Which assessment is harder? Some limits of statistical linking

Tom Benton and Joanna Williamson (Research Division)

Introduction

Equating methods are statistical processes whose purpose is to put scores from different assessments onto the same scale. A key application of equating is to determine equivalent scores when candidates for the same qualification can take alternate versions of certain assessment components. For example, candidates who are in different time zones or who take the same qualification at a different time of year may sit different versions of a written examination component, and it is necessary to know which scores represent the same level of achievement on the different assessment versions so that no candidate is disadvantaged. Definitions of equating stress that equating is for adjusting between alternate versions of assessments targeting the same content at the same level, with the aim that scores from the different versions can be used "interchangeably" (Kolen & Brennan, 2014, p. 2).

The statistical processes used in equating have also, however, been extended to compare pairs of assessments that do not meet these strict criteria. There is often great interest in the comparability of assessment scores from related assessments targeting the same construct at different levels, from parallel qualifications targeting the same subject at the same level, and from assessments of the same qualification type that assess different subjects. The use of equating methods and close variants to statistically "link" assessments for such comparisons has a different conceptual basis to equating in the clear sense that statistical adjustments cannot make the scores from a Physics exam and a History exam "interchangeable". There are however high-stakes situations in which such scores are in fact interpreted interchangeably (e.g., school league tables, or a university place conditional on achieving three A Levels at grades AAB), providing ample motivation for asking whether certain assessments are "too hard" or "too easy" in comparison with others. Despite careful debate over the basis for statistical linking and the precise conclusions that can and cannot be drawn (e.g., Mislevy, 1992, pp. 21-26; Newton, 2010), including in the literature on inter-subject comparability (e.g., Bramley, 2011; Coe, 2008; Newton, 2012), it can be tempting

to apply equating methods and conclude that they have provided a definitive answer regarding whether a qualification is harder or easier than others.

The purpose of this article is to explore how accurately various equating methods are able to equate between identical assessments. It offers a novel demonstration of some limits of statistical equating by making use of pairs of live assessments that are "cover sheet" versions of each other, that is, identical assessments with different assessment codes. Such pairs occur most commonly where the same assessment is a component of corresponding qualifications of different types (e.g., an IGCSE and O Level in the same subject) or a component of related qualifications of the same type (e.g., IGCSE Combined Science and IGCSE Co-ordinated Sciences). The fact that the assignment of students to particular cover sheet versions is a non-random process means that this context may provide a more realistic evaluation of various equating techniques than others. In particular, the equating methods will have to address issues of differences in the abilities and subject choices of candidates taking different qualifications that occur in real practical situations. At the same time, the evaluation of the equating methods' accuracy is made straightforward by the fact that the true equating relationship is known: since the two assessments in a cover sheet pair are identical, the scores from the two assessments are already on the same scale, and the true equating relationship is the one that maps each score to itself (in mathematical terms, the identity function).

How to link assessments

The outcome of equating two assessments is a statistical transformation or equating function that allows scores from one assessment to be interpreted on the same scale as scores from the second assessment. Two things are needed to generate the statistical transformation or equating function: firstly, locating or collecting some data that links candidate performance on the two assessments, and secondly, making a decision about which definition of "same standard" the equating function should preserve.

In some equating designs, information linking candidate performance on the two assessments is obtained directly, by having a single group of representative candidates take both assessments. Alternatively, candidates may be randomly assigned to sit one or other of the two assessments, and for sufficiently large groups, the groups can be assumed equivalent. In these designs, differences in performance can be interpreted as representing differences in the difficulty of the two assessments, rather than differences in the candidates sitting them.

In high-stakes live assessments such as IGCSEs, O Levels and A Levels, security concerns prevent the pre-testing of assessments, and it is not possible to assign candidates randomly to different live papers. Where the groups sitting each assessment cannot be assumed equivalent, it is – clearly – more challenging to judge comparable standards in the two assessments, as this has to be disentangled from differences in the ability of the two groups. Equating designs for non-equivalent groups require some link between the assessments of interest.

Where available, this link can be achieved via a subset of common items that feature on both assessments (as seen, for example, in tiered GCSE Mathematics). The fact that these items are taken by both candidate groups allows them to function as a reference point or "anchor" for understanding the group differences (hence Non-Equivalent Groups with Anchor Test or NEAT equating design). For many pairs of assessments, however, there is no subset of items forming an internal anchor test, and an external link or anchor must be found. For GCSEs and A Levels in England, prior attainment, at Key Stage 2 and GCSE respectively, is typically used in place of an anchor (see Bramley & Vidal Rodeiro, 2014). In other scenarios, particularly where prior attainment is unavailable, an external link might be identified from common components, that is, assessments taken by candidates from both groups alongside the assessments being equated.

Previous work by Benton (2017) developed a method for going beyond co-components and taking into account all the information linking candidates. The core idea is a summary measure known as the ISAWG (Instant Summary of Achievement Without Grades), a measure of ability that summarises each candidate's performance across multiple assessments on a single scale, whichever assessments they have taken. The ISAWG value for each candidate can be defined informally as "the single number that most accurately reflects the standardised marks they have achieved on whichever assessments they have taken" (Benton, 2017, p. 6). When used to equate between assessments, the ISAWG measure therefore incorporates information about candidates' performance on all co-components (if any exist), but also assessments that are not co-components[1].

