CAMBRIDGE ASSESSMENT

Using statistical equating for standard maintaining in GCSEs and A levels

Tom Bramley & Carmen Vidal Rodeiro

Cambridge Assessment Research Report

22nd January 2014

University of Cambridge
Local Examinations Syndicate

Author contact details:

Tom Bramley
ARD Research Division
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG

Bramley.t@cambridgeassessment.org.uk

http://www.cambridgeassessment.org.uk/

How to cite this publication:

Bramley, T. & Vidal Rodeiro, C.L. (2014). *Using statistical equating for standard maintaining in GCSEs and A levels.* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Contents

1. Introduction

The goal of standard maintaining procedures at GCSE and A level is to produce a set of grade boundaries for the units/components of the examinations that will, when candidates' scores on the units/components are aggregated, produce outcomes (grade distributions) that are comparable to those of previous years. What is meant by 'comparable', and how it can be verified that comparability has been achieved, are of course complex issues that have attracted a great deal of research attention (e.g. Newton et al., 2007; Cambridge Assessment, 2011). The regulator's Code of Practice (Ofqual, 2011) makes it clear that the decisions on where to locate the grade boundaries should draw on a number of sources of evidence (Code of Practice section 6.13). To our knowledge it has never been stated in writing what priority should be given to different sources of evidence, nor what the criterion for correctness is when setting the boundaries.

A *statistical* criterion for correctness might be "The boundaries that would give this cohort of examinees the same distribution of grades on the current examination as on the previous (or a specified reference) examination". In practice we can never achieve the conditions that would allow a comparison with this criterion because different cohorts of examinees take different examinations. But specifying a statistical criterion, even a hypothetical one, would immediately give a rationale for using (or prioritising) a statistical method for arriving at the grade boundaries.

In recent years a statistical method known as the 'comparable outcomes' approach (see later) has become the dominant source of evidence used in setting grade boundaries. The output of the comparable outcomes approach is a target or 'putative' grade distribution for the subset of examinees who have been successfully matched with a measure of prior attainment. At A level this measure of prior attainment is the mean GCSE score; and at GCSE the measure is based on Key Stage 2 (KS2) performance[1].

As far as we are aware, the comparable outcomes method has never been formally linked to any of the established methods in the statistical test equating literature. In this literature (e.g. Kolen & Brennan, 2004; Holland & Dorans, 2006), equating is defined as the task of discovering a transformation of the score scale on one test such that the transformed (equated) scores can be used (i.e. interpreted) interchangeably with those on another test.

One reason for this lack of linking to the equating literature might be that GCSEs and A level examinations have been perceived as too different from the kind of 'tests' (such as US SATs) on which much of the equating theory has been developed. There certainly are many salient differences: the type of items used; the number of components in an assessment; the amount of examinee choice of questions / components available; the use of letter grades for reporting – to name but a few. Another reason might be that the introduction of the comparable outcomes approach occurred when most GCSEs and all A levels had a modular structure, meaning that there were no aggregate raw score scales to equate – grade boundaries derived at unit level determined the Uniform Mark Scale (UMS) scores which were then aggregated to a total UMS score from which the final grades were determined. Alternatively, it may be that in fact the link with equating *has* been made before, but only in confidential internal reports to technical groups such as the Joint Council on Qualifications (JCQ) Standards and Advisory Group (STAG).

We will take the view that most assessment experts and wider stakeholders in education in England have assumed that the exam boards are applying what Newton (2011) has dubbed the 'similar cohort adage' when maintaining standards – namely the default position that if the cohort for an examination seems not to have changed much from a previous cohort then the grade distribution should not differ much either. This relatively relaxed conception allows for statistical input to the setting of grade boundaries, but also for the consideration of other evidence too – such as expert judgment of the quality of work produced – which can form part of a rational argument for setting boundaries which cause the grade distributions to deviate from the default position.

[1] When it was available, Key Stage 3 (KS3) performance was used since KS3 tests were taken only two years prior to GCSE as opposed to five years prior. KS3 tests were phased out in 2008, so since 2011 KS2 has been used as the basis for the prior attainment measure.

However, the increased dominance of the comparable outcomes method, partly because of its suitability as a regulatory tool for ensuring (one form of) comparability both over time and across exam boards, arguably suggests that something closer to the statistical criterion stated above is being used as the de facto standard of correctness when setting boundaries. This means that it is worthwhile to evaluate the comparable outcomes method as a statistical equating method. Furthermore, the fact that currently proposed reforms to both GCSE and A level (Ofqual, 2013) seem likely to produce a return to linear (as opposed to modular) exams means that in many cases there will be a clearly defined aggregate raw score scale that could be statistically equated to the aggregate raw score scale of a previous exam – in which case it may be worth investigating whether other statistical equating methods may be equally or more appropriate than the comparable outcomes method.

In this report we first show how the comparable outcomes method is related to the frequency estimation equipercentile equating method for the non-equivalent groups anchor test (NEAT) data collection design. We then show the results of using a number of different equating methods on a dataset taken from a linear Religious Education (RE) GCSE examination in 2009 and 2010, and compare them with each other and with the comparable outcomes method. Finally we extend the comparison to a tiered Mathematics GCSE to investigate how closely the comparable outcomes method corresponds to statistical equating methods in different circumstances.

2. Statistical equating methods

Equating is deemed necessary because it is accepted that exams can differ in difficulty. If they did not, then there would be no problem – a raw score of 23 out of 80 on one exam would mean the same as a score of 23 out of 80 on another exam. If such exams were graded (like GCSEs and A levels) the same grade boundaries could be used every time. Given the lack of transparency of some statistical equating methods and the difficult conceptual issues they raise even for simple tests it has been suggested that it may even be worth using fixed boundaries and putting up with the consequences for complex multifaceted assessments like GCSEs and A levels (Bramley, 2013).

This section gives a brief summary of how statistical equating is conceptualised, and a short description of some commonly used methods. Many more details can be found in journal articles and the following books (Holland & Dorans, 2006; Kolen & Brennan 2004; von Davier, Holland & Thayer, 2004), from which the descriptions below are based.

As mentioned above, the goal of equating is to find a function that transforms scores on one test such that they can be used interchangeably with those on another test. Several conditions (or desirable properties of equating relationships) of varying stringency are often cited:

- The two tests should be measuring the same thing (same construct);
- The equating transformation should be symmetric (in other words the transformation of scores from test X to test Y should be the inverse of the transformation of scores from test Y to test X);
- It should be a 'matter of indifference' to examinees whether they take test X or test Y;
- Group invariance – the equating transformation should not depend on the group of examinees used to determine the equating relationship (for example, the function should be the same whether it is derived from a group of males or females).

Several different data collection designs are considered in the equating literature. The one most relevant to the statistical criterion above is the 'single group design' where the same group of examinees takes both tests. However, in the context of GCSEs and A levels this design is never implemented[2]. The data collection design that most closely resembles the GCSE and A level context is the non-equivalent groups anchor test (NEAT) design. Group A takes test X, Group B

[2] The subset of examinees resitting an examination cannot be considered to be of equivalent ability.

takes test Y, and both groups take the anchor test V. The exam cohorts[3] from two different years are considered to be non-equivalent (i.e. Groups A and B), rather than randomly equivalent samples from the same population. But what is the anchor test? There are no common items or common papers from one year to another[4] which could be considered as an 'internal anchor'. Nor is there a common reference test taken by both cohorts[5] that could be used as an 'external anchor'. However, there is some information about the relative ability of the two cohorts – their prior attainment. At GCSE this is information from national testing at KS3 or KS2. At A level it is the mean GCSE grade. Neither of these indicators strictly (or arguably even loosely – see discussion) meets the criteria for consideration as an anchor test, first because each cohort will have taken a different KS2/3 test or different GCSEs; and second because these measures were taken a considerable time prior to the exams being equated. However, we will show below that these scores are treated in a way analogous to an anchor test in the comparable outcomes method.

Given a particular equating design, there is a choice in the type of mathematical function used to transform the scores from one test to another. The most general is an equipercentile function which equates scores on test X and Y that are at the same percentile rank in a specified population. Equipercentile functions are non-linear – the equated scores can be 'compressed' at some parts of the scale and 'stretched' at other parts. A linear equating function is more restrictive. It equates scores on test X and Y that are at the same number of standard deviations from the mean in a specified population. A given difference in raw scores will correspond to the same difference in equated scores at all parts of the scale. Further choices can arise in whether and how to carry out any smoothing (either of the raw score distributions, or the equating function). These details are generally beyond the scope of this report, but see the section on kernel equating below for some more information.

One way of equating data from the NEAT design uses the concept of a 'synthetic population'. The scores on the anchor test allow the estimation of the distribution of scores of Group B on test X and Group A on test Y (neither of these was observed). Then an equating relationship can be derived for a synthetic population consisting of a weighted proportion of examinees from each group. Common choices for the weights are 0.5 (equal weight), relative size of group (i.e. giving more weight to the group with more examinees) or 1 and 0 (giving all the weight to one group). Equipercentile equating under this approach is known as the 'frequency estimation' method or 'post-stratification equating' (PSE). Linear equating under this approach is known as the Tucker method after its inventor, Ledyard Tucker.

A different approach to equating with the same NEAT design is known as the 'chained' approach. Scores on test X are equated to the anchor test V in Group A, and scores on test Y are equated to the anchor test V in Group B. Scores on tests X and Y that 'map' to the same score on test V are deemed equivalent. These equatings can be either linear ('chained linear') or equipercentile ('chained equipercentile').

The two approaches do not necessarily generate the same results. There has been a lot of debate about which is better in which circumstances. A distillation of this debate might be that chained methods are more robust when Groups A and B differ in ability, but the frequency estimation method is preferable when they are similar – but this is probably an oversimplification. (See for example Livingston, Dorans & Wright, 1990; Holland, Sinharay, Von Davier & Han, 2008; Wang, Lee, Brennan & Kolen, 2008).

Both these approaches are known as 'observed score equating'. Yet more equating methods exist, using the 'true score' concept from classical test theory, or the 'underlying ability' concept from item response theory. These methods are not considered in this report.

[3] Here 'cohort' simply means the candidates entering for a particular examination, not the entire age cohort.
[4] For simplicity it is here assumed that the 'equating' is from the main examination session in June of one year to the next.
[5] Ofqual (2013) mentions plans to develop a reference test to help with maintaining standards at GCSE.

A recent unified approach to observed score equating is the 'kernel' method developed at ETS (von Davier et al., 2004; von Davier, 2013). It was introduced by Holland and Thayer (1989), who described it as "a new and unified approach to test equating based on log-linear models for smoothing score distributions and on the kernel method of non-parametric density estimation". The Kernel method has a number of advantages over other test equating methods (e.g. linear equating and equipercentile equating). In particular it provides explicit formulas for the standard errors of equating in all data collection designs. The book by von Davier et al. (ibid.) provides a very comprehensive description of the method, which is briefly summarised below.

The first step of the Kernel method of test equating is called *pre-smoothing*. In this stage, a statistical model is fitted to the empirical frequency distribution obtained from the sample data. One way to perform the pre-smoothing is by fitting a polynomial log-linear model to the proportions obtained from the raw data (Holland & Thayer, 2000). It is possible to specify quite complicated log-linear models that cater for the particular features of the data that arises when conducting test equating. There might be some complexities in the data (such as spikes or gaps in the score distribution) that should be accounted for when conducting pre-smoothing. The way that this is carried out is by adding indicator variables in the log-linear model for the particular score values that exhibit irregularities.

In selecting the log-linear model, it is recommended to use a criterion such as the Akaike Information Criterion (AIC) to compare models to each other, and then verify the suitability of the chosen model by assessing the conditional parameters. If the conditional parameters (e.g. means and variances) of the estimated distribution do not deviate too much from the conditional parameters of the observed distribution, then the estimated log-linear model is appropriate to use. If that is not the case then additional parameters may need to be added to accurately model the data. The aim of the pre-smoothing step is to find a model that describes the data 'well enough' with as few parameters as possible.

The next step of the Kernel method is called *continuization*. This step is necessary because it is generally impossible to map one discrete observe-score distribution to another preserving all percentile ranks. The method uses kernel smoothing in this stage. Smoothing is a statistical technique for estimating a function $f(x)$ using data observations, when no parametric model for the function is known. The estimated function is smooth, and the level of smoothness is set by a single parameter, called bandwidth. One of the simplest methods of smoothing, and the one used in this method of test equating, is a kernel smoother, which defines a set of weights $\{W_i(x)\}_{i=1}^{n}$ for each point *x* and defines

$$\hat{f}(x) = \sum_{i=1}^{n} W_i(x) y_i.$$

A kernel smoother in practice represents the set of weights by describing the shape of the weight function via a density function with a scale parameter (the bandwidth) which adjusts the size and the form of the weights. It is common to refer to this shape function as a kernel. The kernel is a continuous, bounded, and symmetric real function which integrates to one. A natural candidate for the kernel is the standard Gaussian density. More details about smoothing and, in particular, about kernel smoothers can be found in Simonoff (1996).

