Cambridge Assessment

Pooling the totality of our data resources to maintain standards in the face of changing cohorts

Conference Paper

Tom Benton

Presented at the 18th annual AEA-Europe conference,
Prague, Czech Republic,
November 2017

Author contact details:

Tom Benton
Assessment Research and Development,
Research Division
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG
UK

benton.t@cambridgeassessment.org.uk

http://www.cambridgeassessment.org.uk

As a department of Cambridge University, Cambridge Assessment is respected and trusted worldwide, managing three world-class examination boards, and maintaining the highest standards in educational assessment and learning. We are a not-for-profit organisation.

How to cite this publication:

Benton, T. (2017, November). *Pooling the totality of our data resources to maintain standards in the face of changing cohorts.* Paper presented at the 18th annual AEA-Europe conference, Prague, Czech Republic.

Table of contents

Introduction

Cambridge Assessment organises thousands of assessments each year involving hundreds of thousands of pupils. For nearly all of these, some decision needs to be made about where grade boundaries, demarking particular levels of achievement from each other, should be positioned. Grade boundaries need to be positioned to ensure comparability over time, so that pupils awarded a particular grade this year can be assumed to have a similar level of ability to those awarded the same grade last year. In particular circumstances there may also be a need to ensure comparability between alternative versions of the same assessment within a given examination session.

A number of sources of evidence may be used to inform the positioning of grade boundaries including both expert judgement and statistical information. The simplest form of statistical information is to assume that the percentage of pupils achieving each grade in a particular assessment should remain stable over time. In some cases, such as where the assessment attracts a large number of entries from the same centres each year, this assumption appears reasonable. In other cases, we may be able to identify particular schools with stable entry numbers over time, and, if there are many such schools, we may feel comfortable assuming that, within this subgroup, the percentage of pupils achieving each grade should be stable. However, identifying a large number of schools with stable entry numbers is not always possible, and in these cases robust statistical recommendations for the positioning of grade boundaries can be hard to generate.

In England the positioning of grade boundaries for GCSEs (examinations taken at age 16), is strongly informed by the achievement of the pupils entering the given GCSE examination in their end of primary school tests, known as key stage 2, taken five years earlier. However, this method of maintaining standards is known to have weaknesses if there are major changes in the abilities of pupils taking particular subjects (see Benton and Sutch, 2014).

This paper proposes a new way of simultaneously making use of all of the data we hold about each candidate's achievement across all available assessments within a given examination session to help maintain standards. Using data from such concurrent assessments to check comparability between awarding organisations offering alternative versions of the same qualifications has been suggested before (see Benton and Sutch, 2014). However, previously documented approaches rely on being able to assign an ability measure to each pupil (namely mean GCSE grade) based upon the grades they have achieved. As such, these approaches rely on grade boundaries having already been set for all assessments so are not immediately applicable to the task of setting grade boundaries in the first place.

The method proposed in this paper addresses this issue and suggests a method of standard maintaining that could be applied to all assessments within a given examination session simultaneously. In particular, it makes use of the fact that different pupils take different combinations of assessments so that we have potential links between most of these. This is illustrated further the network graph displayed in Figure 1. This graph shows a map of all Cambridge Assessment qualifications aimed at 16 year olds that were taken by at least 500

candidates in summer 2017[1]. Each circle represents a qualification and larger circles represent qualifications taken by large numbers of candidates. A line between two circles indicates that at least 200 candidates took both of the two qualifications. The circles towards the bottom of the chart represent those that are only available in the UK, those towards the top right represent qualifications available both within the UK and elsewhere, and those towards the top left represent those that are only available outside the UK. The main thing to note about this chart is the enormous amount of linkage between qualifications. On average each qualification is linked to seventeen others. Thus for any given qualification, we hold a large amount of information about the performance of candidates elsewhere. Note that because Figure 1 shows only whole qualifications[2], and excludes any qualifications entered by less than 500 candidates, it actually underrepresents the scale of the information at our disposal. The challenge then is how to make use of all of this information to inform the positioning of grade boundaries.

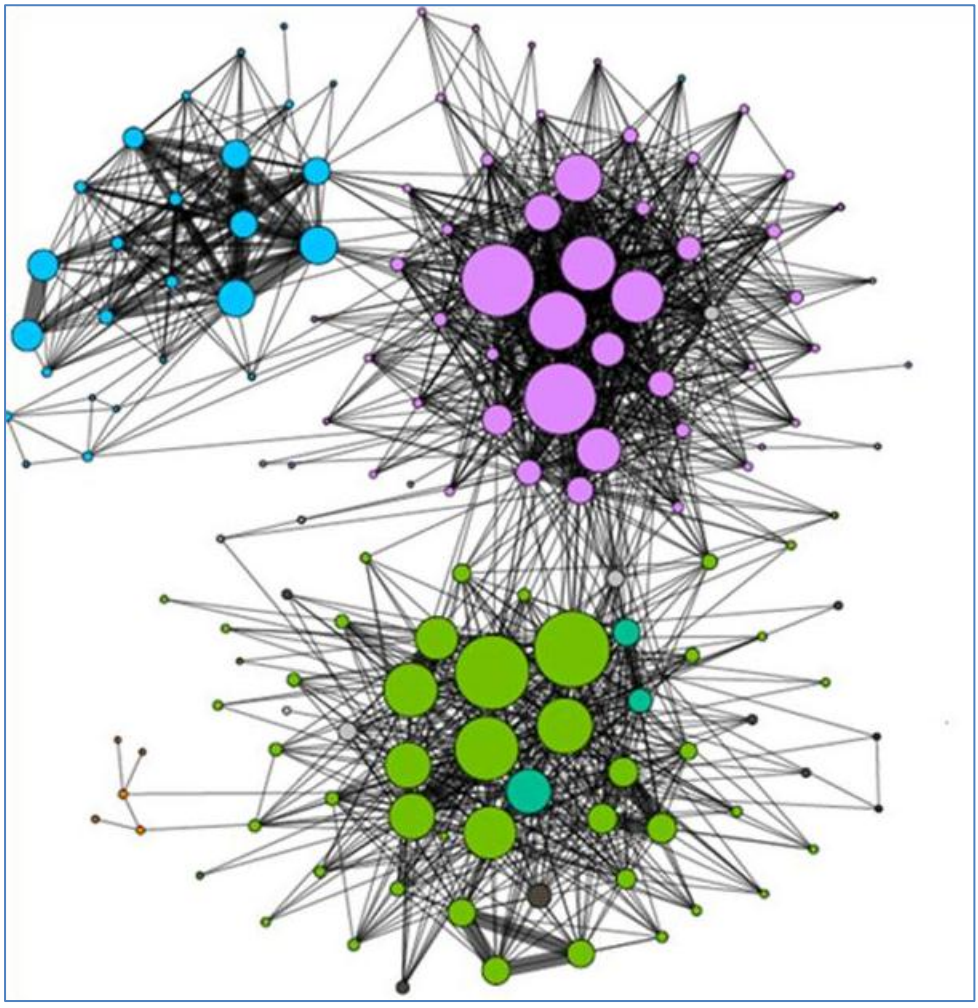

Figure 1: A network graph of Cambridge Assessment qualifications aimed at 16-year olds in summer 2017

[1] For example, by "qualification" we might mean a GCSE in a particular subject (e.g. GCSE Mathematics).

