A peer reviewed, open-access electronic journal: ISSN 1531-7714

Simultaneous linear equating for scenarios with optional test versions or across multiple alternative anchors

Tom Benton, *Cambridge University Press and Assessment*

Abstract: This paper proposes an extension of linear equating that may be useful in one of two fairly common assessment scenarios. One is where different students have taken different combinations of test forms. This might occur, for example, where students have some free choice over the exam papers they take within a particular qualification. In this scenario we wish to transform scores on the all the various test forms to a common scale so that scores can be used interchangeably. A second scenario is where, perhaps for reasons of security, we use more than one anchor test between test forms with different students exposed to different anchors. In this second scenario, we wish to equate test forms using all of the data from different anchor versions in a coherent manner. The method proposed in this paper to address the above scenarios may be particularly useful in instances where the alternative of using item response theory is problematic.

Keywords: equating, linking, scaling

Introduction

Imagine the following situation. A qualification requires students to take any two out of a possible four tests. Each student's final score is calculated based upon a sum of the scores on the two tests they have completed. However, tests cannot be assumed to be of equal difficulty and so test scores should be transformed to a common scale (i.e., equated) before this summation takes place. How should this be done?

One answer to the above problem would be to fit an item response theory (IRT) or Rasch model to the data and use this to find a suitable transformation. However, such models require multiple assumptions that may or may not be correct. For example, such models would typically assume that all tests measure the same unidimensional construct. They also assume particular functional forms that may not fit the data. For example, the Rasch model assumes that all items have the same slope for the item response function. Finally, for more complex IRT models (i.e., anything other than the Rasch model) sample sizes of 500 or more students for each test version are often recommended (Reise & Yu, 1990).

In the scenario described above, there are no immediately obvious classical approaches to equating that could be used. Specifically, whilst there are many classical approaches to equating scores on *two* tests that have been done by the same group of people, these do not easily generalise to the kind of situation described above. For example, if we imagine our four tests are labelled tests A, B, C and D, then we have five approaches to equating scores on tests A and B. We could:

1. Use a single-group form of equating using only those that took tests A and B together.
2. Use chained equating of test A to test B via test C. That is, first equate tests A and C based on students that have done both and then equate tests C and B based on those students that have completed this pair. Finally, identify scores on tests A and B that correspond to the same score on test C as being equivalent.
3. Use chained equating of test A to test B via test D.
4. Use chained equating of test A to test B via first test C and then test D. That is equate, A to C, C to D, and finally D to B.
5. Use chained equating of test A to test B via first test D and then test C.

Which of the above approaches is most appropriate would depend upon the number of students that have completed each pair of tests. Alternatively, we might use all the different possible options and somehow combine the results to give a consensus equating function (Holland & Strawderman, 2011). However, the specifics of how this should be done are not obvious. This means we do not have a single overarching way of equating all our different tests to a common scale.

This paper describes how we can linearly equate a large number of tests to each other in a single analysis in a way that uses evidence from all the possible ways in which any two tests can be linked. The mathematics underlying the method builds very easily upon the Kelly method (Kelly, 1971). The Kelly method is most commonly used in the context of comparing the difficulty of different examination subjects such as whether Mathematics is harder than Physics rather than in formal equating scenarios. However, the algorithm it uses could potentially be used to equate different test forms. The usual application of the Kelly method identifies additive adjustments for each test such that, when they are applied, the mean adjusted score on any test equals the mean adjusted score on all tests taken alongside it. This approach is very similar to a very simple method of equating known as mean equating. In mean equating, test "form X is considered to differ in difficulty from form Y by a constant amount along the score scale" (Kolen & Brennan, 2004, p. 30). The Kelly method extends this idea to assume that every test form in a large set differs from the others by a constant amount. Furthermore, these differences between pairs of tests must be logically consistent. For example, if form X is two score points harder than form Y, and form Y is one score point harder than form Z, then form X must be three score points harder than form Z. As such, the Kelly method could be called simultaneous mean equating as it considers all possible pairs of test forms at the same time and ensures that the results are logically coherent. We will describe this method in more detail later in the paper.

This paper describes an extended version of the Kelly method that we will refer to as simultaneous linear equating. Rather than simply identifying additive adjustments, it identifies a linear transformation for the set of scores on each test such that, across all possible pairwise comparisons of any combination of two tests:

- The mean transformed score on any test equals the mean transformed score across all tests taken alongside it.

- When we average values across data on pairs of tests taken by the same group of students, the (weighted[1]) geometric mean of the standard deviations of transformed scores on any test equals the (weighted) geometric mean of the standard deviations of the transformed scores of all tests taken alongside it.

These ideas will be illustrated in more detail later in the paper. Note that the idea presented in this paper is similar, but not identical, to the idea of Average Marks Scaling (AMS) sometimes used for tertiary examinations in Western Australia (see Howie et al., 2008; TISC, 2015).

The method described in this paper is an extension of linear equating. In traditional linear equating, "scores that are an equal (signed) distance from their means in standard deviation units are set equal" (Kolen & Brennan, 2004, p.31). As such, the method assumes that a linear transformation is sufficient to account for differences in the difficulty of two test forms across every point on the score scale.

The simplest case of linear equating is in the single group design where we have two test forms (form X and form Y) taken by the same group of students. In this case, equating form Y to form X involves identifying a linear transformation for form Y scores to make the mean and standard deviation equal to the mean and standard deviation of form X. In other words, it achieves exactly the same goals as those listed for simultaneous linear equating in the bullets above. The only difference is that, since we have only one pair of tests, there is no need to average means and standard deviations across pairs. As such, simple linear equating is a special case of simultaneous linear equating.

Another special case of simultaneous linear equating is chained linear equating (Livingston, 2014). In chained linear equating we assume that form X and form Y have been taken by different sets of students. We also assume that both sets of students took an anchor test. This type of data collection is known as the non-equivalent groups with anchor test (NEAT) design. In the NEAT design, we can linearly equate forms X and Y by the following procedure. First find a linear transformation for the scores on the anchor test so that, so for students that also took form X, the mean and standard deviation of the anchor match the mean and standard deviation of form X. Apply this same transformation to the anchor test scores for the students that took form Y. Finally, identify a linear transformation to the form Y scores so that the mean and standard deviation will match the mean and standard deviation of the (now transformed) anchor test scores. Thus, once all the transformations have been applied, for the set of students that took form X and the anchor, the means and standard deviations match. Furthermore, for the set of students that took the anchor and form Y, the means and standard deviations also match. Thus, chained linear equating achieves identical goals (and will produce identical score equivalencies) to simultaneous linear equating. As such, chained linear equating can be seen as a special case of simultaneous linear equating.