In the Kernel method of equating, the smoothed versions of the two population distributions are used in the equating function. The more current proposal for implementing kernel equating (von Davier et al. (2004)) is to use bandwidths that vary according to the sample data, meaning that the kernel continuization function works like a smoother when pre-smoothed data are rough. The bandwidth selected for an extremely smooth distribution tends to be fairly small (e.g. 0.5 or 0.6), while the bandwidth of a distribution that retains only the mean and variance of the discrete distribution needs to be large (e.g. greater than 10 times the standard deviation of the distribution).

The last step in the Kernel equating method is the *evaluation of the equating transformation*, via the computation of the standard error of equating (SEE, see below). The method provides a

general expression for the standard error that is derived using the δ-method[6]. The SEE is strongly affected by sample size and is always lower for smoothed data than for raw data, regardless of whether the pre-smoothing model was correct.

According to von Davier et al. (2004) there are two major areas where the Kernel equating method can be viewed as an improvement on other methods. Firstly, it will often have smaller standard errors and is less subject to sampling variability. Thus, it is well suited to applications where sample sizes are small. Secondly, it is a consistent system that develops equating functions and their estimated standard errors in a similar way across all of the commonly used data collection designs. Until recently, the requirement for specialist software was the main drawback to using the method, but with the release of the free R package *kequate* (Andersson, Branberg & Wiberg, 2013), this is no longer an issue.

Errors of equating can be systematic or random. Random error arises when Groups A and B are considered to be random samples from some population, and the interest is in quantifying the variability of sample equating results from the population result. In most equating studies the samples are not random, and when the context is GCSEs or A levels they are arguably not even samples but rather the complete data. Nonetheless, since the equating error is reported for each raw score, there is still an intuitive appeal to the idea that there is more equating error in the parts of the score range where the data was sparse, and it is useful to be able to see the extent of that error.

Systematic error can only be defined when the 'correct' result is known or defined in some way, which is usually only in the context of simulation studies, though there are exceptions (e.g. Holland et al. 2008). There can be a trade-off between systematic and random error – for example a method with some systematic error (or bias) might still be preferable to a method with no systematic error if its random error was substantially lower. Equating methods are usually evaluated with respect to both types of error.

3. The comparable outcomes method

Despite it having been used as a source of evidence in setting grade boundaries in live high-stakes examinations since 2001, for a long time there has been little or no publicly available documentation at a technical level giving the rationale or statistical details of how the comparable outcomes method works. There is a description of the method in Benton & Lin (2011), but not at the level of explicitly stating assumptions and statistical formulas. A more precise technical description presented as an algorithm with a worked example is given in Taylor (2013), but this document is not yet published.

In practice the comparable outcomes method is used both to maintain standards within boards over time and to maintain comparability between boards at a given time. In this paper we consider only its function of maintaining standards within boards over time, because here the conditions are most similar to those required for statistical equating (namely tests constructed to the same specification). Ignoring the special problems that can arise when syllabuses change, and taking a simplified version of the algorithm that is actually used[7] we can see the parallels with the frequency estimation equipercentile equating method in Table 3.1 below.

[6] The δ-method is a general approach for computing standard errors. It takes a function that is too complex for analytically computing the standard error, creates a linear approximation of that function, (using a Taylor series expansion) and then computes the standard error of the simpler linear function. Details about the method and how to apply it can be found, for example, in Oehlert (1992) or Davison (2003).

[7] In practice the current test is related to a weighted average of more than one previous year's test, and there is a statistical adjustment made to 'allow' for inflation of KS2 scores over time (see Taylor, 2013).

Table 3.1: Comparison of comparable outcomes method with the frequency estimation equipercentile equating method for a linear GCSE or A level.

	Comparable Outcomes	Frequency Estimation
Group A and B	Two different exam cohorts, A=current, B=previous	Samples of test-takers from different populations
Test X and Y	Complete exams (multiple components) designed to same specification, X=current, Y=previous	Test forms constructed to same specification
Anchor test V	Measure of prior attainment	Score on concurrent set of items internal or external to X and Y
Synthetic population weights	All weight on current cohort (Group A)	Analyst's choice
Explicit (publicly stated or acknowledged) assumptions	Value-added relationship between prior attainment and exam score is the same for Groups A and B.	For both X and Y, the conditional distribution of total score given each anchor score is the same in Groups A and B. The anchor test is representative of the tests to be equated in content and difficulty.
Implicit assumptions	1. It is only of interest to equate X to Y for the current cohort (A). 2. Prior attainment scores are already equated. 3. For Y, the conditional distribution of total score given each prior attainment score is the same in Groups A and B.	Percentile and percentile rank functions satisfactorily continuize the discrete score distribution.
Known quantities	Cumulative grade distribution at each prior attainment score for Group B on test Y; distribution of prior attainment scores for Group A; cumulative distribution of scores of Group A on test X.	Score distribution at each anchor test score for Group A on test X and Group B on test Y.
Estimated quantities	Cumulative grade distribution for Group A on test Y	Score distribution for Group B on test X and Group A on test Y (unless one of the synthetic population weights is zero).
Estimator*	$\sum_v G_B(\tilde{y}\|v)h_A(v)$	$\sum_v g_B(y\|v)h_A(v)$
Output	'Putative' cumulative grade distribution for Group A on test X	Equating function mapping raw scores on test X to test Y equivalents, sometimes with standard errors of equating at each test X score point.
Use of output	Find raw scores on test X that give closest cumulative percentage to putative distribution at key grade boundaries.	Equated scores on test X can now be used as if they had come from test Y (e.g. reported on the same scale).
Nature of mapping	Integer key grade boundaries on test X to integer grade boundaries on test Y.	Integer raw scores on test X mapped to non-integer equated scores on test Y scale.

* where:
- the summation is over anchor test scores / prior attainment measures *v*;

- $G_B(\tilde{y}|v)$ is the cumulative proportion at grade $\tilde{y}$ on test Y for a given prior attainment measure *v* in Group B;
- $g_B(y|v)$ is the proportion at score *y* on test Y for a given anchor test score *v* in Group B;
- $h_A(v)$ is the proportion at score *v* on the anchor test / prior attainment measure in Group A.

The differences between the estimators used in these methods are:
i) the cumulative frequency *G*[8], rather than the frequency *g* is used for the comparable outcomes method (this makes no difference mathematically).
ii) the test Y grade (here symbolised ỹ) rather than the test Y score y is used in the comparable outcomes method.

Thus it can be seen that the two methods are structurally very similar, especially in the quantities that are estimated from the data. The main difference is that the comparable outcomes method is only seeking to 'equate' at the points in the score distribution corresponding to the key grade boundaries, whereas equating methods map every score point. At GCSE these are A, C and F, and from now on this report focuses on GCSE. A more minor difference is that in the comparable outcomes method the grade boundaries are always set at integer points on the raw score scale – thus by definition the test X raw scores at A, C and F map to the test Y raw scores at A, C and F. In equating methods the definitions of the symmetric equating functions apply to continuous variables, so some kind of continuization procedure is necessary to make the discrete test scores continuous, and the equated scores are thus not necessarily (or even usually) integers.

For the purposes of this report it was necessary to treat the comparable outcomes method as an equating method that could map every score on test X to an equivalent score on test Y. This was done using the following steps:

- The grade boundaries on test X at A, C and F were determined as the integer score points giving a cumulative percentage at that grade as close as possible to the 'putative' value.
- The other arithmetically determined grade boundaries on test X were found by applying the usual rules for calculation of such boundaries (see Appendix 2 of Ofqual, 2011).
- The mapping from test X to test Y at all other score points was obtained by linear interpolation (using zero and maximum marks as limits). That is, if p_X and q_X represented adjacent grade boundaries on test X, and p_Y and q_Y represented the same adjacent grade boundaries on test Y, then an (integer) score *s* between p_X and q_X on test X would map to a (possibly non-integer) score on test Y of
$$p_y + \frac{(q_Y - p_Y)}{(q_X - p_X)} \times (s - p_X)$$

4. Data

4.1 GCSE Religious Studies

For the first analyses, a single dataset was compiled from the OCR's linear GCSE in Religious Studies (syllabus code 1931, option A), from 2009 and 2010. The aggregate raw score scale ran from 0 to 168. Candidates were matched with their KS3 levels using the National Pupil Database (NPD). Candidates who had been scaled (i.e. their marker had been scaled) were dropped, as were any candidates without a matched KS3 score. A pseudo 'anchor test score' for use in the NEAT design equating methods was created from the average KS3 levels in Maths, English and Science as shown in Table 4.1 below.

[8] By tradition these cumulative proportions count down from the top rather than the more usual practice of counting up from zero.

Table 4.1: Anchor test scores created from KS3 levels.

Anchor score	Mean KS3 level	2009 Percent (N=23263)	2010 Percent (N=26472)
0	<3	0.11	0.17
1	=3	0.74	0.90
2	>3 to <3.5	0.46	0.48
3	3.5 to <4	1.23	1.56
4	=4	2.79	3.10
5	>4 to < 4.5	3.84	3.89
6	4.5 to <5	5.57	5.90
7	=5	9.80	9.42
8	>5 to < 5.5	11.76	11.25
9	5.5 to <6	12.41	12.35
10	=6	13.20	13.54
11	>6 to < 6.5	11.84	12.71
12	6.5 to <7	10.31	9.66
13	=7	9.53	8.45
14	>7 to <7.5	6.03	6.18
15	>=7.5	0.40	0.43

Table 4.2: Prior attainment categories used in comparable outcomes method.

ks3decile	Mean KS3 level	2009 Percent (N=23263)	2010 Percent (N=26472)	Equivalent anchor test scores
high 1	>=6.334	26.27	24.73	12-15
2	6.001 to 6.334	11.84	12.71	11
3	5.668 to <6.001	13.20	13.54	10
4	5.334 to <5.668	12.41	12.35	9
5	5.001 to <5.334	11.76	11.25	8
6	4.668 to <5.001	9.80	9.42	7
7	4.334 to <4.668	5.57	5.90	6
8	4.001 to <4.334	3.84	3.89	5
9	3.334 to <4.001	4.02	4.66	3-4
low 10	<3.334	1.31	1.54	0-2

Table 4.3: Descriptive statistics for RE total test scores and total anchor test scores.

Year	Variable	N	Mean	Std Dev	Minimum	Maximum	Correlation
2009	RE_tot	23263	108.00	30.95	0	168	0.679
	anc_tot	23263	9.44	2.84	0	15	
2010	RE_tot	26472	107.74	32.62	0	168	0.683
	anc_tot	26472	9.35	2.89	0	15	

For the comparable outcomes method, the mean KS3 scores were assigned to the categories actually used in that method (taking the cut-off values from Table 9 of Pinot de Moira (2008)). There are 10 categories referred to as 'deciles' even though they are not strictly deciles. Table 4.2 shows the categories and their relative frequencies. (Note that the 'deciles' would not be expected to contain equal proportions of candidates for a specific exam entry – the categorisation applies to the entire GCSE cohort).

Table 4.3 shows that the 2010 RE cohort had slightly lower mean prior attainment, but also was slightly more spread out in terms of prior attainment than the 2009 cohort. The 2010 and 2009 RE score distributions had similar means, but the 2010 distribution was more spread out (higher standard deviation).

4.2 GCSE Maths

The second set of analyses used OCR's tiered linear GCSE in Mathematics (syllabus code J512) from 2009 and 2010. The matching with KS3 and creation of pseudo-anchor test scores were carried out in the same way as for the RE. This particular pair of years was chosen because from the actual grade boundaries it appeared that there had been a large change in the difficulty of the exam on both tiers, with the 2010 exam being harder (lower boundaries) than the 2009 exam. We were also interested to compare the different equating methods in cohorts that were more restricted in range of ability than the RE cohort. The aggregate Maths raw score scale on both tiers ran from 0 to 200, and the descriptive statistics are shown in Tables 4.4 and 4.5.

Table 4.4: Descriptive statistics for Maths Foundation tier total test scores and total anchor test scores.

Year	Variable	N	Mean	Std Dev	Minimum	Maximum	Correlation
2009	Maths_tot	15878	118.64	38.84	0	194	0.762
	anc_tot	15878	6.09	2.20	0	13	
2010	Maths_tot	15976	106.93	36.20	0	190	0.730
	anc_tot	15976	5.90	2.21	0	13	

Table 4.5: Descriptive statistics for Maths Higher tier total test scores and total anchor test scores.

