table A4c Relative grade difficulties in unit of grade for Board C derived using the PCM.

	Difficulty (unit of grade)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	0.21	0.10	0.01	-0.03	-0.03	-0.10	-0.23
English Lit	-0.03	-0.09	-0.09	-0.03	0.00	-0.07	-0.18
French	0.10	0.11	0.07	0.01	-0.02	0.00	0.01
Geography	0.12	0.10	0.07	0.05	0.05	0.01	-0.07
German	0.07	0.05	0.03	0.00	-0.04	-0.06	-0.09
History	0.03	0.02	0.02	0.02	0.01	0.00	0.01
Maths	0.05	-0.01	-0.04	0.00	-0.03	0.00	-0.02
Maths App	-0.07	-0.10	-0.06	0.03	0.06	0.03	0.00
Maths Meth	0.07	0.00	-0.04	-0.02	-0.04	0.00	-0.12
Add Science	0.25	0.23	0.18	0.05	-0.03	0.00	0.01
Biology	-0.05	-0.01	0.03	0.01	-0.04	-0.04	-0.04
Chemistry	0.22	0.16	0.10	-0.01	-0.07	0.00	0.04
Fur Add Science	0.32	0.25	0.09	0.00	0.05	0.13	0.31
Physics	0.14	0.12	0.10	0.07	0.02	0.01	-0.03
Science	-0.01	-0.03	-0.03	-0.09	-0.06	-0.07	-0.41
Spanish	0.06	0.00	-0.04	-0.01	0.01	-0.02	-0.06

Table A4d Relative grade difficulties in unit of grade for Board D derived using the PCM.

	Difficulty (unit of grade)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	0.05	-0.02	-0.07	-0.10	-0.03	-0.10	-0.12
English Lit	0.18	0.06	-0.05	-0.08	-0.04	-0.03	0.03
French	-0.14	-0.10	-0.06	-0.08	-0.11	-0.06	0.03
Geography	-0.15	-0.15	-0.11	-0.03	0.05	0.03	-0.01
German	-0.41	-0.32	-0.28	-0.14	-0.09	-0.12	-0.18
History	-0.07	-0.05	-0.07	-0.10	-0.11	-0.08	-0.05
Maths	-0.11	0.04	0.18	0.30	0.27	0.05	-0.10
Maths App	0.13	0.12	-0.01	-0.22	-0.17	-0.13	-0.21
Maths Meth	0.07	-0.04	-0.05	-0.10	-0.22	-0.40	-0.55
Add Science	0.67	0.49	0.28	0.09	0.13	0.06	-0.14
Biology	0.10	0.16	0.08	-0.10	-0.05	-0.06	-0.15
Chemistry	-0.23	-0.02	-0.08	-0.22	-0.11	-0.10	-0.14
Fur Add Science							
Physics	0.24	0.15	0.09	-0.10	-0.06	-0.06	-0.21
Science	0.41	0.20	0.06	-0.03	0.13	-0.07	-0.40
Spanish	0.10	0.04	0.04	0.01	0.00	0.02	-0.02

Table A5a Change in grade outcomes using the existing inter-board screening with mean GCSE score for Board A.

	Change (%)							
	G	**F**	**E**	**D**	**C**	**B**	**A**	**A***
English	0.02	0.07	-0.40	-0.69	1.10	-0.68	0.34	0.22
English Lit	0.09	-0.06	-0.39	-0.24	0.66	-0.42	0.25	-0.06
French	0.09	0.30	0.12	-0.65	-0.98	0.84	0.73	-0.47
Geography	0.05	-0.11	-0.05	0.19	0.28	-0.29	-0.26	0.10
German	-0.01	-0.01	-0.28	-0.25	-0.49	0.07	0.38	0.60
History	-0.01	0.03	-0.16	-0.09	-0.79	-0.18	0.45	0.77
Maths	-0.38	-1.01	-1.18	2.43	-1.40	0.77	-0.53	0.52
Maths App	-0.14	0.05	-0.14	-0.16	0.85	-0.22	-0.01	-0.13
Maths Meth	-0.03	-0.17	-0.08	0.73	-1.09	-0.54	-0.31	1.32
Add Science	0.05	0.20	-0.05	-0.07	-0.36	-0.11	0.02	0.18
Biology	0.00	0.04	0.05	-0.11	-0.16	-0.10	-0.43	0.72
Chemistry	0.01	0.06	-0.05	-0.53	-0.38	0.92	0.70	-0.75
Fur Add Science	0.11	0.06	-0.66	-1.01	0.82	1.46	1.12	-1.93
Physics	0.01	0.05	0.00	-0.33	-0.23	0.52	-0.64	0.60
Science	-0.03	-0.07	-0.23	-0.36	0.49	0.11	-0.06	0.15
Spanish	0.06	-0.10	-0.27	0.08	0.02	-0.18	0.04	0.36

Table A5b Change in grade outcomes using the existing inter-board screening with mean GCSE score for Board B.

