Cambridge Assessment

Comparing the simplified pairs method of standard maintaining to statistical equating

Research Report

Tom Benton
Euan Cunningham
Sarah Hughes
Tony Leech

13th October 2020

Author contact details:

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

benton.t@cambridgeassessment.org.uk

http://www.cambridgeassessment.org.uk

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

Cambridge Assessment is committed to making our documents accessible in accordance with the WCAG 2.1 Standard. We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, contact our team: ARD@cambridgeassessment.org.uk.

If you need this document in a different format contact us telling us your name, email address and requirements and we will respond within 15 working days.

How to cite this publication:

Benton, T., Cunningham, E., Hughes, S. & Leech, T. (2020). *Comparing the simplified pairs method of standard maintaining to statistical equating.* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Contents

Introduction

In the context of examinations, the phrase “maintaining standards” usually refers to any activity designed to ensure that it is no easier (or harder) to achieve a given grade or above in one year than in another. Various methods by which comparative judgement (CJ) can be used in standard maintaining have been suggested over the years. In England, particular attention has been paid to a method suggested by Bramley (2005), which uses the Bradley-Terry model to analyse the results of a CJ study using scripts from two different test versions. The results of this method produces a measure of performance (a CJ “measure”) for each script based on which other scripts it was deemed superior to, and which it was deemed inferior to, over a number of pairwise comparisons[1]. Crucially, these CJ measures are located on the same scale for each of the two different tests, thus providing a mechanism to map the original (non-CJ) scores from one test on to equivalent scores on the other. This mapping process has been used primarily for the purpose of identifying scores on new versions of a test that can be deemed equivalent to the grade boundary scores (or cut-scores) that have been set on the previous version of the test. A large-scale trial of this method has been run by Ofqual, and is described fully in Curcin et al. (2019).

Building on this earlier CJ work, Benton (2019) recently proposed a method now referred to as *simplified pairs* that allows us to map scores between test versions, but without the intermediary step of fitting a Bradley-Terry model. The removal of this intermediary step brings with it the notable benefit of eliminating the need for each script to be included in several comparisons. Indeed, such a need is central to previous methods, like that of Bramley (2005), where, in order for the method to be employed effectively, it is essential that each script be included in enough comparisons to allow the calculation of reliable CJ-based measures of scripts quality. According to Verhavert et al (2019), this is likely to require at least 10 comparisons per script. In contrast to this, by allowing that each script can be included in only a single comparison, the simplified pairs method offers a significant practical advance on these previous CJ methods: it enables us to include a substantially larger number of scripts in the study – thus increasing the scope of the study to generate accurate results – without imposing any additional time burden upon the judges who are taking part. Indeed, Benton (2019) suggested we may halve the amount of time required from judges whilst achieving the same level of accuracy as previous CJ approaches.

While the simplified pairs method thus has clear potential to improve the CJ process, its accuracy needs to be established. Accordingly, the primary aim of the research reported here was to evaluate its accuracy experimentally in a context where the relative difficulty of the two test forms being equated was known in advance from statistical equating. However, before presenting the results of this experiment we first provide a detailed description of how a simplified pairs study and associated statistical analysis is conducted.

[1] For Bramley’s study, judgements were collected as ranks and converted to pairs for analysis.

1. The method of simplified pairs

The aim of any simplified pairs study is to calibrate the scores[2] (that is, the marks) from two assessments onto the same scale. To simplify the subsequent description, we will call these two assessments form A and form B. Furthermore, throughout our description we will assume we want to identify the scores on form B that are equivalent to some existing grade boundaries (or cut scores) on form A.

Designing a simplified pairs study

The data from a simplified pairs study consists of the decisions from a number of judgements about which of a pair of scripts is superior allowing for the difficulty of the tasks they were each responding to. In every case the pair should consist of one script from form A and one from form B. Each decision is made by an expert judge (e.g. a subject expert) reviewing the two scripts side by side.

Our aim throughout analysis will be to explore the relationship between the scores awarded to the scripts from each form in a pair and the probability of form B being judged superior to form A. More specifically, we will want to answer the following key question:

> *Suppose a script on form A has been awarded a score of x_A. How many marks would a script from form B need in order to have a 50% chance of being judged superior?*

We denote the answer to the above question as x_B. For example, if x_A is the grade boundary on form A, then x_B would be the estimated grade boundary on form B.