Year	Variable	N	Mean	Std Dev	Minimum	Maximum	Correlation
2009	Maths_tot	11767	119.96	36.56	9	200	0.735
	anc_tot	11767	10.74	1.96	1	15	
2010	Maths_tot	11228	107.00	41.53	0	199	0.754
	anc_tot	11228	10.46	2.04	1	15	

Comparing tables 4.4 and 4.5 with Table 4.3 for RE it can be seen that each tier on the Maths exam did have a narrower range of prior attainment (smaller anchor test SD), and that in terms of overall level of ability the Maths Higher tier cohort was more similar to the RE cohort. Despite the smaller ranges of ability, the correlations of Maths scores with anchor scores on both tiers were higher than for RE. This could be explained if the Maths GCSE construct is more similar than the RE GCSE construct to whatever construct (general academic ability?) is represented by the combined KS3 score – which is plausible since the KS3 score comprises maths, English and science but not RE. Also of course it would seem likely that the GCSE RE scores would contain more error attributable to marker variance given the less objective nature of the mark schemes, which would tend to lower their correlation with any other variable. However, Benton & Sutch (2013) found that GCSE RE grades correlated more highly with mean GCSE grade than GCSE maths grades did, which does not support this supposition.

5. Equating results

The equating analyses below were carried out using the open source statistical software R (R core team, 2013) with the packages *equate* (Albano, 2013) and, for the kernel methods, *kequate* (Andersson et al., 2013). The graphs in this section showing the results of the equating have the 2010 (test X) score on the x-axis, and the difference between the equated scores and the 2010 scores on the y-axis. Negative values mean that the 2010 test was easier at this score point (i.e. you would need to reduce scores to equate them to 2009) and vice versa.

5.1 RE dataset

Figure 1 below shows the results for the comparable outcomes method, the frequency estimation method with all weight on the 2010 population, the chained equipercentile method, and the (linear) Tucker method. The vertical reference lines show the grade boundaries in 2010 according to the comparable outcomes method (not the actual boundaries).

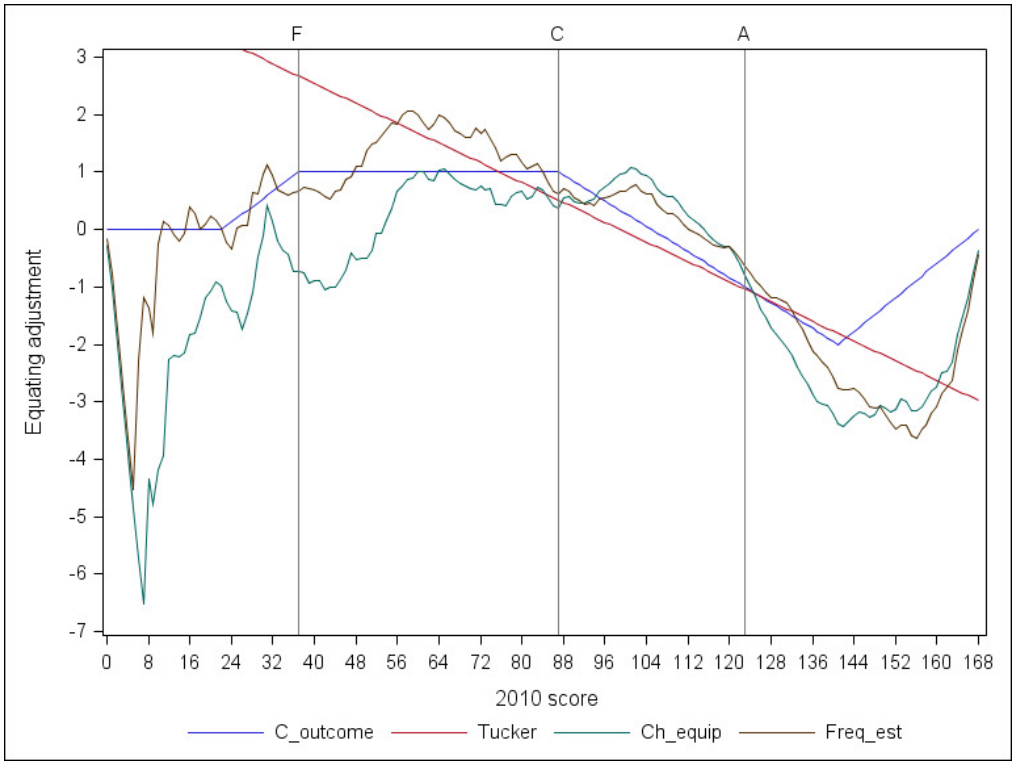

Figure 1: GCSE RE dataset, comparison of four equating methods.

All four methods gave very similar results, especially in the range of marks from slightly below C to slightly above grade A. As anticipated, the frequency estimation equipercentile method was

closest to the comparable outcomes method at the three key grade boundaries[9]. The three non-linear methods were closer to each other than to the linear Tucker method at lower scores, and there were noticeable differences between the two equipercentile methods at scores below grade C, with the frequency estimation method consistently producing a more positive equating adjustment – i.e. implying that the 2010 test was 'harder' with this equating method than with the chained method. The large equating adjustments implied by the equipercentile methods at very low scores were caused by the scarcity of data at that part of the mark range and probably reflect equating error (see later graph).

Figure 2 below shows the results for the frequency estimation method, the chained equipercentile method, and the kernel versions of the frequency estimation and chained equipercentile methods.

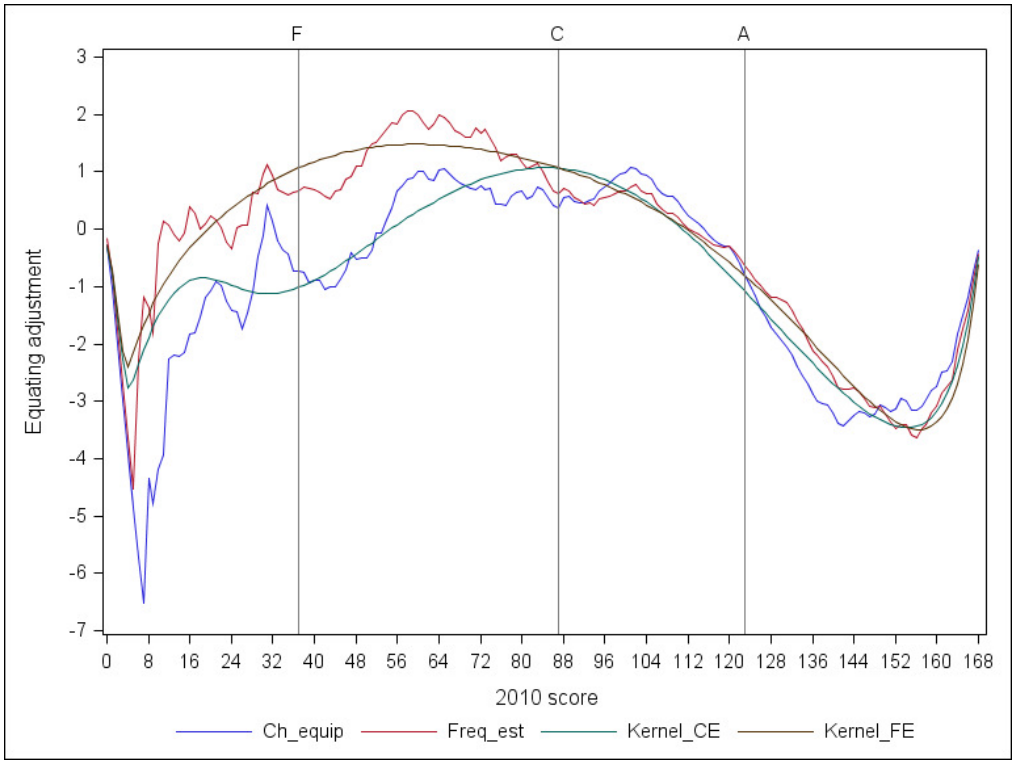

Figure 2: GCSE RE dataset, comparison of equipercentile methods with kernel equivalents.

It can be seen from Figure 2 how the kernel approach smoothed the respective results for chained equipercentile equating and frequency estimation equipercentile equating. In particular, the kernel approach reduced the adjustments implied at low scores and in general gave a more aesthetically appealing result.

Figure 3 below shows that two possible variations on the frequency estimation method had little practical significance for this dataset. One variation was to use the relative proportions of candidates in the 2009 and 2010 cohorts as synthetic weights (as opposed to putting all the weight on the 2010 cohort). The other was to define the 'anchor test' using exactly the same categories as used in the comparable outcomes method – i.e. with 10 categories as opposed to 16.

[9] The chained equipercentile method was actually slightly closer at A, but because the comparable outcomes method by definition maps integer to integer at the key boundaries the frequency estimation method can be up to 0.5 marks different at these boundaries (though it can differ by more elsewhere).

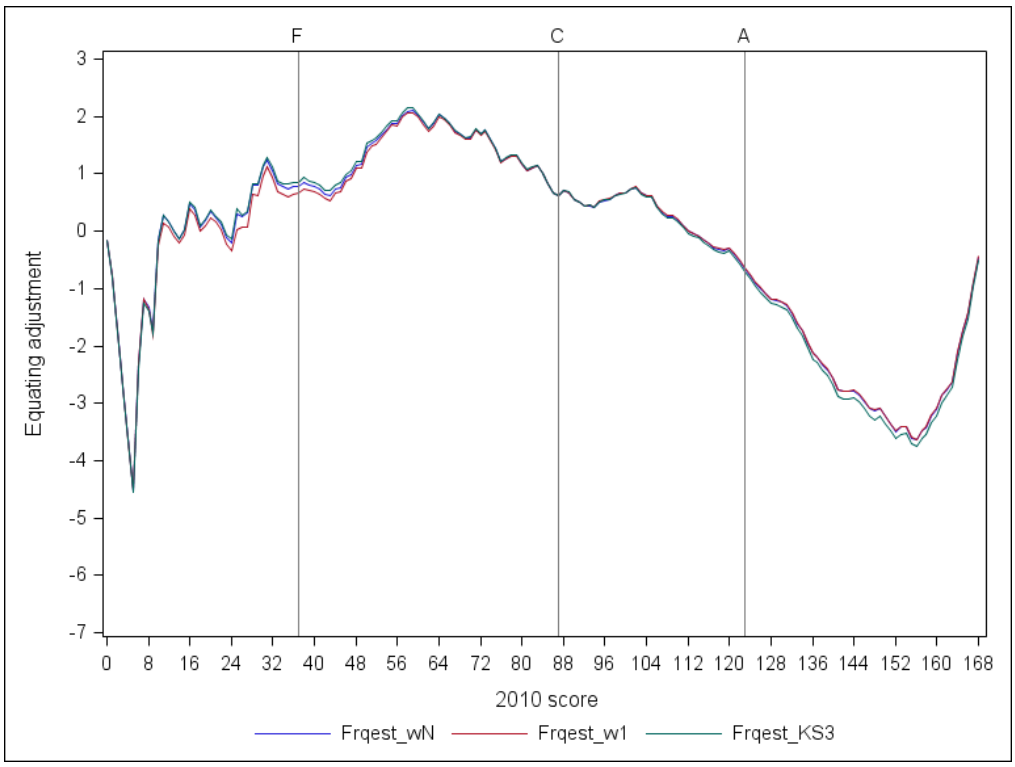

Figure 3: GCSE RE dataset, comparison of three variants of the frequency estimation method. (Key: _wN uses the relative weights of the 2009 and 2010 cohort sizes, _w1 has all weight on 2010, _KS3 uses the same categorisation as the KS3 deciles used in the comparable outcomes method).

Varying the sample size

To compare the robustness of the equating results from the different methods with respect to variations in sample size, two further datasets were created: a 'medium' dataset with around 1800 candidates from each year, and a 'small' dataset with around 700 candidates from each year. Although these might in some contexts both be considered small datasets, the sizes were chosen to reflect Ofqual's categorisation of entry sizes for 'tolerances' when using the comparable outcomes method: there is no set tolerance for N<=500; a tolerance of 3% for 500<N<=1000; a tolerance of 2% for 1000<N<=3000; and a tolerance of 1% for N>3000. The tolerance refers to the acceptable deviation (at grades A and C for GCSE) between the cumulative percentage outcome for the matched candidates (i.e. those who could be matched with their prior attainment score) and the 'putative' percentage outcome generated by the comparable outcomes method for those candidates. Deviations outside tolerance need to be justified to the regulator.

The medium and small datasets were created by randomly sampling centres from each year and including all candidates from those centres, then varying the number of centres sampled and repeating until the resulting numbers were in the ranges that would attract a 2% and 3% tolerance respectively. It was deemed better to sample centres rather than examinees given that from one year to another it is more realistic to assume that centres will start or stop following the syllabus and entering their candidates for the exam. However, no consideration was given to the actual proportion of 'churn' from one year to the next so it is likely that the 2009 and 2010 sub-groups created by this process were more different than would occur normally in exams with cohorts of this size. But equating results are supposed to be the same regardless of sub-population (the population invariance condition mentioned earlier) so it was still of interest to explore the extent to which equating results fluctuated.

