OXFORD REVIEW OF EDUCATION, 2018
VOL. 44, NO. 4, 494–513
https://doi.org/10.1080/03054985.2018.1430562

Routledge
Taylor & Francis Group

Inter-subject comparability of examination standards in GCSE and GCE in England

Qingping He, Ian Stockford and Michelle Meadows

Office of Qualifications and Examinations Regulation, Coventry, UK

ABSTRACT

Results from Rasch analysis of GCSE and GCE A level data over a period of four years suggest that the standards of examinations in different subjects are not consistent in terms of the levels of the latent trait specified in the Rasch model required to achieve the same grades. Variability in statistical standards between subjects exists at both individual grade level and the overall subject level. Findings from this study are generally consistent with those from previous studies using similar statistical models. It has been demonstrated that the alignment of statistical standards between subjects based on the Rasch model would likely result in substantial change in performance standards of the examinations for some subjects evidenced here by significant changes in grade boundary scores and grade outcomes. It is argued that the defined purposes of GCSE and A level qualifications determine how their results should be interpreted and reported and that the existing grading and results reporting procedures are appropriate for supporting these purposes.

KEYWORDS

GCSE; GCE A level; examination standards; inter-subject comparability; partial credit model (PCM)

1. Introduction

1.1. The issue of inter-subject comparability in England

Examinations are frequently compared based on a range of characteristics or attributes. These may be characteristics of the examinees, the measurements of examinees provided by the examinations, or the properties of the examinations themselves (see Coe, Searle, Barmby, Jones, & Higgins, 2008; Elliot, 2013; Lockyer & Newton, 2015; Newton, Baird, Goldstein, Patrick, & Tymms, 2007; Newton, He, & Black, 2017; Opposs, 2015). For example, in the context of examinations used in academic qualifications such as the General Certificate of Secondary Education (GCSE, taken by students aged 16) and General Certificate of Education Advanced level (GCE A level, taken by students aged 18) in England, the characteristics for comparison may include, among others, the level of specific skills and knowledge required for achieving a specific grade, the quality of teaching, the amount of teaching time dedicated to the course of study, and the ability of candidates in specific subject areas (frequently characterised by their prior attainment). There have been attempts to compare some of these characteristics over time for the same examination, similar examinations from

CONTACT Qingping He Qingping.He@ofqual.gov.uk

different examining boards, and even examinations that test different subject areas (see Coe, 2008; Coe et al., 2008; Lockyer & Newton, 2015). The comparison of examination standards for different subjects is referred to as inter-subject comparability, which is the focus of the present study.

Inter-subject comparability of standards in GCSEs and A levels in England has been a matter of debate for a long time (see Coe, 2008; Coe et al., 2008; Lockyer & Newton, 2015; Newton, 2012, 2015; Newton et al., 2007). This subject has been studied extensively, involving the use of both judgemental and statistical approaches to conceptualise and quantify inter-subject comparability (Coe et al., 2008; Lockyer & Newton, 2015). Coe et al. (2008) provided a comprehensive review of the various methods that have been used to investigate comparability. These include subject pairs analysis, common examinee linear models, latent trait models (e.g. Rasch models), reference tests, value added models (including multilevel modelling), and comparative judgement (see also Lockyer & Newton, 2015).

Results from analyses using a variety of statistical models conducted by Coe and co-workers and other researchers suggested that there has been a consistent pattern in the relative 'difficulty' of examinations in GCSE and GCE A level subjects, with some subjects shown to be consistently 'harder' or 'easier' than others (see Coe et al., 2008). Disagreement between researchers remains, however, regarding the various assumptions involved in the different statistical models, the interpretation of the characteristics of the examinees on which the comparisons were made, the interpretation of the differences between subjects, and the implications of the results obtained. Coe et al. (2008) discussed the major issues with statistical methods used for studying inter-subject comparability. These, among others, include the strong assumptions made about the data being analysed (e.g. the unidimensionality assumption of the underlying latent trait shared by the examinees and assessed by the different examinations required by the Rasch model) which are seldom met by real data, unrepresentativeness of samples used, imperfect data–model fit, and different results from different statistical models for the same dataset (see also Lockyer & Newton, 2015).

1.2. Aligning standards between subjects based on statistical analysis and its implications

There has been debate about the implications and potential consequences of incomparability of examination standards between subjects (see Coe, 2008; Coe et al., 2008; Lockyer & Newton, 2015; Newton, 2015). Coe (2008) suggested that subject standards may need to be statistically aligned when grades from different subjects are used for specific purposes (see also Coe et al., 2008), particularly when they are used interchangeably or as equivalent currencies in situations such as admissions to certain university courses and use of examination results as part of school accountability measures. Coe also indicated that in some countries such as Australia and Scotland, incomparability of statistical standards between subjects is taken into account for specific uses (also see Lamprianou, 2009). Less attention has however been paid to the implications of aligning statistical standards between subjects based on inter-subject comparisons for the performance standards of the examinations which are related to subject-specific grade criteria in the context of GCSEs and GCEs.

The defined purposes of GCSEs and GCEs suggest that grades be interpreted as the levels of skills and knowledge achieved at specific performance levels (see Ofqual, 2014a, 2015). The performance standards at different levels are currently articulated through grade

descriptions for individual subjects which represent a source of evidence used during awarding (standards setting). It may arguably be assumed therefore that the performance standards in the examinations and the grade criteria or subject performance standards are aligned. If the statistical standards based on inter-subject comparability studies were to be aligned for different subjects that are graded based on subject-specific grade requirements, the consequence would be a mismatch between the grade criteria and the performance standards of the examinations (e.g. subjects which are either too 'easy' or too 'hard' based on inter-subject statistical comparisons) and a change in the distribution of grades.

1.3. Aims of study

The main aims of this research were:

- to generate new evidence on the relative difficulty of examinations in different GCSE and GCE A level subjects over time based on Rasch analysis; and
- to gain an understanding of the impact of aligning statistical standards between subjects based on Rasch analysis on performance standards and grade outcomes for individual subjects.