[2] In order to achieve a qualification candidates will generally have to complete more than one assessment.

The proposed method is as follows:

1. Within each year create a single measure of ability for all candidates that is on the same scale regardless of which assessments any given pupil has taken.
2. Having created this measure, by looking at achievement across all assessments together, we can find sufficient centres[3] with stable entries over time to calibrate it between years.
3. Now that we have a calibrated measure of achievement that is comparable over time, we can use it to inform the positioning of grade boundaries in all qualifications.

Although the main practical application of this work will be in determining the positioning of grade boundaries, this paper actually addresses the more general problem of equating. That is, rather than simply finding a cut-score of this year's test that is comparable to the cut-score on last year's test, we will actually look to identify the score on this year's test that is equivalent to *each* score on last year's test.

Introducing the ISAWG

The first step in our proposed method is to create a single measure of ability. This measure should summarise each pupil's achievement across numerous assessments on a single scale regardless of which ones they have taken. The hope that a useful measure of this type may be created comes from Charles Spearman's very old (and much debated) theory of general ability (or "g"). Spearman's 1904 paper proposed that "all branches of intellectual activity have in common one fundamental function (or group of functions), whereas the remaining or specific elements of the activity seem in every case to be wholly different from that in all the others." (Spearman, 1904, page 284). In other words, although different tests may measure slightly different skills, all of them should relate to each candidate's "fundamental function" (or "g") which in turn should give an idea of how well they are likely to perform on different tests. Following Thorndike (1994) we will use some form of simple one-factor model to capture a "working definition of g" and hope that, despite the variety of available assessments, it will remain the case that "the great preponderance of the prediction that is possible from any set of cognitive tests is attributable to the general ability that they share." (Thorndike, 1994, page 150). The resulting ability measure will be called the ISAWG, which stands for Instant Summary of Achievement Without Grades.

The ISAWG is defined for each candidate to be the single number that most accurately reflects the standardised marks they have achieved on whichever assessments they have taken. More formally the ISAWG is defined as follows. Let y_{ij} be the standardised score[4] of the *i*th candidate on the *j*th assessment within a particular session. Now for each candidate

[3] Usually schools but potentially including other educational institutions

[4] By standardised score we mean that the raw scores are linearly rescaled so that, across all candidates with available marks for the assessment (at the time of calculation), they have a mean of zero and a standard deviation of one.

and for each assessment we define $ISAWG_i$, β_j and α_j so that any candidate's score on any assessment can be estimated by

$$\widehat{y_{ij}} = \alpha_j + \beta_j ISAWG_i$$

Note that the same value of the ISAWG must apply across all of the assessments the candidate has taken. The parameters β_j and α_j will determine the expected relationship between the ISAWG and standardised scores on the *j*th assessment. Broadly speaking, the parameter α_j relates to ease of achieving a high ranking amongst the candidates taking assessment *j*. The β parameters say something about how strongly the scores on the jth assessment relate to candidates' other scores more widely. Each of $ISAWG_i$, β_j and α_j are estimated to ensure that the estimates of standardised scores (that is, the $\widehat{y_{ij}}$) are as accurate as possible. Specifically they are chosen to minimise the overall sum of squared errors (that is, squared differences between y_{ij} and $\widehat{y_{ij}}$) across all candidates and all assessments within a particular session. Within this process, the scale of the ISAWG is fixed to have a mean of zero and a standard deviation of one across all candidates in total.

In fact, the definition of the ISAWG given above is equivalent to a long standing form of statistical analysis – Principal Components Analysis. In technical terms the ISAWG is just the first principal component calculated for each candidate but applied to a data set including missing values (as no candidates take all of the available assessments offered by Cambridge Assessment within a single session). The precise method used to calculate the ISAWG is known as alternating regression and was proposed, in the context of chemistry, by De Ligny et al (1981). It works as follows.

1. Initially set all the α parameters equal to 0 and all the β parameters equal to 1.

2. For each individual pupil, estimate a value for $ISAWG_i$ by fitting a linear regression model (through the origin) of $(y_{ij}-\alpha_j)$ on β_j. This process is completed by the simple closed form solution:

$$ISAWG_i = \frac{\overline{\beta_j y_{ij}} - \overline{\beta_j \alpha_j}}{\overline{\beta_j^2}}$$

   Where the bars (e.g. $\overline{\beta_j y_{ij}}$) relate to averaging values over all assessments (j) taken by an individual pupil (i).

3. Now that we have estimates for each $ISAWG_i$, for each individual assessment we update the α and β parameters by a linear regression of standardised scores (y_{ij}) on the current estimates of $ISAWG_i$ for each assessment. This process is also completed by the following two closed form solutions

$$\widehat{\beta_j} = \frac{\overline{ISAWG_i y_{ij}} - (\overline{ISAWG_i})(\overline{y_{ij}})}{\overline{ISAWG_i^2} - (\overline{ISAWG_i})^2}$$

   And

$$\widehat{\alpha_j} = \overline{y_{ij}} - \widehat{\beta_j}\overline{x_i}$$

Where the bars in this case (e.g. $\overline{ISAWG_i y_{ij}}$) relate to averaging values over all candidates (i) taking a particular assessment.

4. Repeat steps 2 and 3 until the overall sum of squared errors across all candidates and all assessments is no longer substantially decreasing.

5. Apply a linear transformation to the ISAWG estimates so that they have a mean of 0 and a standard deviation of 1. Then adjust the α and β parameters to reflect this transformation.

Using alternating regression to generate the ISAWG has two particular advantages. Firstly, it allows the ISAWG to be calculated very quickly. Completing the above iterative procedure for several hundred thousand candidates within a session who have cumulatively taken several million assessments takes less than 10 minutes using an ordinary desktop computer. Secondly, it is easy to calculate each individual candidate's ISAWG either restricted to or excluding particular assessment components (subsequently referred to simply as components) using the formula in step 2. This can be useful when we wish to compare how a candidate has performed in one specific assessment against their performance across all of the others.

Using the ISAWG in equating

In this paper our interest is in the possible practical use of the ISAWG in equating. More specifically, having a general measure of ability that relates to all our various assessments could potentially help to ensure that a common standard of difficulty is applied to grades across all of them. Alternatively, it could be used to help to maintain standards between years.