Table 5.1: Descriptive statistics for the RE medium and small datasets.

	Year		N	Mean	Std Dev	Min	Max	Correlation
Medium	2009	RE_tot	1842	114.64	28.72	0	168	0.647
		anc_tot	1842	9.96	2.75	0	15	
	2010	RE_tot	1768	110.54	35.88	0	168	0.740
		anc_tot	1768	9.29	2.89	0	15	
Small	2009	RE_tot	671	110.70	30.26	10	167	0.666
		anc_tot	671	9.47	2.99	0	15	
	2010	RE_tot	714	99.68	33.75	13	166	0.695
		anc_tot	714	8.39	3.14	0	14	

Comparing Table 4.3 with Table 5.1 it can be seen that the 2009 and 2010 cohorts in the medium and small datasets were more different from each other than in the full dataset. Figures 4 to 6 show, respectively, the comparable outcomes, frequency estimation equipercentile and chained equipercentile equating results for the full, medium and small datasets. The vertical reference lines are the same as in the earlier figures.

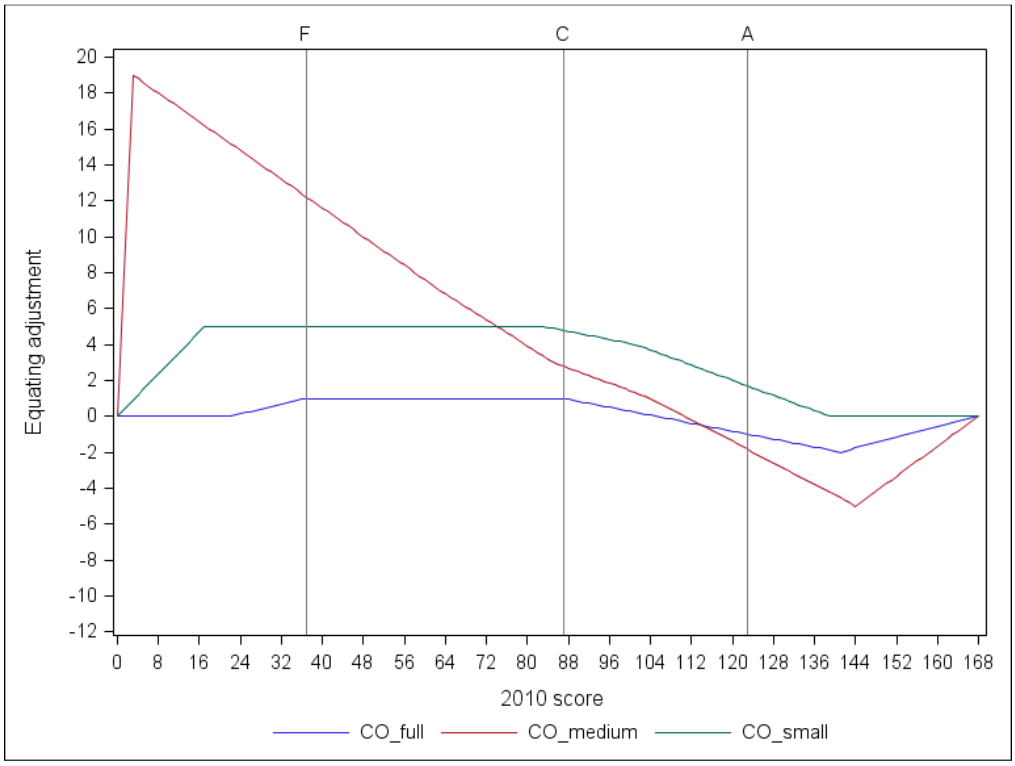

Figure 4: GCSE RE datasets – comparison of comparable outcomes results.

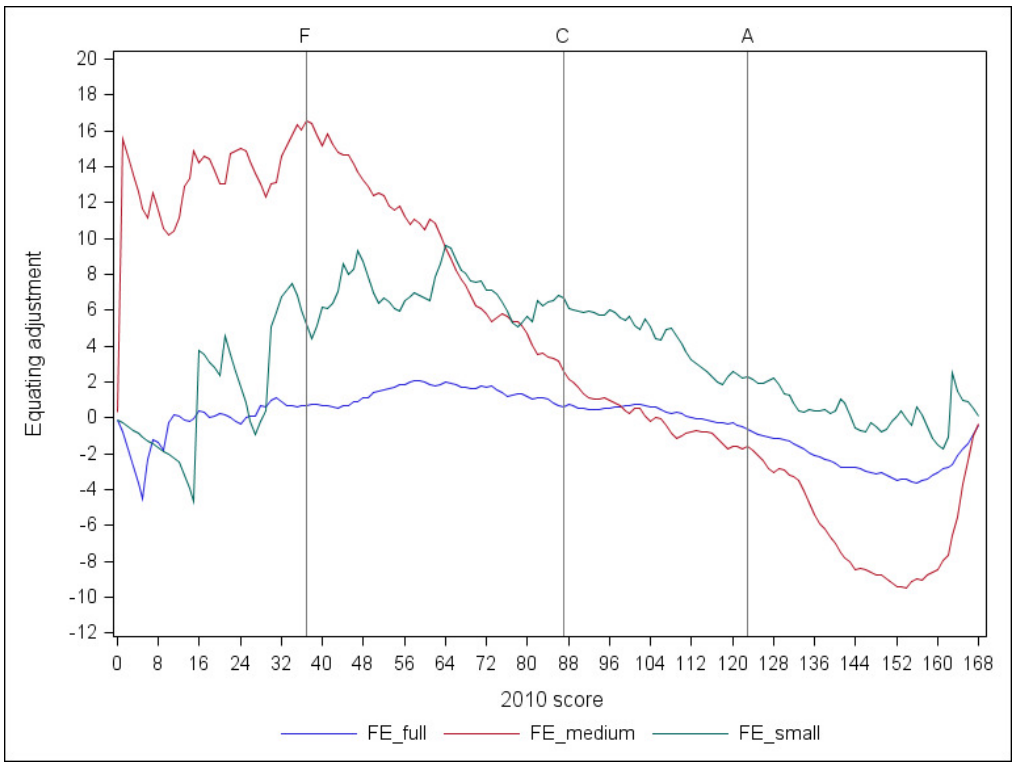

Figure 5: GCSE RE datasets – comparison of frequency estimation equipercentile results.

Figure 4 shows that with the comparable outcomes (CO) method, the medium dataset gave a result that was within 2 marks of the full dataset at grade C and within 1 mark at grade A. In the 2010 (full) dataset there were around 0.87% of candidates on each mark around the C boundary, and around 1.2% on each mark around the A boundary. So both results would have been just within the allowable tolerance of 2%. The small dataset gave a result that was within 4 marks at C and within 3 marks at A, both just outside the allowable tolerance of 3%. The discrepancies were much larger at F and below – a reflection of the fact that there was very little data (even in the full dataset) at this part of the mark range, and hence little reliable information for statistical methods to distinguish the score scales at these points (see the graphs of standard errors of equating in Figures 11 to 14).

The frequency estimation (FE) results in Figure 5 were similar to the CO results, particularly at the grade boundaries (as would be expected), giving the same result at grade A but with slightly larger differences among the three sample sizes at grade C presumably reflecting the different rounding/continuization elements in the methods.

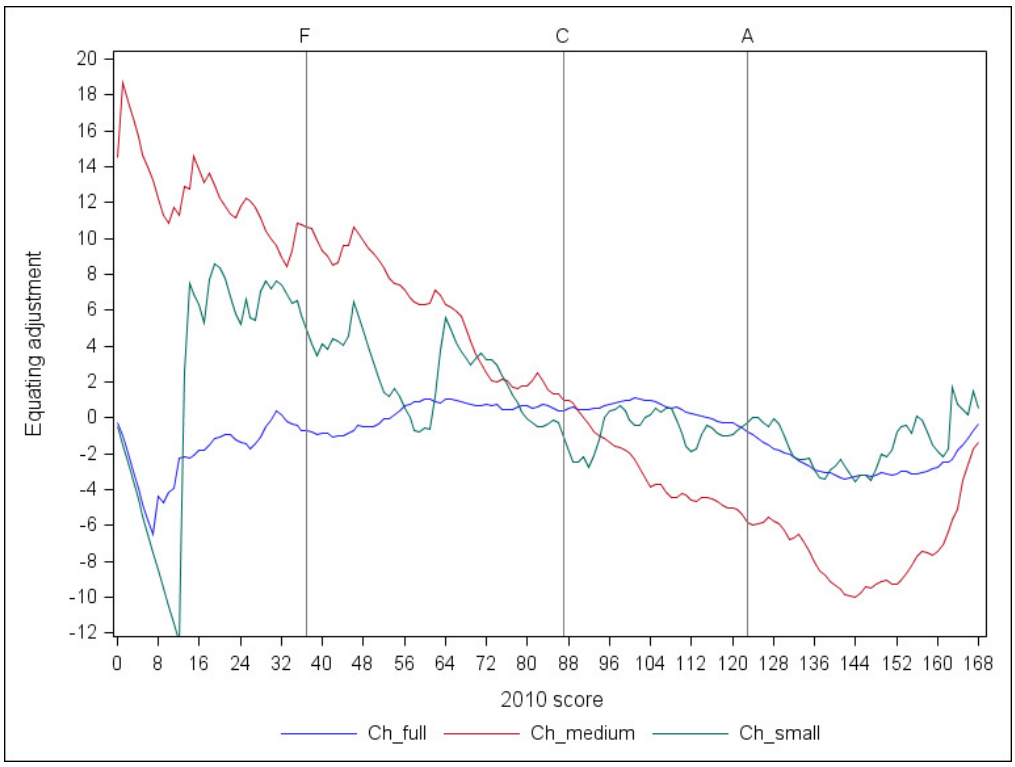

Figure 6: GCSE RE datasets – comparison of chained equipercentile results.

Interestingly, the differences among the three sample sizes were less at grades F and C with the chained equipercentile (Ch) method shown in Figure 6, and at grade A the small dataset gave the same result as the full dataset. However there was a much larger discrepancy at grade A for the medium dataset, weakening the suggestion that the chained method might be generally preferable.

Figures 7 and 8 show equivalent results to Figures 5 and 6, but using the kernel approach. The smoothing made the results in the three sample sizes more similar to each other at grades F and C, but less so at grade A, where the smoothing for the chained method exacerbated the discrepancy for the medium dataset.

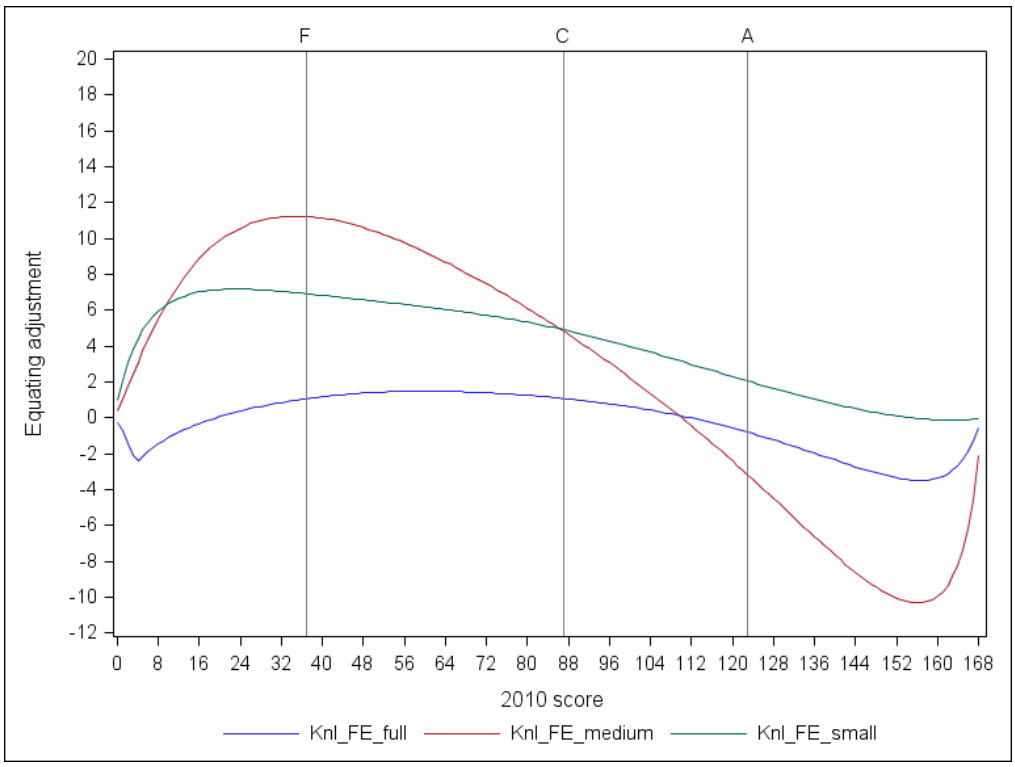

Figure 7: GCSE RE datasets – comparison of kernel frequency estimation results.