2. The partial credit model and its use for studying inter-subject comparability

The unidimensional Rasch model was originally developed to analyse psychological and educational tests composed of dichotomous items (see Rasch, 1960/1980) and has been extended subsequently for analysing tests composed of polytomous items. These extended Rasch models include Andrich's Rating Scale Model (RSM), Masters' Partial Credit Model (PCM), and other models (see Andrich, 1978; Masters, 1982; Muraki, 1992; Wright & Masters, 1982). The PCM, which is to be used in the present study, states that, for a polytomous item with a maximum available score of m (the number of score categories minus 1), the probability $P(\theta, x)$ of an examinee with ability (latent trait) θ scoring x on the item can be expressed as:

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

where δ_k is the location of the kth category score on the latent trait continuum (or category threshold or difficulty [see Andrich, 2015]). Recently, the Rasch PCM model and item response theory (IRT) models have been used for investigating the comparability of standards in different subject examinations (see Bramley, 2011; Coe, 2008; Coe et al., 2008; Korobko, Glas, Bosker, & Luyten, 2008; 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 for 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 the examinations to be analysed here together define a shared construct which is closely related to the constructs being measured by the individual examinations and that difference in

difficulty overall and at individual grades reflects difference in standards between the exams. However, it has to be noted that the latent trait specified in the Rasch model is operationally defined by the examinations included in the analysis. Coe (2008) interpreted such a trait as 'general academic ability' of the individual examinees which will also be adopted in this study.

3. Data collection and analysis

3.1. Data collection

Candidates are classified or graded into eight performance categories (grades) for GCSEs (from A* to G and U for unclassified, with A* representing the highest level of performance in a qualification) and six performance categories for GCEs (from A* to E and U for unclassified). For this study, candidate-level data which include some basic background information and GCSE and A level grades in individual subjects from 2010 to 2013 were collected from the UK Department for Education's National Pupil Database (NPD). These grades were used to perform Rasch analyses as described below. For GCSEs, subjects with entries fewer than 5000 were excluded from the analysis in order to obtain accurate estimates of model parameters. For A levels, subjects with entries fewer than 1000 were excluded. Further, candidates taking fewer than three GCSE subjects and, for A level subjects, candidates taking fewer than two were excluded in order for the results to be more accurate and reliable. Final sample sizes vary from 4811 for Vocational GCSE Applied Business to 548,045 for English (including English Language) for the GCSE subjects and from 659 for Critical Thinking to 65,347 for Mathematics for the A level subjects.

Based on results from the Rasch analysis, a selection of GCSE and A level subjects ranging from 'easy' to 'hard' as defined by their Rasch 'difficulty' measures were identified for further investigation into the impact of aligning statistical standards between subjects on performance standards. Candidate-level data which included subject-level grades and uniform mark scale (UMS) marks were collected from the exam boards that provide general qualifications in England for this purpose. (The UMS mark is a scaled score used to ensure the comparability of raw marks from different examination series. See http://www.aqa.org.uk/exams-administration/about-results/uniform-mark-scale for a detailed explanation.)

3.2. Data analysis

When analysing the GCSE and A level data for individual examination series using the Rasch model, the comparability of performance standards of the examinations for similar qualifications from different exam boards (inter-board comparability) is assumed. Further, for the same qualification for a specific year, candidates' grades from different exam boards are combined to produce the grade distribution for the whole cohort. To facilitate the analysis, the 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, A*→ 8 for GCSEs, and U→ 0, E→ 1, D→ 2, C→ 3, B→ 4, A→ 5, A*→ 6 for A levels. The data were analysed using the software WINSTEPS which implements the PCM by treating each exam as an item and the numerical grade received by a candidate a score on the item. Further, the average of the ability measures for all candidates included in the analysis was set to zero for the individual dataset. The category

parameters of the subjects from different exam series may be compared directly if it can be assumed that the ability distribution of the candidates from different exam series is the same.

It is to be noted that since not every candidate took all the subjects included in the study, the analysis involved missing data. Missing data always present problems for analysis and results interpretation. Using simulations, Bramley (2016) demonstrated that the existence of non-random missing data could produce biased estimates of subject difficulty for one of the statistical approaches used for studying inter-subject comparability. However, it is to be noted that 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 examination data analysed here, the nature of missing data is different from that in Bramley's analysis as 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 nature of the missing data in our study is also different from that of the missing data produced from a conventional test where the test takers are generally required to answer all items in the test. Here, the students were required to take a set of subjects, and the missing data could be seen to be missing by design from a data collection perspective. The Rasch analysis presented here would be in some sense similar to concurrent calibration of items using data collected from a test equating design that involves the use of multiple common item non-equivalent groups (CINE) (see Kolen & Brennan, 2014). The effect of missing data on estimates of subject difficulties derived using the PCM is likely to be minimal in the present study.

4. Relative difficulty of GCSE and GCE A level examinations

4.1. Unidimensionality and model fit

Although findings from research suggest that the impact of misfit to an IRT model on its application varies (see Tendeiro & Meijer, 2015; Zhao & Hambleton, 2017), the data being analysed must meet the model assumptions and fit the model sufficiently well for the results to be appropriately interpreted. The unidimensionality assumption of Rasch models can be investigated using factor analysis of row scores or the residuals of person scores (differences between observed scores and Rasch model predicted scores) (see He, Anwyll, Glanville, & Opposs, 2014; Linacre, 2015; Reckase, 2009; Reeve & Fayers, 2005; Smith, 2002; Yen, 1993). The degree to which the test data fit the model can be evaluated using model fit statistics for both items and persons. Frequently used fit statistics include some of the residual based fit statistics such as unweighted mean squares (outfit MNSQ) and weighted mean squares (infit) (see Linacre, 2015; Wright & Masters, 1982; Wu & Adams, 2007). Both infit and outfit have an expected value of 1 when the data fit the Rasch model. Views on the acceptable values for infit and outfit MNSQs vary, depending on the purpose of the analysis (see Coe, 2008; Coe et al., 2008; Linacre, 2002; Tan & Yates, 2007; Wong, McGrath, & King, 2011). Linacre (2002) suggested that when model fit statistics are above 2.0, the measurement system would be distorted. This value of 2.0 has been used to judge whether an item or person fits the Rasch model sufficiently well in the present study. This may have resulted in some of the subjects judged not to fit the Rasch model in other studies being included in the present study.

Initial analysis of the datasets suggested that, for the GCSE data, the fit statistics for category 0 (grade U) for a substantial number of subjects were over 2.0. Some subjects with high misfit values had disordered category thresholds. For some of these subjects, the

bottom two categories (grades U and G) were also disordered. Both disordered thresholds and disordered categories reflect the violation of the measurement construct and inappropriate functioning of the items as would be expected by the measurement model. For the A level data, results from initial analysis indicated that most of the subjects (except Chinese) had fit statistics at both subject level and category level within 2.0. To resolve the problems of disordered categories and disordered thresholds and large misfit of the bottom categories for the GCSE exams, the bottom grade U was treated as missing and excluded from the analysis. Grade G was treated as the new bottom grade. Inspection of person fit statistics suggested that about 7% of GCSE candidates and 10% of A level candidates had fit statistics over 2.0. In order to obtain more accurate estimates of item parameters (which are the focus of this study), candidates with fit statistics over 2.0 were excluded from the analysis.