The ISAWG itself is based upon all of the components that each candidate takes within a given examination session. As such, when used in equating, the link between different assessment components is provided by scores from the other components (referred to from now on as the co-components) taken alongside those that are being equated. When viewed in this way it is clear that the ISAWG does not represent the only way in which such information could be used. In fact, these attempts at using information from co-components can be seen as part of a much wider area of research into how data from covariates might be used within equating. One approach to equating in this situation, described by Anderson et al (2013) as the non-equivalent groups with covariates (NEC) design, is based upon using log-linear models to capture the relationship between the covariates of interest and the score distribution on each component. However, because this technique begins with a full enumeration of the numbers of candidates with each possible combination of values on each covariate and upon the components of interest, it is not suitable for situations involving large numbers of covariates. Nor does it suggest how we might deal with different missing data patterns within co-components.

In the past, methods using co-components (or covariates) in equating have usually taken a frequency estimation or (equivalently) weighting approach to equating. This means that they first attempt to make the two sets of candidates taking the two assessments being equated match in terms of their covariate distributions, and then use traditional equating as would be applied to equivalent groups. However, it is possible that chained equating methods using the ISAWG or other ability measures as if they were an anchor might lead to more accurate results than a frequency estimation approach[5]. This is potentially particularly true if there are large differences in ability between groups. This paper will provide some further research into this issue.

This paper examines the following research questions:

- To what extent does the ISAWG provide an accurate method of equating and how does this compare to other possibilities such as using data from primary school tests (key stage 2)?
- Should the data from the ISAWG be applied using frequency estimation or chained equating?
- If we assume that the distribution of the ISAWG for a large population of candidates must be stable across sessions, can we then use it to effectively maintain standards even for components taken by relatively small numbers of candidates?

[5] See Table 4 from von Davier and Chen (2013) for an example of this.

Part 1: Examining different methods of equating using co-components

Method

Initial work evaluated the effectiveness of the ISAWG for equating. The data used for analysis consisted of all component scores within GCSE units taken by at least 10,000 candidates in June 2015. The following steps were then used for research.

Identifying existing pairs suitable for equating and defining a criterion equating function

The analysis identified 16 pairs of GCSE components within the data set where at least 90 per cent of candidates taking either assessment took both, where the scores on the assessments displayed a correlation of at least 0.7 and where the maximum scores available on the two assessments were similar. The criteria used to choose these pairs of assessments ensured that calculating how one of the assessments equates to the other was at least a sensible exercise. Table 1 gives some information about the sixteen pairs of components used in analysis. As can be seen, some of these pairs contained two exams in different subjects (e.g. Biology and Chemistry). Since these pairs tended to occur where both components led to a common qualification (combined science) it is reasonable to imagine that we might want to ensure grade boundaries represent equivalent levels of difficulty on the two tests. However, such analyses should probably be referred to as linking rather than equating (Kolen and Brennan, 2004). Having said this, in order to keep the language as simple as possible, we will refer to all attempts to link standards between components as equating (rather than linking) for the duration of this paper.

For each pair, unsmoothed equipercentile equating using the full set of candidates taking both tests was used to define a criterion or "true" equating function between the two assessments.

Table 1: Pairs of components used in analysis

Form X	Form Y	Maximum score on X	Maximum score on Y	N candidates taking both
Art 1	Art 2	100	100	10359
Biology 1	Chemistry 1	60	60	12176
Biology 2	Chemistry 2	60	60	47548
Biology 3	Chemistry 3	60	60	46892
Business 2	Business 3	60	90	12853
Computing 2	Computing 3	45	45	19677
IT 1	IT 2	60	60	10164
PE 2	PE 3	60	60	10030
Science 3	Science 5	75	85	15209
Science 4	Science 6	75	85	33155
Science 8	Science 9	75	85	31125
Biology 6	Biology 7	75	85	19952
Chemistry 6	Chemistry 7	75	85	19558
Physics 6	Physics 7	75	85	19262
Math 1	Math 2	100	100	35863
Math 3	Math 4	100	100	20601

Splitting data from pairs into two groups

So far, for each pair of assessments, it was possible to equate using a single group design. That is, the same group of candidates had taken both assessments. In order to evaluate the possible value of co-components for linking in general, it was necessary to convert this data into a form that would be familiar in a NEC (non-equivalent with covariates) design. This required splitting the data set into two groups (P and Q) so that in group P only the scores on form X were stored and in group Q only the scores on form Y were stored with information from co-components retained within each group. The challenge was to attempt to reconstruct the criterion equating functions defined above using this reduced data set.

The split of data into groups P and Q was not done completely at random. In particular, in order to ensure that the scenarios presented a suitably difficult equating challenge it was decided to set up the data so that candidates in group Q had higher scores on average than those in group P. This was done by first calculating the total score across both assessments in the pair and standardising by subtracting the mean and dividing by the standard deviation. Candidates were then assigned to groups P or Q with the probability that a candidate was assigned to group Q being defined by:

$$P(Pupil\ assigned\ to\ group\ Q) = Max\left(0, Min\left(1, 0.5 + Standardised\ Score * \left(\frac{0.3}{4}\right)\right)\right)$$

The above procedure, similar to one described in von Davier and Chen (2013), ensured that, on average, there was a difference of roughly 0.3 standard deviations between total scores for candidates in group P and scores for those in group Q. This should be large enough to make equating challenging but is slightly less than a difference of 0.5 standard deviations which was described by Kolen and Brennan (2004, p286-7) as 'especially troublesome'. The above formula also ensured that roughly equal numbers of candidates were assigned to groups P and Q.

Attempting to equate forms across groups

For each pair, using the data stored in groups P and Q a number of different methods were trialled to reconstruct the equating function between the two assessments. The following methods were trialled:

1. **Full ISAWG**. Using the full set of data across both group P and group Q, including both the assessments being equated and all possible co-components, calculate an ISAWG measure for each student. This measure is then used in the place of an anchor test.
2. **ISAWG based on strong major co-components.** First the major co-components are identified as those taken by at least 500 candidates alongside at least one of assessments being equated. Next any of these co-components that only displayed a weak correlation with the components being equated are removed[6]. Next calculate

[6] Specifically, any co-components where the correlation with the form X assessment was more than 0.1 below the maximum correlation between any major co-component and the form X assessment.

an ISAWG based on these co-components only (and not the assessments being equated). This measure is then used in place of an anchor.

3. **Key stage 2**. The majority of pupils in England complete assessments in English and Maths at the end of primary school (aged 11). Since data on these assessments is available for most candidates, use the data from these tests to equate between assessments.
4. **Single co-component**. For each pair, identify the major co-component taken alongside those of interest that displays the highest correlation[7] with the form X assessment[8]. Use this co-component alone as an anchor.
5. **Assume equivalence**. This method simply ignores the information contained in co-components and assumes that the candidates in group P are directly comparable to those in group Q. This method was included to provide some sense of scale in the results by displaying the accuracy of a worst plausible method.