	Change (%)							
	G	**F**	**E**	**D**	**C**	**B**	**A**	**A***
English	0.16	-0.12	-0.06	0.05	-1.55	-0.26	0.93	0.77
English Lit	-0.09	-0.16	-0.28	-0.86	-2.04	2.11	0.73	0.54
French	0.20	0.16	-0.48	-0.39	2.02	-1.58	-1.82	1.76
Geography	-0.02	0.33	0.12	0.29	1.08	-1.17	-0.62	0.31
German	0.03	0.13	0.97	-0.43	0.02	0.00	0.68	-1.47
History	0.03	0.03	-0.06	-0.04	0.14	0.25	0.16	-0.59
Maths	-0.28	-0.83	-0.48	4.41	-6.23	3.19	-1.17	0.52
Maths App	-0.01	-0.31	0.13	0.19	-0.37	-0.65	-0.06	0.99
Maths Meth	0.26	0.27	-0.41	-1.28	-1.40	2.11	-1.44	1.76
Add Science	0.28	0.30	0.89	-1.26	0.08	0.32	-0.20	-0.37
Biology	0.02	-0.01	-0.02	-0.03	0.06	0.09	0.90	-1.01
Chemistry	0.00	-0.06	0.10	0.58	0.27	-1.25	-1.05	1.44
Fur Add Science	0.01	0.52	2.11	4.02	-0.37	-4.52	-3.44	1.80
Physics	0.00	-0.07	0.08	0.54	0.41	-0.97	0.68	-0.66
Science	-0.06	0.26	0.69	-0.91	0.06	-0.17	0.10	-0.03
Spanish	0.46	0.40	0.20	0.34	2.93	-1.17	-2.21	-1.24

Table A5c Change in grade outcomes using the existing inter-board screening with mean GCSE score for Board C.

	Change (%)							
	G	**F**	**E**	**D**	**C**	**B**	**A**	**A***
English	-0.12	-0.08	0.52	0.74	-0.40	0.77	-0.25	-0.95
English Lit	0.22	0.42	0.55	-0.31	-0.71	1.20	-0.23	-0.81
French	-0.24	-0.62	-0.15	0.84	0.96	-0.78	-0.34	0.38
Geography	-0.41	-0.40	-0.15	0.20	-0.34	1.04	0.65	-0.51
German	-0.07	-0.18	-0.17	0.50	0.93	-0.01	-0.62	-0.33
History	-0.02	-0.01	0.06	-0.19	0.34	0.17	-0.25	0.16
Maths	0.22	0.51	0.50	-0.77	0.84	-0.89	0.13	-0.13
Maths App	0.35	0.38	-0.38	-1.12	-0.81	0.98	0.11	0.13
Maths Meth	0.06	-0.09	0.38	-0.52	1.51	-1.30	1.36	-1.18
Add Science	-0.48	-0.94	-0.97	2.00	1.46	-0.68	-0.12	0.02
Biology	-0.03	-0.13	-0.24	0.26	1.13	-0.11	-0.61	-0.27
Chemistry	-0.11	-0.18	-0.19	0.68	1.74	-1.89	-0.83	0.86
Fur Add Science	-0.29	-0.47	0.56	0.42	-1.98	-1.35	-1.05	4.16
Physics	-0.05	-0.11	-0.23	-0.37	0.40	-0.05	0.68	-0.18
Science	0.41	0.03	-0.04	2.29	-1.55	-0.92	0.05	-0.42
Spanish	-0.14	0.09	0.68	-0.51	-0.73	0.83	0.08	-0.32

Table A5d Change in grade outcomes using the existing inter-board screening with mean GCSE score for Board D.