Analysis of variances suggested that the total variances of the datasets that can be accounted for by the Rasch model are about 78% for the GCSEs and 83% for the A levels (see Table 1). 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 1.4 for GCSEs and 1.1 for A levels, suggesting that these contrasts are of relatively equal importance in explaining the variance unexplained by the Rasch model and therefore it may be assumed that no meaningful second dimension could be constructed for the original numerical grades. Therefore, these datasets could be essentially treated as unidimensional (see Linacre, 2015; Pae, 2012). Some of the other statistics showing how well the Rasch model functioned in establishing the measurement scale for the various datasets listed in Table 1 include person separation index defined as the ratio of the standard deviation of person measures and their average standard error of estimation, person reliability which is related to the separation index and is defined as 1 minus the square of the ratio of person average measurement error to the standard deviation of person measures, and the average item point-measure correlation between the observations on an item and the corresponding person measures. These indicators suggested that the Rasch model functioned reasonably well overall.

4.2. Examination characteristic curves and subject difficulty

An important concept in Rasch modelling is the item characteristic curve (ICC) which has important applications in developing Rasch measurement scales. ICC shows the relationship between the expected score on the item from a person with ability θ and is defined as:

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

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

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

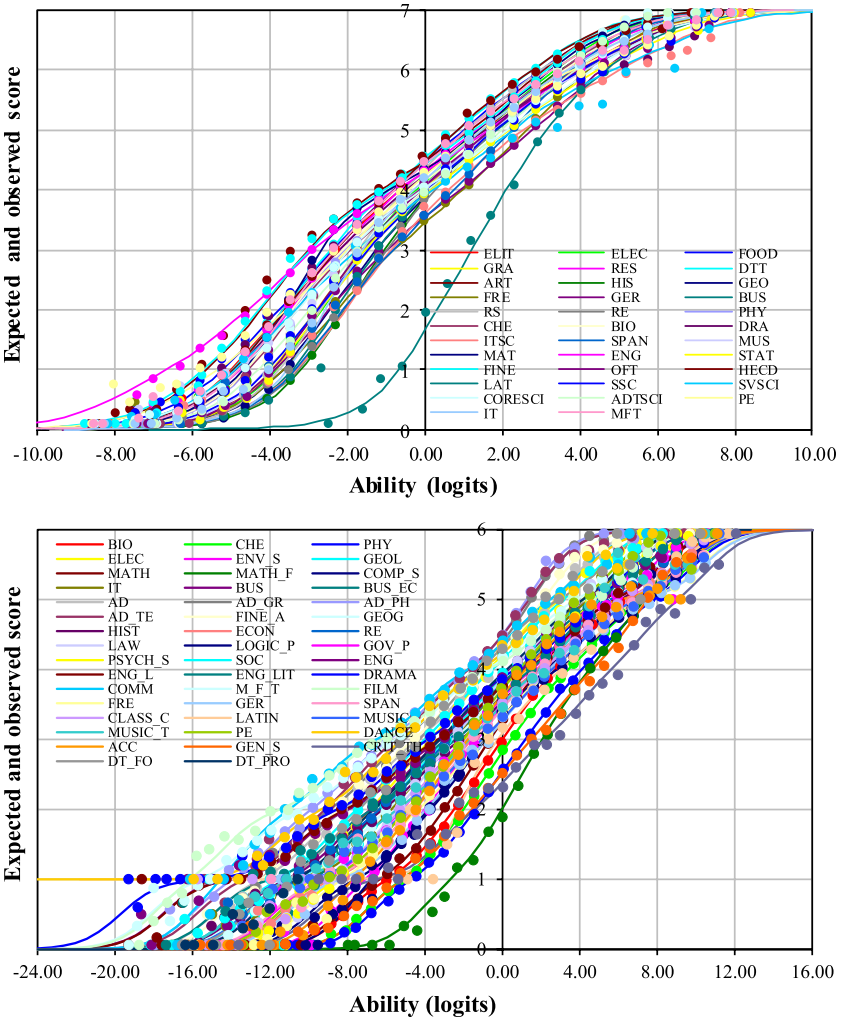

Figure 1. Comparison of the distributions of model expected ICCs and the observed distributions of scores (numerical grades) for the 35 GCSE (top) and 47 A level (bottom) subjects from the 2013 examination series.

The graphs in Figure 1 show the model predicted ICCs and observed distributions of scores (numerical grades) for the 35 GCSE and 47 A level subjects from 2013. The curves on the left are for subjects which are generally 'easy' and those on the 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 for both GCSEs and A levels spread a slightly wider range of ability for the lower grades than for the upper grades, suggesting there is a degree of differentiated relative difficulty.

4.3. Difficulty of individual grades

Since each score category (grade) in a subject is characterised by its own difficulty parameter and modelled using the Rasch model, it is possible to compare individual grades between

the subjects. However, as the category parameter δ_k cannot be interpreted as the difficulty of the category k, an alternative definition of category difficulty based on the ICC has been proposed (see Linacre, 2015; Wu & Adams, 2007). For this definition, the difficulty of a score in category k of the item d_k is the ability at which the expected score on the ICC is $k - 0.5$:

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

This definition is similar to the definition of the item difficulty for dichotomous items and has been adopted in this study. The average of the category parameters can be used to characterise the overall difficulty D of the item (Wu & Adams, 2007):

$$D = \frac{1}{m}\sum_{k=1}^{m}\delta_k \tag{4}$$

Figure 2 compares the difficulty of individual grades between the 35 GCSE and 47 A level subjects from 2013 with the subjects ordered by the overall difficulty of the item (subject). The subjects in the left of the graph are generally easier at all individual grades based on their Rasch difficulty measures, while those in the right are more difficult. For GCSEs, Art, Fine Art, and English are among the easiest subjects, while Spanish, German, French, and Short Course IT are among the most difficult. Mathematics is easier than Biology, Chemistry, and Physics which are of average difficulty. Latin is considerably more difficult than most of the other subjects in almost the entire ability range. Although the distributions of grade difficulties are generally consistent with the distribution of the overall subject difficulty, there is considerable variability in difficulty between the subjects at individual grades. For example, the overall difficulty of Geography is 0.80 logits which is about 1.26 logits higher than the overall difficulty of English. The grade difficulty of English at A* is 5.43 logits which is 0.63 logits higher than the grade difficulty of 4.80 for Geography. That is at A*, English is harder than Geography in terms of the level of the Rasch ability required to achieve this grade. However, at grade E, the grade difficulty of English is –5.51 logits while that of Geography is –3.04 logits. Therefore, under this definition, English is considerably easier than Geography at grade E.