Since our aim is to relate the scores on the two test forms to the probability of the script from form B being judged superior, we should design the study such that a range of scores from both test forms are included. However, in common with other applications of CJ to standard maintaining we exclude scripts with scores of zero as if two candidates have answered nothing correctly then there is no basis for judging between them. Similarly, we exclude scripts with scores equal to the maximum available as if candidates have answered everything correctly then there is no obvious basis for judging between them. To date, most CJ studies relating to standard maintaining tend to restrict attention to scripts where between 15 and 95 per cent of the maximum available score has been achieved.

It is also important to ensure that a wide range of score differences between the form A and form B scripts are included across different pairs. Typically, we would ensure that the score differences between form A and form B scripts in the same pair range from at least -20 to +20 per cent of the maximum available score. Pairs with very different scores are not compared as this is (usually) a waste of effort (we're confident of what the result would be without getting expert judges to spend their energy on the matter). However, it is important to include a sufficiently wide range to accurately estimate the relationship between mark differences and the probability of a script from one version being judged superior to a script from another.

Whereas CJ standard maintaining studies involving construction of a latent scale usually use around 40 scripts from each test form, a simplified pairs study can include several hundred

[2] Throughout this article, by "scores" we mean the raw total of the item scores across a whole assessment.

scripts from each test form. An idealised simplified pairs study will use each script just once. This represents the best use of judge time as we get the maximum amount of new information from every judgement they make. Nonetheless, the simplified pairs methodology presented here can be used even if we need to repeatedly reuse the same scripts across different pairs. However, in reusing scripts, the efficiency gains of using simplified pairs over other approaches to using CJ for standard maintaining will be much reduced.

Similarly, note that although the data is analysed in pairs it need not be collected this way. As demonstrated by Bramley (2005), we may instead ask judges to rank order a number of scripts within a pack and convert these rankings back into pairs. The methodology presented below is largely unaffected by this alternative data collection method except requiring a small adjustment to the way in which confidence intervals are calculated. This will be discussed later.

Figure 1 shows the full details of an example data set from a simplified pairs study. Form A and form B were two Sociology assessments with a maximum available score of 75. A total of 289 pairs were judged by 5 separate judges. The design of the study ensured a reasonably even spread of scores within the score range from 15 to 68 within each form. The total scores of the scripts within a pair differed by up to 15 in either direction. As can be seen, broadly speaking, the Form B script was more likely to be chosen in pairs where it had the higher score.