	Change (%)							
	G	**F**	**E**	**D**	**C**	**B**	**A**	**A***
English	-0.04	-0.08	0.60	1.05	-1.58	1.08	-0.73	-0.30
English Lit	-0.24	0.04	0.72	0.83	-0.67	-0.03	-0.63	0.26
French	0.10	0.12	0.07	1.30	1.27	-1.33	-1.89	0.36
Geography	0.44	0.60	0.27	-1.32	-1.80	0.90	0.73	0.04
German	0.28	0.51	1.49	0.18	-0.07	-0.35	-0.69	-1.43
History	-0.01	-0.18	0.46	0.99	0.49	-1.00	-1.01	-0.26
Maths	-1.76	-3.26	-3.40	-2.74	2.36	5.24	1.46	-0.36
Maths App	-0.47	-0.84	1.59	3.95	-0.40	-0.95	-0.20	-1.77
Maths Meth	-0.79	0.64	-0.37	1.86	1.92	3.60	-1.19	-5.50
Add Science	-0.84	-1.36	-1.20	0.61	-2.00	3.86	2.23	-0.79
Biology	-0.07	-0.17	0.02	1.14	-0.62	1.04	0.82	-2.15
Chemistry	0.05	0.04	0.44	1.62	-0.48	0.69	-0.14	-2.26
Fur Add Science								
Physics	-0.03	0.04	-0.09	1.48	-0.75	0.12	2.59	-3.36
Science	-1.13	-0.55	-0.33	1.81	-1.86	3.83	0.17	-0.50
Spanish	-0.20	0.05	-0.66	0.84	0.25	-0.73	0.95	-0.33

Table A6 Correlations between Rasch ability, mean GCSE, mean GCSE excluding the subject itself (mean GCSE1) and individual subjects.

	Mean GCSE	Ability	Eng	Eng Lit	Fren	Geog	Germ	His	Maths	Maths (App)	Maths (Meth)	Add Sci	Bio	Chem	Fur Add Sci	Phy	Sci	Span
Mean GCSE	1	0.98	0.88	0.87	0.82	0.91	0.81	0.90	0.90	0.92	0.93	0.90	0.90	0.90	0.87	0.89	0.91	0.80
Ability	0.98	1	0.87	0.86	0.83	0.91	0.82	0.88	0.88	0.91	0.92	0.90	0.90	0.89	0.87	0.88	0.90	0.80
Mean GCSE 1			0.79	0.79	0.71	0.85	0.71	0.82	0.79	0.87	0.88	0.82	0.85	0.85	0.78	0.84	0.81	0.65

Table A7a Change in grade outcomes using the existing inter-board screening with Rasch ability for Board A.

	Change (%)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	0.06	-0.43	-0.70	1.07	-0.60	0.37	0.22
English lit	-0.07	-0.42	-0.31	0.67	-0.33	0.30	-0.06
French	0.30	0.12	-0.64	-0.97	0.86	0.71	-0.49
Geography	-0.11	-0.06	0.16	0.26	-0.25	-0.23	0.09
German	-0.01	-0.29	-0.24	-0.48	0.07	0.38	0.59
History	0.01	-0.15	-0.06	-0.78	-0.22	0.46	0.78
Maths	-0.86	-1.06	2.28	-1.63	0.62	-0.55	0.55
Maths App	0.04	-0.15	-0.16	1.02	-0.29	-0.08	-0.08
Maths Meth	-0.31	0.00	0.93	-0.94	-0.67	-0.37	1.43
Add Science	0.21	-0.05	-0.15	-0.42	-0.03	0.06	0.19
Biology	0.04	0.05	-0.11	-0.19	-0.09	-0.43	0.73
Chemistry	0.06	-0.05	-0.53	-0.44	0.93	0.73	-0.74
Fur Add Science	0.08	-0.61	-1.01	0.71	1.39	1.16	-1.86
Physics	0.05	0.00	-0.33	-0.27	0.52	-0.61	0.61
Science	-0.07	-0.35	-0.39	0.50	0.16	-0.01	0.16
Spanish	-0.10	-0.27	0.07	0.04	-0.16	0.02	0.34

Table A7b Change in grade outcomes using the existing inter-board screening with Rasch ability for Board B.