An important theoretical objection to equating assessments via co-components or ISAWG measures is the defensibility of comparing assessments in one subject using data from assessments designed to measure candidates' abilities in different subjects or qualifications. Besides assessing different content, factors that can undermine comparisons include differences in teaching and levels of student motivation, and whether an assessment is compulsory or the result of student choice. Data on candidates' achievement in different assessments can be used in such a way that the most relevant information is prioritised over less relevant information (e.g., by restriction to related subjects, or to similar qualification types, or prioritisation of co-components according to correlation with assessment scores and candidate numbers), but the concern is a valid one, and has been extensively debated. The theoretical basis for pursuing a measure such as ISAWG is Spearman's (1904) theory of general ability or "g", which would suggest that "although different tests may measure slightly different skills, all of them should relate to each candidate's 'fundamental function' (or 'g')" (Benton,

1 The technical procedure for calculating ISAWG is equivalent to carrying out Principal Components Analysis on a data set including all of the assessments offered by Cambridge International and OCR in a single session – with missing values included, since no candidate takes all available assessments – and taking the first principal component for each candidate. Although other research has also investigated how to incorporate information from covariates into equating (e.g., Andersson et al., 2013; Wiberg & Branberg, 2015), the ISAWG is uniquely well suited for equating using very large sets of covariates (in this case, assessments) with highly variable missing data patterns (see Benton, 2017, p. 8).

2017, p. 6). This, in turn, should provide a reasonable basis for estimating the candidate's likely achievement on other assessments. In the context of setting grade boundaries or cut scores for standard maintaining, previous work has shown that co-component and ISAWG equating methods are promising (Benton, 2017). Importantly, in the context of setting grade boundaries, equating outcomes can be and are considered alongside multiple other sources of evidence, including expert judgements about question papers and candidate scripts, and sometimes alternative types of statistical evidence.

Once data linking the assessments to be equated has been identified, there are multiple ways to define an equating relationship. One widely used approach is equipercentile equating, in which scores from Test X and Test Y are considered equivalent if they represent the same percentile rank for the specified population. Equipercentile equating allows for non-linear relationships between Test X and Test Y: for example, the equating function may indicate that Test X is easier than Test Y at the very top and bottom of the score range, but not in the middle of the range. Equipercentile equating requires more data than some other methods, but is less restrictive in its assumptions and requirements and hence is suitable for the type of assessments considered in this article, where lack of data is not a problem.

Equating percentile ranks for the complete and non-equivalent groups taking Test X and Test Y would of course not account for any differences in group ability, and there are two main approaches to dealing with this. In frequency estimation (FE) equipercentile equating, the candidate groups for Test X and Test Y are first weighted so that they are equivalent in terms of their anchor test score distributions (i.e., to create equivalent groups, so far as we are able, for which we have both Test X and Test Y scores). Using the weighted data, the score distributions for Test X and Test Y are created. These are then used to equate percentile ranks. As an alternative to FE, the chained method equates percentiles first from Test X to the anchor test within the Test X candidates, then equates percentiles from the anchor test to Test Y within the Test Y candidates.

Method

Several equating methods were investigated by exploring how accurately they equated between pairs of identical assessments with different cover sheets. Each method produced an estimated equating function linking the scores between the two assessments in each pair (Test X and Test Y). These equating functions were evaluated by comparison to the true equating function, which, for all pairs, was the identity function. For each Test X score, the difference between the estimated equated score and actual equivalent score (equal to the Test X score) was calculated. The differences between the estimated equated scores and actual equivalent scores were then summarised in terms of the cumulative percentage of candidates achieving each score or above. This information indicated the differences in pass rates that would result from cut scores at any chosen point, allowing the equating errors to be interpreted in terms of their impact rather than just magnitude.

Five equating techniques were investigated: four versions of equipercentile

equating with an anchor test or measure, and, as a contrasting but widely used approach, Rasch equating[2]. To support fair comparison between the equating methods, each method was restricted to considering the same set of possible co-components. For each cover sheet pair, the set of usable co-components was defined to be the 20 largest components (by joint *N* also taking Tests X and Y) taken by at least 100 and at least 5 per cent of candidates taking Test X, and also at least 100 and at least 5 per cent of the candidates taking Test Y.

The details of the five methods were as follows:

1. Single co-component (FE)

Components were equated by choosing a single co-component to use as an anchor test, in frequency estimation (weighted) equipercentile equating. The co-component chosen from the set of (up to) 20 usable co-components was that with the highest minimum correlation with Test X and Test Y. Single co-component equating was investigated due to its simplicity as well as good performance in prior equating studies (Benton, 2017).

2. Single co-component (chained)

Components were equated using a single co-component as an anchor, in chained equipercentile equating. As in Method 1, the co-component selected was that which had the highest minimum correlation with Test X and Test Y. It was considered important to test chained as well as frequency estimation methods, since frequency estimation is recommended only when groups are "reasonably similar" (Kolen & Brennan, 2014, p. 146), and there is evidence that chained methods may be more successful when the abilities of Test X and Test Y candidates in fact differ meaningfully (Benton, 2017).

3. ISAWG (FE)

Components were equated using candidates' ISAWG measure in place of an anchor test score in frequency estimation equipercentile equating. The ISAWG measure was recoded into integers 0-19 before equating was carried out (since nearly all equating methods expect integer anchor test scores).

