Cambridge Assessment

Evaluating the 'similar items method' for standard maintaining

Conference Paper

Tom Bramley

Presented at the 19th annual AEA-Europe conference,

Arnhem/Nijmegen, the Netherlands,

November 2018

Author contact details:

Tom Bramley
Assessment Research and Development,
Research Division
Cambridge Assessment
The Triangle Building
Shaftesbury Road
Cambridge
CB2 8EA
UK

Bramley.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:

Bramley, T. (2018, November). *Evaluating the 'similar items method' for standard maintaining.* Paper presented at the 19th annual AEA-Europe conference, Arnhem/Nijmegen, the Netherlands.

Abstract

The aim of the research reported here was to get some idea of the accuracy of grade boundaries (cut-scores) obtained by applying the 'similar items method' described in Bramley & Wilson (2016). In this method experts identify items on the current version of a test that are sufficiently similar to items on previous versions for them to be treated as pseudo-anchor items. It could be useful in any international testing context using similar item types and under similar test development constraints (no pre-testing and no item re-use) to GCSEs and A levels in England.

Study 1 aimed to discover: i) the extent to which the equated grade boundary depends on *which* items are identified as similar; and ii) the extent to which it depends on *how many* items are identified as similar. Study 2 attempted a direct comparison with established methods for equating tests taken by non-equivalent groups. This was achieved by constructing a scenario in which *all* the similar items came from the *same* previous version (which is not the case in the intended application of the method). In this scenario the method can be directly compared with methods where common items form an internal anchor test.

Study 1 found that in the ideal case where the 'similar' items were in fact identical, roughly 20% of items or marks were enough to give a cut-score that was within 1 score point of the average (across different combinations of a fixed number of similar items). As expected, the fewer similar items, the greater the variability. There was a small amount of bias in the method – some inherently arising from using integer cut-scores on different versions of the test, some arising out of the equating method used to define equivalent cut-scores.

Study 2 found that the when the similar items method was applied to the scenario where all items came from the same previous test (a standard common-item equating scenario), it gave very similar outcomes to IRT true-score equating. However, when applied in the 'one item at a time' way intended for real scenarios where similar items might come from different tests it was vulnerable to distortions created by outlying items. This problem can be diagnosed by inspecting empirical item characteristic curves and equating functions implied by individual items. Increasing the smoothing of the empirical item characteristic curves improved the accuracy of the equating from the similar items method.

Introduction

The aim of the research reported here was to get some idea of the accuracy of grade boundaries obtained by applying the 'similar items method' (SI method[1]) described in Bramley & Wilson (2016).

The SI method has the following assumptions and requirements:

- The goal is to set a grade boundary (or boundaries) on a component of an examination for which item level data (ILD) for a reasonably large number of candidates is available.
- The same grade boundaries have been set on one or more previous versions of the same component, for which item level data (ILD) for a reasonably large number of candidates is also available.
- We are reasonably confident that the grade boundaries have been set in the correct place on the previous versions.

[1] In that work two methods were described: the current paper is based exclusively on the second method.

- Empirical item characteristic curves (EICCs) have been produced for every item on the current and previous components. These are (tabulations of) smoothed curves plotting the mean score on the item for each possible score on the component overall.
- Experts have identified one or more items on previous versions that they are willing to treat as identical to item(s) on the current version for the purposes of linking standards. These pseudo-anchor items are called 'similar items' in this paper. Note that the similar items do not all have to come from the same previous version.

The SI method works as follows:

1. For each similar item, find the expected item score corresponding[2] to the grade boundary on the previous version.
2. Find the score on the current (new) component corresponding to that score for each similar item.
3. Take the average of the component scores obtained in step 2 as the estimate of the grade boundary on the current component.

Study 1

Study 1 aimed to discover: i) the extent to which the equated grade boundary depends on *which* items are identified as similar; and ii) the extent to which it depends on *how many* items (or how many marks-worth of items) are identified as similar.

These questions were investigated by a resampling approach: repeatedly applying the SI method in circumstances where the 'correct' answer was known, and noting the bias and error of the method. Bias is quantified as the extent to which the average boundary across replications differs from the correct boundary, and error is quantified as the standard deviation (SD) of the boundary across replications.

Data

An initial dataset was prepared by merging the three components of an OCR A level Chemistry[3] examination taken in June 2017.
Paper 1 contained 47 items and was out of a total of 100 marks.
Paper 2 contained 48 items and was out of a total of 100 marks.
Paper 3 contained 27 items and was out of a total of 70 marks.
Data was retained for the 18807 candidates who took all three components and scored >0 on all of them.

Method

After converting item scores that were missing to zero, the 122 items were calibrated together on the full sample using the Rasch partial credit model (Masters, 1982) and the software RUMM2020 (Andrich et al, 2003). The Test Characteristic Curve[4] (TCC) for the full test was generated and used to find the ability estimates corresponding to the option-level grade A and E boundaries (198 and 63) set in June 2017.

[2] Corresponding' here means the point on the y axis of the EICC plot where a vertical line from the grade boundary on the x-axis intersects the smoothed EICC. Likewise in step 2 'corresponding' means the point on the x-axis of the EICC plot where a horizontal line from the expected item score (identified in step 1) on the y-axis intersects the smoothed EICC.
[3] Specification code H432. Question papers and mark schemes are available from https://www.ocr.org.uk/qualifications/past-papers/
[4] This is a tabulation or plot of expected test score against ability.

For the purpose of this study, the items from Paper 3 were treated as the 'current' test with an unknown boundary that needed to be set. The TCC for this set of items was generated and the raw scores corresponding to the ability estimates obtained above were found to be 47.13 and 11.06 respectively. (Note that these differ from the actual Paper 3 boundaries of 46 and 13 used in June 2017, which is not too surprising since the components are not operationally aligned by Rasch equating, but by other means).

The 122 items were then split randomly into four mutually exclusive groups, to simulate four 'previous versions' of tests containing potential similar items. All of these groups contained some items from Paper 3.
The full cohort was split into five dummy cohorts, but not at random. This is because the SI method would not be necessary if we were prepared to assume that successive cohorts were randomly equivalent – we could just use equipercentile equating (for example) to set boundaries. Instead, five approximately equally-sized groups were formed by sorting the data by centre number, then selecting the first n1 centres needed to obtain a sample of 18,807/5 ≈ 3,760, then the next n2 centres needed, and so on.

Tabulations were prepared of the EICCs of the first four of these dummy cohorts on the four previous versions, and of the fifth dummy cohort on the current version (the Paper 3 items) [5]. Nominal A and E grade boundaries on the four previous versions were derived by calculating the 'correct' boundaries as above using the TCCs for these four tests. However, as previously, this naturally generated non-integer scores. Since using the SI method in practice would entail using integer grade boundaries, these were rounded to the nearest whole number, as shown in Table 1 below.

Table 1. Descriptive statistics, un-rounded and rounded grade boundaries on the four previous versions (PV) and the Paper 3 items, calculated via Rasch model TCCs.

Test	# Items	Out of	N	Mean	SD	A	E	A	E
PV 1	30	66	3755	39.77	13.15	48.40	14.47	48	14
PV 2	30	63	3756	39.94	11.50	46.62	17.85	47	18
PV 3	31	71	3698	44.57	12.91	54.61	17.33	55	17
PV 4	31	70	3819	40.74	13.61	48.37	13.34	48	13
Paper 3	27	70	3779	39.20	13.94	47.13	11.06	47	11

It is important to note that the unrounded boundaries on the four previous versions sum to the correct total boundaries of 198 and 63, whereas the rounded boundaries sum to 198 and 62. This is because the E boundary was rounded down on three of the four previous versions. This shows that the SI method can of necessity introduce its own biases, independent of the number of similar items that are identified. The extent to which it will have an impact depends on the relative number of similar items coming from rounded-down or rounded-up previous versions.

The SI method was then applied to derive a grade boundary on Paper 3, varying the number of similar items used in the equating from 1 to 27. For 1, 2, 25, 26 or 27 similar items, all the possible combinations were considered. For the rest, a random sample of ≈1000 was taken to avoid running out of computer memory. Figure 1 and Table 2 below show how the Paper 3 boundaries varied with number of similar items.

[5] Some example SAS code for creating a smooth EICC is given in the appendix.

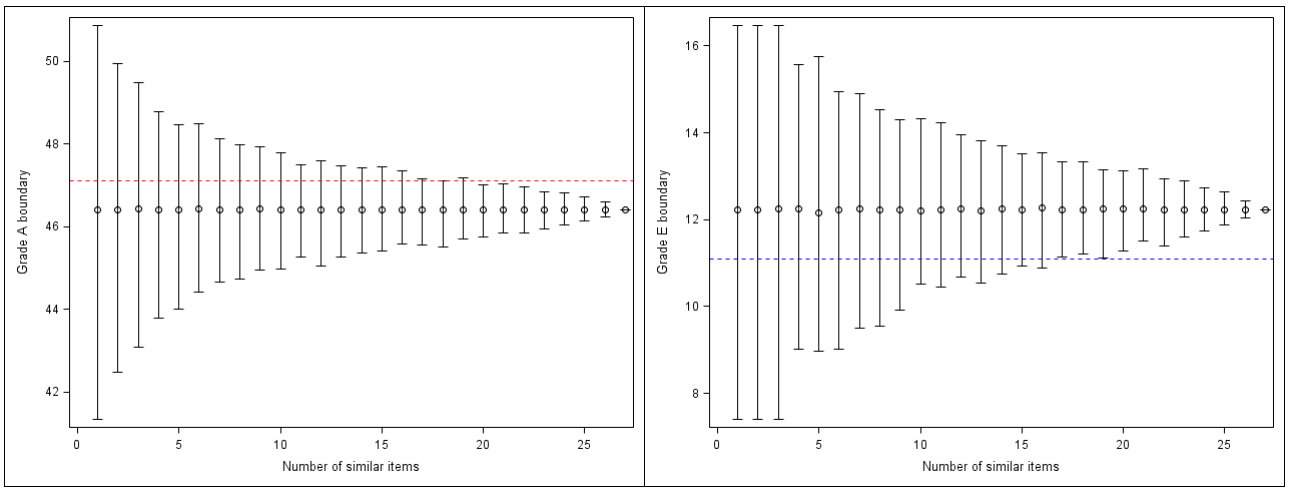

Figure 1. Distribution of (unrounded) grade boundaries on Paper 3 for each number of similar items used in the equating. Left – grade A, Right – grade E. Key: Black circles are the mean, bars show min and max. Red/blue dashed lines are the correct values of 47.1 and 11.1.