	Change (%)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	-0.19	-0.24	-0.02	-1.32	-0.09	0.92	0.74
English Lit	-0.17	-0.33	-0.95	-1.99	2.26	0.71	0.50
French	0.16	-0.49	-0.45	2.05	-1.55	-1.79	1.72
Geography	0.28	0.07	0.27	1.11	-1.13	-0.57	0.34
German	0.15	0.99	-0.49	0.16	0.08	0.56	-1.56
History	0.03	-0.05	-0.05	0.13	0.28	0.18	-0.59
Maths	-0.75	-0.47	4.42	-6.28	3.14	-1.18	0.53
Maths App	-0.36	0.11	0.18	-0.28	-0.33	0.03	0.72
Maths Meth	0.21	-0.53	-1.65	-0.98	2.69	-1.48	1.45
Add Science	0.32	0.95	-1.13	0.08	0.19	-0.25	-0.37
Biology	-0.02	-0.04	-0.06	0.09	0.06	0.94	-1.00
Chemistry	-0.06	0.08	0.55	0.31	-1.27	-1.04	1.44
Fur Add Science	0.53	2.14	4.12	-0.25	-4.59	-3.67	1.86
Physics	-0.07	0.07	0.51	0.42	-0.97	0.69	-0.64
Science	0.26	0.71	-0.91	0.15	-0.23	0.06	-0.04
Spanish	0.40	0.20	0.34	2.88	-1.20	-2.20	-1.17

Table A7c Change in grade outcomes using the existing inter-board screening with Rasch ability for Board C.

	Change (%)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	-0.14	0.36	0.62	-0.15	0.80	-0.24	-0.94
English Lit	0.32	0.40	-0.32	-0.56	1.30	-0.25	-0.83
French	-0.62	-0.16	0.82	0.95	-0.78	-0.32	0.40
Geography	-0.40	-0.15	0.19	-0.38	1.03	0.70	-0.51
German	-0.19	-0.16	0.53	0.90	-0.03	-0.64	-0.33
History	-0.05	0.03	-0.20	0.35	0.22	-0.21	0.18
Maths	0.42	0.46	-0.74	0.93	-0.79	0.16	-0.13
Maths App	0.32	-0.43	-1.21	-0.96	1.12	0.25	0.13
Maths Meth	-0.02	0.26	-0.55	1.21	-1.33	1.49	-1.16
Add Science	-1.03	-1.10	1.97	1.65	-0.58	-0.13	0.01
Biology	-0.13	-0.23	0.28	1.13	0.11	-0.68	-0.46
Chemistry	-0.19	-0.22	0.63	1.84	-1.68	-0.90	0.69
Fur Add Science	-0.52	0.43	0.37	-1.76	-1.13	-1.02	3.94
Physics	-0.11	-0.24	-0.38	0.51	0.12	0.60	-0.37
Science	-0.11	0.02	2.30	-1.52	-0.76	0.03	-0.42
Spanish	0.07	0.67	-0.48	-0.74	0.80	0.11	-0.29

Table A7d Change in grade outcomes using the existing inter-board screening with Rasch ability for Board D.

	Change (%)						
	F	**E**	**D**	**C**	**B**	**A**	**A***
English	-0.04	0.73	1.11	-1.63	0.88	-0.78	-0.29
English Lit	0.09	0.84	0.99	-0.76	-0.29	-0.72	0.27
French	0.15	0.11	1.32	1.23	-1.46	-1.90	0.43
Geography	0.67	0.36	-1.16	-1.68	0.69	0.50	0.01
German	0.52	1.48	0.07	-0.19	-0.38	-0.53	-1.32
History	0.02	0.51	1.00	0.49	-1.18	-1.27	-0.34
Maths	-2.62	-3.15	-2.81	1.68	4.28	1.15	-0.42
Maths App	-0.50	1.80	4.30	-0.85	-1.70	-0.52	-1.50
Maths Meth	1.11	0.10	1.85	1.54	2.98	-1.41	-5.36
Add Science	-1.15	-0.88	1.33	-1.84	2.64	1.94	-0.76
Biology	-0.15	0.11	1.48	-0.52	0.58	0.56	-1.96
Chemistry	0.06	0.56	2.01	-0.28	0.15	-0.57	-2.03
Fur Add Science							
Physics	0.05	-0.03	1.80	-0.47	-0.35	2.21	-3.19
Science	0.06	1.44	2.29	-3.00	2.44	-0.43	-0.57
Spanish	0.10	-0.60	0.83	0.17	-0.78	0.97	-0.33

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

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