4. ISAWG (chained)

Components were again equated using the separately calibrated ISAWG measure in place of an anchor test, but within chained equipercentile equating. In contrast to the standard ISAWG measure (Benton, 2017), the ISAWG measures used in Methods 3 and 4 were separately calibrated for each cover sheet pair. The calculation of ISAWG values was restricted to using scores from that pair's set of (up to) 20 usable co-components as well as the two components themselves. Note that the two components in the pair being equated were *not* themselves treated as cover sheet versions of each other. That is, the ISAWG calculation considered these two components as entirely separate assessments, while, if they

2 In contrast to the four equipercentile equating procedures, which are observed-score equating methods, the Rasch equating method is a form of item-response theory (IRT) true-score equating in which scores from Test X and Test Y are considered equivalent when they correspond to the same level of underlying ability construct (see Kolen & Brennan, 2014, pp. 175, 213).

were among the relevant co-components, all other cover sheet pair equivalences remained "known" to the ISAWG calculation.

5. Rasch equating

For each cover sheet pair, Test X, Test Y and the corresponding (up to) 20 usable co-components were first analysed as a (up to) 22 "item" test using a polytomous extension of the Rasch model (an Extended Nominal Response Model fitted in R, using the package Dexter (Maris et al., 2021)). This allowed raw scores from Test X and Test Y to be related to a single unidimensional ability scale. Scores on Test X and Test Y were then linked so that for each Test X score, the equated Test Y score was the Test Y score corresponding to the same point on the ability scale.

Data set and description of equating context

Equating was carried out on pairs of IGCSE and O Level components taken by Cambridge International candidates in summer 2018. Each pair consisted of identical assessments with different cover sheets, taken by non-overlapping groups of candidates. Analysis was restricted to pairs with at least one co-component suitable for equating, defined to be an assessment component taken by at least 100 and at least 5 per cent of the Test X candidates, and by at least 100 and at least 5 per cent of the Test Y candidates. Analysis was further restricted to pairs in which each assessment was taken by at least 3000 candidates who also took at least one co-component, and where the total available marks were at least 50, to avoid the results reflecting the difficulties of equating with too few candidates or too few marks. Individual qualifications were included only once. Where multiple cover sheet pairs from the same qualification met the conditions for inclusion, the pair with the higher number of available marks was retained, and if multiple pairs still remained, the pair of components with the smallest difference in raw mark means was retained.

Eight pairs of assessments met the above conditions, covering subjects from English as a Second Language (ESL) to Mathematics (Table 1). All the assessments were externally assessed written examinations. The first pair of ESL components (ESL 1) assessed both reading and writing, while the second pair (ESL 2) assessed writing only. Both the Maths 1 and Maths 2 component pairs consisted of two-hour written tests, but belonged to different mathematics qualifications. All pairs had at least five usable co-components. Where candidates are non-randomly assigned to assessment versions, the candidate groups taking each version may differ substantially – particularly if taught in different school systems – and the differences in mean assessment scores shown in Table 1 reflect this. For some pairs the mean marks achieved in Test X and Test Y were extremely close, for others there was a moderate difference, and for two pairs the difference was very large.

Table 1: Description of component pairs investigated.

Component pair	Max. mark	Test X N	Test Y N	Number of usable co-components (capped at 20)	Difference between Test X and Test Y mean scores (as per cent of maximum mark)
Business	80	5375	10 788	7	0.87
Computing	75	7627	4222	9	2.35
Economics	90	5898	9469	7	4.14
ESL 1 (Reading and Writing)	90	5906	9031	20	0.97
ESL 2 (Writing only)	60	3103	24 889	5	12.93
History	60	3088	12 033	18	6.57
Maths 1	80	4726	6263	7	5.28
Maths 2	104	3518	8104	10	20.55

As noted in the description of equating methods, the co-component selected for single co-component equating (Methods 1 & 2) was that (from the set of usable co-components) which had the highest minimum correlation with Test X and Test Y scores. Table 2 shows that these minimum correlations were generally high. The pairs for which the single best co-component had the lowest correlations with scores were History (0.52 with Test X) and ESL 2 (0.62 with Test Y). For History, the single chosen co-component was also only taken by a relatively small minority of those taking Tests X and Y. For Business, Computing, Economics and the two Maths pairs, correlations between the single best co-component and component scores were all 0.8 or higher. Correlations between component scores and the separately calibrated ISAWG measures used in Methods 3 and 4 were also high. The lowest correlation occurred for ESL 2 (0.72), and for Business, Computing, Economics and the two Maths pairs, correlations between the ISAWG measures and component scores were again particularly high (around 0.9). Part of the reason for these high correlations is, of course, that the components themselves (that is, Tests X and Y) contribute to the calculation. The use of multiple co-components also avoids potential loss of data by restricting to a single co-component.

Table 2: Correlation of component scores with single co-component anchor measures and ISAWG. Number of students with available score on the single co-component are also shown.