Table 2. Descriptive statistics for distribution of (unrounded) grade boundaries on Paper 3 for each number of similar items used in the equating.

# Similar items	Grade	# Combinations	Mean	SD	Min	Max
1	A	27	46.41	1.92	41.34	50.86
	E	25[6]	12.23	2.53	7.40	16.47
2	A	351	46.41	1.31	42.48	49.93
	E	350	12.23	1.84	7.40	16.47
3	A	992	46.43	1.06	43.08	49.47
	E	992	12.24	1.42	7.40	16.47
4	A	1027	46.40	0.86	43.80	48.77
	E	1027	12.25	1.24	9.01	15.57
5	A	963	46.41	0.79	44.02	48.45
	E	963	12.15	1.10	8.97	15.75
6	A	1002	46.43	0.68	44.42	48.48
	E	1002	12.23	0.95	9.01	14.95
7	A	995	46.40	0.65	44.66	48.11
	E	995	12.24	0.88	9.50	14.90
8	A	994	46.40	0.57	44.74	47.98
	E	994	12.23	0.79	9.55	14.54
9	A	1006	46.43	0.51	44.94	47.94
	E	1006	12.22	0.72	9.92	14.30
10	A	951	46.40	0.49	44.96	47.79
	E	951	12.21	0.65	10.53	14.33
11	A	1003	46.40	0.44	45.27	47.49
	E	1003	12.22	0.60	10.44	14.23
12	A	962	46.41	0.41	45.04	47.60
	E	962	12.24	0.56	10.69	13.96
13	A	980	46.40	0.37	45.27	47.46
	E	980	12.21	0.52	10.54	13.81
14	A	1038	46.40	0.36	45.37	47.42

[6] On two of the items at grade E the smoothed EICC did not intersect with the required value for interpolation.

# Similar items	Grade	# Combinations	Mean	SD	Min	Max
	E	1038	12.25	0.51	10.75	13.70
15	A	958	46.40	0.33	45.42	47.44
	E	958	12.22	0.44	10.93	13.52
16	A	1014	46.41	0.31	45.57	47.35
	E	1014	12.26	0.41	10.89	13.55
17	A	956	46.41	0.28	45.56	47.16
	E	956	12.23	0.38	11.13	13.33
18	A	991	46.40	0.27	45.51	47.11
	E	991	12.22	0.36	11.21	13.32
19	A	955	46.40	0.24	45.70	47.19
	E	955	12.24	0.33	11.12	13.15
20	A	1049	46.41	0.22	45.74	47.00
	E	1049	12.24	0.30	11.28	13.13
21	A	947	46.40	0.20	45.84	47.03
	E	947	12.24	0.27	11.52	13.16
22	A	1024	46.41	0.18	45.86	46.97
	E	1024	12.22	0.24	11.40	12.95
23	A	954	46.41	0.15	45.94	46.85
	E	954	12.23	0.21	11.61	12.89
24	A	1011	46.41	0.13	46.03	46.81
	E	1011	12.24	0.18	11.74	12.73
25	A	351	46.41	0.10	46.13	46.72
	E	351	12.23	0.14	11.89	12.65
26	A	27	46.41	0.07	46.24	46.60
	E	27	12.23	0.10	12.05	12.43
27	A	1	46.41	.	46.41	46.41
	E	1	12.23	.	12.23	12.23

The figures and table show that, as might be expected, the error (variability) decreases as the number of similar items increases, but that the bias (systematic error) remains constant.

The SI method as described in the introduction takes a simple (unweighted) average of the boundaries implied by each similar item to arrive at a final estimated boundary. If the items differ in the number of marks they are out of, it might be considered more reasonable to use a weighted average – i.e. to give more weight to similar items worth more marks. The equivalent information to Table 2 but using a weighted average is given in Table A1 in the appendix. The bias was almost identical; the error was slightly lower at grade A and slightly higher at grade E.

An alternative way of accounting for the variability in mark tariff is to display the results by number of similar marks (rather than number of similar items). This is shown in Figure 2 below and Table A2 in the appendix.

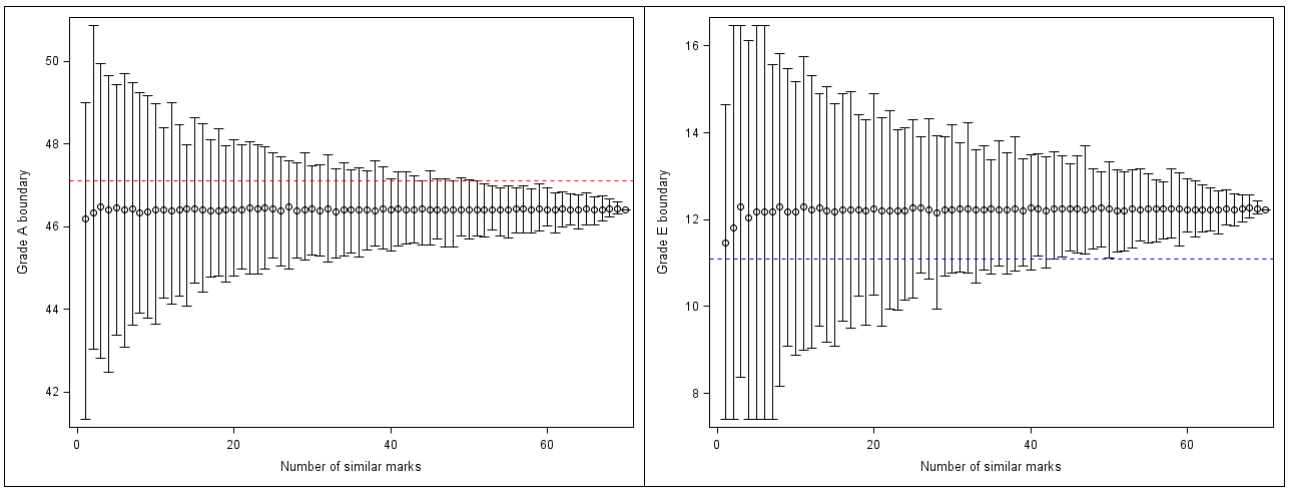

Figure 2. Distribution of (unrounded) grade boundaries on Paper 3 for each number of similar marks used in the equating. Left – grade A, Right – grade E. Key: Black circles are the mean, bars show min and max. Red/blue dashed lines are the 'correct' values of 47.1 and 11.1.

Considering first the practical impact of the error on the grade boundaries (and for the moment ignoring the bias), the table below shows the proportion of replicates where the rounded boundary was within ±1 mark of the rounded mean (i.e. 46 at A and 12 at E).

Table 3: Percentage of replicates where the rounded A or E boundaries were within 1 mark of the rounded mean A or E boundaries of 46 and 127.

# Similar items	% within 1 (grade A)	% within 1 (grade E)
1	59.26	44.00
2	74.07	56.29
3	80.44	68.25
4	89.19	74.59
5	90.86	82.87
6	95.61	87.63
7	96.28	90.96
8	97.69	92.66
9	98.61	95.43
10	99.05	97.37
11	100	98.50
12	100	99.17
13	100	98.98
14	100	99.52
15	100	99.90
16	100	99.90
17	100	100

Table 3 shows that there is 'diminishing returns' to increasing the number of similar items: with 4/5 similar items the A boundary was within 1 mark of the mean 90% of the time, and with 7 similar items the E boundary was. In terms of number of similar marks, a randomly chosen 5 items on paper 3 were worth 13.1 marks on average (SD=3.1), and a randomly chosen 7 items were worth 18.3 marks (SD=3.5). This suggests a ballpark value of 20% for either percentage of similar items

[7] Note that in Table 3 the boundaries are treated independently – they do not show the proportion of times both A and E were simultaneously within 1 mark of the rounded mean.

or percentage of similar marks[8] could be used in practice as a basis for being reasonably confident of getting to within a mark of the boundary that would be obtained by applying the SI method if every item had a similar (identical) counterpart on previous tests.

Explaining the bias

The long-run average of the SI method gave values of 46.4 and 12.2 for the A and E boundaries, whereas the correct values according to the Rasch true-score equating were 47.1 and 11.1. One potential contributor to the bias was mentioned above – the fact that the grade boundaries on the previous versions were taken as whole numbers (which they would be in practice). However, any bias arising from this should have roughly cancelled out at grade A, and at grade E should have worked in the *opposite* direction to that observed, since it would have led to lower values for expected item scores on the three out of four tests that were rounded down, and hence a lower value for the grade E boundary than the correct value.

Another possibility is that the bias arose from using the Rasch model to derive the expected scores on the previous versions and the target test (Paper 3). Misfit to the Rasch model could have introduced inaccuracies here. To investigate this, interpolations of smoothed EICCs for the full dataset, using the overall A and E boundaries of 198 and 63, were used to generate the expected scores instead. (This is effectively using a non-parametric IRT approach to calibrate the items.) Table 4 shows the 'correct' boundaries on the previous versions and Paper 3 items by this approach compared to the previous Rasch-based approach.