To summarise, equating is an important process as it transforms scores from different tests so that they can be directly compared. However, current non-IRT approaches to equating are limited by the fact that, in some circumstances, they do not make use of all the data from candidates taking different combinations of tests. The goal of this paper is to present the simultaneous linear equating method which is easy to use and makes use of all available data in a consistent manner.

Method for simultaneous linear equating

This aim of this section is to provide a clear description of the how simultaneous linear equating can be completed. This begins with a description of the method behind the existing Kelly method as, once this is understood, it is fairly easy to work out how to extend the method to allow simultaneous linear equating.

[1] Weighted by the number of students that have taken each pair. This will become clearer in the worked examples in this paper.

This section focuses on *how* to complete the calculations rather than the steps required to derive the formulae (i.e., *why* they work). Readers who are interested in the mathematical derivation of the Kelly method should refer to Lawley's appendix to Kelly (1971).

It is important to note that, although the algebraic formula can make the Kelly method look daunting, the method is not particularly hard to apply. The starting point for calculations is simply a table showing for each pair of tests taken together:

- The number of students that took both tests.
- The mean score achieved on each of the tests by the students that took both.
- The standard deviation of the scores on each test for the students that took both.

With the above information in place, the remaining steps can be easily completed in any statistical software package (including Microsoft Excel). To help show how this is done, we not only present all the algebraic formulae used in calculations but also show how they are applied step by step to a given data set. We begin with a description of the illustrative data set.

Illustrative example

To help illustrate the approach we begin the information in Table 1. This table shows data from four simulated tests labelled A, B, C and D (simulation parameters will be provided later). Each test has a maximum score of 50. We imagine that every student has taken two of these tests. The table shows the number taking each pair as well as the means and standard deviations of scores on each test in the pair for those students taking both. Note that, in order to make the following stages of calculations easier to follow, the data for each pair is shown twice with the tests in each pair shown in both orders.

A brief review of the data in Table 1 indicates that test A may be the easiest of the four as students tend to have higher mean scores on this test than any of those taken alongside it. Similarly, test D may be the hardest of the four as students tend to have lower mean scores on this test than the tests taken with it. We can also see that, in total, tests A, B and C were taken by larger numbers of students than test D.

The steps for the usual Kelly method

The aim of the Kelly method is to find additive adjustments to the scores on each test, so that, when these are applied, the mean score on any test will equal the mean score on other tests taken alongside it. An algebraic approach to solving this problem was provided in Lawley's appendix to Kelly (1971). Here, we record the steps required for the Kelly method.

1. First create an N matrix recording the numbers of students taking each pair of tests. Specifically, it should be a square matrix with as many rows and columns as there are tests in the analysis (four in our example). We denote the number of students entering both test i and test j together as n_{ij}. Let us also denote the value in the ith row and jth column of the N matrix as N_{ij}. The off-diagonal values in the N matrix are defined as follows:

$$N_{ij} = n_{ij} \text{ if } i \neq j \quad (1)$$

The diagonal values of the N matrix are defined as minus the sum of the off-diagonal values in the same row. In other words,

$$N_{ii} = -\sum_{j \neq i} n_{ij}. \quad (2)$$

The above equations mean that the values in each row and each column of the N matrix will sum to zero. For further details of the N matrix in our illustrative example, see Table A1 in the appendix.

Table 1. Summary of scores on tests taken by pairs of students

Test 1	Test 2	N taking pair	Mean score on test 1	Mean score on test 2	SD of scores on test 1	SD of scores on test 2
A	B	300	26.17	25.34	7.81	7.99
A	C	250	26.21	24.81	7.81	8.18
A	D	10	26.20	24.40	9.69	9.23
B	A	300	25.34	26.17	7.99	7.81
B	C	100	26.79	25.27	8.66	8.88
B	D	60	27.12	25.77	8.11	9.35
C	A	250	24.81	26.21	8.18	7.81
C	B	100	25.27	26.79	8.88	8.66
C	D	180	26.10	25.30	7.90	7.91
D	A	10	24.40	26.20	9.23	9.69
D	B	60	25.77	27.12	9.35	8.11
D	C	180	25.30	26.10	7.91	7.90

2. Now create a difference matrix (which we will denote D) which contains weighted sums of the differences in means. The D matrix is a single column matrix that is created by taking a sum of the differences of the mean scores on each test relative to the mean scores on each test taken alongside it multiplied by the number of students taking the pair. In algebraic terms, we define μ_{ij} as the mean score on test i for those students that also took test j. Note that, under this definition, $\mu_{ij} \neq \mu_{ji}$. Then, the ith element of the D matrix (D_i) is defined as follows:

$$D_i = \sum_{j \neq i} n_{ij}(\mu_{ij} - \mu_{ji}). \quad (3)$$

For example, in our illustrative data, the first element of the D matrix would be calculated as 300*(26.17-25.34)+250*(26.21-24.81)+10*(26.20-24.40)=617. The full D matrix for our example is shown in Table A2 in the appendix.

3. Remove the first row and column from the N matrix. In effect, this will mean we are ultimately defining the Kelly adjustment for the first test (e.g., test A) to be zero and that all other Kelly adjustments are defined relative to this.

4. Similarly, remove the first value from the D matrix.

5. Finally, invert the (reduced) N matrix and multiply it by the D matrix (without the first value). In algebraic terms, the Kelly adjustments (denoted by α for the vector of adjustments or by α_i for the ith test) are given by the following formula:

$$\alpha = N_{\_}^{-1} D_{\_}. \quad (4)$$

Where $N_{\_}$ refers to the N matrix with the first row and column removed and $D_{\_}$ refers to the D matrix with the first value removed. The inverted (reduced) N matrix is illustrated in Table A3 in the appendix.

The results of applying the steps above to the data in Table 1 (i.e., the final Kelly adjustment parameters) are shown in Table 2. Note that, for clarity, we have also added a Kelly adjustment of zero for test A back into the table.

Table 2. Kelly adjustment parameters (rounded to 2 decimal places)

Test	Kelly adjustment
A	0.00
B	0.67
C	1.57
D	2.26

The values in Table 2 indicate that tests A to D are ordered in increasing difficulty with test D slightly more than 2 points harder than test A on average.

Note that, as intended, if we add the values in Table 2 to the scores in the respective tests, the weighted sum of the differences in scores on pairs of test becomes zero. For example, based on the values in Table 1 and Table 2, the overall summed difference between adjusted scores on test B and scores on other tests would become:

300*(25.34+0.67-26.17-0.00)+100*(26.79+0.67-25.27-1.57)+60*(27.12+0.67-25.77-2.26).