To begin with, for each method, groups P and Q were weighted to be equivalent. For methods 1 and 2 this was facilitated by first replacing the ISAWG with deciles of achievement[9]. Once the data within each group was weighted, unsmoothed equipercentile equating was used to estimate the equating function. As an alternative, chained equating via the ISAWG was also applied. The weighted[10] mean absolute difference across the score range between these equating functions and the criterion function (defined earlier) was calculated.

Results

The accuracies of the different equating techniques based on weighting (also sometimes referred to as frequency estimation) are shown in Table 2. For each of the sixteen pairs of assessments being equated, the method with the lowest weighted mean absolute error of equating is highlighted. As can be seen, the method based on using the full ISAWG led to the most accurate results on average. This may be due to the fact that since it uses data from all co-components as well the assessment being equated, it retains a much larger amount of data for analysis than the other methods. On average, this method yielded an equating function just under one and a half marks away from the criterion equating.

The method based upon using an ISAWG derived from only strong major co-components displayed a very similar level of overall performance. The simpler method based upon using

[7] Analysis was also performed using the co-component with the largest number of matched candidates. However, this led to considerably worse performance and so, for brevity, is not displayed in this analysis.

[8] The correlation with one assessment (form X) was chosen for convenience. The correlations with form Y for individual co-components were usually fairly similar.

[9] This is necessary as most standard equating software expects anchor scores to be represented by whole numbers. Work by Benton and Yin (2011), in the context of using mean GCSE scores to maintain standards at A level, has previously shown that deciles tend to retain sufficient information for linking.

[10] For this calculation weight is defined by the known distribution of scores on the form X assessment. More weight is given to errors of equating at scores achieved by large number of candidates.

a single, highly correlated co-component, for equating also performed relatively well – we will return to this point further below.

Using key stage 2 data led to higher errors of equating. This is not particularly surprising as the median correlation between key stage 2 achievement and assessment scores was only 0.49. In contrast the median correlation derived from an ISAWG based on strong co-components was 0.73. In noting the errors of equating from key stage 2, it is worth remembering that this experiment was deliberately set up to be challenging by creating a fairly large difference in ability between groups. As such, in practice, the accuracy of key stage 2 in equating across years is likely to be better than suggested below. In addition, it can be seen from Table 2 that using key stage 2 data does lead to a noticeable improvement in accuracy over simply assuming that the two groups of candidates are equivalent.

Table 2: Weighted mean absolute error of equating for each method based on weighting the groups (frequency estimation)

Assessments (X-->Y)	Full ISAWG	ISAWG (strong co-components)	Key Stage 2 levels	One co-component	Assume equivalence
Art 1 --> Art 2	2.70	2.66	3.94	1.78	5.04
Biology 1 --> Chemistry 1	1.10	1.03	1.86	1.03	2.35
Biology 2 --> Chemistry 2	0.40	0.56	1.77	0.83	2.98
Biology 3 --> Chemistry 3	0.64	0.62	2.33	1.26	3.35
Business 2 --> Business 3	1.31	2.40	2.54	2.95	4.05
Computing 2 --> Computing 3	1.23	1.63	2.30	1.66	2.67
IT 1 --> IT 2	0.83	0.82	1.74	1.09	2.89
PE 2 --> PE 3	1.21	1.51	1.39	1.51	1.92
Science 3 --> Science 5	2.77	2.15	3.66	2.25	4.54
Science 4 --> Science 6	0.93	0.60	2.83	0.77	3.93
Science 8 --> Science 9	1.53	1.54	2.84	1.70	4.08
Biology 6 --> Biology 7	0.89	0.80	2.70	1.03	3.55
Chemistry 6 --> Chemistry 7	1.13	1.00	3.47	1.38	4.20
Physics 6 --> Physics 7	1.13	0.83	3.11	1.12	4.00
Math 1 --> Math 2	2.66	3.34	3.28	3.59	5.33
Math 3 --> Math 4	2.50	2.13	4.11	1.56	5.66
Median	1.17	1.27	2.77	1.44	3.96
Mean	1.43	1.48	2.74	1.59	3.78
Min	0.40	0.56	1.39	0.77	1.92
Max	2.77	3.34	4.11	3.59	5.66

Having considered the accuracy of equating methods similar to frequency estimation, Table 3 shows the accuracy of methods based on using the derived ability measures within chained equating. Using chained equating had a dramatic effect on accuracy. Most importantly, for every equating technique considered, there was a noticeable drop in equating error. This is particularly true for the method using the ISAWG based on strong co-components where the mean error dropped below one mark.

Table 3: Weighted mean absolute error of chained equating techniques

Assessments (X-->Y)	Full ISAWG	ISAWG (strong co-components)	Key Stage 2 levels	One co-component
Art 1 --> Art 2	2.32	1.98	2.64	2.13
Biology 1 --> Chemistry 1	0.87	0.50	1.10	0.49
Biology 2 --> Chemistry 2	0.15	0.17	1.03	0.43
Biology 3 --> Chemistry 3	0.37	0.21	1.56	0.71
Business 2 --> Business 3	0.91	1.47	1.55	1.92
Computing 2 --> Computing 3	0.85	0.96	1.75	1.01
IT 1 --> IT 2	0.66	0.35	1.16	0.98
PE 2 --> PE 3	0.90	1.04	1.04	1.04
Science 3 --> Science 5	2.35	1.36	2.56	1.55
Science 4 --> Science 6	0.69	0.32	1.95	0.39
Science 8 --> Science 9	1.23	1.06	1.96	1.12
Biology 6 --> Biology 7	0.54	0.35	1.97	0.46
Chemistry 6 --> Chemistry 7	0.74	0.51	2.72	0.76
Physics 6 --> Physics 7	0.71	0.34	2.31	0.50
Math 1 --> Math 2	2.41	1.46	1.78	1.98
Math 3 --> Math 4	2.03	1.28	2.90	0.96
Median	0.86	0.74	1.86	0.97
Mean	1.11	0.83	1.87	1.03
Min	0.15	0.17	1.03	0.39
Max	2.41	1.98	2.90	2.13

It is worth noting that, using chained equating, using a single co-component for equating led to similar accuracy to using the Full ISAWG. At first glance this is a little surprising but is less so once we inspect some further information about the single co-component used in each case as shown in Table 4. In particular, as can be seen, for many of the pairs in this experiment a single co-component can be found that is available for large number of candidates and has a high correlation with those being equated.