Table 4. Un-rounded and rounded grade boundaries on the four previous versions and Paper 3 items using EICCs from the full cohort on all the items. (PV=previous version).

		Rasch		Rasch rounded		EICCs		EICCs rounded	
Test	# Items	A	E	A	E	A	E	A	E
PV 1	30	48.40	14.47	48	14	48.65	15.02	49	15
PV 2	30	46.62	17.85	47	18	47.34	16.80	47	17
PV 3	31	54.61	17.33	55	17	53.53	18.84	54	19
PV 4	31	48.37	13.34	48	13	48.48	12.33	48	12
Paper 3	27	47.12	11.06	47	11	46.33	12.93	46	13

Table 4 shows that the different approaches did indeed imply different grade boundaries on the various sub-sets of items. Further investigation revealed that this was because there was a slight tendency for the higher tariff items to misfit the Rasch model, as shown in the Bland-Altman[9] plots in Figure 3 below. At grade A the Rasch model predicted a higher score than observed and vice versa at grade E on these higher tariff items, particularly on Paper 3.

[8] On components of this length. But 70 marks is a reasonably typical value for a GCSE or A level written component.
[9] These plot the difference between two measurements against their average (mean), as described in Altman & Bland (1983).

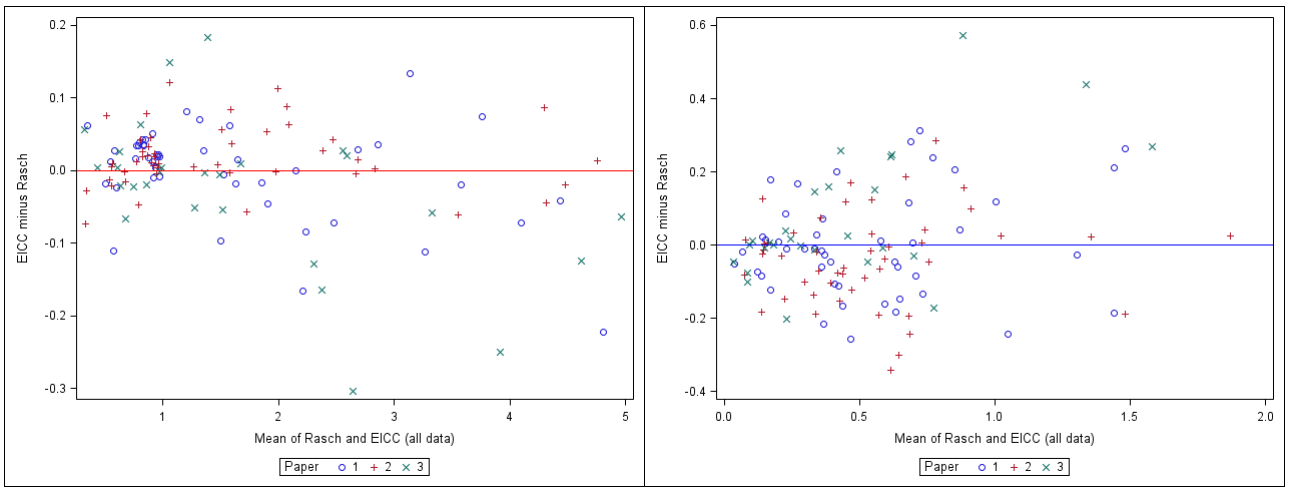

Figure 3: Plot of difference between EICC and Rasch item expected scores against mean. Left – grade A, right – grade E.

Repeating the previous analysis using the values for the grade boundaries based on the EICC values in Table 4 gave the result in Figure 4 below.

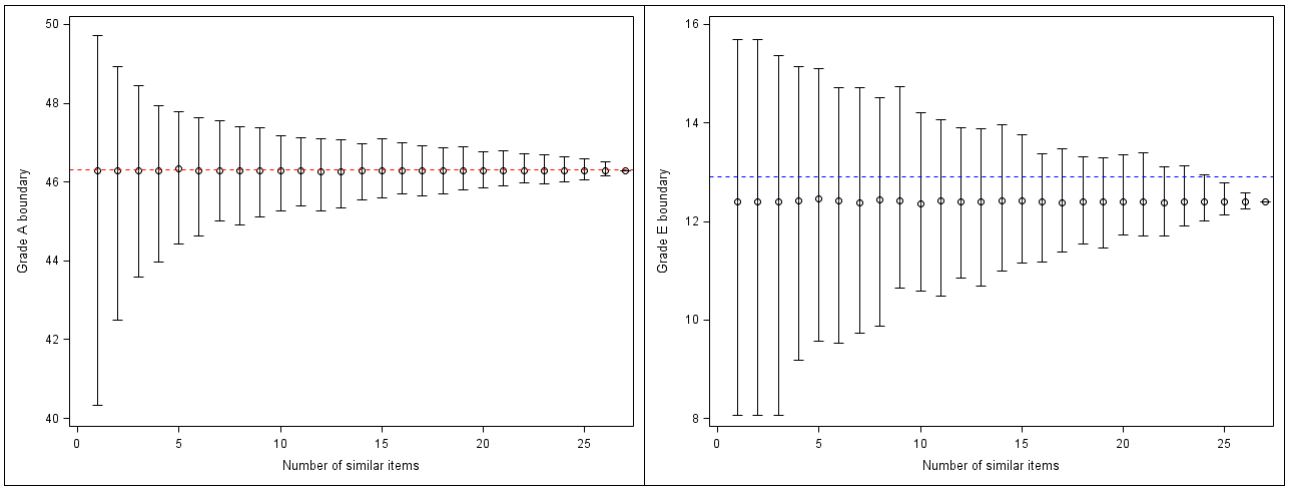

Figure 4. Distribution of (unrounded) grade boundaries on Paper 3 for each number of similar items used in the equating. Left – grade A, Right – grade E. Key: Black circles are the mean, bars show min and max. Red/blue dashed lines are the new 'correct' values of 46.3 and 12.9.

Figure 4 shows that the bias was almost entirely removed at grade A, and reduced at grade E. The standard deviations of the replications at each number of similar items were also slightly reduced in both cases, for example with 5 similar items from 0.79 to 0.61 at A, and from 1.10 to 0.99 at E. (Table A3 in the appendix gives the data underlying Figure 4, analogous to Table 2 above). The effect on the proportion of replicates within ±1 mark of the rounded mean (cf Table 3) is shown in Table 5.

Table 5: Percentage of replicates where the rounded A or E boundaries were within 1 mark of the rounded mean A or E boundaries of 46 and 13.

# Similar items	% within 1 (grade A)	% within 1 (grade E)
1	81.48	40.00
2	83.76	58.00
3	88.87	69.04
4	94.50	76.49
5	99.08	81.55
6	99.52	85.12
7	99.69	85.57
8	100	89.36
9	100	91.62
10	100	91.55
11	100	95.82
12	100	95.60
13	100	96.55
14	100	97.32
15	100	98.20
16	100	99.10
17	100	99.59
18	100	100

The number of similar items needed to ensure 90% of replicates within ±1 mark of the rounded mean was reduced from 5 to 4 at grade A, but increased from 7 to 9 at grade E. This suggests that the ballpark figure of 5 similar items or 20% similar marks suggested previously is reasonable, especially when there is plenty of data around the boundary on the previous versions (as at grade A here).

Study 2

The aim of study 2 was to compare the SI method with established equating methods, in order to see whether its results were broadly similar. As in study 1, the Paper 3 items were taken as the test on which equated scores were needed ('Test Y' in standard equating terminology) and the items from the artificially created test PV1 (henceforth Test 1) in study 1 were used as the test being equated to ('Test X' in standard equating terminology). There were 30 items on Test 1, worth 66 marks in total, and 27 items on Paper 3 worth 70 marks in total (see Table 1). There were 6 items in common between the two tests, worth 12 marks in total - around 17-18% of the total marks.

The 'criterion' equating for comparison purposes was taken as the equipercentile equating (unsmoothed, with linear interpolation) of Paper 3 to Test 1 using the entire cohort and a single group design. Thus the criterion equating was based on scores from 18,807 candidates who took both tests[10]. A second criterion equating ('criterion2') was taken also using a single group design but restricting to the 7,534 candidates who had been nominally assigned to Test 1 or Paper 3 (cohorts 1 and 5 in study 1). Table 6 below shows that these cohorts were of slightly different ability based on their scores on the common items.

[10] Of course, these were not actually taken as separate test forms.

Table 6: Scores of the two cohorts to be equated on the tests and the common items.

		Test score		Common item score	
Cohort	N	Mean	SD	Mean	SD
Test 1	3755	39.77	13.15	8.25	2.60
Paper 3	3779	39.20	13.94	8.61	2.54

The dataset used for equating contained the scores of these two cohorts on the Test 1 and Paper 3 items respectively. The following standard equating methods were used (see Kolen & Brennan (2004) for descriptions of these methods):

- Chained equipercentile equating using the common items as an internal anchor test;
- Frequency estimation equipercentile equating using the common items as an internal anchor test;
- IRT true score equating based on separate calibration of the items in the two tests using the Graded Response IRT model and the Stocking-Lord method of aligning the score scales;
- IRT observed score equating based on separate calibration of the items in the two tests using the Graded Response IRT model and the Stocking-Lord method of aligning the score scales;
- Rasch true score equating based on separate calibration of the items in the two tests using the Rasch Partial Credit model and the Stocking-Lord method with a fixed slope to align the score scales;
- Rasch observed score equating based on separate calibration of the items in the two tests using the Rasch Partial Credit model and the Stocking-Lord method with a fixed slope to align the score scales[11].