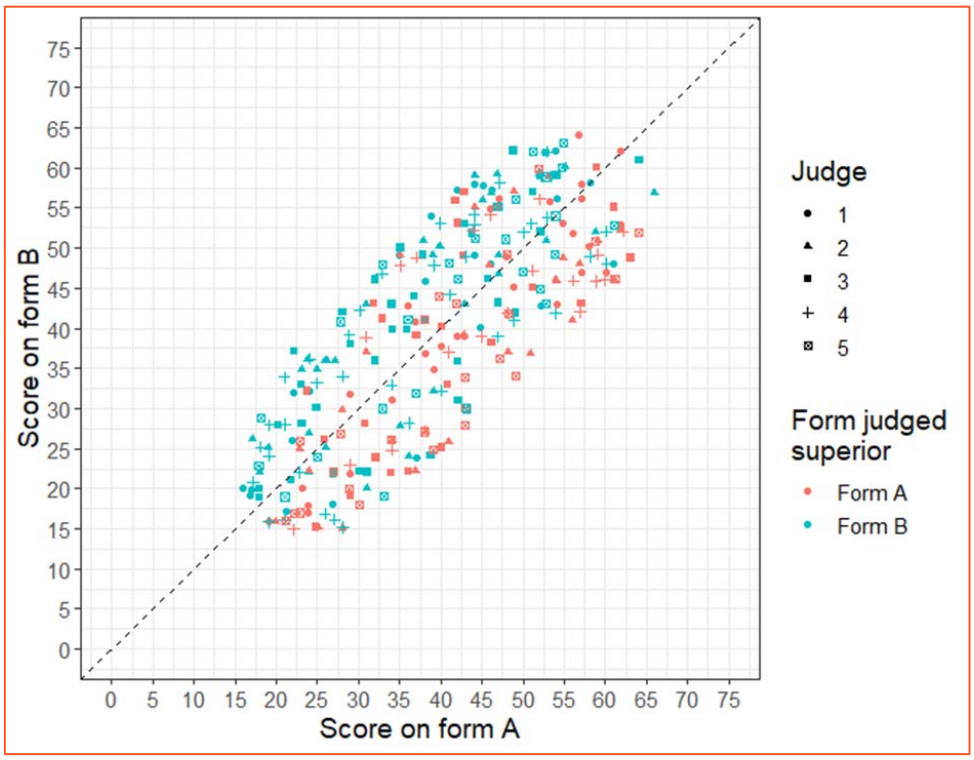

Figure 1: Example data from a simplified pairs study. All scores were whole numbers but some jitter has been added to the plot to allow all the pairs in the data set to be seen. Dashed line represents line of equality.

Estimating the overall difference in the difficulty of two tests

The main aim of analysis is to find, for each score on form A (denoted x_A), a score on form B (denoted x_B) representing an equivalent level of performance. The simplest model to approach this question is to assume that there is a fixed difference in the difficulty of the two tests such that for any x_A:

$$x_B = x_A + d \qquad (1)$$

Thus, in this simplest form of simplified pairs analysis, the goal reduces to finding the value of *d*. That is, the number of additional score points needed on a form B script for it to have at least a 50% chance of being considered superior to a form A script. Note that *d* can be either positive (suggesting form B is easier than form A) or negative (suggesting form B is harder than form A). The idea of a fixed difference between equivalent score points is also used in the method of mean equating used in formal test equating (Kolen & Brennan, 2004).

From equation (1) it can easily be seen that $d = x_B - x_A$. Thus, the analysis can focus entirely on the difference in the marks. A simple way to approach this is shown in Table 1. The table shows the number of pairs in the study with each score difference, the number where form A was judged superior and the number where form B was judged superior. Cells are highlighted in red to highlight score differences where the form A script was judged superior more often than not, and cells in blue highlight score differences where the form B script was usually judged superior. Viewing the data in this way immediately suggests that form B was roughly 2 marks harder than form A; at most score differences greater than -2 form B was usually chosen, whereas at most score differences below -2 form A was usually chosen.

Table 1: Numbers of pairs where form A and form B judged superior at different score differences

Score difference (B-A)	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Number of pairs	10	8	8	7	8	9	11	18	8	10	8	10	9	6	8	9	7	11	9	10	10	9	10	16	11	10	10	7	7	8	7
Number where form A judged superior	9	7	4	4	5	8	5	11	3	9	3	8	5	3	3	3	5	4	4	3	0	3	2	8	2	0	3	1	1	3	0
Number where form B judged superior	1	1	4	3	3	1	6	7	5	1	5	2	4	3	5	6	2	7	5	7	10	6	8	8	9	10	7	6	6	5	7

A more formal method of analysing this data is to use logistic regression. For the ith pair of scripts judged by the jth judge we denote the difference between the score awarded to the form B script and that awarded to the form A script as d_{ij}. We set $y_{ij} = 0$ if the judge selects the form A script as superior and $y_{ij} = 1$ if they select the form B script. The relationship between y_{ij} and d_{ij} is then modelled using the usual logistic regression equation:

$$P(y_{ij} = 1) = \{1 + exp(-(\beta_0 + \beta_1 d_{ij}))\}^{-1} \qquad (2)$$

Now we need to find the value of d_{ij} where $P(y_{ij} = 1) = 0.5$. That is, the score difference associated with a 50% chance of the form B script being judged superior. If we denote the estimated coefficients of the logistic regression model as $\widehat{\beta_0}$ and $\widehat{\beta_1}$ then by setting the left hand side of equation (2) to 0.5 and rearranging, the estimated difference in difficulty is given as:

$$\hat{d} = \frac{-\widehat{\beta_0}}{\widehat{\beta_1}} \tag{3}$$

The standard error of $\hat{d}$ can be easily derived using the multivariate delta method. Specifically, we define the above as a function of the parameters:

$$\hat{d}(\widehat{\beta_0}, \widehat{\beta_1}) = \frac{-\widehat{\beta_0}}{\widehat{\beta_1}} \tag{4}$$