Table 4: Single co-components used use for this method of equating

Form X	Form Y	Further anchor description	N with form X	Correl with form X	N with form Y	Correl with form Y
Art 1	Art 2	English Literature 1	253	0.670	311	0.566
Biology 1	Chemistry 1	Physics 1	6029	0.712	5975	0.712
Biology 2	Chemistry 2	Biology 3	14500	0.824	15888	0.823
Biology 3	Chemistry 3	Chemistry 2	14762	0.823	15576	0.831
Business 2	Business 3	Computing 1	358	0.709	359	0.690
Computing 2	Computing 3	Computing 1	9573	0.596	9734	0.590
IT 1	IT 2	Physics 3	361	0.642	410	0.647
PE 2	PE 3	PE 1	5028	0.484	4976	0.504
Science 3	Science 5	Math 2	644	0.633	753	0.642
Science 4	Science 6	Science 8	5827	0.814	5844	0.842
Science 8	Science 9	Science 6	6031	0.823	5627	0.840
Biology 6	Biology 7	Chemistry 7	8896	0.795	9152	0.830
Chemistry 6	Chemistry 7	Biology 7	9018	0.808	9028	0.838
Physics 6	Physics 7	Chemistry 7	9017	0.803	9049	0.841
Math 1	Math 2	Chemistry 1	501	0.727	467	0.714
Math 3	Math 4	Chemistry 6	713	0.737	800	0.769

Summary

The results of this section have shown that data from co-components can indeed be used to reconstruct the correct equating relationship between two assessments. Indeed, the results in Table 3 show that the best approaches tended to yield equating functions that were only around 1 mark on average from their true values. In particular, for the purposes of equating over time using the full ISAWG (discussed next), it is important to note at this stage, that this provides a workable approach to equating within a given examination session. Furthermore, if we are using a frequency estimation equating method, the full ISAWG provides the most accurate of the available anchors.

The results also suggest that chained equating may provide a more accurate procedure for equating between co-components than a weighting or frequency estimation approach. This supports the current approach taken within England in the context of cross-tier equating[11]. However, it should be noted that reproducing these results is dependent upon the way in which candidates are split into the separate groups taking different test forms. Within this analysis, this was done on the basis of their ability (actual achievement) in the components being equated. Weighting the data by co-components is likely to under-compensate for these differences and so approaches based on chained equating are likely to be more successful[12]. Similar comments might be made for cross-tier equating as entry tier is likely to relate to candidates' abilities in the assessments being equated. However, if the groups were rather split based upon general ability rather than ability in the specific subject we would almost certainly find that frequency estimation techniques were more effective. Specifically, if we had deliberately split the data into groups P and Q based on the ISAWG rather than total scores on the test being taken, then the frequency estimation method based upon the ISAWG would broadly reverse the exact procedure used to create the groups in the first place and, for that reason, would provide the most accurate results. As such, while these results suggest that chained equating is worthy of attention, it cannot be definitely concluded that it should always be used in place of frequency estimation.

[11] See https://ofqual.blog.gov.uk/wp-content/uploads/sites/137/2017/03/Awarding-and-Comparable-Outcomes-maths-meeting-2017-03-07.pdf (downloaded on 6th November 2017).

[12] Although it is still likely to under-compensate for differences between groups, albeit to a lesser extent than frequency estimation.

Part 2: Exploration of equating across sessions using the ISAWG and calibration via common centres

Method

The next stage was to evaluate the possibility of using the ISAWG in maintaining standards over time. This could potentially be achieved using common centres to calibrate the ISAWG between sessions. Since the ISAWG can be calculated for all candidates taking any assessments it is ideal for such a procedure as very large numbers of candidates will be available for the calibration. This fact gives the ISAWG a unique advantage over all of the other methods described in the previous section if we wish to use it to assist with equating over time.

To test this idea we split the entire data set within the June 2015 session into two and then treated the two groups as if they were two separate sessions.

For the purposes of this part of analysis it is important that we include assessments with relatively small entries. This is because, for large entry assessments, techniques based directly upon common centres can usually be applied with confidence for each individual assessment in turn and so there is no need to rely upon the ISAWG. With this in mind, all GCSE components taken by at least 500 candidates in June 2015 (a total of 305 components) were included in analysis.

The entire data set of candidates was split into two with one half taken to represent session 1 and the other half taken to represent session 2. As with the earlier analysis, to make the situations where equating was required more challenging, this was not done completely at random[13]. Rather, it was devised so that for each assessment session 2 would tend to contain more able candidates than those in session 1. This might occur if high-achieving centres increased their entries between years and low-achieving centres decreased their entries. To simulate this, we first calculated the mean GCSE grade achieved across all candidate entries within each centre for Cambridge Assessment GCSEs taken in June 2015. These centre values were then standardised by subtracting the mean and dividing by the standard deviation. Finally, the probability that a candidate was assigned to session 2 was defined by the equation below:

$$P(Pupil\ assigned\ to\ session\ 2) = Max\left(0, Min\left(1, 0.5 + Standardised\ Mean\ Centre\ Grade * \left(\frac{0.8}{4}\right)\right)\right)$$

Note that this equation defines the probability of being assigned to either session entirely by achievement at centre level. Thus, within a centre, allocation to notional "sessions" was completely at random but within high achieving centres a greater proportion of candidates were assigned to session 2 than to session 1. The coefficient of 0.8 in the above equation was sufficient to ensure that on average across components in analysis there was a

[13] For completeness, appendix A shows results from a simulation where candidates are split into sessions completely at random.

difference of 0.25 standard deviations between component scores in session 1 and component scores in session 2[14].

The ISAWG was estimated separately within each session. Note that "weak" components were not removed from the estimation of the ISAWG as the same ISAWG was to be used to equate across sessions for all components. As such, there was no obvious single correlation that could be used to differentiate those that were "weak" from those that should have been retained.

Note also that the ISAWG in each session is defined by default to have a mean of 0 and a standard deviation of 1. However, since session 2 tends to contain higher ability candidates than session 1, these scores will not be comparable across sessions. To overcome this, we first need to calibrate the ISAWG across sessions. This was done using a combination of common centres and linear equating. To begin with all centres with at least 50 candidates allocated to each session and where the number of candidates differed by no more than 30 per cent between sessions were identified. Within the 576 centres identified in this way, the groups of candidates assigned to each session were assumed to be equivalent and linear equating was used to transform the ISAWG in session 2 to the same scale as that used in session 1. Over 60,000 candidates per session were included in this step. Figure 2 shows how the mean (pre-calibration) ISAWG within centres varied between sessions and where the equating line was placed. As can be seen, the same centres tended to have lower (uncalibrated) ISAWGs in session 2 as they were being compared to a higher ability set of candidates. The equating line provided a transformation that addressed this issue.

[14] Although, these values vary considerably between components. The difference between scores between session 1 and session 2 varied between -0.13 and +0.77 standard deviations.