For A levels, Communication Studies, Film Studies, Graphics (Art and Design), Photography (Art and Design), Media Studies, and Textiles (Art and Design) are among the easiest subjects, while STEM (science, technology, engineering and mathematics) subjects, Modern Foreign Languages, General Studies, and Critical Thinking are among the hardest subjects. Further Mathematics and Latin are the most difficult subjects. Similar to the GCSEs, the grade difficulties exhibit greater variability between the subjects than the subject level overall difficulties.

It is clear from Figure 2 that the difficulty gaps between two adjacent grades for individual subjects vary across the range of the grades for both GCSEs and A levels. For example, for GCSE English, the differences in difficulty between adjacent grades are 2.67 (A*–A), 2.26 (A–B), 2.41 (B–C), 1.82 (C–D), 1.78 (D–E), and 2.62 logits (E–F) respectively. These differences would suggest that different relative progress, or more specifically, the amount of the latent trait specified by the Rasch model, would be required for progressing from one grade to another at different grades. In general, for GCSEs, the gap between grade D and grade E is the smallest, while that between A* and A is the largest. For A levels, the gap between A* and A is among the smallest for the easy subjects but similar to the gap between D and E

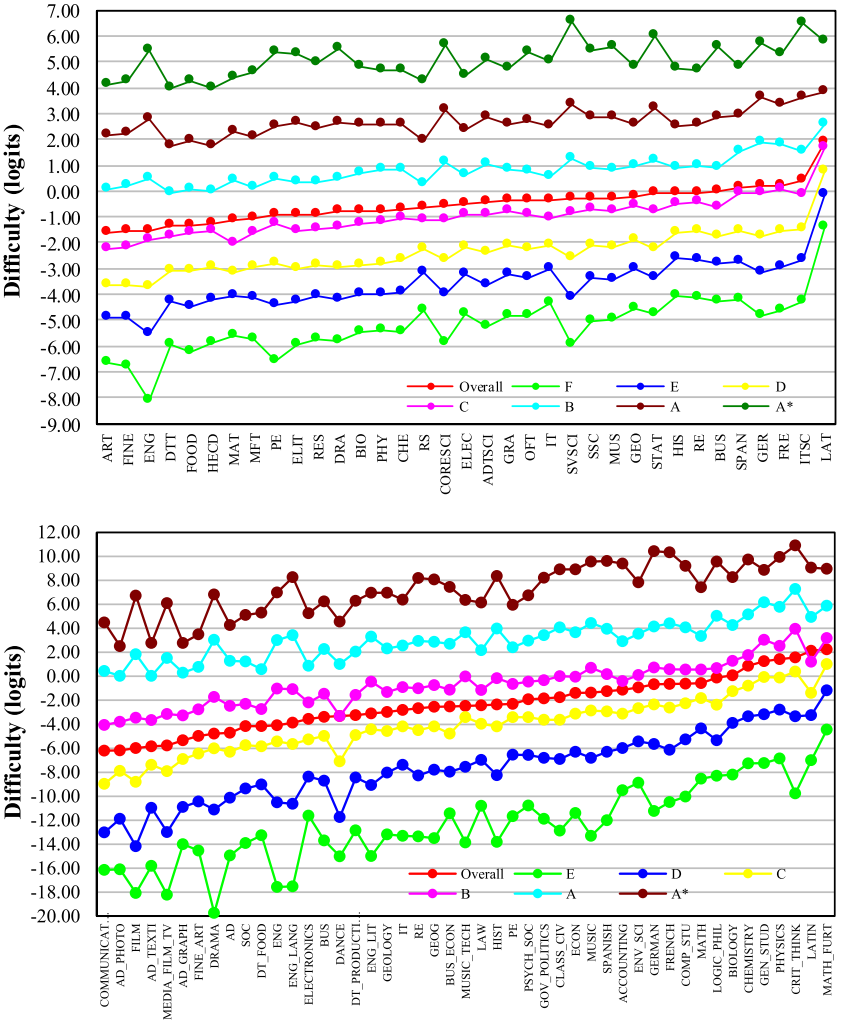

Figure 2. Comparasion of the overall difficulty and the difficulty at individual grades for the 35 GCSE (top) and 47 A level (bottom) subjects from the 2013 examination series.

for the harder subjects. This is different from GCSEs where the gap between A* and A is the largest. For the mid-difficulty subjects, the gaps between B and C and between C and D are the smallest. The gap between D and E is the largest for almost all subjects. The grade gaps also become smaller for more difficult subjects.

4.4. Relative grade difficulty

The measurement scale established using the Rasch modelling approach is linear (i.e. the same difference between two points on the Rasch ability continuum has the same meaning), while the ordered numerical grade scale is not. Here, we have assumed that the numerical grade scale is linear but we have used the Rasch grade difficulty measure to compare the relative difficulty between subjects at individual grades. At a specific grade k for a specific subject, the difference between the grade difficulty d_k of this subject and the mean difficulty

of all subjects at this grade is defined as the relative difficulty $d_{k,R}$ of this grade (relative to the mean of all subjects at the same grade):

$$d_{k,R} = d_k - \frac{1}{N_S}\sum_{i=1}^{N_s} d_{ki} \tag{5}$$

where N_S is the total number of subjects included in the analysis, and d_{ki} is the difficulty of subject i at grade k. If $d_{k,R}$ is negative, the subject concerned at this specific grade is easier in relation to subjects of average difficulty at this grade; if, on the other hand, it is positive, the subject is more difficult at this grade.

It may be more intuitive to express relative grade difficulty in the unit of grade. Although the difference in difficulty between two adjacent grades is not a constant, an average grade gap Δ (logits) can be defined across all grades and subjects as:

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

where N_G is number of grade gaps (5 for GCSEs—between A and F, and 4 for A levels—between A and E.), $d_{i,E/F}$ is the difficulty of grade F for GCSEs and E for A levels and $d_{i,A}$ is the difficulty of A for both GCSEs and A levels. Dividing the relative grade difficulty $d_{k,R}$ by the average grade gap in logits gives the relative grade difficulty $d_{k,RG}$ in the unit of grade:

$$d_{k,RG} = \frac{d_{k,R}}{\Delta} \tag{7}$$

Figure 3 shows the distribution of the relative grade difficulty in units of grade for both the GCSE and GCE A level subjects from the 2013 examination series. As can be seen, the relative grade difficulties vary within a subject. Although for most subjects, the signs of the relative grade difficulties are consistent across the grades (either negative or positive), there are a few subjects for which the relative grade difficulties have both positive and negative values. For example, for GCSE Core Science the relative grade difficulty is positive for grades A*–B but negative for grades C–F. Similarly, for A level English Language, the relative grade difficulty is positive for grades A* and A but negative for grades B–E.