Component pair	Method 1 & 2 co-component with Test X		Method 1 & 2 co-component with Test Y		Correlation of ISAWG with Test X, Test Y	
	Correlation	N	Correlation	N	Test X	Test Y
Business	0.80	5373	0.82	10 774	0.91	0.92
Computing	0.80	7619	0.81	4221	0.88	0.91
Economics	0.80	5885	0.80	9457	0.93	0.93
ESL 1	0.73	5880	0.73	9004	0.83	0.82
ESL 2	0.62	3100	0.69	24 871	0.88	0.72
History	0.52	337	0.62	1500	0.82	0.83
Maths 1	0.91	4726	0.89	6262	0.94	0.92
Maths 2	0.88	3518	0.88	8089	0.90	0.96

A first indication of the differences between Test X and Test Y candidate groups was given by the differences in component scores reported in Table 1. To allow a closer look at any between-group differences, Figure 1 compares the standardised difference between Test X and Test Y candidates' component scores[3] (on the x axis) with the standardised difference in their ISAWG measures[4] (on the y axis). Figure 1 demonstrates, firstly, that there was a high level of agreement between the two measures in terms of which candidate group was higher performing. Both the direction and size of the ability difference indicated by the standardised score difference was generally reflected by the standardised ISAWG difference. The largest discrepancy was for ESL 2, where component scores indicated a standardised difference of -1 between candidate groups, whereas the ISAWG measure indicated a difference of -0.67 standard deviations. Secondly, Figure 1 highlights that for some of the cover sheet pairs, the difference between Test X and Test Y candidate groups was rather large, confirming that this real-world equating context included a high level of challenge. As a rule of thumb, Kolen and Brennan (2014, p. 301) note that equating can be "especially troublesome" where group differences are larger than 0.5 standard deviations. Figure 1 shows that the difference between candidate groups in History was around this threshold, while for ESL and Maths 2 the standardised score differences were around double this threshold.

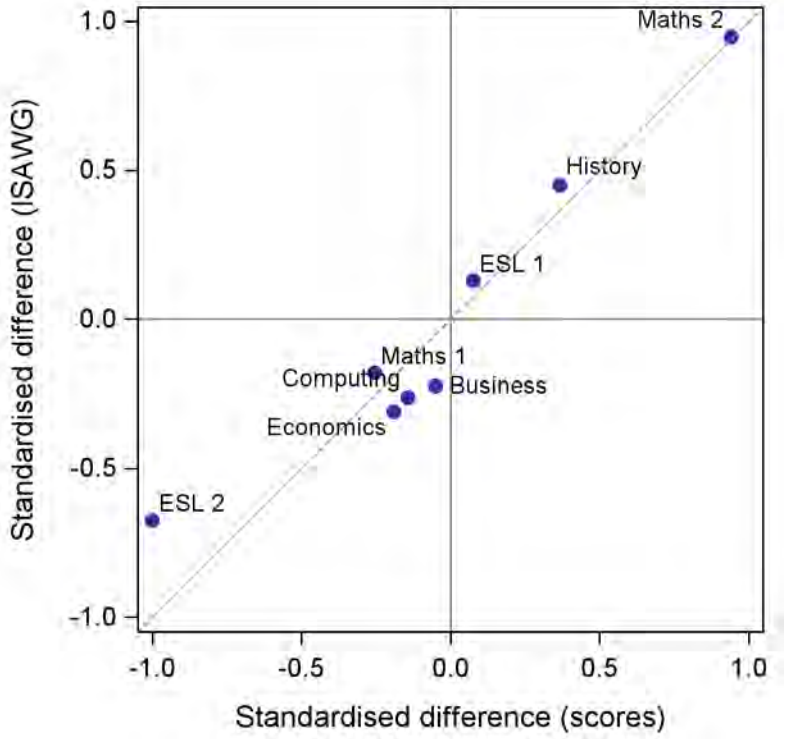

Figure 1: Comparison of standardised component differences.

3 The standardised score difference was calculated by subtracting the mean Test Y score from the mean Test X score and dividing by the pooled standard deviation of Test X and Test Y scores.

4 The standardised ISAWG difference was calculated by subtracting the mean ISAWG of Test Y candidates from the mean ISAWG of Test X candidates and dividing by the pooled standard deviation of Test X and Test Y candidates' ISAWG measures, using the ISAWG values separately calibrated for that specific Test X Test Y component pair.

Findings

How accurate was equating?

Figure 2 shows the equating outcomes for each cover sheet pair and equating method. For each score on Test X of the pair (shown on the x axis, as a percentage of the test's maximum mark), the graphs show the difference between the estimated equivalent Test Y score and actual equivalent Test Y score (equal to the Test X score itself). This provides a visual summary of how closely each estimated equating function resembled the correct (identity) function. It allows the accuracy of the different methods to be compared through the score range: in the Computing pair, for instance, the graph shows that the Rasch equating method over-estimated the equivalent Test Y scores much more than other equating methods for Test X scores between 5 per cent and 25 per cent. For higher Test X scores, on the other hand, the Rasch equated scores had similar accuracy to those estimated from the other methods.

The patterns of equating error shown in Figure 2 (that is, the deviations from the correct equating function – the identity function – as plotted on the y axis) varied by pair. For Business, equating errors were small and highly consistent between the different methods. Equating errors were also consistently small in Economics, although Figure 2 shows some separation between the two single co-component methods (which produced very similar results), and the ISAWG and Rasch methods. Equating errors were still consistently within a small range for Maths 2, although here there was more variation between the equating methods. The correlations between Business, Economics and Maths 2 scores and their anchor measures (both co-component scores and ISAWG) were all high, which would tend to support equating accuracy. The large difference in Test X and Test Y candidate abilities for Maths 2 was an apparent challenge to overcome, but this pair was nevertheless equated very accurately.