If this calculation is completed using unrounded values (rather than the rounded values in Table 2), it is equal to zero.

Extending the Kelly method to allow linear equating

The Kelly method as originally suggested only provides an additive adjustment for each test score. This is only sufficient if we believe that the difference in the difficulty of each pair of tests is constant across the entire score scale. This is the same as the usual assumption of mean equating.

The Kelly method will not provide an adequate means of equating if differences in the difficulty of different test forms to varies along the score scale. This might be seen by the differences in the standard deviations of scores on tests taken by the same sets of candidates. All else being equal, a lower standard deviation indicates that students have a better chance of achieving at or above low scores on the test whilst being less likely to achieve the highest scores. That is, a test with a lower standard deviation might be easier at the bottom end of the score scale and harder at the top.

To address this, we need to our method to include a multiplicative adjustment to scores as in addition to the additive adjustment. This can easily be addressed by (essentially) applying the algebraic approach of the Kelly method twice. First to identify the multiplicative adjustment and then subsequently to identify the additive adjustment. This creates a method of simultaneous linear equating.

The multiplicative adjustment is estimated by following the same steps as above but applied to the differences in the logs of the standard deviations, rather than the differences in the means. Our real focus is on the ratios of standard deviations to one another. However, the log of a ratio is equal to the difference in the logs of the two quantities. Thus, finding multiplicative adjustments to make a set of quantities have the same (geometric) mean, is equivalent to finding additive adjustments to make the logs of the quantities of interest have the same (arithmetic) mean. As such, we can use essentially the same process as we use in the usual Kelly method. The steps are as follows.

1. Remove any pairs from the data (i.e., Table 1) where the standard deviation of scores on either test is zero or where it cannot not be calculated (e.g., if there is only a single observation for the pair). In our example, and many other practical cases, no pairs will need to be removed.

2. Construct the N matrix as for the original Kelly method but based on the reduced data (if any pairs have been removed).

3. Create the L matrix to denote the weighted sum of differences in log standard deviations. In algebraic terms, we define σ_{ij} as the standard deviation of scores on test i for those students that also took test j. Again, note that, under this definition, $\sigma_{ij} \neq \sigma_{ji}$. Then, the ith element of the L matrix (L_i) is defined as follows:

$$L_i = \sum_{j \neq i} n_{ij}\left(\log\left(\sigma_{ij}\right) - \log\left(\sigma_{ji}\right)\right). \quad (5)$$

For example, in our illustrative data, the first element of the L matrix would be calculated as 300*(log(7.81)-log(7.99))+250*(log(7.81)-log(8.18))+10*(log(9.69)-log(9.23))=-17.9. The full L matrix for our example is shown in the appendix in Table A4.

4. If we denote the multiplicative coefficients we are looking for as β then $\log(\beta)$ can be estimated by multiplying the inverse of the N matrix defined earlier (without the first row and column) by the L matrix (without the first value). In algebraic terms we write this as:

$$\log(\beta) = N_{\_}^{-1} L_{\_}. \quad (6)$$

Where $N_{\_}$ refers to the N matrix with the first row and column removed and $L_{\_}$ refers to the L matrix with the first value removed. The β coefficients for the illustrative example are shown in Table 3. Note that, for clarity, we have also added a Kelly multiplicative adjustment of 1 for test A back into the table.

Table 3. Kelly multiplicative parameters for the illustrative example (calculations shown to 4 decimal places)

Test	$\log(\beta)$ value	β value
A	0.0000	1.0000
B	-0.0123	0.9877
C	-0.0539	0.9475
D	-0.0749	0.9279

To finish the process all that is needed is to apply these multiplicative coefficients to the original set of pairwise comparisons (e.g., Table 1) and then to recalculate the additive coefficients in the light of these as before.

In our illustrative example, we can update Table 1 as shown in Table 4. All that has changed is that all the means and standard deviations have been multiplied by the relevant β values from Table 3. Note that we label the adjusted means as "Interim adjusted" since they will be altered again by the additive coefficients identified next. In contrast, the adjusted standard deviations are at their final values as they will not be affected by the purely additive adjustments applied subsequently.

Note that across the pairs in Table 4 with the same value for "test 1", the weighted geometric mean of the standard deviations on test 1 equals the geometric mean of the standard deviations of test 2. For example, across all pairs where in Table 4 where test 1 is "A" the geometric means of the standard deviations are given by[2]:

$$(7.81^{300} * 7.81^{250} * 9.69^{10})^{\frac{1}{300+250+10}} = (7.89^{300} * 7.75^{250} * 8.56^{10})^{\frac{1}{300+250+10}} = 7.84$$

[2] In practice, to avoid numerical problems with very large numbers, the calculation usually requires taking logs, calculating the weighted mean, and then taking the exponential of the result.

Note that if the geometric means of standard deviations are equivalent, then the geometric means of variances will also be equivalent. As such, it does not matter whether we measure dispersion using standard deviations or variances. This is an advantage of the technique being defined in terms of a geometric rather than an arithmetic mean.

More easily, it can be seen in Table 4, that, for the pairs taken by the largest numbers of students, the adjusted standard deviations on the two tests are extremely close to one another.

Table 4. Pairwise comparisons of tests in illustrative example after applying multiplicative adjustments (results shown to 2 decimal places)

Test 1	Test 2	N taking pair	Interim adjusted mean score on test 1	Interim adjusted mean score on test 2	Adjusted SD of scores on test 1	Adjusted SD of scores on test 2
A	B	300	26.17	25.03	7.81	7.89
A	C	250	26.21	23.51	7.81	7.75
A	D	10	26.20	22.64	9.69	8.56
B	A	300	25.03	26.17	7.89	7.81
B	C	100	26.46	23.94	8.55	8.41
B	D	60	26.79	23.91	8.01	8.68
C	A	250	23.51	26.21	7.75	7.81
C	B	100	23.94	26.46	8.41	8.55
C	D	180	24.73	23.47	7.49	7.34
D	A	10	22.64	26.20	8.56	9.69
D	B	60	23.91	26.79	8.68	8.01
D	C	180	23.47	24.73	7.34	7.49

Having updated our table of differences, we now repeat steps 1 to 5 from the original Kelly method but using Table 4 as the original input. This provides the additive (α) coefficients. The final set of linear adjustments (both multiplicative and additive) are shown in Table 5.