In addition to these standard methods, three variants of the SI method were used:

- The SI method but combining (i.e. summing) the EICCs of the common items to a single common item[12] worth 12 marks;
- The SI method without weighting;
- The SI method weighting the equated scores corresponding to each common item by the item maximum mark.

Results of study 2

In the graphs below, the equating methods are compared with each other by plotting the difference between the equated Paper 3 score from each method and the criterion equate against the Test 1 total score. To avoid cluttering the plots the IRT and Rasch observed score methods and the frequency estimation equipercentile method are not included. It can be seen from Figure A1 in the appendix that the IRT and Rasch observed score methods were both similar to their corresponding true score method (the similarity was greater for Rasch than IRT). Figure A2 in the appendix shows that the frequency estimation method gave similar results to the chained equipercentile method, being slightly closer to the criterion at the lower end of the score scale and further away at the higher end.

[11] The IRT and Rasch methods used the R packages *mirt* (Chalmers, 2012) for item calibration and code written internally for the Stocking-Lord equating. Note that the *mirt* package uses marginal maximum likelihood (MML) for estimation whereas RUMM uses conditional maximum likelihood (CML).

[12] This is not possible in the intended application of the SI method, where the similar items will have come from different previous versions, but since it was possible here and corresponds more closely in spirit to the IRT/Rasch methods it was included.

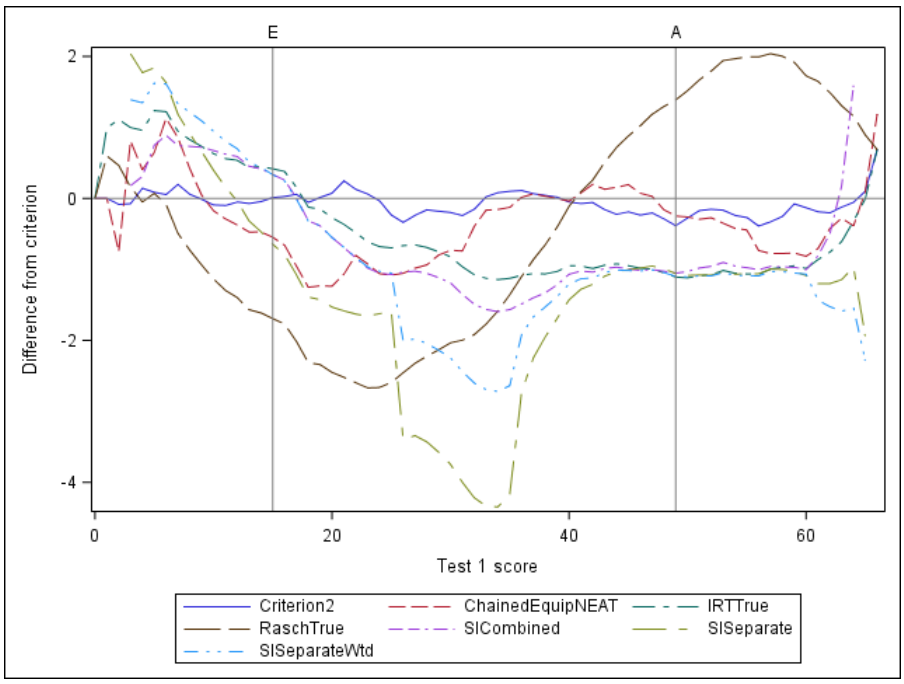

Figure 5: Comparison of different equating methods. (Vertical lines indicate approximate integer boundaries on Test 1 (see Table 4).

There are several noteworthy features of Figure 5. First, the blue line for Criterion2 shows that the 'missing' ~20k candidates from cohorts 2,3, and 4 did not affect the results – the Criterion2 results were very close to the criterion at all Test 1 scores. Second, the Rasch method yielded equated scores that were consistently lower than the others for low Test 1 scores (<20) and consistently higher than the others for high Test 1 scores (>40). Third, the SI method using the combined anchor item (test) gave very similar results to the IRT true score method, though slightly further from the criterion for most of the score range. Most striking however, is the dip in the graph in the middle of the score range for the two SI results based on average equatings from each item separately, the effect being much more pronounced for the unweighted average.

Further investigation showed that this anomaly was caused by a single common item 'com1' worth only 1-mark. The EICCs for this item are shown in Figure 6.

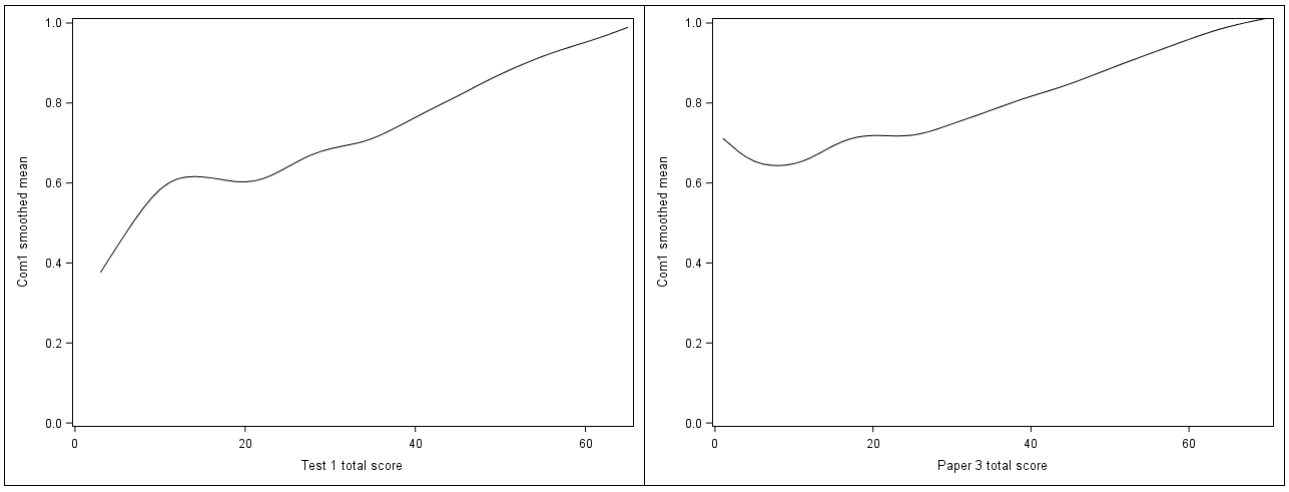

Figure 6. EICCs for common item 1(com1) in Test 1 (left) and Paper 3 (right).

This was an easy item on both tests with a flat (or even negatively sloping) EICC in the lower part of the score range. The EICC of this item on Test 1 had a negative slope in the range 11-20, but in fact this did not matter too much because the corresponding item score on Paper 3 did not intersect the EICC at all so the application of the method just yielded missing data here. The problem was mostly caused by the fact that when the item scores from Test 1 did begin to intersect the EICC on Paper 3, the equated scores were much lower than those implied by the other common items (see Figure 7). The flatness of the Paper 3 EICC in the score range 19-25 meant that a small increase in total score on Test 1 at around a score of 37 led to a big jump in the equated score on Paper 3, as shown in Figure 7. It was at this point that the results from the SI method started to move back towards those of the other methods (Figure 5).

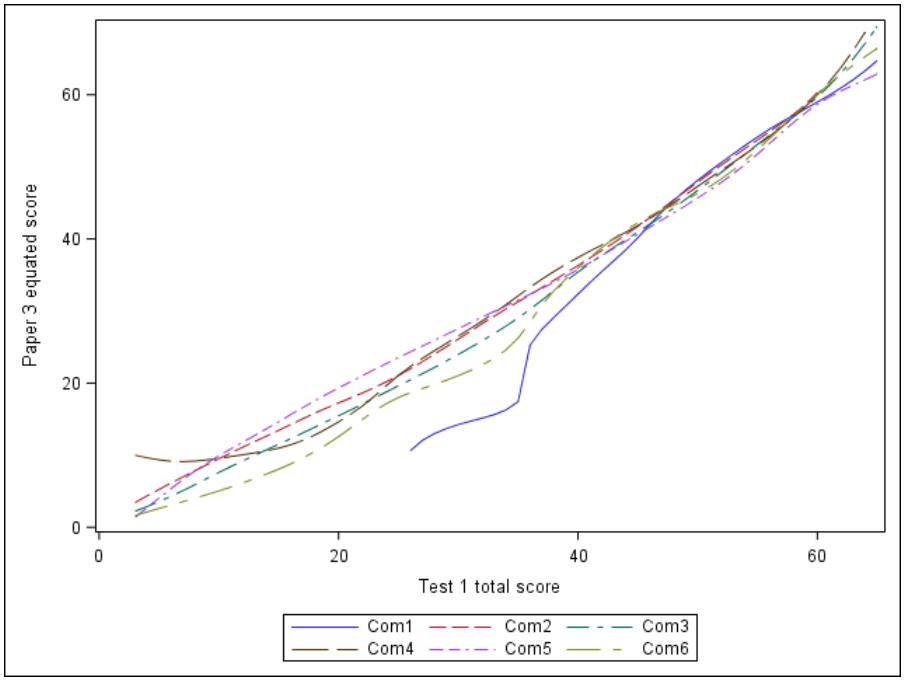

Figure 7. Equating relationship based on each common item using the SI method.

Figure 8 shows the results of repeating the comparison of equating methods, but not including com1 as one of the common items for the item-based methods. There was a slight improvement for the IRT and Rasch methods (see Figure A1 in the appendix), but a big improvement for the similar item methods using the items separately (both weighted and unweighted).