For Computing and Maths 1, pairs which had high score-anchor correlations and fairly small group differences, equating errors were small except for deviations in specific methods towards the lower end of the score range. In Computing, the larger errors occurred with the Rasch equating method, and in Maths 1, the larger errors occurred in both the Rasch and ISAWG methods.

Equating errors were slightly larger for the ESL 1, History, and particularly ESL 2 component pairs, consistent with the fact that these three pairs showed the lowest score-anchor correlations (Table 2). In the case of History and ESL 2 there were also fairly large ability differences between the Test X and Test Y candidate groups, further increasing the level of equating challenge.

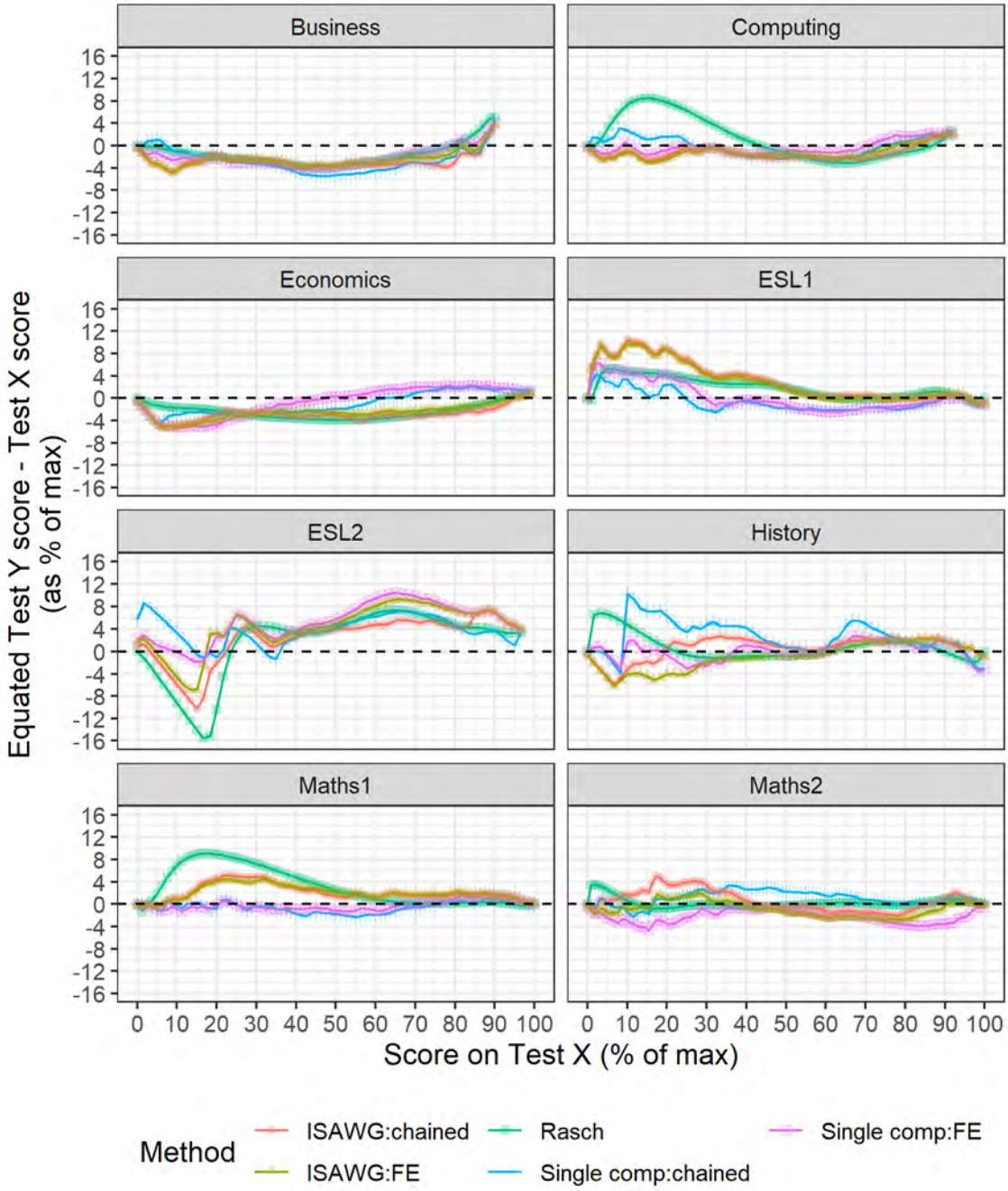

Figure 2: Equating errors, by method, against Test X score.

To summarise the size of equating errors, Table 3 reports the weighted mean absolute error of equating for each pair of equated components under each method. Weighting by the number of (Test X) candidates at each score gives priority to those parts of the score range with higher numbers of candidates. For each component pair, the lowest overall equating error (using this definition) is highlighted, which highlights that the equating method achieving the highest accuracy varied between pairs. Comparing the results for the different component pairs within each method, however, shows that each method produced its highest levels of equating error for the ESL 2 pair. Prior to running the analysis, we expected chained equating to outperform the FE method in cases where there was a large difference in group means – namely ESL 2, History and Maths 2 (see Figure 1). However, for History this was not the case, with the FE method providing more accurate results for both the ISAWG and single co-component approaches.