Table 5. Final coefficients from simultaneous linear equating for our illustrative example (rounded to 2 decimal places)

Test	α value	β value
A	0.00	1.00
B	0.98	0.99
C	2.88	0.95
D	4.05	0.93

At the end of the process, the final coefficients allow linear equating between any pair of tests. Specifically, if we denote a particular total score on the ith test as t_i, scores on different tests are deemed equivalent (i.e., $t_i \equiv t_j$) if they satisfy the following equation:

$$\alpha_i + \beta_i t_i = \alpha_j + \beta_j t_j \quad (7)$$

In our illustrative example, since the α and β coefficients for test A are 0 and 1 respectively, the coefficients for the other tests can be used to convert scores to equivalent scores on test A. For example, a

score of zero on test D is equivalent to a score of 4.05 on test A whereas a test D score of 25 maps to a score on test A of 27.3 (=4.05+25*0.93). Like any application of linear equating, we will also need to decide how to handle equated values that fall outside the normally allowable range of scores. In practice, we would usually truncate these to be within the allowable range.

Equating accuracy for illustrative example

We now turn our attention to whether the method actually provides accurate answers. To do this we take advantage of the fact that the data for the illustrative example was actually simulated from a graded response model (GRM; Reise & Yu, 1990). The GRM was used as, with suitably chosen parameters, it provides a formal way of simulating data that approximately meets the requirements for linear equating to be appropriate. That is, it allows us to create test forms where a linear transformation of test scores will approximately address the true differences in form difficulty across the score scale.

For our example, each test was simulated to consist of a single item with a maximum score of 50 with a slope parameter of 5 and GRM thresholds evenly spaced across the ranges defined in Table 6. The slope parameter of 5 was chosen as it was found to lead to correlations between test scores of just around 0.90 – a plausible level of correlation for tests of this length. As can be seen from Table 6, whilst the highest scores were equally difficulty to achieve in each test, scores at the lower end of the range on each successive test version represent higher levels of achievement.

Table 6. GRM threshold parameters used to simulate data for illustrative example

Test	Minimum GRM threshold	Maximum GRM threshold
A	-3.6	3.0
B	-3.4	3.0
C	-3.2	3.0
D	-3.0	3.0

Using the known parameters in Table 6, it was possible to identify an IRT true score equating relationship between each of the tests (Kolen & Brennan, 2004, p. 176).

For each test version, the true equating relationships of each test with test A are shown in Figure 1. Figure 1 shows that, except at the very bottom of the score range, the true equating relationship is approximately linear. Scores at the top of the range on each test are approximately equivalent to one another without adjustment. However, scores at the bottom of the range of tests B, C and D represent higher levels of achievement than on test A and should be adjusted upwards to be made equivalent.

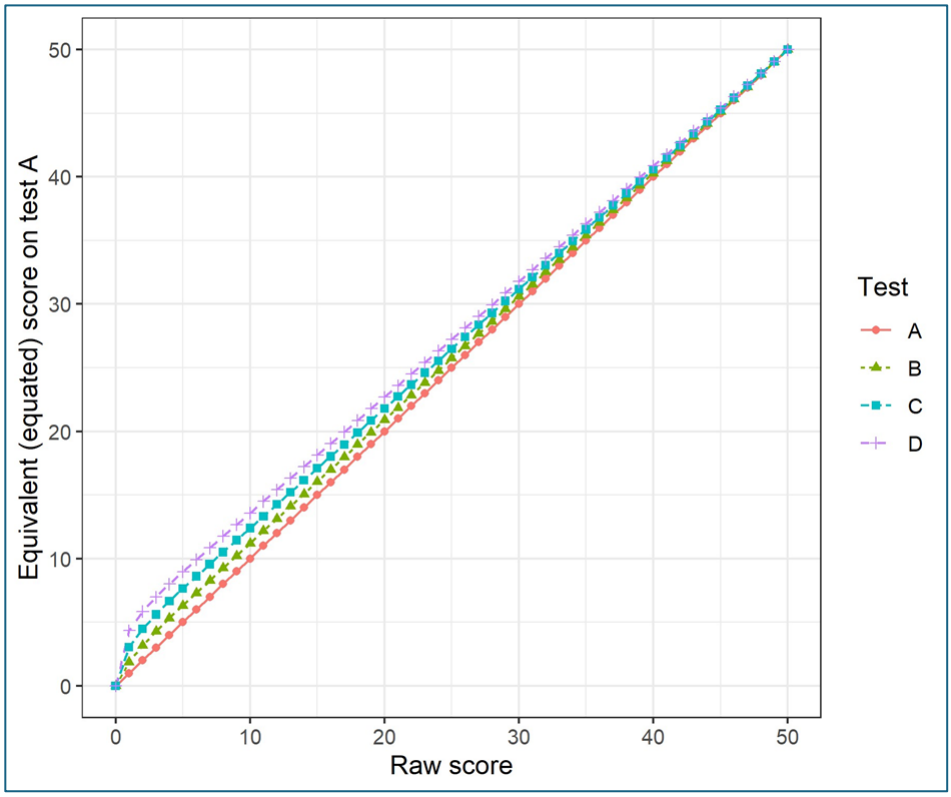
Scores for students taking different pairs of tests (Table 1) were simulated separately based on students having normally distributed abilities with a standard deviation of 1 and with mean ability varying according to the pair of tests being taken (students entering harder pairs of tests were simulated to be of slightly higher ability)[3]. Simultaneous linear equating was applied to this simulated data as described in the previous section. The accuracy of this analysis is shown in Figure 2.

From Figure 2 it can be seen that simultaneous linear equating provided a very accurate equating method in this scenario. Except for at the very bottom of the score range, for each test, the approach is accurate to within about 0.5 score points. Linear equating is less accurate at the extreme bottom end of the score range.

[3] Specifically, the mean abilities for students taking different pairs of tests were simulated as follows. Students taking pair A&B had a mean ability of -0.2, A&C had a mean of -0.05, A&D had a mean of 0.0, B&C had a mean of 0.0, B&D had a mean of 0.05, and C&D had a mean of 0.1. In every pair, the standard deviation of ability was set to 1.0.

However, this is unsurprising given that the true equating relationships are noticeably non-linear at this point. Nonetheless, this short trial still provides some confidence that simultaneous linear equating will provide a reasonable idea of the relative difficulty of different tests.

Figure 1. True IRT equating relationships between each test and test A



Equating accuracy across replications

To further explore equating accuracy, the simulation described above was repeated 500 times. In each replication, the test scores of students taking each pair of tests was simulated as before. Simultaneous linear equating was applied and the resulting mapping of scores on each test version to scores on test A were stored.