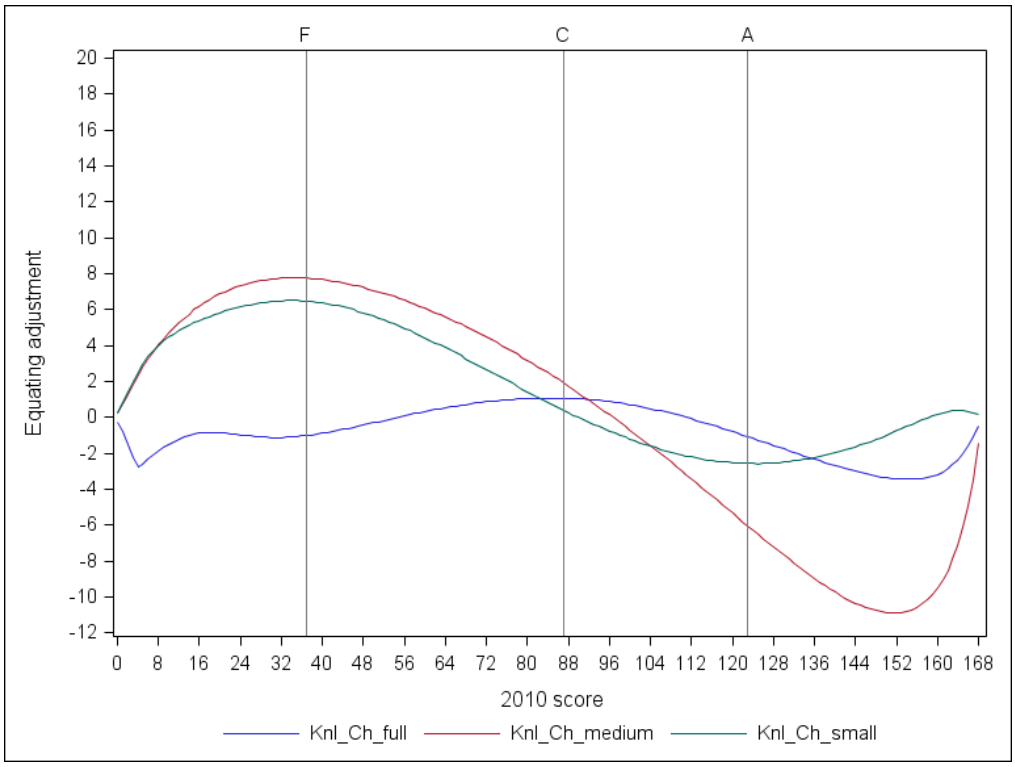

Figure 8: GCSE RE datasets – comparison of kernel chained results.

As a final illustration of the equating, Figures 9 and 10 show the comparison of the five equating methods for the medium and small datasets.

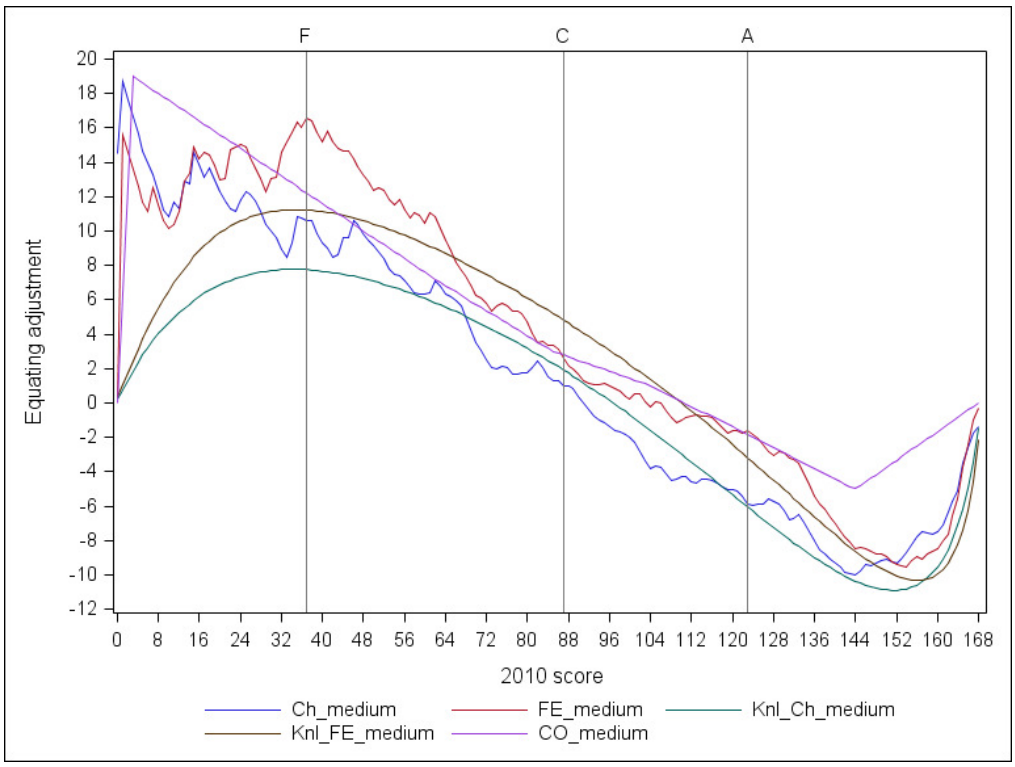

Figure 9: GCSE RE medium dataset – comparison of equating methods.

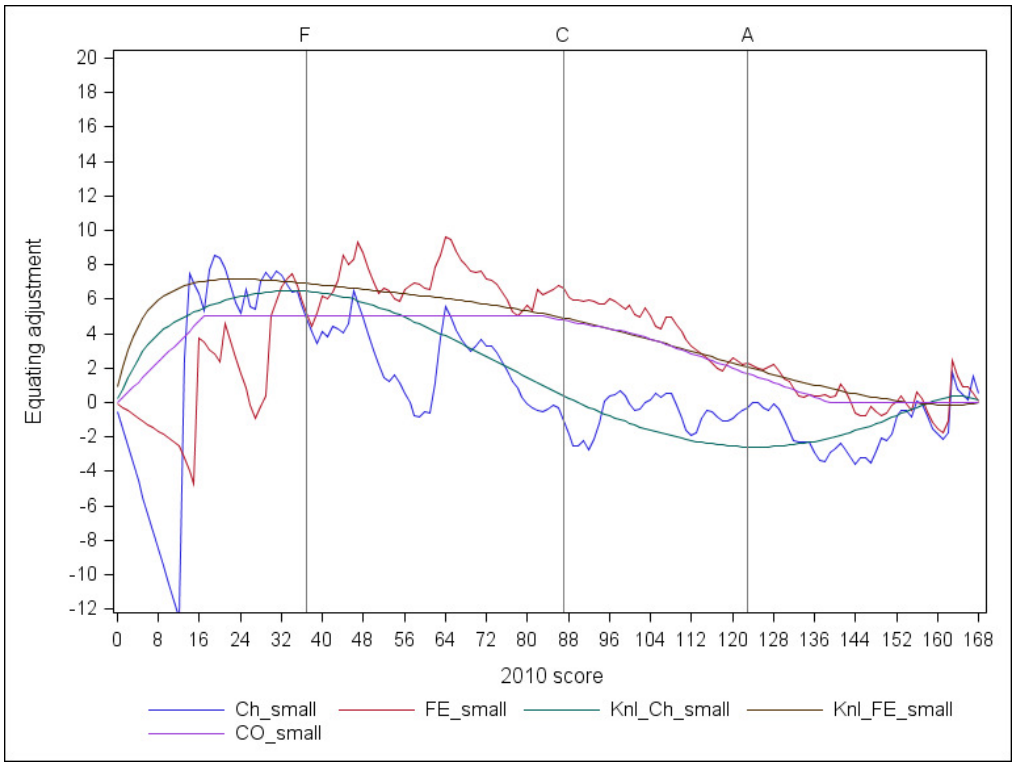

Figure 10: GCSE RE small dataset – comparison of equating methods.

Table 5.2 below puts all the results together to show what the grade boundaries on the 2010 exam would have been, assuming each method was used to map 2010 raw scores to the 2009 scale and then applying the original 2009 grade boundaries. Note that the comparable outcomes (CO) boundaries result from applying the usual interpolation rules[10] for arithmetically determined boundaries, whereas the boundaries from the equating methods do not (that is, it is possible for the grade B boundary, for example, not to be half-way between grade A and C).

Table 5.2: Summary of 2010 RE grade boundaries implied by each method.

	Full dataset					Medium dataset					Small dataset				
Grade	CO	Ch	FE	KCh	KFE	CO	Ch	FE	KCh	KFE	CO	Ch	FE	KCh	KFE
G	21	24	22	23	22	3	11	12	16	14	17	16	20	17	16
F	**37**	**39**	**38**	**39**	**37**	**23**	**26**	**24**	**31**	**28**	**33**	**31**	**32**	**32**	**31**
E	53	54	53	55	53	43	45	38	47	44	49	53	46	49	48
D	69	70	69	70	69	63	64	60	65	61	65	66	63	67	64
C	**87**	**88**	**88**	**87**	**87**	**85**	**88**	**85**	**86**	**83**	**83**	**91**	**82**	**88**	**83**
B	105	105	105	105	105	104	110	105	108	104	101	105	100	107	102
A	**123**	**123**	**123**	**124**	**123**	**124**	**128**	**124**	**130**	**126**	**120**	**123**	**120**	**125**	**120**
A*	141	143	142	142	142	144	149	148	150	149	139	142	139	141	139

Key: CO=comparable outcomes, Ch=Chained equipercentile equating, FE=frequency estimation equipercentile equating with all weight on 2010 cohort, KCh=Kernel Ch, KFE=Kernel FE.

Figures 11, 12 and 13 show the standard errors of equating (SEE) for the four equating methods in the full, medium and small datasets respectively. The same pattern is observed in all three datasets – namely that the chained method had a larger error than the frequency estimation method, and the kernel (smoothed) methods had smaller errors than the unsmoothed methods. The advantage of the smoothing was particularly noticeable for the small dataset. Also the SEEs were considerably larger where there was less data – for example around the F boundary compared to the C and A boundaries. Figure 14 compares the frequency estimation SEE from the three datasets, showing the increase in error as the sample size reduced.

[10] In 2009 the grade D boundary was set 1 mark below what the usual rules would imply (allowed by the procedures in place at the time in certain circumstances). For consistency, the same principle was applied to all comparable outcomes results in Table 5.2.

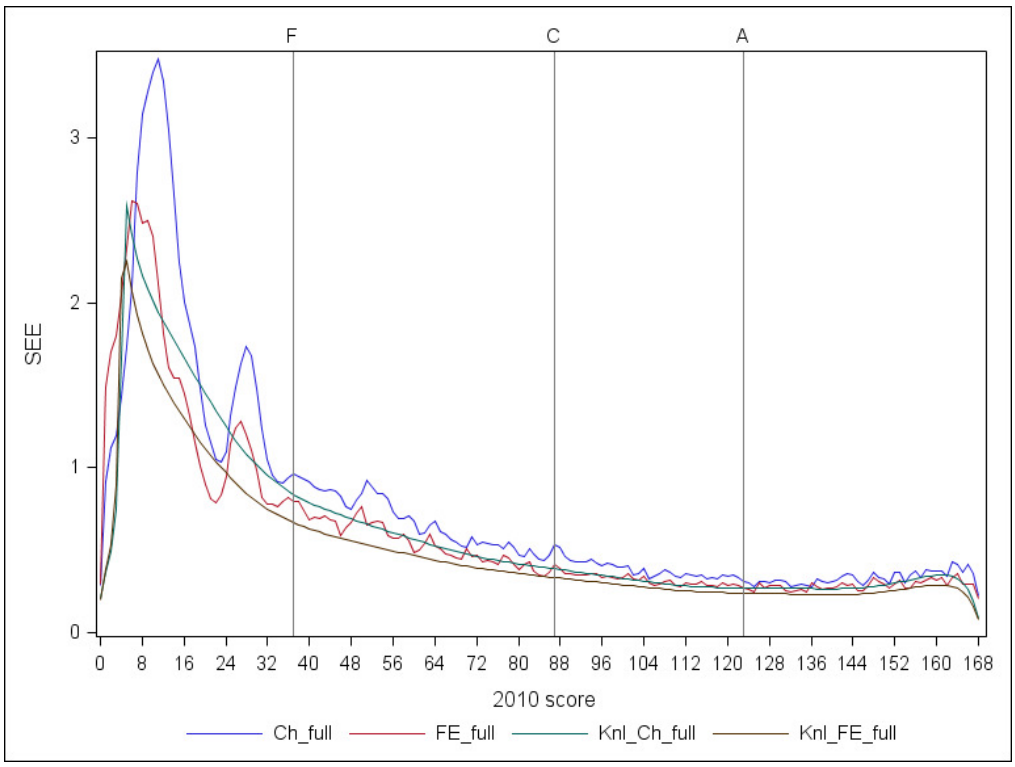

Figure 11: GCSE RE full dataset, standard errors of equating for each method.