At a specific grade, for the GCSE subjects, except Latin, the most difficult subjects are about half a grade harder than the subjects of mean difficulty, and the easiest subjects are about a half grade easier. The most difficult subjects are therefore about one grade harder than the easiest subjects. GCSE Latin is over two grades harder than the easiest subjects at lower grades. For the A level subjects, except for Further Mathematics, the most difficult subjects (the STEM subjects and Modern Foreign Languages) are nearly two grades harder than the easiest subjects.

4.5. Variation of subject and grade difficulties over time

Figure 4 shows the overall difficulties of the GCSE and A level subjects and the difficulties at a number of grades from 2010 to 2013 (the subjects are ordered by the overall subject difficulty for the 2010 examination series). It is to be noted that, since the data for each examination series were analysed using the Rasch model separately, the difficulty measures for

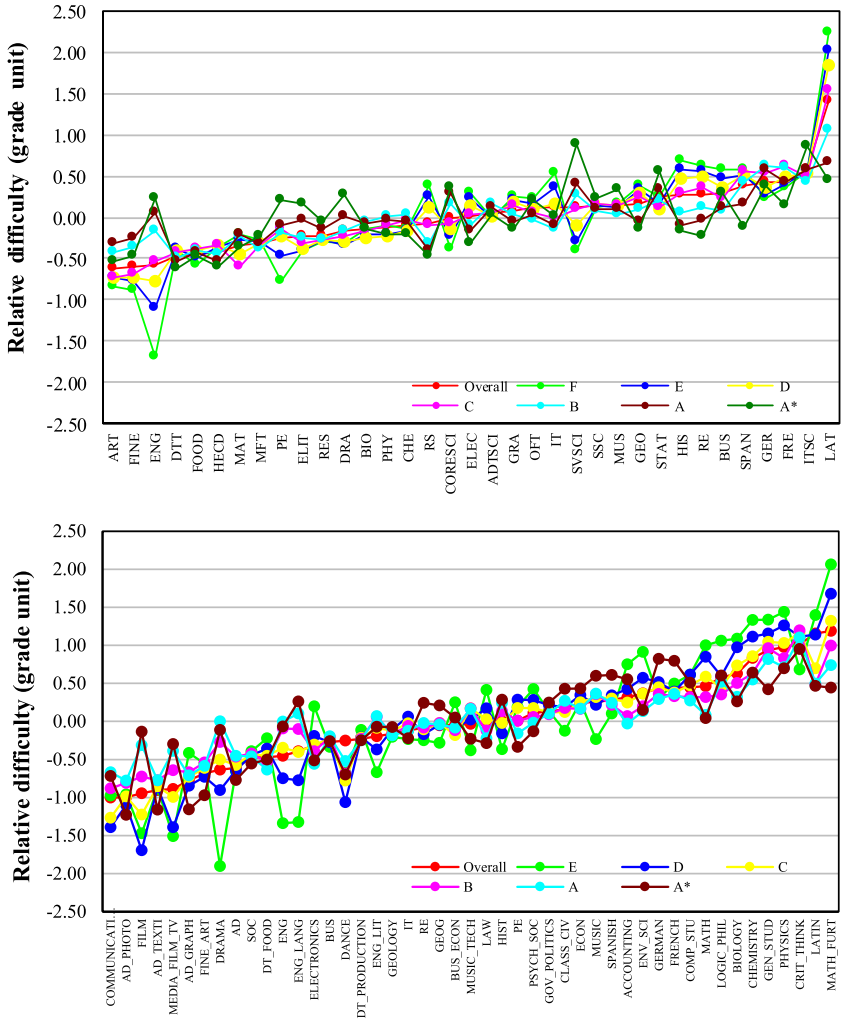

Figure 3. Relative grade difficulties in the unit of grade for the 35 GCSE (top) and 47 A level (bottom) subjects from 2013.

different years are not necessarily on the same scale as a result of the use of a sub-sample rather than the whole GCSE population in the calibration. A systematic shift in difficulty parameter over time does not necessary suggest a change in actual difficulty over time. Instead, the focus here is on the patterns of the distributions of average difficulties between subjects which are generally consistent over the four-year period. This is not unexpected due to the strong use of statistical evidence, albeit a different methodology to that presented here, to maintain standards over time in these subjects. For GCSEs, subjects such as Art, Fine Art, Physical Education, and Child Development (Home Economics) are generally among the easier subjects, while Statistics, Modern Foreign Languages and Latin are among the more difficult subjects. For A levels, subjects such as Film Studies, Media Studies, Graphics, among others, are consistently easy, while subjects such as Mathematics, the Sciences, and Modern Foreign Languages are consistently hard. The grade difficulties of GCSE subjects at grades A and C (which are judgemental grades where students' work is evaluated by experts)