The final row of Table 3 shows the mean error of each method across the eight data sets. On this measure the Rasch approach was slightly more accurate than the alternatives, although the difference between them was very small.

Table 3: Weighted mean absolute errors of equating (as per cent of component max mark). For each pair, the lowest overall equating error is highlighted.

Component pair	Single comp (FE)	Single comp (chained)	ISAWG (FE)	ISAWG (chained)	Rasch
Business	2.99	4.08	2.86	3.24	2.88
Computing	1.06	1.75	1.56	1.94	2.40
ESL 1	1.56	1.68	0.43	0.61	0.53
ESL 2	7.83	5.52	6.91	4.46	5.58
Economics	1.69	1.63	2.30	2.93	2.89
History	1.15	2.49	1.17	1.48	0.95
Maths 1	0.90	0.82	2.02	1.70	1.27
Maths 2	2.72	0.86	1.51	1.12	0.36
All	2.49	2.35	2.34	2.19	2.11

Figure 3 demonstrates how the equating errors shown in Figure 2 would affect pass rates, by comparing the cumulative percentage of candidates reaching actual and equated scores. For each cover sheet pair, the Figure 3 x axis shows the cumulative percentage of Test Y candidates above a given score, for example, 40 represents the score which 40 per cent of Test Y candidates achieved or exceeded. The y axis shows the difference between this percentage and the percentage of Test Y candidates who reached or exceeded the corresponding equated score. The top left cell of Figure 3 shows that for the pair of Business assessments, the "pass rate" at a cut score achieved by 50 per cent of the Test Y cohort would have been around 10 percentage points higher using the equated cut scores from the single co-component (chained equipercentile) method shown in blue. This corresponds to the fact that the single co-component (chained) method resulted in equated scores for Business that were lower than actual scores throughout most of the score range (see top left cell of Figure 2).

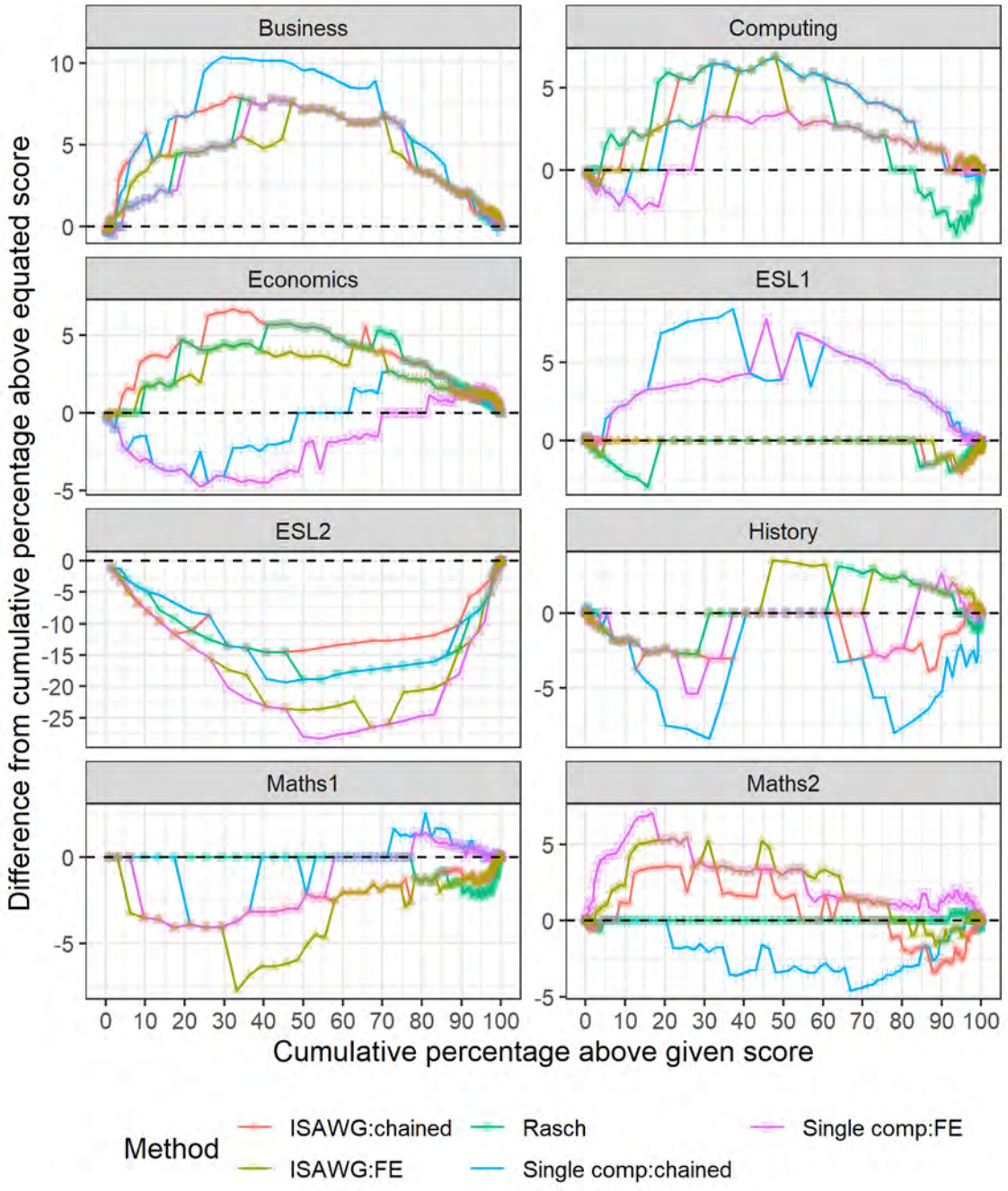

Figure 3: Differences between cumulative percentages of Test Y candidates reaching actual and equated scores.