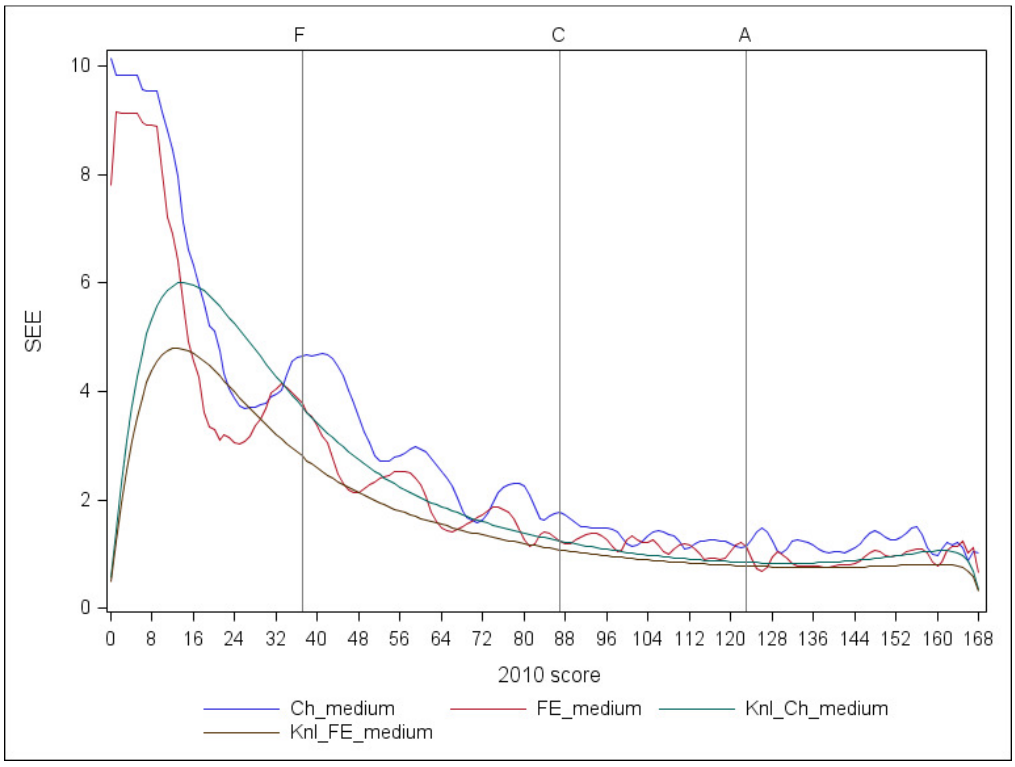

Figure 12: GCSE RE medium dataset, standard errors of equating for each method.

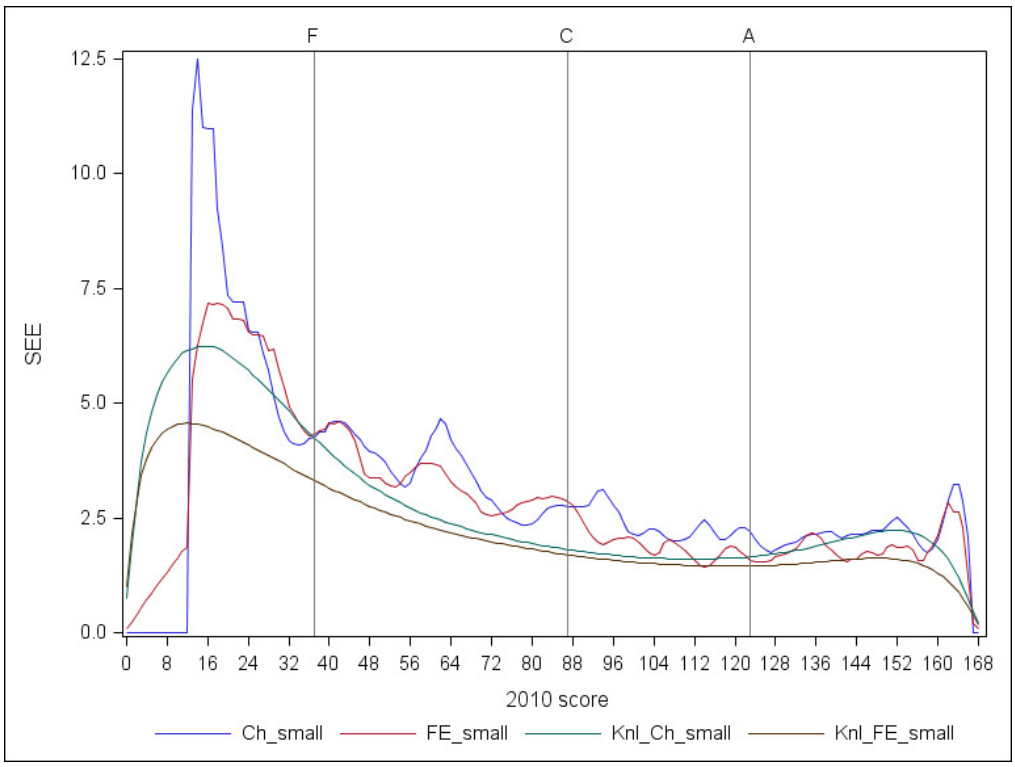

Figure 13: GCSE RE small dataset, standard errors of equating for each method.

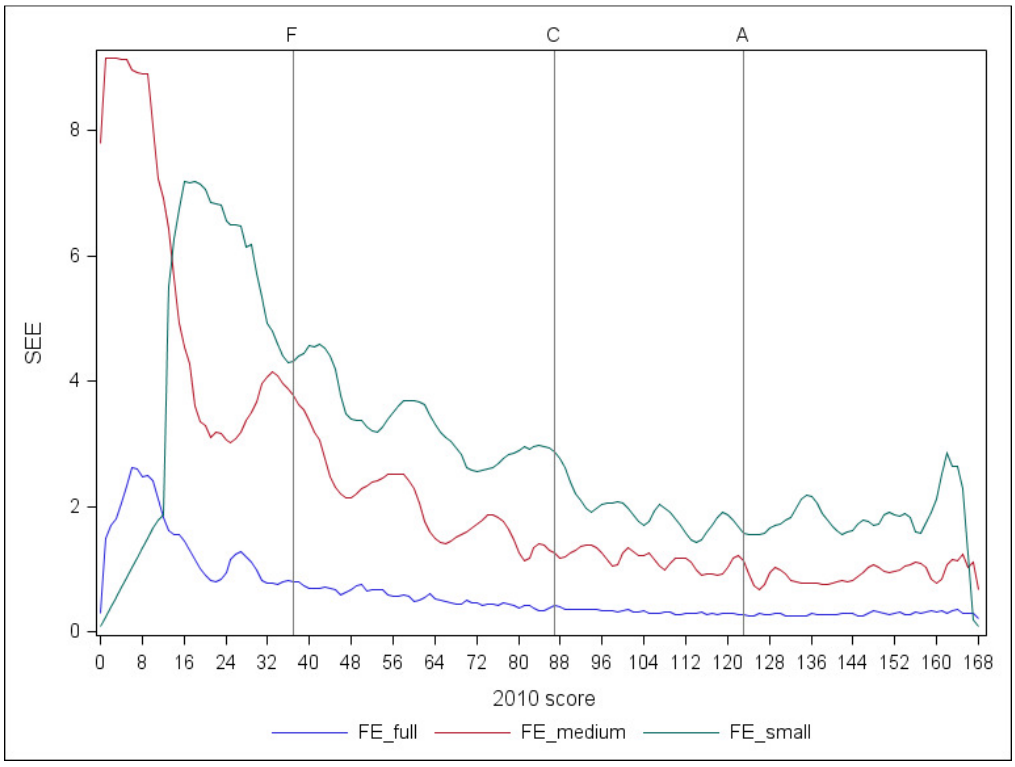

Figure 14: GCSE RE datasets, comparison of standard errors from frequency estimation equipercentile method across sample sizes.

The Ofqual tolerances for deviations from the putative distribution generated by the comparable outcomes method are based on a nominal 75% confidence interval (Benton & Lin, 2011). Table 5.3 shows the 75% confidence interval for equated scores from the frequency estimation method at the A, C and F boundaries in the three datasets, and the corresponding change in cumulative percentage[11] outcome in those datasets.

Table 5.3: 75% confidence intervals around RE equated scores at the A, C and F boundaries (FE method).

	Full		Medium		Small	
Grade	75% CI (marks)	Cuml. %	75% CI	Cuml. %	75% CI	Cuml. %
F	± 0.9	± 0.2	± 4.3	± 0.7	± 5.0	± 0.9
C	± 0.5	± 0.4	± 1.4	± 1.1	± 3.3	± 2.7
A	± 0.3	± 0.3	± 1.3	± 1.5	± 1.8	± 2.1

Table 5.3 shows that at A and C (where tolerances are applied) the medium and small datasets were within the ±2% and ±3% current limits respectively – suggesting that these values are reasonable. However, it should be noted that the SEEs reported here were based on bootstrap estimates assuming simple random sampling. The more complex approach of balanced replicated resampling (Benton & Lin, ibid.) attempts to take account of the hierarchical structure of the data and would be likely to produce larger SEEs. It is also noteworthy that for the medium and small datasets at grade F, and for the small dataset at grade C, there would still be a fairly wide range of mark points (at least seven) on which to choose the boundary and remain within tolerance, with the potential for accumulated 'benefit of the doubt' decisions to allow grade inflation over a long period.

Summary of findings from the RE datasets

- The previous demonstration of the structural similarity between the comparable outcomes method and the frequency estimation equipercentile equating method with all weight on the current cohort suggested that the results from those two methods should have been very similar. This was observed – in all three datasets they did not differ by more than one mark at the key grade boundaries of A, C and F. They did differ by more than that elsewhere, because the method for deriving the arithmetically determined boundaries enforces a linear relationship between scores on each test for the comparable outcomes method, whereas this is not the case for the frequency estimation method.
- With the full dataset (N≈25,000 per cohort) the results from all the equating methods were very similar, not differing by more than one mark at grades A and C and not differing by more than two marks at grade F.
- With the full dataset there was a negligible difference between the equating results when the weight given to the 2010 cohort was the relative proportion of candidates (as opposed to 1). There was also a negligible effect of changing the 'anchor test' categories so that they exactly matched the KS3 decile categories.
- As expected from theory, the standard errors of equating were lower for the frequency estimation equipercentile method than for the chained equipercentile method, and were lower for the smoothed 'kernel' methods than for the corresponding unsmoothed methods.
- Treating the comparable outcomes result from the full dataset as the 'true' result, in the medium and small datasets the chained equipercentile result was closer than the comparable outcomes result 3 times out of 6, further away 2 times out of 6 and equally far away once. This suggests that it there may be circumstances where a chained method would be preferable. In any case it might be prudent to use both methods and investigate further if there is a significant discrepancy between the results.
- The ranges of fluctuation in the cumulative outcome at the boundaries (corresponding to 75% confidence intervals around equated scores) were within the tolerance ranges applied by Ofqual for the respective sample sizes, suggesting that these tolerances are reasonable.

[11] Calculated by taking the average of the percentage of the 2010 cohort on the boundary and one mark either side, and multiplying this by the confidence interval.

However, the standard errors of equating may have been underestimated by the method used in this study.

– The range of equated scores within tolerance at the F boundary in the medium and small datasets and at the C boundary in the small dataset would still have allowed a lot of choice in where to set the boundary, suggesting that the practical utility of equating methods, including the comparable outcomes method, may not be that great for samples below 1,000 where tolerances are allowed. One practical way of dealing with this might be to allow the boundary-setting procedures to be changed in cases where there relatively many more candidates at the grade D boundary than the grade F boundary to allow grade D to be set by the comparable outcomes (or other equating) method and then extrapolating to lower grades with arithmetical rules.