Then the gradient *G* of this function is given by:

$$G = \frac{d\hat{d}(\widehat{\beta_0}, \widehat{\beta_1})}{d\boldsymbol{\beta}} = \begin{pmatrix} \frac{-1}{\widehat{\beta_1}} \\ \\ \frac{\widehat{\beta_0}}{\widehat{\beta_1}^2} \end{pmatrix} \tag{5}$$

Then, using the delta method, the standard error of $\hat{d}$ is given by

$$SE(\hat{d}) = \sqrt{G^T \boldsymbol{V}(\boldsymbol{\beta}) G} \tag{6}$$

Where $\boldsymbol{V}(\boldsymbol{\beta})$ is the 2x2 parameter covariance matrix from the original logistic regression[3].

A graphical illustration of this process is shown in Figure 2 below. The x-axis in this chart represents a score difference between form A and form B. The blue line and points capture the proportion of observations at each score difference where the form B script was judged as being superior to the form A script it was compared to. As such, this line is simply another way of representing the data in Table 1. The solid red line shows how a smooth curve can be fitted through the jagged set of observations from individual score differences using logistic regression. The score difference where this line intersects with a probability of 0.5 of the form B script being judged superior provides our estimate of *d*. In the example presented here, the estimate of *d* is -2.6; indicating that form B is just under 3 marks harder than form A on average. The vertical solid purple line is used to mark this point on the horizontal axis.

The dashed red lines show the 95 per cent confidence interval for the fitted logistic regression. Looking at where these confidence intervals intersect with a probability of 0.5 would provide one way of putting a confidence interval around the estimated difference. In an extreme example, if the estimated relationship (the solid red line) is very weak it is possible that either (or both) of these confidence intervals will never intersect with 0.5. We refer to this situation as “flatlining”

[3] All standard statistical software packages will produce this matrix as part of the output of logistic regression. The square roots of the diagonals of this matrix provide the standard errors of the regression parameters.

and this would indicate a complete failure of the CJ method in the sense that we would not produce any reliable confidence interval for the estimated difference in difficulty. Traditional approaches to using CJ in awarding can also “go wrong” in some sense in that the CJ measures of script quality can be found to either have low reliability or a low correlation with script marks (Bramley & Gill, 2010). However, some judgement is required to determine what is meant by “low”, and in either of these cases the analysis within the traditional approach can still be run and will produce a finite confidence interval for results. In contrast, within simplified pairs, the possibility of flatlining gives a clearer answer to the question of when the relationship between awarded marks and CJ outcomes is too weak to be of any use at all.

Although using the confidence intervals from the logistic regression provides one way of generating confidence intervals for this estimate, the delta method tends to identify a confidence interval with a narrower range. As such, this latter method is more efficient. The dashed purple lines in Figure 2 shows the confidence intervals derived using the delta method. In this example, they are very close to where the logistic regression confidence intervals intersect with a probability of 0.5; though this needn’t always be the case.