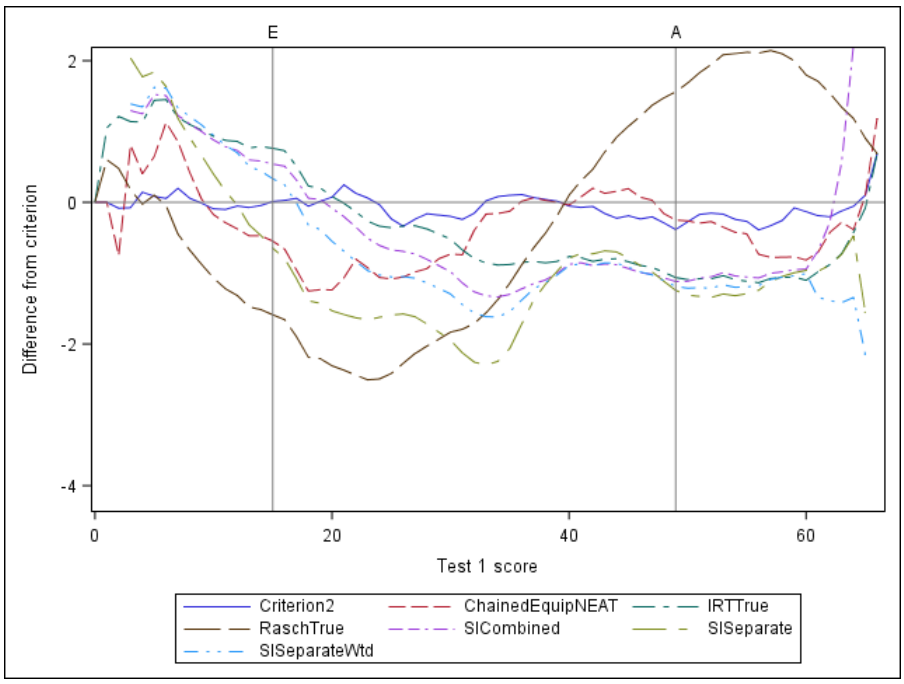

Figure 8. Comparison of different equating methods, excluding com1 from IRT, Rasch and SI methods.

The above results suggested two further questions:

- Could the similar items method (as a common-item equating method) be improved by increasing the smoothing of the EICCs?
- Could a standard anchor test equating method be made to work item-by-item?

Effect of increasing the smoothing

The EICCs for Test 1 and Paper 3 were re-generated, increasing the smoothing parameter in PROC TRANSREG from 50 to 70, which ensured that all EICCs were monotonically increasing. This did not alter the fact that 'com1' was an outlier in the sense of giving a different equating result to the other pseudo anchor items, as shown in Figure 9. However, Figure 10 (cf Figure 8) shows that the overall equating with this item excluded was considerably improved for all three variants of the similar items method, with the 'combined' and 'weighted' methods now appearing better than all other methods except criterion2.

Equating item-by-item with a standard equating method

The frequency estimation equipercentile method was applied item by item. A plot of the equating functions implied by each item separately did not identify 'com1' as an outlier (see Figure 11), so this item was not excluded when calculating the unweighted and weighted averages of the equating functions. Since this method essentially modifies the results of random groups equipercentile equating to account for different score distributions on an anchor test/item, it was of interest to compare the outcomes with those from random groups equipercentile equating – i.e. not using any common items and simply assuming that the Test 1 and Paper 3 groups were of equal ability. Figure 12 shows that all three variations of equipercentile equating that made use of information about the common items were an improvement on the random groups equating, but that doing it item-by-item was noticeably worse than the usual way of combining the items into a single anchor test. Weighting by maximum mark had no noticeable effect (unlike the SI method).

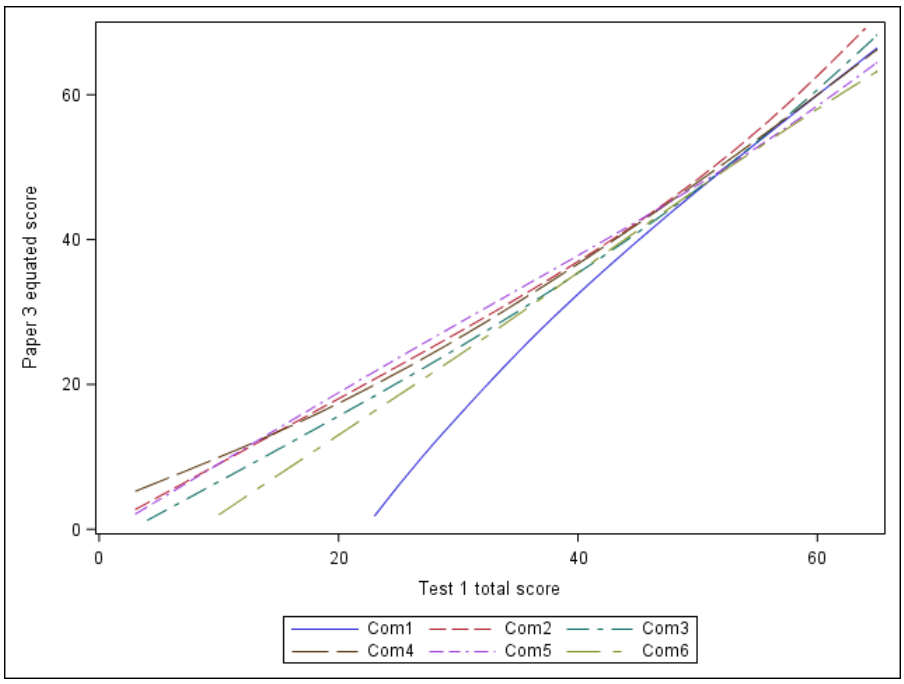

Figure 9. Equating relationship based on each common item using the SI method with smoother EICCs.

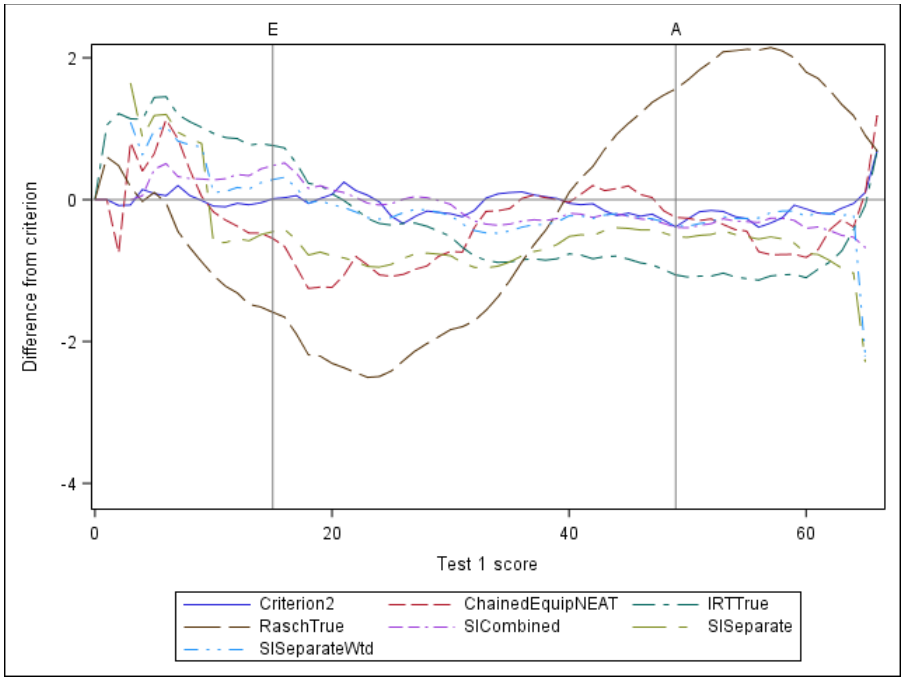

Figure 10. Comparison of different equating methods, (excluding com1 from IRT, Rasch and SI methods) and using smoother EICCs.

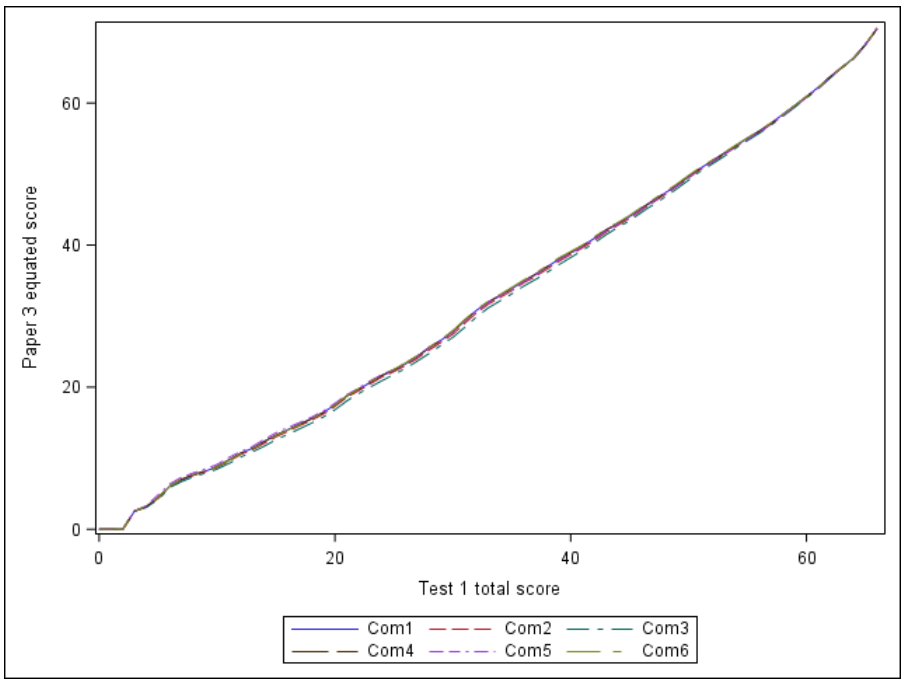

Figure 11. Equating relationship based on each common item using the frequency estimation equipercentile equating method.