5.2 Maths datasets

Figures 15 and 16 show the equating outcomes in the Foundation and Higher tier from the same five methods used with the RE. Again it can be seen that, as expected, the comparable outcomes method was very close to the frequency estimation method and its smoothed (kernel) equivalent at the key grade boundaries (F and C on Foundation tier; D, C and A on Higher tier). On the Higher tier the chained method and its smoothed equivalent were clearly different from the comparable outcomes method and frequency estimation methods, implying a smaller equating adjustment and hence underestimating the difference in difficulty (on the assumption that the comparable outcomes method gives the correct picture). On the Foundation tier the results from all the methods were more similar and in fact the rules for linear interpolation of intermediate boundaries between grades C and F meant that the comparable outcomes result diverged from the frequency estimation results more than it did from the chained results. Tables 5.4 and 5.5 are the equivalent for the maths tiers to Table 5.2 for the RE, showing the effective grade boundaries on 2010 that would have applied if 2010 scores had been equated to 2009 and then graded according to the 2009 boundaries.

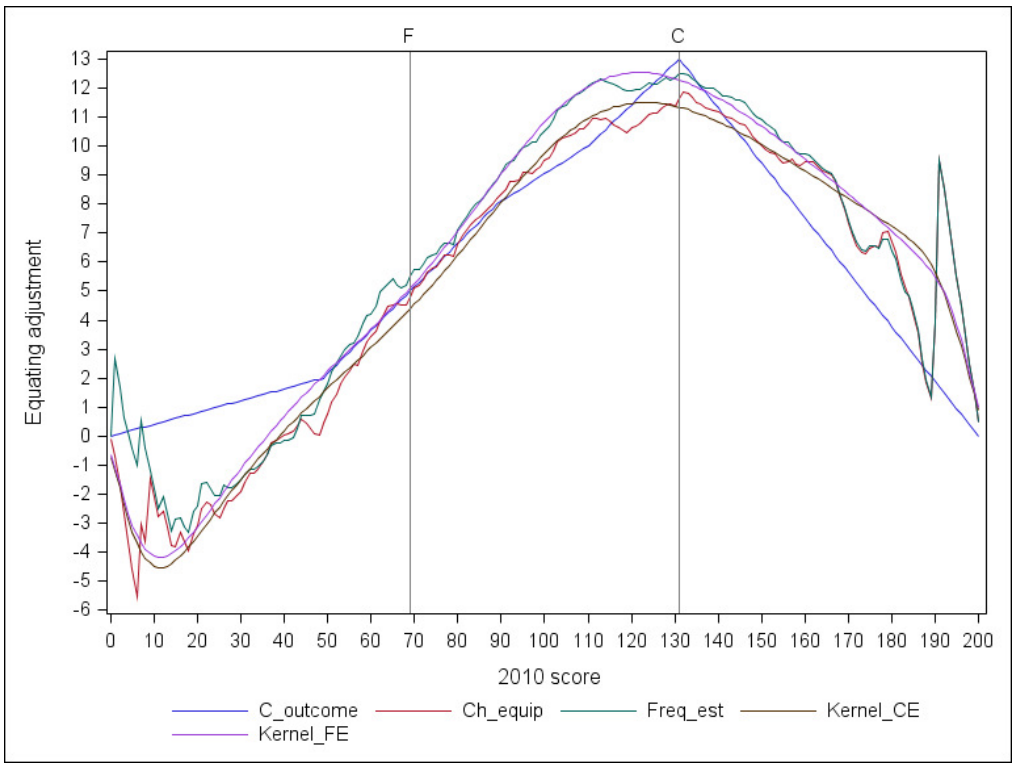

Figure 15: GCSE Maths Foundation tier dataset, comparison of equating methods.

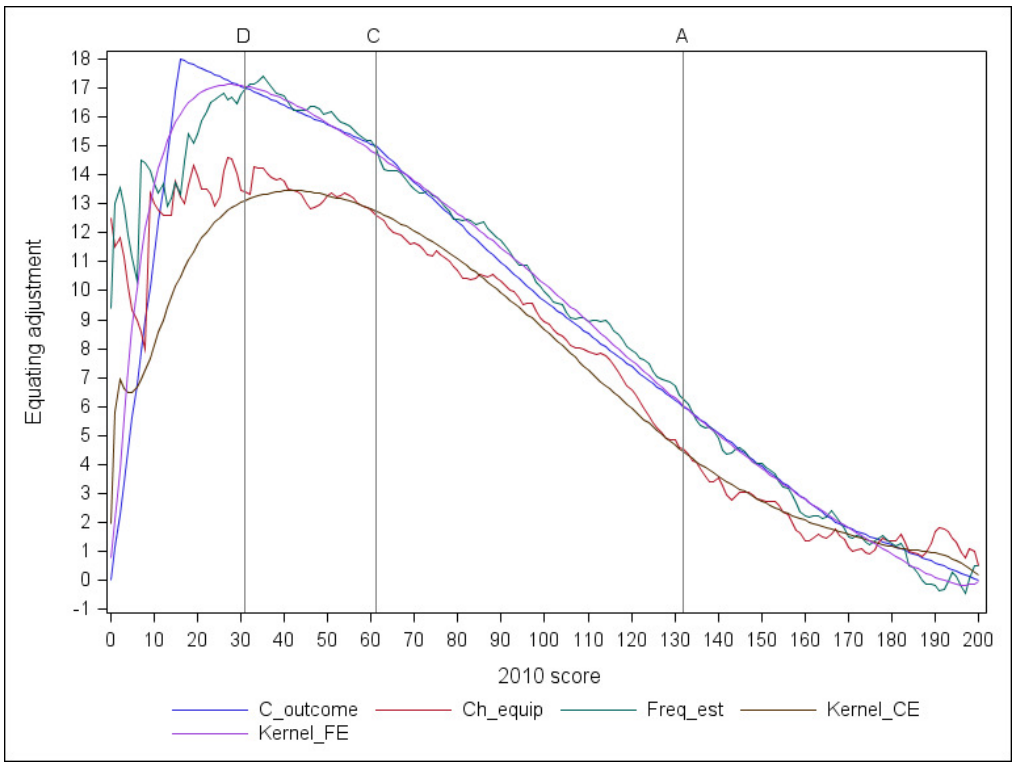

Figure 16: GCSE Maths Higher tier dataset, comparison of equating methods.

Table 5.4: Summary of 2010 Foundation tier Maths boundaries implied by each method.

Boundary	CO	Ch	FE	KCh	KFE
G	49	51	50	50	49
F	**69**	**70**	**69**	**70**	**69**
E	89	89	89	90	89
D	110	110	109	110	109
C	**131**	**133**	**132**	**133**	**132**

Table 5.5: Summary of 2010 Higher tier Maths boundaries implied by each method.

Boundary	CO	Ch	FE	KCh	KFE
U	0	0	0	0	0
E	16	21	19	22	18
D	**31**	**34**	**32**	**35**	**31**
C	**61**	**64**	**62**	**64**	**62**
B	97	98	97	99	97
A	**132**	**134**	**132**	**134**	**132**
A*	167	168	167	168	167

Key: CO=comparable outcomes, Ch=Chained equipercentile equating, FE=frequency estimation equipercentile equating with all weight on 2010 cohort, KCh=Kernel Ch, KFE=Kernel FE.

Figures 17 and 18 show the standard errors of equating from the four methods. As with the RE the errors were lower for the smoothed than the unsmoothed methods, and (with only a few exceptions) lower for the frequency estimation methods compared with their equivalent chained methods. In practical terms the errors were roughly the same for all methods within the mark ranges covered by the key grade boundaries, with the possible exception of grade D on the Higher tier where the kernel frequency estimation method had a noticeably lower SEE.

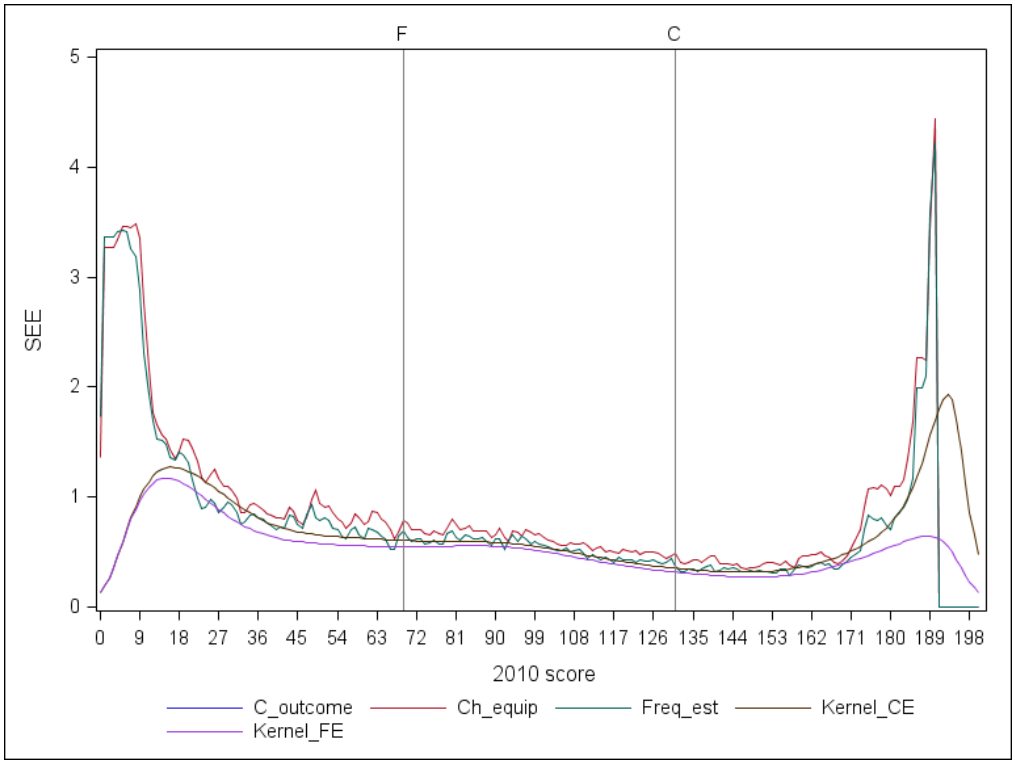

Figure 17: GCSE Maths Foundation tier dataset, comparison of SEE among equating methods.

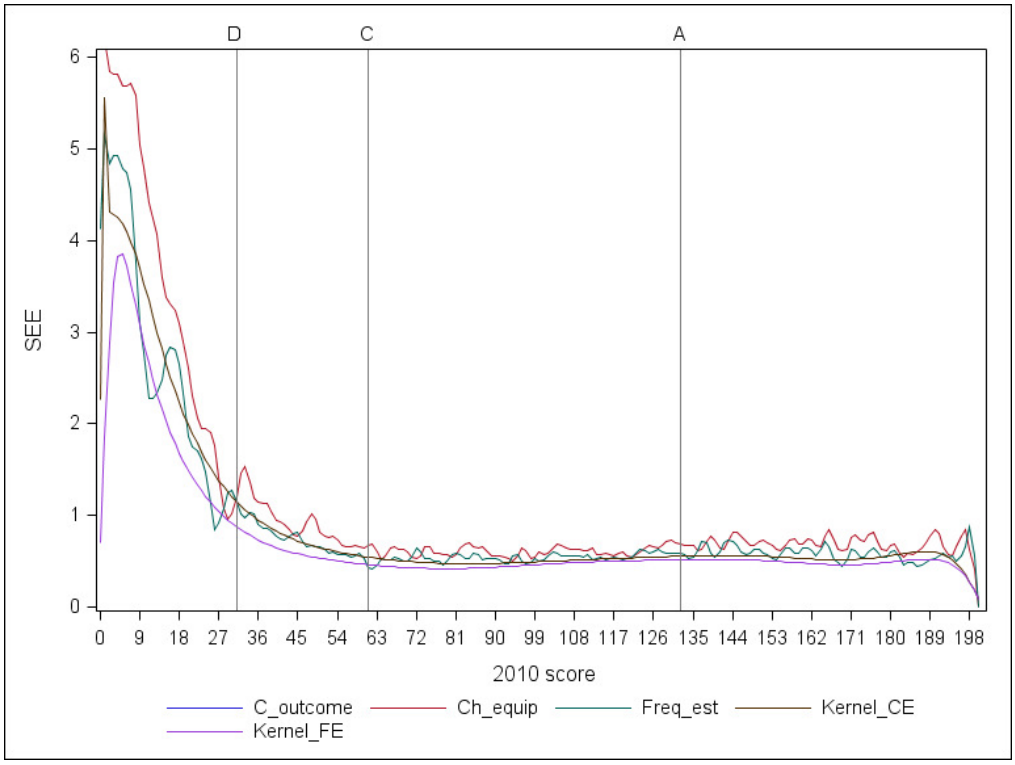

Figure 18: GCSE Maths Higher tier dataset, comparison of SEE among equating methods.