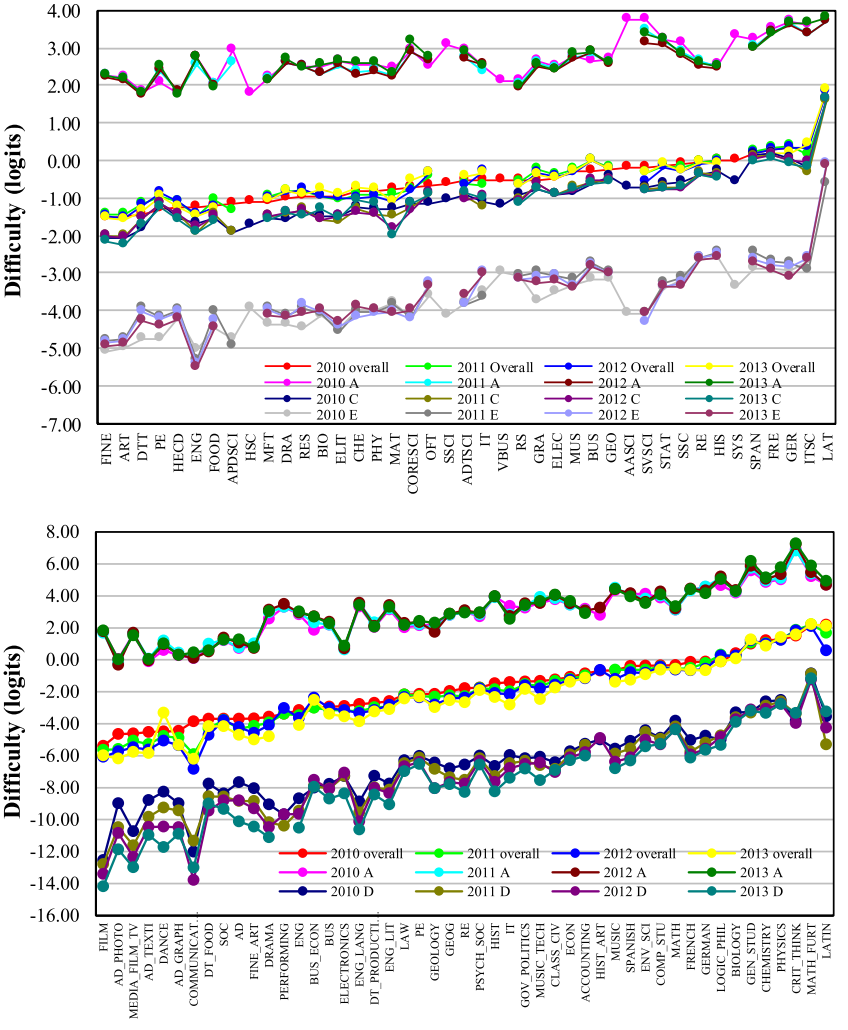

Figure 4. Variation of the over subject difficulties and grade difficulties for the GCSE (top) and GCE A level (bottom) subjects from 2010 to 2013.

and E (which is a non-judgemental grade derived arithmetically) and A level subjects at grades A (a judgemental grade) and D (a non-judgemental grade) from 2010 to 2013 also remain relatively stable over the four year period of study.

4.6. Potential causes of differences in difficulty between subjects

It has been shown that there are considerable differences in Rasch difficulty between subjects at both individual grade level and the overall subject level. Accepting the underlying assumptions of the Rasch model and the level of model–data fit, these differences may be caused by many factors. Detailed discussion of those factors is beyond the scope of this paper. However, Coe (2008) and Coe et al. (2008) discussed a range of potential causes for such differences. These include differences in grading severity (or leniency), level of subject

demand, allocation of teaching time and other resources, motivation of students, efficiency and effectiveness of teaching and learning, and others (also see Lockyer & Newton, 2015; Newton, 2012).

5. Impact of aligning statistical standards

This section explores the impact of aligning statistical standards based on results from the Rasch analysis presented above.

5.1. Impact on examination performance standards

The relative grade difficulty $d_{k,RG}$ represents the proportion of average grade width in scaled score unit (e.g. UMS marks) to be adjusted if standards were to be aligned. If it is negative, the subject is relatively too easy and the boundary score should increase. If it is positive, the subject is relatively too difficult and the corresponding boundary score should decrease. GCSEs and GCEs adopt a standards-based results reporting system to support their defined purposes and the grades awarded to candidates should be interpreted as the levels of attainments in individual subjects. There are well established grade descriptions for both GCSEs and GCEs for the judgemental grades which represent a source of evidence used during awarding. In addition, there are also expectations from users of grades regarding the level of performance that grades represent. A shift in grade boundary scores will likely imply a different performance standard from those established (officially or unofficially) which would impact on the interpretation of grades.

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

$$b'_k = b_k - wd_{k,RG} \tag{8}$$

Equation (8) can be used to investigate the impact of aligning statistical standards on grade distributions. When the UMS mark is used for a qualification, the grade interval at subject level is, by definition, 10% of the maximum available uniform marks. When raw scores are used, the grade interval is the mean of the subject-level grade bandwidths.

5.2. Impact on subject grade distributions

A number of GCSE and A level subjects were investigated further for the impact of aligning statistical standards between subjects on grade distributions. For A level subjects, the effect on A* was not examined. This is due to the non-linear rules for determining those candidates that achieve an A* in specifications using UMS. Instead, the impact on the combination of A* and A was examined.

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

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

5.2.1. GCSE subjects

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

5.2.2. GCE A level subjects

Table 3 shows the original grade distributions for four A level subjects: English Language, Further Mathematics, Physics, and German from the 2013 examination series. Percentage

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

Subject	Number of candidates	Grade distribution (%) and grade boundary change (% of max UMS marks)						
			A*+A	B	C	D	E	U
English Language	24,600	Original (Ind.)	12.76	29.76	35.76	17.77	3.52	0.42
		New (Ind.)	14.76	23.86	26.13	19.35	8.65	7.25
		Change (Ind.)	2.00	−5.90	−9.63	1.58	5.13	6.83
		Original (Cum.)	12.76	42.52	78.28	96.05	99.57	99.99
		New (Cum.)	14.76	38.62	64.75	84.10	92.75	100.00
		Change (Cum.)	2.00	−3.90	−13.53	−11.95	−6.82	0.00
		Boundary shift	−1.02 (A)	1.01	4.02	7.74	13.24	
Mathematics (further)	13,642	Original (Ind.)	56.27	21.61	11.73	5.92	3.00	1.46
		New (Ind.)	67.45	22.16	6.47	3.03	0.78	0.11
		Change (Ind.)	11.18	0.55	−5.26	−2.90	−2.22	−1.35
		Original (Cum.)	56.27	77.88	89.61	95.54	98.54	100.00
		New (Cum.)	67.45	89.61	96.08	99.11	99.88	99.99
		Change (Cum.)	11.18	11.73	6.47	3.57	1.35	0.00
		Boundary shift	−7.42 (A)	−9.98	−13.21	−16.77	−20.64	
Physics	35,781	Original (Ind.)	30.94	23.29	19.41	14.28	9.00	3.07
		New (Ind.)	47.80	22.94	17.57	9.91	1.64	0.14
		Change (Ind.)	16.86	−0.35	−1.85	−4.37	−7.36	−2.93
		Original (Cum.)	30.94	54.24	73.65	87.93	96.92	100.00
		New (Cum.)	47.80	70.75	88.31	98.22	99.86	100.00
		Change (Cum.)	16.86	16.51	14.67	10.29	2.93	0.00
		Boundary shift	−7.13 (A)	−8.30	−10.30	−12.60	−14.41	
German	4032	Original (Ind.)	41.69	26.22	17.96	9.65	3.82	0.67
		New (Ind.)	50.72	24.98	15.33	6.85	1.74	0.40
		Change (Ind.)	9.03	−1.24	−2.63	−2.80	−2.08	−0.27
		Original (Cum.)	41.69	67.91	85.86	95.51	99.33	100.00
		New (Cum.)	50.72	75.69	91.02	97.87	99.60	100.00
		Change (Cum.)	9.03	7.79	5.16	2.36	0.27	0.00
		Boundary shift	−2.93 (A)	−3.63	−4.49	−5.18	−3.03	