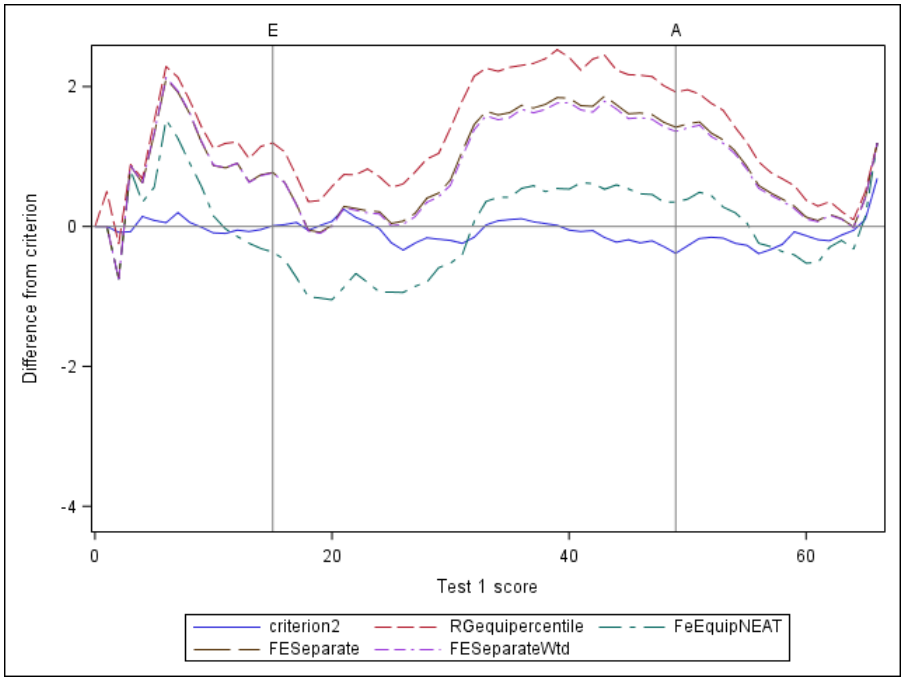

Figure 12. Comparison of different equipercentile equating methods.

For completeness, Tables 7 and 8 show the results from all the different equating methods that were tried in this study. Table 7 shows the average absolute difference across the score range between the equating method and the correct value (as defined by the 'criterion' equate). The left columns show the raw average, and the right columns show this average weighted by the score distribution on Test 1 (i.e. giving more weight to the parts of the score range where more people had scored on Test 1). Table 8 focuses on the two grade boundaries – the left columns show grade A and the right show grade E. In both tables the equating methods are ordered from best to worst.

Table 7 shows that the unweighted SI method item by item with no items excluded was the worst method, but that excluding the outlying item and increasing the smoothing made a big improvement, to the extent that, even item by item, the SI method with a weighted average was better than all the standard equating methods (both item- and test-based), being on average less than half a score point away from the correct result across the score range. The relatively low position of the random groups equipercentile equate and the Rasch based methods in both Table 7 and 8 shows that, in this particular dataset at least, common item equating was necessary and that the Rasch model was not able to capture all the variability in item response functions.

Table 7: Average absolute difference (across the score range) of each method from the criterion.

Unweighted		**Weighted by Test 1 score distribution**	
Method	**AbsDiff**	**Method**	**AbsDiff**
Criterion2	0.141	Criterion2	0.165
SICombined_Smooth_NoCom1	0.268	SICombined_Smooth_NoCom1	0.257
SISeparateWtd_Smooth_NoCom1	0.313	SISeparateWtd_Smooth_NoCom1	0.277
SICombined_Smooth	0.360	SICombined_Smooth	0.387
ChainedEquipNEAT	0.518	ChainedEquipNEAT	0.426
FeEquipNEAT	0.537	FeEquipNEAT	0.500
SISeparateWtd_Smooth	0.604	SISeparateWtd_Smooth	0.636
SISeparate_Smooth_NoCom1	0.713	SISeparate_Smooth_NoCom1	0.649
IRTObs_NoCom1	0.769	IRTTrue_NoCom1	0.803
IRTTrue_NoCom1	0.797	IRTObs_NoCom1	0.839
IRTObs	0.818	IRTTrue	0.910
IRTTrue	0.821	SICombined_NoCom1	0.938
SICombined_NoCom1	0.900	IRTObs	0.963
SICombined	0.902	SICombined	1.040
FESeparateWtd	0.919	FESeparateWtd	1.074
FESeparate	0.954	SISeparateWtd_NoCom1	1.097
SISeparateWtd_NoCom1	1.054	FESeparate	1.125
SISeparate_NoCom1	1.229	SISeparate_NoCom1	1.294
SISeparate_Smooth	1.249	SISeparate_Smooth	1.313
SISeparateWtd	1.258	SISeparateWtd	1.368
RGequipercentile	1.381	RaschObs_NoCom1	1.400
RaschObs_NoCom1	1.447	RaschObs	1.402
RaschTrue_NoCom1	1.449	RaschTrue_NoCom1	1.483
RaschTrue	1.477	RaschTrue	1.486
RaschObs	1.478	RGequipercentile	1.627
SISeparate	1.623	SISeparate	1.816

Table 8: Absolute difference (at the grade boundaries) of each method from the criterion.

Grade A		Grade E	
Method	**AbsDiff**	**Method**	**AbsDiff**
ChainedEquipNEAT	0.249	SICombined_Smooth	0.011
FeEquipNEAT	0.349	Criterion2	0.012
SISeparateWtd_Smooth_NoCom1	0.372	IRTObs	0.036
Criterion2	0.384	SISeparateWtd_Smooth	0.282
SICombined_Smooth_NoCom1	0.391	SISeparateWtd_Smooth_NoCom1	0.282*
SICombined_Smooth	0.451	SISeparateWtd	0.334
SISeparateWtd_Smooth	0.466	SISeparateWtd_NoCom1	0.334
SISeparate_Smooth_NoCom1	0.524	SICombined	0.336
SISeparate_Smooth	0.686	IRTObs_NoCom1	0.360
IRTObs_NoCom1	1.042	FeEquipNEAT	0.364
SICombined	1.052	IRTTrue	0.417
SISeparate	1.058	SISeparate_Smooth	0.452
IRTTrue_NoCom1	1.062	SISeparate_Smooth_NoCom1	0.452
SISeparateWtd	1.098	SICombined_Smooth_NoCom1	0.480
IRTObs	1.100	SICombined_NoCom1	0.539
IRTTrue	1.107	ChainedEquipNEAT	0.548
SICombined_NoCom1	1.118	SISeparate	0.647
SISeparateWtd_NoCom1	1.186	SISeparate_NoCom1	0.647
RaschObs	1.240	IRTTrue_NoCom1	0.763
SISeparate_NoCom1	1.243	FESeparateWtd	0.765
FESeparateWtd	1.360	FESeparate	0.773
RaschTrue	1.388	RGequipercentile	1.196
RaschObs_NoCom1	1.416	RaschTrue_NoCom1	1.588
FESeparate	1.419	RaschTrue	1.693
RaschTrue_NoCom1	1.561	RaschObs_NoCom1	1.813
RGequipercentile	1.922	RaschObs	1.928

*Note that at low scores the excluded common item did not contribute to the equating in the SI method so the result is the same.

Conclusions from study 2

The study has shown that in conditions where the similar items are in fact identical items, application of the SI method can give results that are broadly in line with other established equating methods. In the special case where all the similar items come from the same previous version of a test and can hence be treated as an internal anchor (the 'SIcombined' method), the results were very close to those obtained from IRT true score equating. This is not too surprising, given that the SI method essentially uses non-parametric IRT to produce its ICCs, and these curves are more likely to be similar to those from the more flexible Graded Response Model than from the more restrictive Partial Credit Model.

However, in its intended application, the similar items would come from different tests and so either the unweighted ('SIseparate') or the weighted ('SISeparateWtd') method would be needed. This study has shown that both these methods can be seriously affected by similar items whose EICCs are flat in part of their range, or that give a substantially different equating result to other similar items. This has two implications for use of the method in practice:

- It is important to evaluate the contribution of each individual similar item to the equating outcome;
- It is advisable to produce equating functions over the full range of scores (instead of just focussing on grade boundaries) because this will help to identify discontinuities and other anomalies. Inspection of plots like Figure 7 will help to identify outliers.

As might be expected, giving more weight to items worth more marks when calculating the average improved the accuracy. Less expected was the finding that increasing the smoothing parameter of the EICCs created a substantial improvement in the accuracy of equating (with this particular data). Further exploration could investigate whether there is an optimum degree of smoothing beyond which results start to deteriorate again, and whether the smoothing parameter needs to be set at the level of the item, the test, or the pair of tests to be equated. It is clearly desirable to have an a priori rationale for setting the smoothing parameter – such as the lowest value for which the EICCs are strictly increasing – because this reduces the amount of human input required and (perhaps) reduces the likelihood of capitalising on chance.

It is worth noting that the aim of this research was not to show that the SI method is better than other standard equating methods, merely that it can 'take its place at the table' as a legitimate means of equating two tests. It would doubtless have been possible to make tweaks and modifications that would have increased the accuracy of some of the other standard equating methods that were explored in this study.