6. Discussion

This paper has shown how the comparable outcomes method used by the English exam boards and regulator to maintain standards is conceptually very similar to one particular method from the statistical equating literature – namely the frequency estimation equipercentile equating method. The main difference is that the former is only applied at certain points of the score distribution – the 'key' grade boundaries. A more minor difference is that the comparable outcomes method is used to identify integer (whole number) values for the grade boundaries whereas in statistical equating the equating function maps integer scores on one test to decimal scores on another test. This difference means that the comparable outcomes 'result' can differ from the frequency estimation result, but by no more than one mark (score point).

In the five datasets analysed here, this was found to be the case – the comparable outcomes result and the frequency estimation result never differed by more than one mark at a key boundary. This is despite the fact that there was the potential for more discrepancy since the frequency estimation method we tried used a more fine-grained (16-category) pseudo-anchor test than the comparable outcomes 10-category indicator of prior attainment.

This similarity raises several questions for discussion:

- If a statistical equating method is being used to set grade boundaries, does this imply that the boards and the regulator endorse a statistical definition of standard maintaining? If they did, would this be desirable – and should it be the subject of a more public debate?
- If a statistical equating method is being used to set grade boundaries should there be a more explicit evaluation of the plausibility of the various assumptions that must hold for the method to give an accurate result? In cases where the assumptions can be shown not to hold, or to be implausible, what are the implications for how grade boundaries should be set?
- Given that there is a large number of different statistical equating methods, each of which may be more effective in certain circumstances than others, should a single method be used across the board for all GCSEs and A levels?

First of all it should be recognised that it is debatable whether the comparable outcomes method can be considered to be an equating method, despite its apparent similarity to the frequency estimation method. For the purposes of this report, we only considered the method as a way of maintaining standards in the same examination year on year. In this case, the 'measuring the same construct' condition for equating clearly applies to the tests being equated, if not to the anchor test. However, in practice a second and no less important purpose of the comparable outcomes approach is to try to ensure comparability across examination boards. The prediction matrix (the distribution of exam grades conditional on prior attainment category) is based on an amalgamation of all boards' results. Although it is reasonable to assume that the boards' exams in a given subject at a given level are broadly measuring the same thing, clearly they are not all designed to exactly the same specification so there is some blurring of this equating requirement.

Furthermore, it is clear that the prior attainment score is very different from an actual anchor test of the kind used in equating, which is usually either an internal anchor (subset of items common to both tests) or an external anchor (separate test) but in both cases designed to be representative of the full tests in both content and difficulty, and taken at the same time as the tests to be equated (see Kolen & Brennan, 2004, p19 and p271-272; Sinharay & Holland, 2007). The prior attainment score, in stark contrast, is based on the aggregate of scores on tests in different subjects taken two or more years previously, and these are not the same tests for the two cohorts whose tests are to be equated. There is therefore a strong assumption that the level-setting process on the KS3 or KS2 test (for GCSE standard maintaining), and the grade boundary setting process on GCSE (for A level standard maintaining) has in fact maintained standards. Alternatively, if prior attainment categories are created from KS2/KS3/mean GCSE results as true deciles (or other quantiles) then there is the assumption that at the population level, prior attainment does not change from year to year.

Newton (2010) bemoaned the tendency for statistical equating to be idolised and every other 'weaker' form of linking evaluated in terms of how well it lives up to the equating ideal ("deficit rhetoric" was Newton's phrase). Newton's framework for 'thinking about linking' distinguished three perspectives: i) phenomenal, where comparability is defined in terms of the outcomes from learning that equally graded students have in common; ii) causal, where comparability is defined in terms of antecedent factors like prior attainment, effort, teaching quality; and iii) predictive, where comparability is defined in terms of potential – the likelihood of future success that students with the same grade have in common. Use of the framework clarifies the kinds of inferences that can be drawn about students with the same grade. In this framework, the use of prior attainment as the (sole) linking construct would be a 'specific causes- causal' definition of comparability (Newton ibid.). However, the price to pay for using this definition in the GCSE and A level context is that it would only seem to support inferences of the kind 'Students with a grade A this year have (on average) the same level of prior attainment as students with a grade A last year (or from a different board).' We are not sure that this is the definition of comparability that the wider public adopts when making inferences from exam grades. We would suggest that in fact something closer to our statistical definition in the introduction is what the majority of stakeholders are applying when they interpret grades, even if not in explicitly statistical terms. That is, there is the understanding that a student with the same level of attainment should get the same grade on different exams in the same subject – either over time within board or across boards. If this is right then it is worth evaluating the comparable outcomes method as though it were a statistical equating method.

Given the need for transparency in the system, there is probably an a priori case for using observed score equating methods, rather than IRT or true score methods. But perhaps some consideration should be given to whether the frequency estimation method (i.e. the comparable outcomes method) is always preferable to the chained method. For example, Kolen & Brennan (2004, p298) recommend frequency estimation when the two groups being equated are of similar ability. This is most likely to happen when the cohorts are stable (i.e. no large change in entry size, relatively little 'flux' from schools switching boards). Guo, Oh & Eignor (2013) also showed that if the two tests being equated were of similar difficulty and the (observable) conditional distributions of test score given anchor score are the same for both groups then the frequency estimation assumptions are met and the method is appropriate (although arguably not that necessary the more similar in difficulty the tests are!). As noted previously, some research comparing equating methods has found that the chained method has less bias than the frequency estimation method when there are larger group differences, so in circumstances where there appears to have been a relatively large change in prior attainment it may be worth checking whether a chained method gives a different result and if so, widening the allowable tolerance in the direction of the chained result. Note that it would be fairly easy to modify the comparable outcomes method so that it used chained equating to generate a 'putative' grade distribution and hence identify scores corresponding to the key grade boundaries.

However, chained equating may be less acceptable in principle because it so clearly requires that the anchor and the tests being equated are measuring the same thing, which is implausible in this context. Using the chained method would produce changes in grade distribution in line with changes in prior attainment regardless of the correlation between prior attainment and exam scores, whereas the frequency estimation method produces changes in the grade distribution in line with changes in prior attainment to the extent that prior attainment and exam scores are correlated. The use of prior attainment may be more easily justifiable as a 'covariate' used to adjust for systematic differences in ability (e.g. Bränberg & Wiberg, 2011) – this is an area for further investigation.

The likely return to linear examinations at GCSE and A level will also make it possible to use smoothing methods[12] which can reduce random equating error – von Davier (2013, p612) suggests that this can be effective in samples smaller than 20,000, which would include most A levels and

[12] Smoothing requires a bivariate distribution of test and anchor scores to smooth and as noted earlier, there is no single aggregate raw score distribution for a modular examination.

many GCSEs. The kernel approach, which integrates pre-smoothing of the distributions and optimising the continuisation of the discrete raw scores, would seem to be the most principled way to do this. However, this would only work for equating within boards across years.

The return to linear examinations will also allow scope for reconsidering the rules for the calculation of intermediate boundaries. Since it will be possible in principle to equate the full raw score scales, it would be possible to set the intermediate boundaries at their nearest equated score point rather than by linear interpolation between the key boundaries. With the current approach the standard maintaining method is more accurate for examinees at the key boundaries than elsewhere in the distribution, which could possibly be seen as unfair.

In summary, we have shown that, when considered as a method for maintaining standards within boards over time, the comparable outcomes method is equivalent to the frequency estimation equipercentile equating method, applied at a small number of points in the score range corresponding to the key grade boundaries. The following suggestions are intended as starting points for discussion about how within-board standard maintaining could be done for reformed GCSEs and A levels, on the assumption that they will be mostly linear examinations producing a single aggregate raw score scale.

1. Acknowledge that the comparable outcomes method is best suited to a statistical definition of standard maintaining and that prior attainment is being used in effect as an anchor test. Use the opportunities of the return to linear exams to explore other statistical equating methods treating prior attainment as an anchor test, in particular investigating whether there would be benefit in smoothing the score distributions, and equating whole mark scales rather than just the key boundary points. Use the outcomes as an explicitly identified 'statistical equating result'.

2. Aim to construct examinations with aggregate raw score scales that have grade boundaries at desirable points (not too high/low/bunched etc.) Modify subsequent examinations in the light of information from the initial ones to move towards these desirable points (target boundaries). Use expert judgment and statistical data from similar items in previous sessions to estimate, before the exam has been taken, any departures from default target grade boundary locations on newly constructed exams. Use the outcomes as an explicitly identified 'test construction result'.

3. Use expert judgment of the quality of candidates' scripts to link mark scales from the two exams, for example by the rank-ordering method (Bramley, 2005) or to link boundary points using the blind comparative direct judgment method described in Benton (2014). Use the outcomes as an explicitly identified 'perceived quality of work' result.

4. Combine the results either according to a weighting formula specified in advance – in other words so that the final result is determined once the results in steps 1 to 3 are known; or by a process of discussion and debate, producing a rationale for the weight given to each result in steps 1 to 3.

A scheme like this would make it more clear to the stakeholders what evidence was informing the standard-maintaining processes, and what weight was being given to it. It would not of course in itself help resolve disagreements in opinion about what evidence should be given more weight but might lead to more productive discussions and ultimately a system that is more widely trusted.

References

Albano, A. (2013). ***Equate:*** *Statistical methods for Test Score Equating.* R package version 1.2.0, URL http://cran.r-project.org/web/packages/equate/index.html.

Andersson, B., Bränberg, K. & Wiberg, M. (2013). ***kequate:*** *The Kernel Method of Test Equating.* R package version 1.3.2, URL http://cran.r-project.org/web/packages/kequate/index.html.

Benton, T., & Lin, Y. (2011). *Investigating the relationship between A level results and prior attainment at GCSE*. Coventry: Ofqual.

Benton, T., & Sutch, T. (2013). *Analysis of use of Key Stage 2 data in GCSE predictions.* Cambridge Assessment interim report to Ofqual dated 06/12/13.

Benton, T. (2014). *Comparing the reliability of standard maintaining via examiner judgement to statistical approaches.* Paper presented at the International Association for Educational Assessment (IAEA) conference, Singapore, 25-31 May 2014.

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement, 6*(2), 202-223.

Bramley, T. (2013). *Maintaining standards in public examinations: why it is impossible to please everyone.* Paper presented at the 15th biennial conference of the European Association for Research in Learning and Instruction (EARLI), Munich, Germany, 27-31 August 2013.

Bränberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement, 48*(4), 419-440.

Cambridge Assessment (2011). Research Matters Special Issue 2: Comparability. http://www.cambridgeassessment.org.uk/Images/109991-research-matters-special-issue-2-comparability.pdf Accessed 06/01/14.

Davison, A. C. (2003). *Statistical Models.* Cambridge: Cambridge University Press.

Guo, H., Oh, H. J., & Eignor, D. (2013). Situations where it is appropriate to use frequency estimation equipercentile equating. *Journal of Educational Measurement, 50*(3), 338-354.

Holland, P. W., & Thayer, D. (1989). *The kernel method of equating score distributions.* Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioural Statistics, 25*, 133-183.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220): ACE/Praeger series on higher education.

Holland, P. W., Sinharay, S., Von Davier, A. A., & Han, N. (2008). An approach to evaluating the missing data assumptions of the chain and post-stratification equating methods for the NEAT design. *Journal of Educational Measurement, 45*(1), 17-43.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices.* (2nd ed.). New York: Springer.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*(1), 73-95.

Newton, P. E. (2010). Thinking about linking. *Measurement: Interdisciplinary Research and Perspectives, 8*(1), 38-56.

Newton, P. E. (2011). A level pass rates and the enduring myth of norm-referencing. *Research Matters: A Cambridge Assessment Publication, Special Issue 2: comparability*, 20-26.

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

Oehlert, G. W. (1992). A note on the Delta Method. *The American Statistician*, 46(1), 27-29.

Ofqual. (2011). GCSE, GCE, Principal Learning and Project Code of Practice. Coventry: Ofqual.

Ofqual (2013). Corporate Plan 2013-2016. Ref Ofqual/13/5310. Coventry: Ofqual. http://ofqual.gov.uk/documents/corporate-plan/ Accessed 20/12/13.

Pinot de Moira, A. (2008). *Statistical predictions in award meetings: how confident should we be?* AQA Research Report RPA_08_APM_RP_013. Guildford: AQA.

R Core Team (2013). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. New York: Springer.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*(3), 249-275.

Taylor, M. (2013). *GCSE predictions using mean Key Stage 2 level as the measure of prior attainment.* Report to Joint Council on Qualifications (JCQ) Standards and Technical Advisory Group (STAG). Revised 26/06/13.

von Davier, A. A. (2013). Observed-score equating: an overview. *Psychometrika, 78*(4), 605-623.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer Verlag.

Wang, T., Lee, W.-C., Brennan, R. L., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item nonequivalent groups design. *Applied Psychological Measurement, 32*(8), 632-651.