The accuracy of simultaneous linear equating was compared to the accuracy of a pragmatic alternative that, in the absence of simultaneous linear equating, might be used in practice. Specifically, each other test was linearly equated to test A via the single route best supported by the available data. Thus, test B was linearly equated to test A purely using the 300 students that took these papers as a pair (see Table 1). In other words, this equating was done using the single group design. Similarly, test C was equated to test A purely using data from the 250 students that took these tests as a pair. Very few students took tests D and A together. Therefore, in this case, chained linear equating of test D to test A was performed using test C as a link. From Table 1 we can see 180 students took tests C and D together, and 250 took tests C and A together meaning this likely provided the strongest single route to equate tests D and A.

The results of this analysis are shown in Figures 3 and 4. Figure 3 shows the bias of each method. This is estimated as the average error (not absolute error) of equating across the 500 replications. The results show that, except at the ends of the score range, each of the two methods were effectively unbiased in the scenario being studied.

Figure 2. Difference between linearly equated scores and true IRT equated scores for the illustrative example data

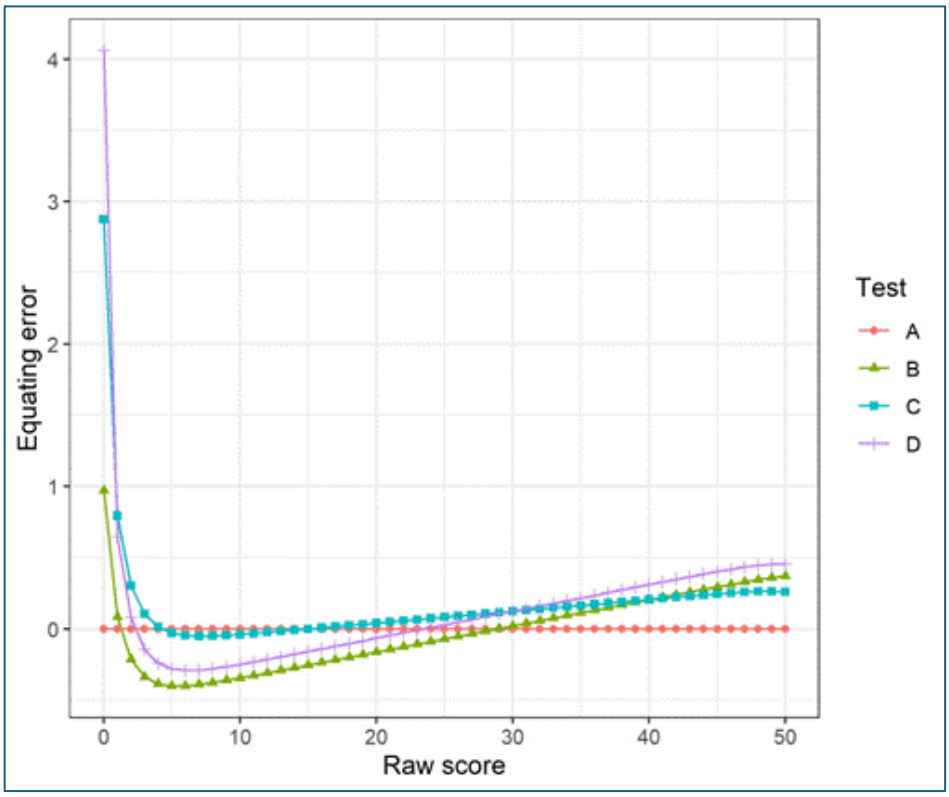


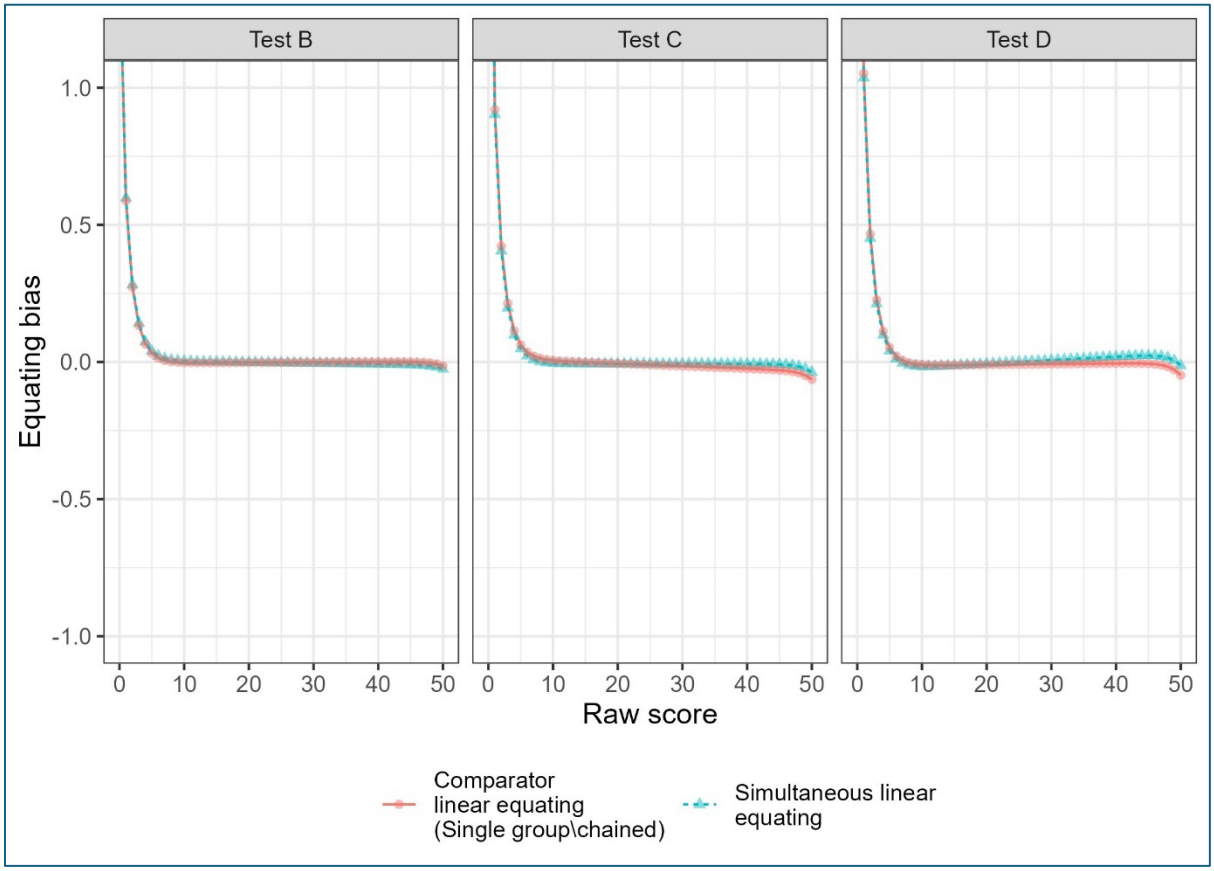
Figure 4 shows the standard error of equating for each method. This is calculated as the standard deviation of the equated scores across replications at each point in the score range. That is, it is a measure of the stability of the method across replications. As can be seen, for each test the simultaneous linear equating led to a lower standard error than relying purely upon equating via a single route.