The cumulative percentage of Test Y candidates achieving equated scores was generally within 10 percentage points of the cumulative percentage at the original score, except for the ESL 2 component pair, where differences for all equating methods exceeded this. Focusing on the cumulative percentages of candidates above a certain score highlights that differences in the "pass rates" at cut scores can become relatively large even when the absolute sizes of equating errors are modest, as seen for the Business components. Conversely, it also emphasises how large equating errors towards the extremes of the score range can have relatively little impact, since there are often few candidates with such scores. To illustrate the impact of equating errors in even more concrete

terms, Table 4 reports differences between the cumulative percentages of Test Y candidates achieving the actual and equated cut scores for key grades. This shows how each method would affect the proportion of candidates reaching the key grades, if candidates were re-graded based on equated score grade boundaries. Table 4 confirms that the equated grade boundaries were exactly equal to the actual grade boundaries for a number of these grades, with consequently no difference between equated and actual pass rates (indicated in the table by "-"). These grade boundaries correspond to those parts of the cumulative distributions shown in Figure 3 where the difference between equated and actual cumulative percentages was zero. The largest differences shown in Table 4 are for the grade C boundary in ESL 2, but differences of 5–10 percentage points are seen in multiple component pairs, and for multiple equating methods.

Table 4: Differences (in percentage points) between pass rates at equated and actual grade boundaries.

Subject	Grade	Single comp (FE)	Single comp (chained)	ISAWG (FE)	ISAWG (chained)	Rasch
Business	A	5.05	10.25	5.05	7.93	5.05
	C	6.21	6.21	4.38	6.21	6.21
Computing	A	3.29	6.87	6.87	6.87	6.87
	C	1.28	1.28	1.28	1.28	-2.56
ESL 1	A	3.52	7.17	-	-	-
	C	3.92	3.92	-	-	-
ESL 2	A	-11.64	-6.68	-11.64	-11.64	-9.21
	C	-27.31	-17.55	-22.27	-13.08	-17.55
Economics	A	-4.49	-2.44	4.05	5.92	4.05
	C	-	2.22	2.22	3.26	3.26
History	A	-	-	3.52	-	-
	C	-	-2.99	-	-2.99	2.66
Maths 1	A	-	-	-2.54	-2.54	-
	C	0.77	0.77	-1.42	-0.75	-1.42
Maths 2	C	3.24	-3.47	3.24	1.63	-
	E	1.17	-3.81	1.17	-	-

Why was the performance of statistical equating so poor in one instance?

As can be seen from the previous sections, the worst equating performance was for ESL 2. As such, it is worth illustrating exactly why statistical equating has not worked in this instance. For simplicity, we will focus upon equating using a single co-component.

ESL 2 was a writing composition task that was taken as part of qualifications assessing English as a Second Language. Note that the group that took the Test X version were all located in one country whereas the group that took Test Y were mainly in another (with a minority scattered across several others). The selected single co-component was a reading comprehension test (also in common across the two assessments). Table 5 shows the performance on the components being equated and this main co-component. As can be seen, while the Test Y group were 0.6 of a standard deviation ahead in reading, they were more than one

entire standard deviation ahead in writing. This mismatch between the two skills resulted in the equating error shown in Table 3 and consequent impact on pass rates shown in Table 4.

This same issue is shown visually in Figure 4. The figure shows the relationship between anchor test scores and scores on the tests being equated in each group. In order to allow the patterns to be seen more easily, the figure is restricted to a random sample of 200 candidates in each group (rather than overloading the chart with almost 30 000 points). Figure 4 shows that for the same performance on the selected anchor test, candidates in the Test Y group tended to perform much better in writing. A chart like this could lead to the misleading impression that Test Y is easier than Test X when in fact the two tests are identical.

This example serves to illustrate how, in the absence of an anchor test that actually measures the same construct as that being equated, no single statistical method can be guaranteed to perform well. The relative performances of two groups of students in a particular subject (e.g. Reading) may not reflect their relative abilities in another (e.g. Writing). As such, any method based upon co-components may occasionally give a misleading picture of the differences between groups.

Finally, it is worth noting that, although we have focused upon the use of a single co-component, neither Rasch analysis nor the ISAWG performed notably better. This demonstrates that accurate equating cannot be achieved simply by making use of more of the same kind of data, nor in altering the way in which analysis is done. At least in this instance, accuracy could only be improved if we had a better external link (ideally measuring writing ability) between the two groups of students.

Table 5: Descriptive statistics relating to equating ESL 2 component pair.