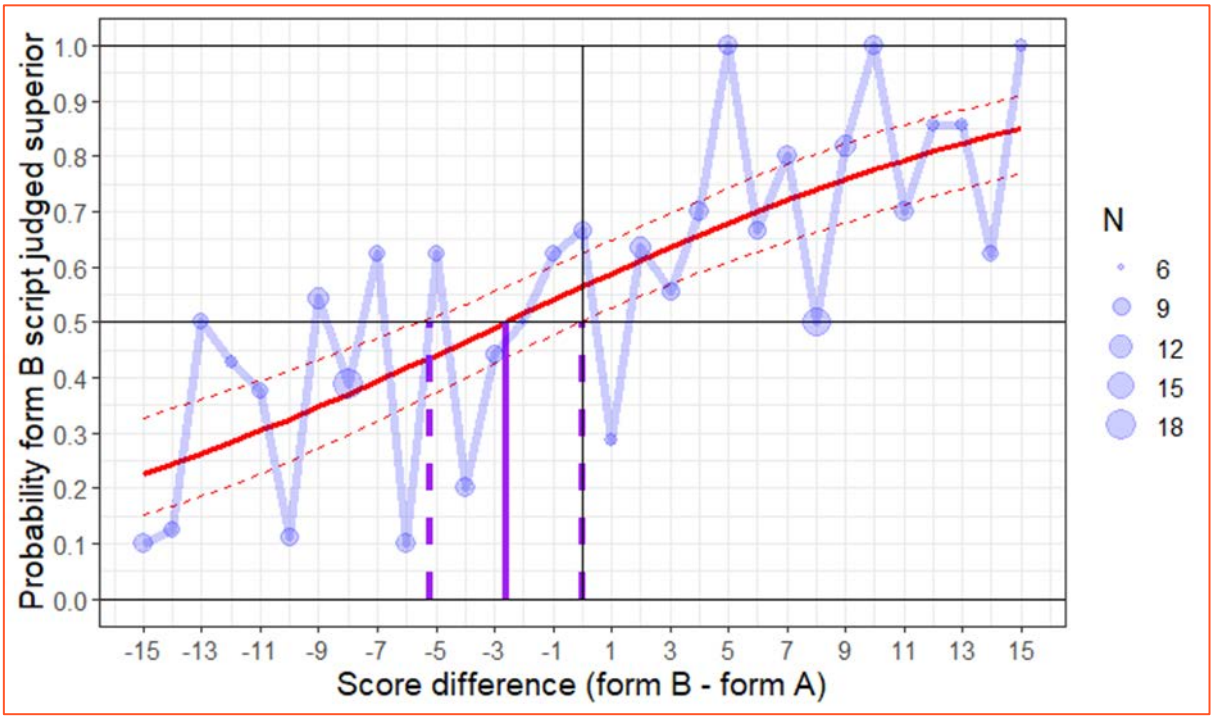

Figure 2: An illustration of the overall simplified pairs method.

Judge Fit

An important step in any CJ exercise of this nature is to check whether the decisions from individual judges look plausible. This is intended as a way of identifying any judges who may potentially have struggled with the task, misunderstood the instructions, or not taken sufficient care in their judgements. The judgements from any judges where the above risks are an issue may be removed from analysis.

In previous uses of CJ in standard maintaining (e.g. Curcin et al, 2019), judge fit is determined with regard to how well their judgements agree with what would be expected given the CJ measures of each script derived from the Bradley-Terry model. In particular, values of INFIT and OUTFIT (Wright and Masters, 1990) are calculated, and judges with values much greater than 1 may be removed and the data re-analysed without their judgements.

In contrast to this, within simplified pairs we focus purely on how the relationship between CJ decisions and the differences in the scores awarded to scripts within a pair varies across judges. As such, the method of exploring judge fit is more closely aligned with the way in which results will be used.

Table 2 shows a range of judge fit statistics from the example simplified pairs study. The table includes the INFIT and OUTFIT statistics common to other CJ studies. However, in this case they show the fit of each individual judge to the main logistic regression model (Figure 2) rather than to a Bradley-Terry model. The values are all reasonably close to 1 and so do not suggest any reason for concern about any of the judges. In reality, there is no particular reason why judges should all "fit" the same logistic regression model. The task they are engaged in involves making a decision about which of two students is superior based on reviewing performances on completely distinct assessments. As such, it requires a judgement from each of them about the relative difficulty of the tasks students have been set. There is usually no pre-defined "correct" answer about the level of difference in difficulty. It is a matter of each judge's opinion – indeed our interest in this opinion is the whole purpose of the exercise. Some of this difference of opinion is shown in Table 2. As can be seen, although pairs of scripts were randomly assigned to judges, one judge (judge 4) chose the form B script as superior more than two thirds of the time, whereas another (judge 1) selected the form B script less than half the time.

With the above in mind, rather than focus on the exact "fit" of each judge, it may be beneficial to simply focus on whether the relationships between mark differences are in the correct direction. That is, the more positive the difference in the scores awarded to the two scripts (i.e. the form B score minus the form A score), the higher we expect the probability of the form B script being selected to be. The strength of this association can be measured by the point biserial correlation. The values of this metric for each judge are shown in Table 2. The correlation coefficients range from 0.24 to 0.53. Given that we have just under 60 observations then this range of results is not unexpected. This can be confirmed by calculating confidence intervals for these correlations, for example using Fisher's R to Z transformation, to get a more precise idea of whether differences between judges are unusually high[4].

[4] In this instance this was done using the tool to calculate confidence intervals for correlations from https://www.psychometrica.de/correlation.html#confidence and using the average correlation from Table 2 of 0.39 and an n of 58.