These results are entirely as expected. Since both equating methods rely on similar assumptions, it is unsurprising that they display similar levels of bias. It is also unsurprising that simultaneous linear equating should have a lower standard error than the comparator approaches. After all, it makes use of all of the available data rather than restricting analysis to convenient subsets.

For these reasons, whilst we could repeat the analyses here with different simulation parameters (e.g., sample sizes), there is no need. Simultaneous linear equating will tend to display the same level of bias as single group or chained linear equating as it is effectively doing the same thing. Furthermore, it will always have lower standard errors than the comparators as it uses more of the data. As such, we can be confident that the approach suggested here will improve accuracy. That said, like any equating method, the more data we have available the more accurate it will become. The exact sample sizes needed for equating will vary depending upon the specifics of the scenario including, in particular, the strength of correlations between different test scores (Kolen & Brennan, 2004, p. 258). In practice, we might estimate the standard errors of equating using resampling techniques and check that they are low enough for results to be trusted. However, as a rough rule of thumb, using Kolen and Brennan's evidence of needing 250 students per test form for linear equating in the random groups design (Kolen & Brenna, 2004, p. 99), and noting that standard errors of equating can be considerably lower in other designs (Kolen & Brennan, 2004, p. 258), we would

recommend that a sample of size of 100 students taking each form is generally sufficient to allow simultaneous linear equating.

Figure 3. Bias of simultaneous linear equating and comparator equating techniques across different points in the score range

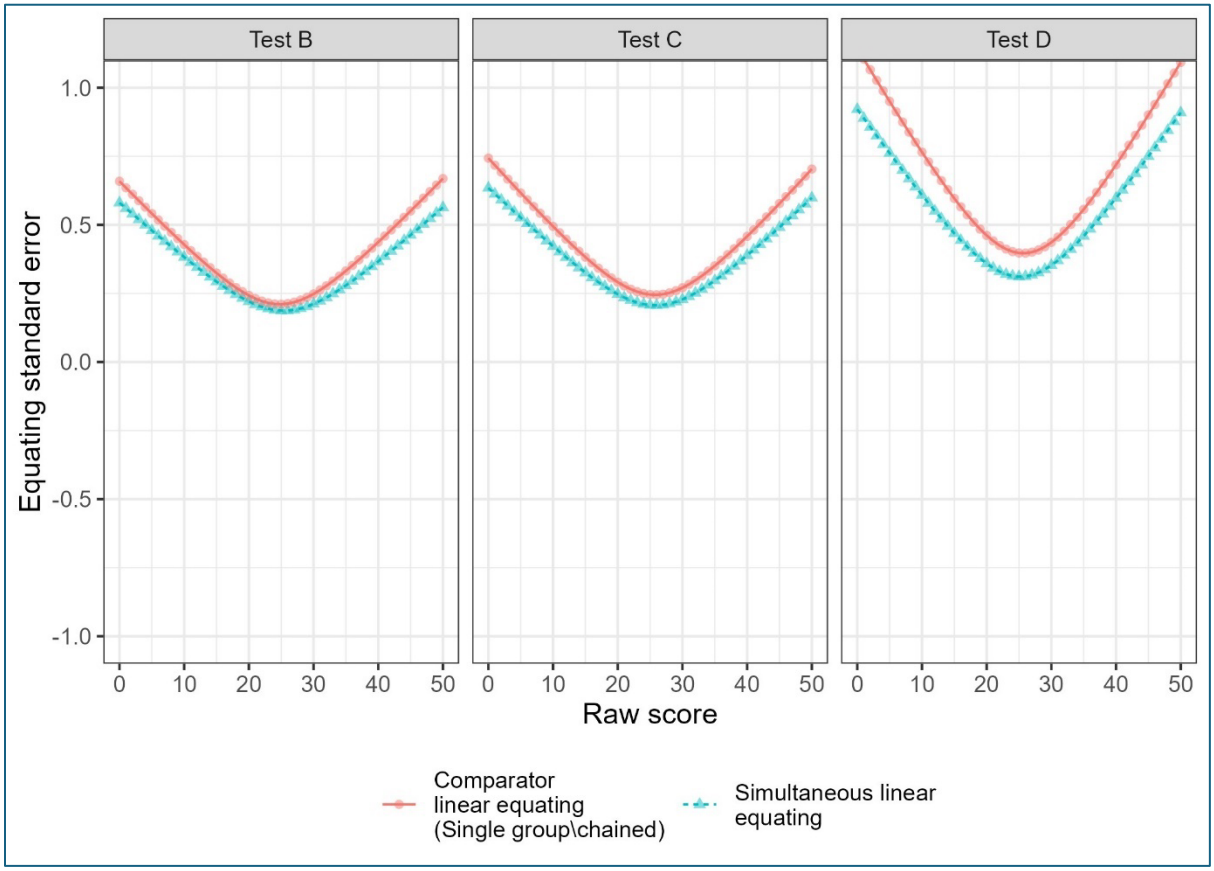


As shown in the example in Figure 4, depending upon the situation, improvements in accuracy may be of a rather small magnitude. Nonetheless, the advantage of simultaneous equating is that it uses all the data in a consistent manner. In contrast, the disadvantage of using a single route of equating is that other sources of evidence, not used in analysis, may give conflicting information and it is not clear how this information should be incorporated. For example, equating tests C to test A and (separately) test B to test A using only the students that did each pair may give a different picture of which scores on test C are equivalent to particular scores on test B than we would get by equating tests C and B directly. Thus, the advantage of simultaneous linear equating is not only the potential improvement in accuracy but also the consistent handling of all of the data rather than restricting to particular subsets.

As mentioned earlier, the method proposed in this paper can be seen as an extension of chained linear equating. Previous research (e.g., Puhan, 2010) has indicated that such approaches can produce good results in terms of having low bias and error. Furthermore, the use of linear equating in this current paper is intended to provide a pragmatic (and reasonably transparent) approach to equating in instances where IRT equating is either not possible or not desirable. Indeed, in the illustrative example, the available sample sizes are lower than those usually recommended for fitting a GRM model. Furthermore, fitting an IRT model would be made problematic by the fact that various hypothetically available test scores did not occur in the data. For

example, for test D, apart from 3 students with a score of 47, no scores greater than 41 appeared within the data.