changes in candidates at individual grades and the changes in the cumulative percentages of candidates as well as the shift in grade boundary scores, represented using a percentage of the maximum available UMS marks after alignment with the Rasch statistical standards for the average of all subjects, are also listed in the table. For English Language, if the statistical standards were to be aligned, cumulative percentage of candidates at A would go up by about 2%, while that of candidates at C would drop by about 13%. For Further Mathematics and German, such alignment would result in over 11% and 9% increases respectively in the cumulative percentage of candidates at A. At grade C, the cumulative percentage of candidates would increase by over 6% for Further Mathematics and 5% for German. For Physics, the cumulative percentage of candidates at grade A would go up by nearly 17%, which is broadly similar to the findings reported by Alton and Pearson (1996) who attempted to produce a grade distribution of A level Physics that was close to the average grade distribution of a set of A level subjects. At grade C, the cumulative percentage of candidates would increase by over 14%, and the UMS boundary mark would decrease by over 10% of the maximum available UMS marks.

6. Further discussion on the implications of aligning statistical standards

As demonstrated above, differences in statistical standards between different GCSE and GCE subjects based on Rasch analysis exist. Coe et al. (2008) and Newton (2015) discussed

potential consequences of incomparability in standards between subjects and possible policy options that could be taken to address inter-subject comparability issues. One of the proposed options which could have a fundamental impact on the qualifications system was to make grades statistically comparable across subjects. Disregarding the operational difficulties in its implementation and the potential consequences, this section examines the appropriateness of its use in the context of GCSEs and GCEs from the following two perspectives:

- the appropriateness of setting standards based on evidence that the Rasch model can provide, i.e. the existence and relevance of a single trait common to all subjects; and
- the impact of the approach considering the defined purposes of GCSE and GCE A level qualifications.

6.1. Interpretation of Rasch ability and its relevance to standards setting in GCSEs and GCEs

Although the GCSE and A level data analysed in this study appeared to meet the unidimensionality requirement of the Rasch model and fit the model reasonably well, the interpretation of the latent trait specified in the Rasch model is not entirely clear. This is because, as indicated before, the construct represented by the data implied from the Rasch analysis is operationally defined by the set of examinations used. This does not involve an actual measurement process in which the construct in relation to the purpose of the test to be measured must be specified and used to guide the development of the test (in this case defined by the complete range of subjects). The Rasch analysis is based on the relative frequencies of candidates receiving different grades in different examinations and the latent trait inferred is likely to be influenced primarily by the subjects that are correlated well and have large entries. Therefore, to a certain degree, the extent of this shared common trait measured by the exams will likely vary between the subjects in relation to the traits which the individual examinations are designed to measure. With respect to the skills and knowledge assessed by the different examinations, although it is likely that some common skills will be assessed by examinations in different subjects, aspects of knowledge and understanding are generally subject specific and can vary considerably between subjects. It is realised that the unidimensionality requirement of the Rasch measurement model does not require a single psychological process that is responsible for performing different tasks but a single underlying pattern that exists in the data. Although the concept of 'general academic ability' has been used to interpret the latent trait specified in the Rasch model the trait is not explicitly specified for the individual exams but inferred from the analysis.

The current process for setting grade standards in GCSEs and GCEs, which is closely related to the purposes of these examinations (see discussion below), involves the use of both experts' judgement of the performance of students on examination tasks and statistical evidence (see Baird, Cresswell, & Newton, 2000; Ofqual, 2014b). Grade descriptions, which are subject specific, set out the standard of work that is expected for the award of key grades and are used to guide the judgemental process. The statistical evidence generated from Rasch analysis that involves a wide range of subjects is likely to be of limited use when awarding a specific subject.

6.2. Purposes of GCSEs and GCE A levels and interpretation of examination results

The purposes of the reformed GCSEs, which are similar to those of the current GCSEs, are set as follows (see Ofqual, 2014a):

- 'to provide evidence of students' achievements against demanding and fulfilling content;
- to provide a strong foundation for further academic and vocational study and for employment; and
- to provide (if required) a basis for schools and colleges to be held accountable for the performance of all of their students'.

For the reformed GCE A levels, the purposes are defined as (see Ofqual, 2015):

- 'to define and assess achievement of the knowledge, skills and understanding which will be needed by students planning to progress to undergraduate study at a UK higher education establishment, particularly (although not only) in the same subject area;
- to set out a robust and internationally comparable post-16 academic course of study to develop that knowledge, skills and understanding;
- to permit UK universities to accurately identify the level of attainment of students;
- to provide a basis for school and college accountability measures at age 18; and
- to provide a benchmark of academic ability for employers'.

Assessments and examinations are an integral part of the qualification. They should be designed to support the defined purposes of the qualification effectively. Any proposed action (or otherwise) regarding inter-subject comparability should therefore be considered in the context of these purposes.

While the aim of a course leading to a qualification should be to help the learners acquire the required knowledge and skills within a specified domain of content and skills, the purpose of the assessment itself is to provide an accurate measurement of the level of attainment or proficiency that a learner has achieved at the end of the course of study in relation to the purpose of the qualification. The purposes of the qualification therefore to a large extent determine how examination results should be interpreted and reported, which in turn will affect the kinds of comparison in standards between different qualifications that one can make validly and effectively. In line with the defined purposes of general qualifications outlined above, the existing results reporting system is standards-based—there are subject content criteria, assessment objectives, and grade descriptions (which may be used to articulate performance standards) defined for individual qualifications. Therefore, grades should represent the levels of skills and knowledge that the candidates have achieved in specific subject areas. Any cross-subject comparison in terms of 'standards' (be it performance-, prior attainment-, demand-, teaching time-, efforts-related etc.) would seem to be of limited meaning with regard to the purposes of the qualifications as they have been defined. The existing examination processes are aimed at producing evidence of validity to support the standards-based interpretation of exam results.

It has been shown that aligning statistical standards between subjects based on the comparison using the Rasch model would likely result in substantial change in examination performance standards and grade distributions for certain subjects. The interpretation of examination results as a measure of 'general academic ability' will not effectively support the defined purposes of these qualifications. The resultant grade distributions for some

subjects would also no longer be appropriate for effectively differentiating the candidates. Such an alignment may invalidate the interpretation of the results from GCSEs and GCEs as implied by their defined purposes.

7. Concluding remarks

Results from Rasch analysis of GCSE and GCE A level data over a period of four years suggest that the standards of examinations from different subjects are not consistent in terms of the levels of the latent trait specified in the Rasch model required to achieve the same grades. There is considerable variability in statistical standards between subjects at both individual grade level and the overall subject level. Findings from this study are broadly consistent with those from studies reported by other researchers working with similar statistical models.