Table 2: Example fit statistics for a simplified pairs study

Judge	Number of pairs judged	Proportion of pairs where form B chosen	INFIT	OUTFIT	Point biserial correlation
1	58	0.45	1.02	1.00	0.41
2	58	0.53	0.93	0.92	0.49
3	58	0.60	1.11	1.16	0.24
4	58	0.67	1.07	1.08	0.27
5	57	0.51	0.86	0.83	0.53

Another way to assess judge fit is the visual approach illustrated in Figure 3, The lines represent the results of fitting separate logistic regressions for each judge. These lines reveal some difference in opinion about the relative difficulty of the two assessments. More reassuringly, they show a positive relationship between mark differences and the probability of form B being selected for each judge. Judges 3 and 4 display the flattest curves which reflects the slightly lower point biserial correlations for these judges in Table 2. In the instance where any judges displayed completely horizontal or negative sloping lines, they could immediately be identified as not having undertaken the exercise as expected and could be removed from analysis.

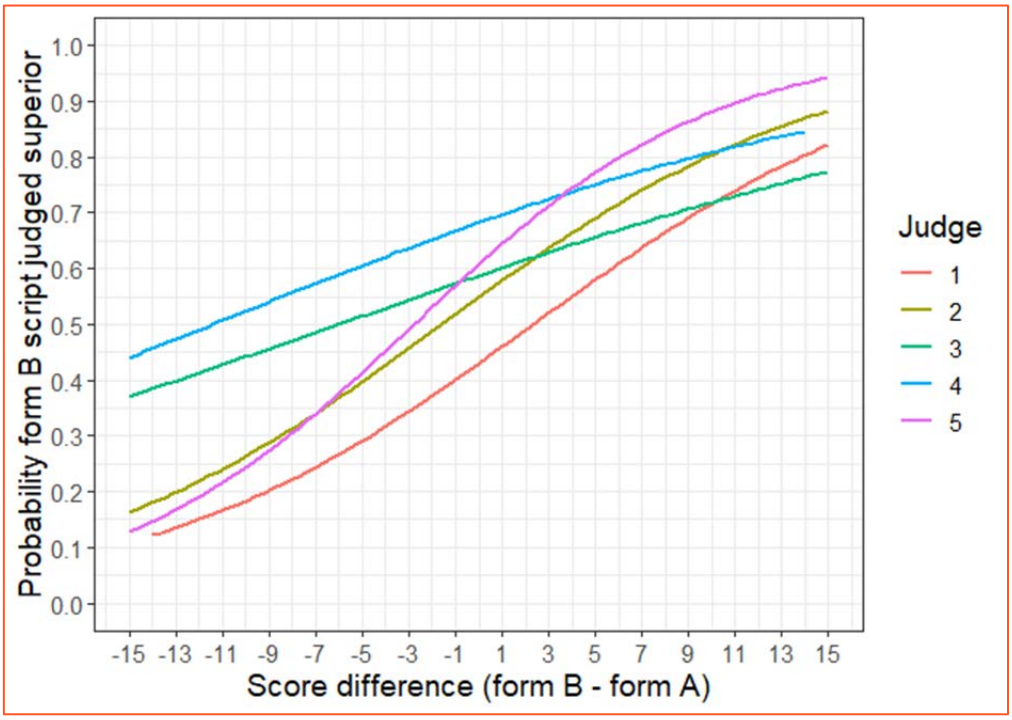

Figure 3: A visual exploration of judge fit using simplified pairs

Finding equivalencies across the entire score range

So far, the method of analysis we have described only identifies a single value estimating the relative difficulty of the two test forms. In many situations, we would like to ascertain far more detail about the relative difficulty of the two tests across the score range. For example, it's possible that two test forms may be of similar difficulty for low ability candidates, but that one test may be harder for candidates at the higher end of the ability range. We now briefly describe how we can explore this kind of detail using data from a simplified pairs study.

The simplest extension to equation 1 would be to look for linear mappings between scores of form A and scores on form B using an equation such as

$$x_B = cx_A + d \tag{7}$$

In the above equation the parameter *d* would now refer to an intercept denoting the score on form B equivalent to a form A score of zero. The parameter *c* would indicate the number of marks on form B that are equivalent to each additional mark on form A.