Figure 4. The standard error of equating (i.e., standard deviation of results across replications) at each point in the score range



Incorporating weights in calculations

In the description of the methodology above, all test entries in any pair have equal weight in calculations. That is, the N matrix is based on the number of students doing any particular pair of tests. This is fine if all students take the same number of tests (e.g., two optional tests, or one test and one anchor) but may have unintended consequences if students have taken different number of tests (for example, if different students are part of different educational programmes with overlapping but different assessment requirements).

In such a scenario, a student entering 10 tests will ultimately have much more weight in calculations than a student entering only 2. We may prefer that rather than each entry having equal weight, each student should have equal weight. This concept is applied both in Lawley's description of the Kelly method (Kelly, 1971) and also in the average mark scaling method in Western Australia (TISC, 2015).

To achieve each student having equal influence, we may weight the data from different students, giving less weight to those students that took greater numbers of tests. Specifically, we define the number of tests entered by student k as m_k and then calculate the weight for each student as $w_k = \frac{1}{m_k - 1}$. Note that students that have only entered one test are not used in calculations.

From here, calculations proceed as before but, rather than n_{ij} representing the number of students taking both test i and test j, it should be calculated as the sum of the weights for the students that take both tests.

Note that, if all students have entered the same number of tests (as is true in the illustrative example), then the inclusion of weights will make no difference to results.

Assumptions of the method

At this point, it is worth noting the assumptions underlying the simultaneous linear equating method. In this respect, the existing method with the greatest similarity is chained linear equating. Much like chained linear equating, the method only relies on comparisons of means and standard deviations of tests that are taken as a pair. In the case of chained linear equating, each pair would consist of a test of interest and an anchor. In simultaneous linear equating, the pairs could consist of any combination of tests (including anchors). Similarly, like chained linear equating, simultaneous linear equating assumes that the differences in the difficulties of different tests can be entirely adjusted for using linear transformations. Like any form of linear equating, this effectively assumes that, if taken by the same population of students, the distribution of scores on the tests being equated would have the same shape (for example, the same skewness and kurtosis).

Finally, both chained equating and simultaneous linear equating fundamentally assume that the equating relationship between two tests is invariant to the population where we wish equating to be applicable. For example, if we assume that if a score of (say) 49 out of 60 on a maths test is equivalent to a score of 53 out of 75 on another maths test for the group of students who took both, then they must also be equivalent for all sets of students. In fact, it is well known that equating relationships are rarely population invariant in practice. According to Kolen (2004) "Equating theory indicates that test form equating is population dependent, except under highly restrictive conditions" (p. 11). However, Kolen (2004) also notes that if test forms have similar content, difficulty and reliability then population invariance holds approximately. Furthermore, an assumption of population invariance is by no means unique to simultaneous linear equating. As well as being part of the underlying assumption of all forms of chained equating, the same assumption is also at play in any application of true score equating based upon the Rasch or any other IRT model.

On a more general level, like chained equating, the effectiveness of simultaneous linear equating is dependent upon the reliability of the test scores used in calculations. In essence, the method relies on each of these scores providing a reliable measure of student ability. Just as chained equating would provide inaccurate results if based on an unreliable anchor, so the mapping between scores on two tests implied by simultaneous linear equating will be less accurate if the (other) tests that provide the link between the two are unreliable. That is, for the technique to work, we require *all* scores used in analysis to be fairly reliable. As with any form of equating, exactly how reliable we require each test to be will depend upon circumstances. In a situation where the groups taking different tests are similar in any case, we can probably get away with using a weak anchor. If we're trying to account for larger differences between groups, a stronger anchor is required.

One implication of the above paragraph is that it is not sensible to apply the technique described in this paper to a data set looking at pairs of performances on individual dichotomous test items rather than whole test scores. This would be rather like performing chained equating via each individual item in an anchor test in turn (and then averaging results) rather than using overall anchor test scores. Since individual item scores are usually much less reliable indicators of ability than overall test scores, such an approach would not be an effective way of using the available data.

An example of using the method to handle multiple anchors

The approach described in this paper may also be useful in scenarios where multiple different anchor tests are available for use in equating. For example, imagine that we have three different test versions that are to be taken by students on three separate occasions. In order to allow scores from the different test versions to be placed on a common scale we include an anchor section in each one. However, perhaps due to concerns over test security, we wish to use different anchor sections with different students. That is, we might be concerned that an anchor section for an earlier test occasion could be leaked to later students, making it unsuitable for use as an anchor. To provide some contingency against this eventuality, we may wish to use multiple alternative anchor sections. For example, we might ensure students taking test version 1 also complete one of anchor section A or anchor section B. Students taking test version 2 also complete anchor section A or anchor section C. Finally, students taking test version 3 also complete anchor section B or anchor section C. We have several possible ways of equating each test version to each other version via an anchor. For example, we can equate test versions 1 and 2, either via anchor section A or anchor section B. There are also other alternatives involving longer chains of equating that go via test version 3.

As with our earlier scenario, what we require is a single method that will coherently summarise all of the evidence (from all of the anchors) to provide a way of placing scores from each of the test versions on the same scale. Table 7 provides an example of what our data set might look like in such a scenario. This is based on simulated data where each main test version has a maximum score of 50 and each anchor section has a maximum score of 10. Note that the anchor scores were simulated as external anchors. However, whether anchors are internal or external does not change the mathematics of the approach. Since the only pairs of tests that we are concerned with consist of a single test version and an anchor, we have only included these pairs in Table 7. For brevity we have also only shown these differences in one direction.

Table 7. Summary of scores on main tests and anchors taken by different groups of students

Test 1 (main test)	Test 2 (anchor)	N taking pair	Mean score on test 1	Mean score on test 2	SD of scores on test 1	SD of scores on test 2
Main V1	Anchor A	100	26.66	5.91	8.60	2.19
Main V1	Anchor B	100	27.02	6.30	7.82	2.23
Main V2	Anchor A	100	26.74	5.96	8.24	2.11
Main V2	Anchor C	100	25.78	6.13	7.84	2.31
Main V3	Anchor B	100	24.98	5.94	8.29	2.36
Main V3	Anchor C	100	26.21	6.10	9.32	2.56

Given the information in Table 7, it is now possible to complete all the steps for simultaneous linear equating described earlier. This would begin with the creation of an N matrix with 6 rows and 6 columns (one row and one column for each main test and for each anchor). Since no students took more than one main test version, and no student took more than one anchor, this matrix would contain zeros in the rows and columns relating to these pairs.