Finally it is worth re-iterating that in spirit the SI method is essentially a judgemental method based on judgements of similarity. It was intended to be used not in situations where there are two tests to be equated (although this research has shown that it can work well here too), but rather in situations where there is a single test on which cut-scores are required that maintain the standards set on several previous versions of the same test. Approximate accuracy and usefulness are all that can reasonably be aimed for when the items are not actually identical.

References

Altman, D. G., & Bland, J. M. (1983). Measurement in medicine: the analysis of method comparison studies. *The Statistician, 32*, 307-317.

Andrich, D., Sheridan, B., & Luo, G. (1997). RUMM2020: Rasch unidimensional measurement models. Computer program. Perth, Australia: RUMM Laboratory Pty.

Bramley, T., & Wilson, F. (2016). Maintaining test standards by expert judgement of item difficulty. *Research Matters: A Cambridge Assessment Publication, 21*, 48-54.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 29. doi:10.18637/jss.v048.i06

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Appendix

Example SAS code[13] used to create a smooth EICC for a single item is given below. Test1 is a SAS dataset containing one row per person and (at least) variables 'test_tot' for test total score and 'item1' for scores on item 1. 'Unsmoothed' is an intermediate SAS dataset containing the raw item mean scores for every observed test score and 'Smoothed' is the output SAS dataset containing the smoothed item mean scores.

```
*Find the item mean score for each possible test score;
proc means data=test1;
 class test_tot;
 types test_tot;
 var item1;
 output out=unsmoothed mean=;
run;

%let smoothpar=50;  *smoothing parameter between 0 and 100;
proc transreg data=unsmoothed noprint;
 model identity(item1) = smooth (test_tot / sm=&smoothpar);
 output out=smoothed;
run;
```

Table A1. Descriptive statistics for distribution of (unrounded) grade boundaries (weighted by max mark) on Paper 3 for each number of similar items used in the equating.

# Similar items	Grade	# Combinations	Mean	SD	Min	Max
1	A	27	46.41	1.92	41.34	50.86
	E	25	12.23	2.53	7.40	16.47
2	A	351	46.43	1.21	43.03	50.24
	E	350	12.28	1.91	7.40	16.47
3	A	992	46.47	0.97	43.46	49.79
	E	992	12.29	1.53	7.40	16.47
4	A	1027	46.44	0.75	44.06	48.67
	E	1027	12.31	1.33	8.57	15.87
5	A	963	46.46	0.66	44.09	48.28
	E	963	12.23	1.21	8.19	15.63
6	A	1002	46.47	0.58	44.60	48.00
	E	1002	12.29	1.05	8.67	15.05
7	A	995	46.43	0.54	44.78	48.00
	E	995	12.32	0.94	9.61	14.92
8	A	994	46.45	0.48	45.08	47.83
	E	994	12.32	0.88	9.42	14.82
9	A	1006	46.48	0.43	45.13	47.76
	E	1006	12.29	0.81	9.99	14.56
10	A	951	46.45	0.40	45.25	47.61

[13] https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_transreg_sect021.htm has more details.

# Similar items	Grade	# Combinations	Mean	SD	Min	Max
	E	951	12.30	0.73	10.33	14.24
11	A	1003	46.46	0.36	45.46	47.52
	E	1003	12.31	0.67	10.16	14.23
12	A	962	46.46	0.34	45.12	47.57
	E	962	12.32	0.66	10.61	14.29
13	A	980	46.46	0.31	45.54	47.25
	E	980	12.28	0.58	10.52	14.19
14	A	1038	46.45	0.29	45.49	47.30
	E	1038	12.32	0.57	10.59	13.85
15	A	958	46.46	0.27	45.64	47.26
	E	958	12.30	0.50	10.80	13.68
16	A	1014	46.47	0.25	45.53	47.20
	E	1014	12.33	0.46	10.96	13.77
17	A	956	46.46	0.24	45.65	47.05
	E	956	12.31	0.45	11.05	13.63
18	A	991	46.46	0.21	45.66	47.00
	E	991	12.30	0.40	11.13	13.50
19	A	955	46.46	0.20	45.89	47.10
	E	955	12.32	0.38	11.08	13.39
20	A	1049	46.46	0.18	45.82	46.96
	E	1049	12.32	0.34	11.15	13.34
21	A	947	46.47	0.16	46.01	46.91
	E	947	12.32	0.30	11.43	13.24
22	A	1024	46.47	0.15	45.95	46.88
	E	1024	12.30	0.27	11.46	13.19
23	A	954	46.47	0.13	46.11	46.80
	E	954	12.32	0.24	11.59	13.07
24	A	1011	46.47	0.10	46.14	46.75
	E	1011	12.32	0.20	11.65	12.95
25	A	351	46.47	0.09	46.20	46.68
	E	351	12.31	0.16	11.80	12.87
26	A	27	46.47	0.06	46.34	46.60
	E	27	12.31	0.11	12.02	12.63
27	A	1	46.47	.	46.47	46.47
	E	1	12.31	.	12.31	12.31

Table A2. Descriptive statistics for distribution of (unrounded) grade boundaries on Paper 3 for each number of similar marks used in the equating.

# Similar marks	Grade	# Combinations	Mean	SD	Min	Max
1	A	8	46.20	2.58	41.34	49.00
	E	7	11.46	2.63	7.40	14.66
2	A	35	46.34	1.72	43.03	50.86
	E	35	11.82	2.14	7.40	16.47
3	A	79	46.47	1.51	42.82	49.93
	E	78	12.29	1.79	8.37	16.47
4	A	142	46.41	1.35	42.48	49.64

# Similar marks	Grade	# Combinations	Mean	SD	Min	Max
	E	141	12.04	1.63	7.40	16.13
5	A	210	46.46	1.19	43.37	49.43
	E	210	12.18	1.65	7.40	16.47
6	A	263	46.40	1.11	43.08	49.70
	E	263	12.18	1.50	7.40	16.47
7	A	280	46.43	1.06	43.63	49.47
	E	280	12.18	1.41	7.40	15.57
8	A	312	46.34	0.96	43.92	49.24
	E	312	12.29	1.39	8.16	15.84
9	A	350	46.34	0.92	43.80	49.16
	E	350	12.17	1.23	9.08	15.49
10	A	364	46.41	0.85	43.65	48.96
	E	364	12.18	1.24	8.89	15.18
11	A	371	46.41	0.83	44.27	48.39
	E	371	12.30	1.18	9.00	15.75
12	A	374	46.38	0.84	44.13	49.01
	E	374	12.23	1.13	9.03	15.31
13	A	373	46.40	0.70	44.33	48.46
	E	373	12.27	1.08	9.55	14.90
14	A	373	46.42	0.70	44.09	47.98
	E	373	12.19	0.97	9.17	15.06
15	A	390	46.42	0.70	44.65	48.62
	E	390	12.19	1.01	9.09	14.66
16	A	382	46.41	0.68	44.42	48.48
	E	382	12.22	0.97	9.66	14.90
17	A	415	46.39	0.61	44.78	48.10
	E	415	12.22	0.96	9.50	14.95
18	A	389	46.38	0.61	44.81	48.37
	E	389	12.23	0.84	10.24	14.41
19	A	410	46.40	0.60	44.66	47.96
	E	410	12.21	0.89	9.57	14.31
20	A	346	46.40	0.58	44.80	48.10
	E	346	12.26	0.79	10.26	14.90
21	A	398	46.41	0.56	44.98	47.99
	E	398	12.20	0.77	9.55	14.36
22	A	381	46.46	0.54	44.85	48.05
	E	381	12.20	0.81	9.93	14.52
23	A	397	46.43	0.50	44.84	47.98
	E	397	12.20	0.71	9.92	14.07
24	A	399	46.44	0.51	44.98	47.93
	E	399	12.20	0.68	10.14	14.11
25	A	393	46.43	0.47	45.25	47.79
	E	393	12.28	0.69	10.19	14.30
26	A	340	46.39	0.46	45.04	47.69
	E	340	12.27	0.65	10.78	13.90
27	A	361	46.47	0.48	44.96	47.59
	E	361	12.23	0.66	10.63	14.33

# Similar marks	Grade	# Combinations	Mean	SD	Min	Max
28	A	347	46.39	0.46	45.24	47.55
	E	347	12.15	0.62	9.94	13.92
29	A	376	46.41	0.43	45.20	47.79
	E	376	12.24	0.61	10.71	13.90
30	A	352	46.42	0.42	45.32	47.47
	E	352	12.22	0.62	10.76	14.19
31	A	353	46.39	0.39	45.29	47.49
	E	353	12.24	0.55	10.80	13.77
32	A	369	46.43	0.41	45.15	47.73
	E	369	12.26	0.53	10.78	14.23
33	A	360	46.36	0.39	45.24	47.40
	E	360	12.21	0.53	10.54	13.60
34	A	342	46.40	0.39	45.29	47.54
	E	342	12.24	0.50	10.84	13.71
35	A	380	46.40	0.39	45.37	47.36
	E	380	12.24	0.50	10.75	13.39
36	A	386	46.41	0.34	45.27	47.42
	E	386	12.22	0.50	10.93	13.81
37	A	394	46.41	0.31	45.43	47.36
	E	394	12.22	0.47	10.76	13.55
38	A	435	46.37	0.36	45.53	47.60
	E	435	12.26	0.48	10.81	13.91
39	A	362	46.43	0.33	45.45	47.44
	E	362	12.21	0.46	10.94	13.41
40	A	389	46.41	0.32	45.42	47.16
	E	389	12.28	0.43	10.84	13.51
41	A	343	46.42	0.31	45.52	47.31
	E	343	12.26	0.41	11.16	13.52
42	A	388	46.41	0.32	45.58	47.32
	E	388	12.20	0.42	10.89	13.45
43	A	380	46.41	0.30	45.61	47.22
	E	380	12.24	0.40	11.09	13.57
44	A	334	46.43	0.30	45.56	47.10
	E	334	12.25	0.41	11.13	13.48
45	A	376	46.42	0.28	45.56	47.35
	E	376	12.26	0.37	11.28	13.29
46	A	408	46.40	0.28	45.69	47.15
	E	408	12.24	0.37	11.24	13.48
47	A	354	46.40	0.26	45.51	47.16
	E	354	12.22	0.37	11.21	13.70
48	A	377	46.41	0.26	45.51	47.11
	E	377	12.24	0.33	11.33	13.18
49	A	361	46.40	0.25	45.78	47.19
	E	361	12.26	0.33	11.38	13.09
50	A	360	46.41	0.25	45.70	47.14
	E	360	12.26	0.34	11.12	13.32
51	A	372	46.41	0.23	45.76	47.11