Components being equated	Total marks available	Test X group		Test Y group		Difference (in overall SDs)
		Mean	Standard deviation	Mean	Standard deviation	
Anchor test (Reading)	50	26.4	7.0	30.8	7.4	0.61
Tests being equated (Writing)	60	35.4	6.4	43.1	7.5	1.11

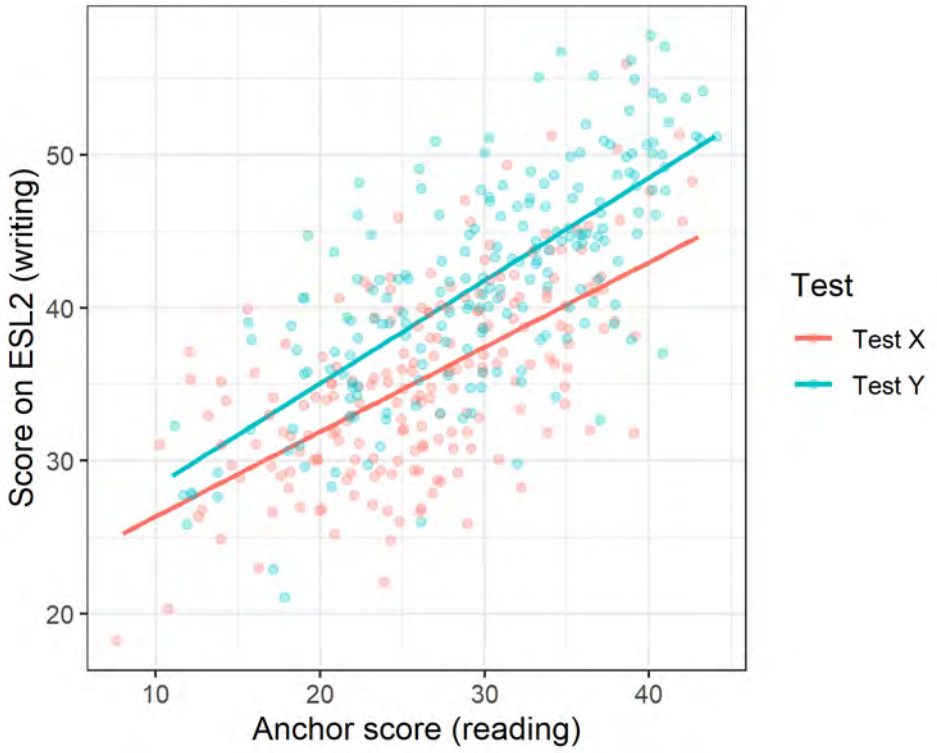

Figure 4: The relationship between performance on the anchor test (reading) and on the ESL 2 tests (writing) being equated for a sample of 200 students from each group. Regression lines are included for each group.

Conclusions

This article reported the results of equating various pairs of identical assessments. While some pairs were equated with very high accuracy by particular methods, the results showed that equating errors with real-world impact (e.g., an increase of 5–10 per cent in the proportion of students achieving a grade A) occurred even where equating conditions were apparently favourable: candidate groups were large, group differences were not extreme, and a very substantial amount of information on candidate performance in co-components was available. No single method consistently produced more accurate results than the others: the most accurate equating method varied by pair, and in fact all methods performed well for at least one component pair.

The results give further evidence that ISAWG and co-component equating methods can offer useful information towards maintaining standards. However, they also emphasise that multiple sources of information should still be considered, to make final boundary decisions.

More broadly, the results are a reminder that if applied uncritically, equating methods can lead to incorrect conclusions about the relative difficulty of assessments. In this equating exercise, Test X was not just written to the same specifications as Test Y, but was in fact identical to Test Y. However, equating between non-equivalent groups using operational data with non-random missingness as an anchor is difficult, even when we have extensive amounts of relevant information on candidates' abilities in other assessments. In the context of this study, the estimated equating relationships between pairs of identical assessments could have produced the paradoxical conclusions that assessments were both "easier" and "harder" in comparison with themselves.

References

Andersson, B., Branberg, K., & Wibery, M. (2013). Performing the Kernel Method of Test Equating with the Package kequate. *Journal of Statistical Software, 55*(6).

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

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

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

Coe, R. (2008). Comparability of GCSE examinations in different subjects: an application of the Rasch model. *Oxford Review of Education, 34*(5), 609–636. https://doi.org/10.1080/03054980801970312

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices* (3rd ed.). Springer.

Maris, G., Bechger, T., Koops, J., & Partchev, I. (2021). *Dexter: Data Management and Analysis of Tests*. https://CRAN.R-project.org/package=dexter

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. ETS Policy Information Center.

Newton, P. (2010). Thinking About Linking. *Measurement: Interdisciplinary Research & Perspective, 8*(1), 38–56. https://doi.org/10.1080/15366361003749068

Newton, P. E. (2012). Making sense of decades of debate on inter-subject comparability in England. *Assessment in Education: Principles, Policy & Practice, 19*(2), 251–273. https://doi.org/10.1080/0969594x.2011.563357

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology, 15*(2), 201–293.

Wiberg, M., & Branberg, K. (2015, Jul). Kernel Equating Under the Non-Equivalent Groups With Covariates Design. *Applied Psychological Measurement, 39*(5), 349–361. https://doi.org/10.1177/0146621614567939