If all the steps are completed correctly then the parameters associated with each test are as shown in Table 8. The proposed transformation could be applied to the summaries of scores in Table 7 (multiply the means and SDs by the relevant β values and add the relevant α values to the resultant means only). Doing this results in the arithmetic mean of the transformed scores on each main test equalling the arithmetic mean of transformed scores on the anchor test. Furthermore, the geometric mean of the standard deviations of the transformed scores on each main test will equal the geometric mean of the standard deviations of the

transformed scores on each anchor. Such checks are made easier by the fact that every pair was taken by the same number of students meaning that no weighting is necessary.

Table 8. Final coefficients from simultaneous linear equating for example with multiple anchors (rounded to 3 decimal places)

Test	α value	β value
Main V1	0.000	1.000
Main V2	-0.198	1.027
Main V3	0.949	0.978
Anchor A	3.408	3.968
Anchor B	4.957	3.470
Anchor C	4.887	3.522

Summary and discussion

Equating is an important part of ensuring fairness and equity in assessment amongst students that have taken different test versions. This paper has described a way of linear equating scores on a set of tests to the same scale that takes account of all the routes by which performance on the tests can be linked. This may be useful in any context where students have taken different combinations of tests, and we wish to place scores from each of them all on a common scale. For example, the technique may be useful for improving test security in a higher education setting. Specifically, we might ask each student to complete an assessment with two test sections and, furthermore, ensure that the test sections taken by students sitting next to each other are different. Alternatively, the technique may be useful in instances where we have multiple anchor tests and different students take different ones and we wish to summarise equating evidence across all of them.

At present, the typical approach to the kinds of scenarios described in this paper is to make use of techniques from item response theory. However, such techniques are not always straightforward to apply. For example, our sample sizes may be too small to confidently apply complex IRT models, whilst simpler models (such as the Rasch model) may not be appropriate for our particular set of items. Furthermore, whilst the purist approach to IRT assumes that each anchor measures the same set of skills as the main test, in practice, due to practical constraints anchor tests may be restricted to consist of only items of a particular type. For example, it may be that it is not possible for an anchor to include extended tasks such as essay writing. With these various considerations in mind, it is good to have an alternative non-IRT approach to equating in our toolkit. Further discussion of why we might prefer not to use IRT can be found in the introduction of Livingston (2014).

Our empirical example showed that the suggested approach is more accurate than relying on a single source of data of equating such as only using data from students that had taken both of a pair of tests or relying upon a single route from chained linear equating. Specifically, as might be expected given that simultaneous linear equating makes use of all available data, it is associated with lower levels of sampling variation than restricting to one particular type of evidence.

The disadvantage of the approach proposed in this paper is that it is restricted to a linear form of equating. This may not always be appropriate particularly if the test forms being equated are of very different levels of difficulty. Addressing this shortcoming could be an area for further research.

Received: 2/9/2024. **Accepted:** 12/22/2024. **Published:** 1/9/2025.

Citation: Benton, T. (2025). Simultaneous linear equating for scenarios with optional test versions or across multiple alternative anchors. *Practical Assessment, Research, & Evaluation*, 30(1). Available online: https://doi.org/10.7275/pare.2087

Corresponding Author: Tom Benton, Cambridge University Press and Assessment. Email: tom.benton@cambridge.org

References

Holland, P. W., & Strawderman, W. E. (2011). How to average equating functions, if you must. *Statistical models for test equating, scaling, and linking*, 89-107.

Howie, S. J., Long, C., Sherman, V., & Venter, E. (2008). *The role of IRT in selected examination systems. Pretoria, South Africa.* Umalusi Council of Quality Assurance in General and Further Education and Training. https://www.umalusi.org.za/docs/research/2009/irt_examination.pdf.

Kelly, A. (1971). The relative standards of subject examinations. *Research Intelligence*, 1(2), 34–38. https://journals.sagepub.com/doi/pdf/10.1177/003452377601600104

Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement*, 41(1), 3-14. https://doi.org/10.1111/j.1745-3984.2004.tb01155.x.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York, NY: Springer.

Livingston S. A. (2014). *Equating test scores (without IRT).* Educational Testing Service. https://www.ets.org/Media/Research/pdf/LIVINGSTON2ed.pdf.

Puhan, G. (2010). A comparison of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement*, 47(1), 54-75. https://doi.org/10.1111/j.1745-3984.2009.00099.x.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, 27(2), 133-144. https://doi.org/10.1111/j.1745-3984.1990.tb00738.x.

TISC (2015). *Average marks scaling.* Tertiary Institutions Service Centre Ltd. Downloaded from https://www.tisc.edu.au/static-fixed/statistics/misc/average-marks-scaling.pdf on 2nd January 2024.

Appendix A. Further details on illustrative calculations

This appendix provides further details of the steps in the application of the Kelly method and simultaneous linear equating for the illustrative example in the main paper. In particular, it provides examples of the matrices that need to be constructed to enable calculations.

Step 1 of the description of the Kelly method explains how we create the N matrix. Table A1 shows what the N matrix would look like for the illustrative example data in Table 1.

Table A1. N matrix for data in the illustrative example.

	A	B	C	D
A	-560	300	250	10
B	300	-460	100	60
C	250	100	-530	180
D	10	60	180	-250

Step 2 of the description of the Kelly method explains how we create the D matrix. Table A2 shows what the D matrix would look like for the illustrative example data in Table 1.

Table A2. D matrix for our illustrative data.

Test	D value
A	617
B	-16
C	-358
D	-243

Step 5 of the description of the Kelly method explains how we need to delete the first row and column from the N matrix and then invert the result. Table A3 shows the result of this process for the illustrative example. In our description the matrix in Table A3 is denoted $N\_^{-1}$.

Table A3. Inverse of the N matrix in the illustrative example after removing the first row and column.

	B	C	D
B	-0.002535589	-0.00091	-0.00126
C	-0.000906834	-0.00282	-0.00225
D	-0.001261462	-0.00225	-0.00592

Many of the same steps are used in simultaneous linear equating. For example, the N matrix and, as such, the $N\_^{-1}$ matrix are identical to those used in the Kelly method. One new step in simultaneous linear equating (step 3 of process for simultaneous linear equating) is the creation of the L matrix relating to differences in the logs of the standard deviations of scores on each test. Table A4 shows the L matrix for our illustrative examples.

Table A4. *L* matrix values for illustrative example. Calculations shown to 4 decimal places.

Test	*L* value
A	-17.9212
B	-4.2097
C	13.8528
D	8.2781

All of the other calculations steps for the illustrative example are shown within the main body of the paper.