The latent trait inferred from the Rasch analysis has been interpreted as 'general academic ability'. However, the relationship between this trait and the traits that individual examinations are designed to assess is not entirely clear and the statistical evidence generated is likely to be of limited use for awarding specific subjects. It has been demonstrated that the alignment of statistical standards between subjects based on comparisons using the Rasch model could result in substantial change in performance standards which are based on subject-specific grade criteria and grade distributions.

The defined purposes of a test should determine how its results should be appropriately interpreted and reported. In line with the given purposes of GCSEs and GCEs, grades in general qualifications should be interpreted as the levels of skills and knowledge achieved in specific subjects. Alignment of statistical standards between subjects based on inter-subject comparisons is likely to invalidate the interpretation of results as implied by the defined purposes of these qualifications. The existing grading and results reporting procedures seem to be appropriate for supporting the defined purposes of GCSEs and GCEs.

It is worth noting that, after considering the evidence from existing research on inter-subject comparability and the arguments for and against aligning statistical standards between subjects (including those presented in this paper), the examinations regulator in England, Ofqual, has recently decided not to take any coordinated action to align standards across the full range of GCSE and A level subjects through grading (see Ofqual, 2016). It has, however, committed to considering the evidence (statistical and judgemental) for one-off adjustments to individual subject standards.

Acknowledgements

The authors would like to thank Paul Newton and Beth Black for their helpful comments on an early draft of the paper. The authors would also like to thank the anonymous referees for their constructive comments on the submitted paper.

Disclosure statement

No potential conflict of interest was reported by the authors. The views expressed in this paper are those of the authors and are not to be taken as the views of the Office of Qualifications and Examinations Regulation (Ofqual).

Notes on contributors

Qingping He is a Research Fellow at the Office of Qualifications and Examinations Regulation. His research interests include assessment development, assessment reliability and validity, test equating, item banking, computer-based testing (including computer adaptive testing), and qualification standards.

Ian Stockford is the Director of Qualifications and Markets at AQA. His contribution to this work was made while Associate Director for Research and Analysis at the Office of Qualifications and Examinations Regulation. His research interests consider assessment quality, qualification design, and delivery.

Michelle Meadows is the Executive Director of Strategy, Risk and Research at the Office of Qualifications and Examinations Regulation. Her research focuses on qualification design, assessment development, examination standards setting and maintenance, and the development of regulatory policy.

References

Alton, A., & Pearson, S. (1996). *Statistical approaches to inter-subject comparability* (Unpublished UCLES research paper).

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

Andrich, D. (2015). The problem with the step metaphor for polytomous models for ordinal assessments. *Educational Measurement: Issues and Practice, 34*, 8–14.

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

Bramley, T. (2011). Subject difficulty—the analogy with question difficulty. In *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: Cambridge Assessment.

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

Coe, R., Searle, J., Barmby, P., Jones, K., & Higgins, S. (2008). *Relative difficulty of examinations in different subjects* (Report for SCORE—Science Community Supporting Education). CEM Centre: Durham University. Retrieved from http://www.score-education.org/media/3194/relativedifficulty.pdf.

Elliot, G. (2013). *A guide to comparability terminology and methods*. Cambridge: Cambridge Assessment. Retrieved from http://www.cambridgeassessment.org.uk/images/130424-a-guide-to-comparability-terminology-and-methods.pdf.

He, Q., Anwyll, S., Glanville, M., & 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.

Kolen, M., & Brennan, R. (2014). *Test equating, scaling, and linking. Methods and practices* (3rd ed.). New York: Springer.

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

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

Linacre, J. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*, 878.

Linacre, J. (2015). *Winsteps® Rasch measurement computer program user's guide*. Beaverton, OR: Winsteps.com.

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

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176.
Newton, P. (2012). Making sense of decades of debate on inter-subject comparability in England. *Assessment in Education, 19*, 251–273.
Newton, P. (2015). *Exploring implications of policy options concerning inter-subject comparability* (ISC Working Paper 6). Coventry: Ofqual.
Newton, P., Baird, J., Goldstein, H., Patrick, H., & Tymms, P. (Eds.). (2007). *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.
Newton, P. E., He, Q., & Black, B. (2017). *Progression from GCSE to A level: Comparative Progression Analysis as a new approach to investigating inter-subject comparability*. Coventry: Ofqual.
Ofqual. (2014a). *GCSE (9 to 1) qualification level conditions and requirements*. Coventry: Ofqual. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/371219/2014-04-09-gcse-qualification-level-conditions-and-requirements-april.pdf.
Ofqual. (2014b). *Setting GCSE, AS and A level grade standards in summer 2014 and 2015*. Coventry: Ofqual. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/451321/2015-08-05-summer-series-gcse-as-and-a-level-grade-standards.pdf.
Ofqual. (2015). *GCE qualification level conditions and requirements*. Coventry: Ofqual. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/423720/gce-qualification-level-conditions-and-requirements.pdf.
Ofqual. (2016). *A policy position for Ofqual on inter-subject comparability*. Coventry: Ofqual.
Opposs, D. (2015). *Inter-subject comparability: International review*. Coventry: Ofqual.
Pae, H. (2012). A psychometric measurement model for adult English language learners: Pearson test of English academic. *Educational Research and Evaluation, 18*, 211–229.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmark Paedagogiske Institute. (Expanded edition, 1980. Chicago: University of Chicago Press.)
Reckase, M. (2009). *Multidimensional item response theory*. New York: Springer-Verlag.
Reeve, B., & 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* (pp. 55–73). Oxford: Oxford University Press.
Smith, E. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*, 205–231.
Tan, J., & Yates, S. (2007). A rasch analysis of the academic self-concept questionnaire. *International Education Journal, 8*, 470–484.
Tendeiro, J., & Meijer, R. (2015). *How serious is IRT misfit for practical decision-making?* (Law School Admission Council Research Report, RR 15-04). Retrieved from http://www.lsac.org/docs/default-source/research-(lsacresources)/rr-15-04.pdf
Wong, H., McGrath, C., & King, N. (2011). Rasch validation of the early childhood oral health impact scale. *Community Dent Oral Epidemiology, 39*, 449–457.
Wright, B., & Masters, G. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.
Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement Solutions.
Yen, W. (1993). Scaling performance assessment: Strategies for managing local item dependence. *Journal of Educational Measurement, 20*, 187–213.
Zhao, Y., & Hambleton, R. (2017). Practical consequences of item response theory model misfit in the context of test equating with mixed-format test data. *Frontiers in Psychology, 8*, 484.