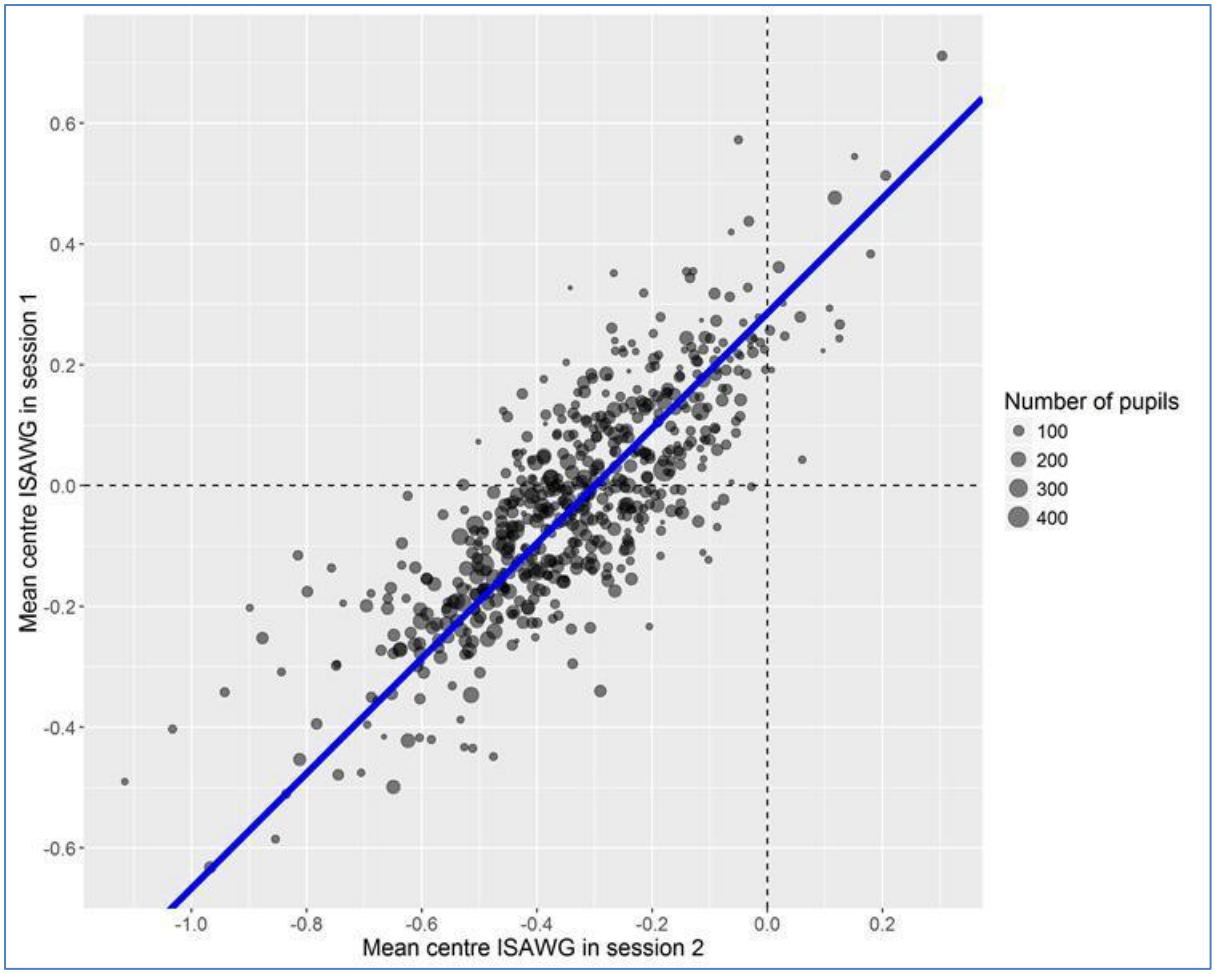

Figure 2: Mean initial ISAWGs in common centres within each session. The blue line shows the results of linear equating between sessions.

Equating from each component's scores in session 1 to the same component's scores in session 2 was completed using a frequency estimation method of equating with the (recalibrated) ISAWG playing the part of the anchor test. Frequency estimation was used, rather than chained equating as, for this experiment, allocation to sessions was achieved using a general measure of centre ability rather than scores within specific subjects. As such, this approach to equating may be more appropriate. In addition, this approach is closer to the way in which external assessment information such as key stage 2 is used in standard maintaining in practice (see Bramley and Vidal Rodeiro, 2014). The resulting equating function was compared to the identity function. This was appropriate as, since each component is being equated to itself, the true equating function must be the identity line.

As an alternative to the ISAWG based approach, common centres equating was applied directly to each component. In this procedure, centres with at least 20 candidates[15] in each session within the component of interest and where the number of candidates varied between sessions by no more than 30 per cent were identified. For candidates within the identified common centres, equipercentile equating between the scores in the different

[15] This number was chosen as it was also small enough to ensure that at least one common centre could be identified in the majority of cases.

sessions was applied directly. The resulting equating function was again compared to the identity line so that the weighted mean absolute error of equating could be calculated. Due to the need to collate results across a very large number of components, these errors were standardised to be presented as a percentage of the maximum available score.

In addition, equating between the notional sessions using key stage 2 was performed. Key stage 2 levels for each pupil were used as an anchor between sessions on the assumption that these will be comparable between sessions[16].

Note that no such common centres meeting the criteria could be found for 40 of the components of interest and so these were removed from the comparison of equating techniques. A further 35 components were removed due to having less than 500 candidates with matching key stage 2 data (in total). This left a total of 230 components in the analysis.

Results

The results of analysis are shown in Figure 3. Each point on this plot represents an attempt at equating for an assessment component. The points in red show the weighted mean absolute error of equating for the direct common centres method. The blue points show the errors of equating for the same set of assessments when the ISAWG approach is applied. As can be seen, for components with fewer than 10000 entries per session, the ISAWG-based approached yielded lower errors of equating. Even for the smallest-entry components included in analysis, the error averaged at only around 1.5 per cent of the maximum available score on the component. In contrast, the average error for small components using the benchmark centres method exceeded 3 per cent of the maximum available score. This is not surprising as for many of these very small entry components only a handful of benchmark centres (perhaps only 1[17]) were available. It should be noted that using linear rather than equipercentile equating for the common centres method led to only a very small reduction in error for these centres[18].

In contrast, for large entry components the common centres method was slightly more accurate than using the ISAWG-based approach. This is unsurprising as, given the way in which this experiment was set up, groups within common centres should be exactly equivalent in terms of ability. This means that, provided we have enough of these, the common centres method will provide extremely accurate results.

[16] Of course, since this is a simulation and all candidates have done the same key stage 2 tests, we know for certain that this assumption holds for this particular experiment.

[17] A total of 38 components had only one identified common centre and 119 had less than five identified common centres.

[18] It also leads to a slightly unfair comparison where one method uses a linear method that reflects the linear nature of the identity equating function and one does not. For this reason these results are not shown in full.