# Similar marks	Grade	# Combinations	Mean	SD	Min	Max
	E	372	12.19	0.31	11.27	13.15
52	A	376	46.42	0.23	45.74	47.04
	E	376	12.21	0.30	11.29	13.09
53	A	399	46.41	0.22	45.92	46.99
	E	399	12.25	0.28	11.36	13.15
54	A	404	46.40	0.20	45.78	46.94
	E	404	12.23	0.28	11.52	13.17
55	A	393	46.41	0.21	45.72	46.99
	E	393	12.26	0.27	11.47	13.06
56	A	393	46.42	0.20	45.85	46.95
	E	393	12.25	0.27	11.50	12.93
57	A	368	46.42	0.18	45.85	46.98
	E	368	12.26	0.24	11.55	12.88
58	A	379	46.41	0.18	45.85	46.90
	E	379	12.25	0.23	11.57	13.16
59	A	358	46.43	0.17	45.89	47.03
	E	358	12.24	0.23	11.40	13.07
60	A	346	46.41	0.16	46.01	46.94
	E	346	12.23	0.23	11.73	12.95
61	A	327	46.40	0.16	45.84	46.82
	E	327	12.23	0.21	11.60	12.89
62	A	327	46.42	0.15	45.99	46.85
	E	327	12.22	0.20	11.72	12.80
63	A	292	46.41	0.14	46.04	46.78
	E	292	12.23	0.18	11.74	12.73
64	A	246	46.41	0.14	45.94	46.78
	E	246	12.24	0.18	11.68	12.67
65	A	221	46.42	0.14	46.03	46.81
	E	221	12.25	0.17	11.88	12.68
66	A	139	46.41	0.14	46.03	46.72
	E	139	12.23	0.14	11.86	12.59
67	A	84	46.41	0.13	46.13	46.74
	E	84	12.25	0.15	11.94	12.58
68	A	35	46.42	0.12	46.22	46.68
	E	35	12.27	0.13	12.05	12.58
69	A	8	46.42	0.10	46.31	46.60
	E	8	12.26	0.10	12.13	12.43
70	A	1	46.41	.	46.41	46.41
	E	1	12.23	.	12.23	12.23

Table A3. Descriptive statistics for distribution of (unrounded) grade boundaries on component 3 for each number of similar items used in the equating (based on overall EICCs rather than Rasch)

# Similar items	Grade	# Combinations	Mean	SD	Min	Max
1	A	27	46.28	1.59	40.33	49.73
	E	25	12.40	2.33	8.06	15.69
2	A	351	46.28	1.08	42.48	48.94
	E	350	12.40	1.70	8.06	15.69
3	A	943	46.28	0.84	43.60	48.45
	E	943	12.40	1.35	8.06	15.38
4	A	1055	46.28	0.73	43.97	47.95
	E	1055	12.41	1.11	9.19	15.14
5	A	981	46.33	0.61	44.43	47.78
	E	981	12.45	0.99	9.57	15.09
6	A	1048	46.28	0.58	44.63	47.62
	E	1048	12.42	0.88	9.52	14.73
7	A	963	46.29	0.53	45.01	47.56
	E	963	12.38	0.79	9.73	14.73
8	A	1015	46.28	0.46	44.93	47.40
	E	1015	12.43	0.72	9.88	14.52
9	A	1062	46.28	0.43	45.13	47.39
	E	1062	12.41	0.66	10.64	14.73
10	A	982	46.29	0.40	45.27	47.17
	E	982	12.35	0.60	10.59	14.20
11	A	957	46.28	0.36	45.39	47.13
	E	957	12.42	0.55	10.49	14.06
12	A	1001	46.27	0.35	45.27	47.10
	E	1001	12.40	0.52	10.85	13.91
13	A	985	46.27	0.32	45.34	47.06
	E	985	12.39	0.50	10.69	13.87
14	A	970	46.29	0.30	45.55	46.98
	E	970	12.41	0.46	11.00	13.95
15	A	997	46.28	0.27	45.60	47.10
	E	997	12.41	0.42	11.16	13.76
16	A	1003	46.29	0.25	45.70	47.00
	E	1003	12.39	0.38	11.18	13.38
17	A	971	46.28	0.24	45.65	46.92
	E	971	12.37	0.35	11.39	13.47
18	A	1010	46.28	0.22	45.71	46.88
	E	1010	12.40	0.33	11.53	13.32
19	A	944	46.29	0.20	45.80	46.91
	E	944	12.40	0.30	11.46	13.29
20	A	1028	46.28	0.18	45.84	46.78
	E	1028	12.41	0.27	11.72	13.36
21	A	1052	46.28	0.16	45.92	46.79
	E	1052	12.40	0.25	11.71	13.39
22	A	997	46.28	0.14	45.97	46.72
	E	997	12.39	0.23	11.70	13.11
23	A	982	46.29	0.13	45.96	46.70

# Similar items	Grade	# Combinations	Mean	SD	Min	Max
	E	982	12.40	0.19	11.91	13.13
24	A	1002	46.29	0.11	46.01	46.65
	E	1002	12.40	0.16	12.00	12.95
25	A	351	46.28	0.09	46.07	46.59
	E	351	12.40	0.13	12.13	12.78
26	A	27	46.28	0.06	46.15	46.51
	E	27	12.40	0.09	12.26	12.58
27	A	1	46.28	.	46.28	46.28
	E	1	12.40	.	12.40	12.40

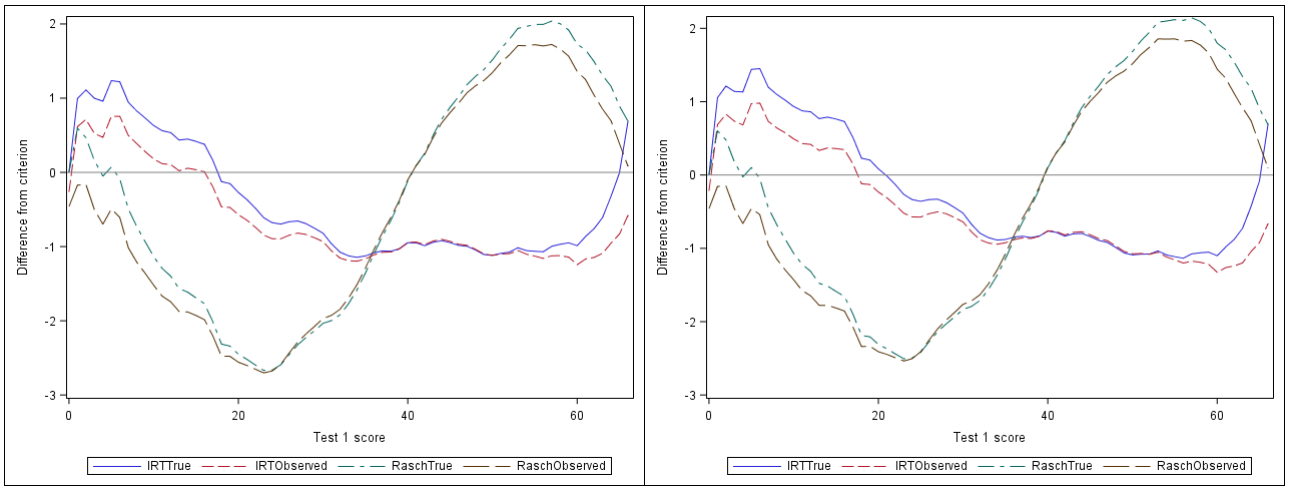

Figure A1. Comparison of IRT/Rasch true score equating with observed score equating, including com1 (left) and excluding com1 (right).

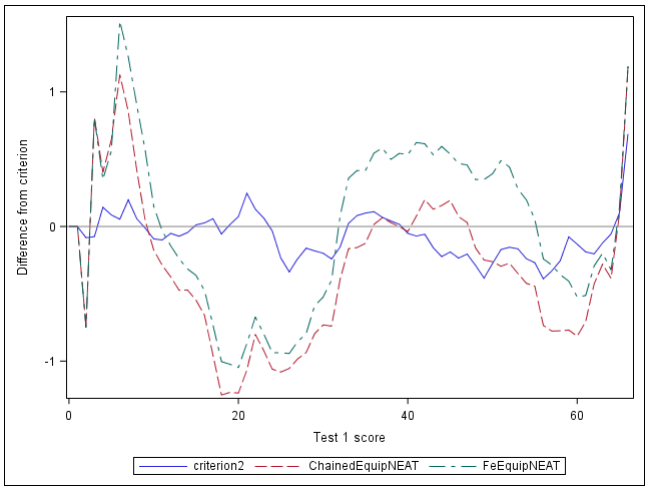

Figure A2. Comparison of frequency estimation equipercentile equating with chained equipercentile equating.