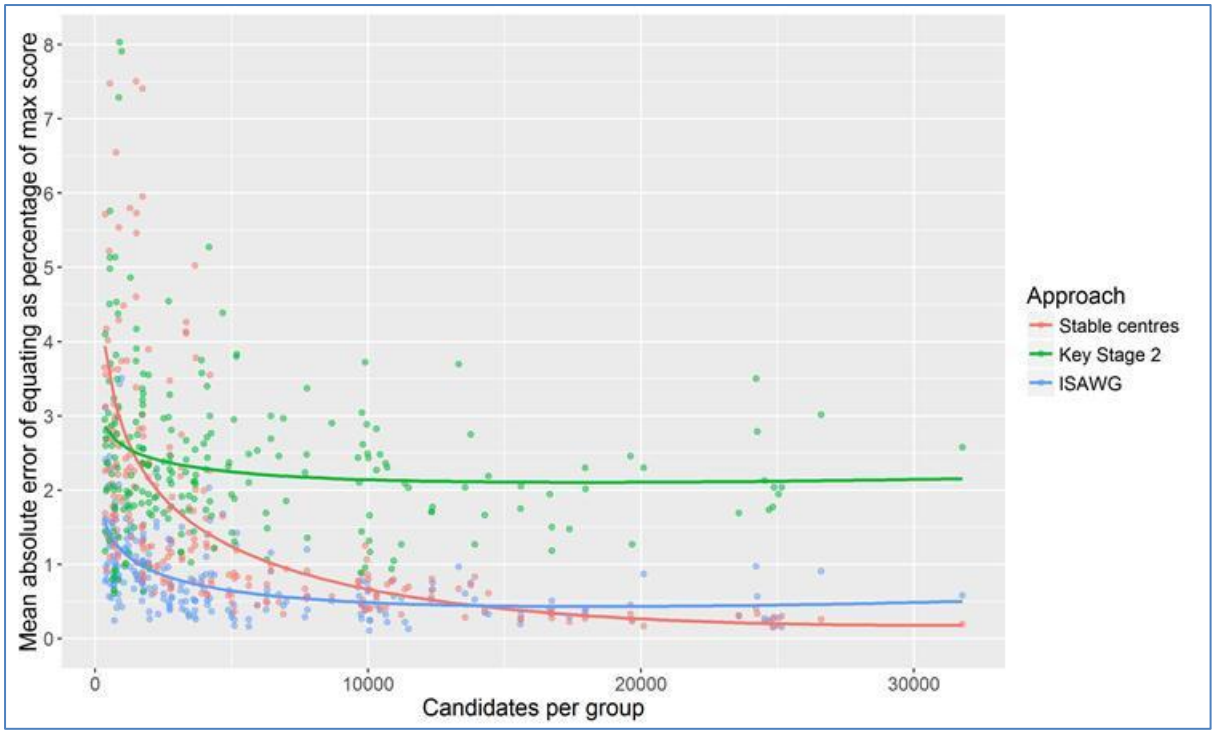

Figure 3: Weighted mean absolute errors of equating for ISAWG and benchmark centres based approaches

Key stage 2 data did not provide a particularly accurate mechanism for equating. This is because we deliberately set up the experiment so that there was a large difference in average abilities between years. As was noted in the introduction, this is exactly the type of scenario where key stage 2 data can fail to fully account for the differences between groups. As such, we would not expect the poor performance of key stage 2 in this experiment to be repeated very often in practice. Nonetheless, this experiment has shown how using the ISAWG can provide a more accurate method for maintaining standards than using key stage 2 data even when we are faced with extreme differences between years.

Discussion

This paper has described the ISAWG as an easy method of summarising pupil achievement regardless of which assessments they have taken. It has also demonstrated its efficacy in equating with the estimated equating functions typically only around 1 mark away from the true values.

If we suspect that the groups taking different assessments may differ in terms of their ability in the specific subject of interest, then chained equating is likely to provide a more accurate method than weighting or frequency estimation techniques. This indicates that, at the very least, results from chained equating should be considered alongside other techniques. As noted elsewhere (Bramley and Vidal Rodeiro, 2014) the main statistical methods used for maintaining standards in England are very similar to frequency estimation. As such, they are likely to share the same weaknesses – namely that they may underestimate the extent of differences in abilities between groups (see Benton and Sutch, 2014, for further evidence of this issue).

The advantage of chained equating will be dependent upon the causal mechanism leading to the assignment of candidates to different assessments and whether this is driven by their ability in the particular subject of interest or by ability (or achievement) more generally. It is certainly possible to simulate scenarios where the performance of frequency estimation techniques is superior to chained methods and, as such, a cautious approach where we gather evidence from both methods may be prudent even if this is inconvenient in meaning that we do not immediately provide a single equating function. It should be noted that, if the differences between groups are small then different methods should provide fairly similar results.

Finally this paper has demonstrated the potential of the ISAWG in equating across sessions. Indeed, as might be expected, for fairly small assessments this method is superior to an approach based purely upon a limited number of stable common centres. For all sample sizes, and with the particular way in which the experiment was set up, it was also more effective than using data from key stage 2 assessments. This suggests that the ISAWG may provide a suitable alternative to both to using key stage 2 and to the use of common centres in maintaining standards.

Mathematically, the calculation of the ISAWG shares some similarities with the Kelly method sometimes used to compare the difficulty of examination grades across subjects[19] (see Bramley, 2014). As such, it is likely to be most effective in terms of ISAWGs derived from different combinations of qualifications being truly comparable when pupils' subject choices are not influenced by their likely future performance in those same subjects (Bramley, 2016). However, complete comparability of all ISAWGs across all pupils is not strictly necessary for the application demonstrated in this paper – only that any weaknesses in the ISAWG are relatively stable over time amongst those pupils entering each assessment. This is because we are using the ISAWG purely to explore differences between cohorts and it is not

[19] With the differences being that the ISAWG begins with raw marks rather than grades and that it includes a slope parameter for each assessment as well as an intercept (difficulty).

something that is reported to individual candidates. In the context of using the ISAWG for equating described in this paper, Part 1 has shown that the ISAWG may still be relatively effective in equating even when assessment choice is directly linked to likely performance. Further work could explore in more detail how different mechanisms regarding the way in which pupils are entered for different assessments impact upon the accuracy of equating using the ISAWG.

References

Anderson, B., Branberg, K., and Wiberg, M. (2013). Performing the Kernel Method of Test equating with the Package kequate, *Journal of Statistical Software*, 55(6), http://www.jstatsoft.org/v55/i06/.

Benton, T. and Sutch T. (2014). *Analysis of the use of Key Stage 2 data in GCSE predictions*. Ofqual, Ofqual/14/5471, Coventry. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/429074/2014-06-16-analysis-of-use-of-key-stage-2-data-in-gcse-predictions.pdf

Benton, T. and Lin, Y. (2011). *Investigating the relationship between A level results and prior attainment at GCSE* (Ofqual Research Report OFQUAL/11/5037). Coventry: Ofqual. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/605906/2011-09-29-investigating-the-relationship-between-a-level-results-and-prior-attainment-at-gcse.pdf

Bramley, T. (2014). Multivariate representations of subject difficulty. *Research Matters: A Cambridge Assessment Publication*, 18, 42-47. http://www.cambridgeassessment.org.uk/Images/174492-research-matters-18-summer-2014.pdf

Bramley, T. (2016). The effect of subject choice on the apparent relative difficulty of different subjects. *Research Matters: A Cambridge Assessment publication*, *22*, 23-26. http://www.cambridgeassessment.org.uk/Images/374638-the-effect-of-subject-choice-on-the-apparent-relative-difficulty-of-different-subjects.pdf

Bramley, T & Vidal Rodeiro, C. (2014). *Using statistical equating for standard maintaining*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. http://www.cambridgeassessment.org.uk/Images/182461-using-statistical-equating-for-standard-maintaining-in-gcses-and-a-levels.pdf

De Ligny, C.L., Nieuwdorp, G.H.E, Brederode, W.K., and Hammers W.E. (1981). An application of factor analysis with missing data, *Technometrics*, 23(1), 91-95.

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

Spearman, C. (1904). “General Intelligence,” Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201-292.

Thorndike, R.L. (1994). g, *Intelligence*, 19, 145-155. http://dx.doi.org/10.1016/0160-2896(94)90010-8.

von Davier, A. A., and Chen, H. (2013). *The Kernel Levine Equipercentile Observed-Score Equating Function*. Research Report. ETS RR-13-38. ETS Research Report Series. https://www.ets.org/Media/Research/pdf/RR-13-38.pdf.

Appendix A: Results of simulation of equating across a completely random split into sessions

This appendix repeats the analysis from part 2 but this time with candidates split into sessions completely at random so that there is no systematic tendency for the candidates in session 2 to be of higher ability than those in session 1. Because of the nature of the simulated split, it was no longer necessary to rely on common centres or key stage 2 in order to perform a simple form of equating. Instead, equating based upon assuming the equivalence of all candidates between sessions provided a sensible alternative to the method based on the ISAWG[20]. Analysis was based on the same set of components as those described in section 2. A total of 305 such components were retained in analysis as, since we were not relying on common centres or key stage 2, none were removed on grounds of being unable to find suitable matching data. As before, in each case, each equating method was evaluated by the accuracy with which the estimated equating functions matched the identity equating line. The results of analysis are shown in Figure A1.

As before, for components with a very large number of entrants, because the assumption of equivalence is likely to be met, both methods displayed a similar level of accuracy. However, for components with between 1,000 and 5,000 entrants per group the method based on the ISAWG had lower error. This is because it can adjust for cases where the random split of candidates happens to lead to a difference in ability between candidates taking a particular component in different sessions.

For components with very small entry sizes, the performance of the two methods was again similar. This might be because the additional estimation of parameters for the ISAWG-based method[21] adds some slight instability to the method which might counteract the possible benefits in terms of adjusting for differences in ability between the two groups. Note that, given the way this simulation was set up, the differences between groups should be small in each case. This means that the design of this particular simulation does not allow the ISAWG-based method to demonstrate its full potential. As such, the fact that the ISAWG method still generally performs at least as well and often better than assuming equivalence is encouraging.

[20] For consistency, the ISAWG itself was still calibrated between simulated sessions using common centres. Of course, nearly all large centres will now meet the criteria to be defined as common centres so that the majority of data will be retained for this step.

[21] In particular the principal components coefficients for each individual assessment.

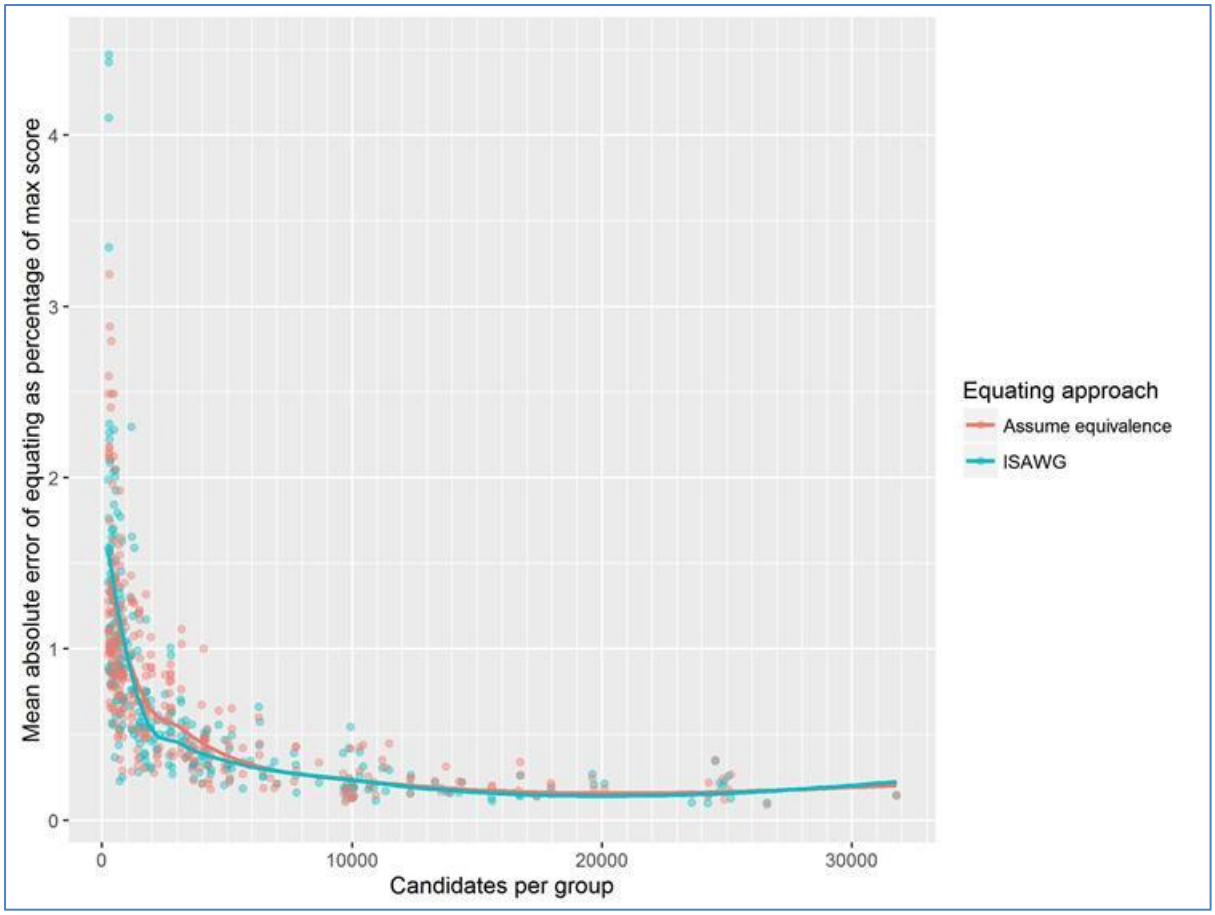

Figure A1: Weighted mean absolute errors of equating for ISAWG-based method and method based on assuming equivalence for simulation study where candidates are split into sessions completely at random.