The trouble with the above formulation is that, in practice, it may yield implausible results, particularly at the extreme ends of the score distribution. For example, negative values of *d* would indicate some scores on form A relating to negative scores on form B – an impossibility. In practice, the above approach to mapping often also leads to the confidence intervals around x_B being unhelpfully wide at each end of the range.

To overcome this issue, we apply a simple transformation to the scores:

$$x_B^* = ln\left(\frac{x_B}{max_B - x_B}\right) \tag{8}$$

$$x_A^* = ln\left(\frac{x_A}{max_A - x_A}\right) \tag{9}$$

Where max_A and max_B are the maximum available scores on form A and form B respectively. The transformation in equations 8 and 9 are similar to the log-odds transformations familiar from logistic regression and transform the score range from being from zero to the maximum available score to being from minus to plus infinity. The above transformations can easily be used in the context of simplified pairs because scripts with scores of either zero or the maximum are not included in the data. The above transformation could not be used in the context of equating more generally.

Note that to get from transformed scores back to scores on the raw scale we simply use the transformation:

$$x_B = \frac{max_B \exp(x_B^*)}{1 + \exp(x_B^*)} \tag{10}$$

With the above transformation in place, we now seek for a mapping between (transformed) form A scores and (transformed) form B scores of the following type.

$$x_B^* = cx_A^* + d \tag{11}$$

Note that equation 11 implies that scores of zero on form A will map to scores of zero on form B and that scores equal to the maximum on form A will map to scores equal to the maximum on form B. This assumption is familiar within equating methods such as true score IRT equating and circle-arc equating (Kolen & Brennan, 2004, Livingston & Kim, 2009).

To estimate the parameters of the equation 11 in practice we again use a logistic regression equation. Specifically, we now model the probability of form B being selected using the equation

$$P(y_{ij} = 1) = \left\{1 + exp(-(\beta_0 + \beta_1 x^*_{Aij} + \beta_2 x^*_{Bij}))\right\}^{-1} \quad (12)$$

Where x^*_{Aij} and x^*_{Bij} are the (transformed) form A and form B scores for the ith pair of scripts seen by the jth judge. For each possible value of x^*_A we wish to find the value of x^*_B such that $P(y_{ij} = 1) = 0.5$. This is given by the following formula

$$\widehat{x^*_B} = \frac{-\widehat{\beta_0} - \widehat{\beta_1} x^*_A}{\widehat{\beta_2}} \quad (13)$$

The standard error of $\widehat{x^*_B}$ can again be derived using the multivariate delta method. Specifically, we define the above as a function of the parameters and the fixed value of x^*_A:

$$\widehat{x^*_B}(\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}, x^*_A) = \frac{-\widehat{\beta_0} - \widehat{\beta_1} x^*_A}{\widehat{\beta_2}} \quad (14)$$

Then the gradient of this function with respect to the model parameters is given by:

$$G = \frac{d\widehat{x^*_B}(\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}, x^*_A)}{d\boldsymbol{\beta}} = \begin{pmatrix} \frac{-1}{\widehat{\beta_2}} \\ \frac{-x^*_A}{\widehat{\beta_2}} \\ \frac{\widehat{\beta_0} + \widehat{\beta_1} x^*_A}{\widehat{\beta_2}^2} \end{pmatrix} \quad (15)$$

Then, using the delta method, the standard error of $\hat{d}$ is given by

$$SE(\widehat{x^*_B}) = \sqrt{G^T \boldsymbol{V}(\boldsymbol{\beta}) G} \quad (16)$$

Where $\boldsymbol{V}(\boldsymbol{\beta})$ is the 3x3 parameter covariance matrix from the logistic regression defined in equation 12.

Using the above formulae, we can derive a confidence interval for x_B^* at every possible score value on form A. The transformation in equation 10 converts this confidence interval to a score range on form B itself. A plot showing how the form A scores relate to scores on form B for the example data set is shown in Figure 4. As can be seen, this more detailed analysis suggests that form B may be harder than form A at the lower end of ability whilst being of relatively similar difficulty towards the top of the range. Having said this, the confidence intervals also suggest that a flat difference in difficulty of around 3 marks (as suggested by the earlier analysis and displayed by the thick (double) red line in Figure 4) is plausible across the vast majority of the score range. Specifically, except at the very top and bottom of the score range, a constant difference of 3 marks sits within the confidence bands for the non-linear method.

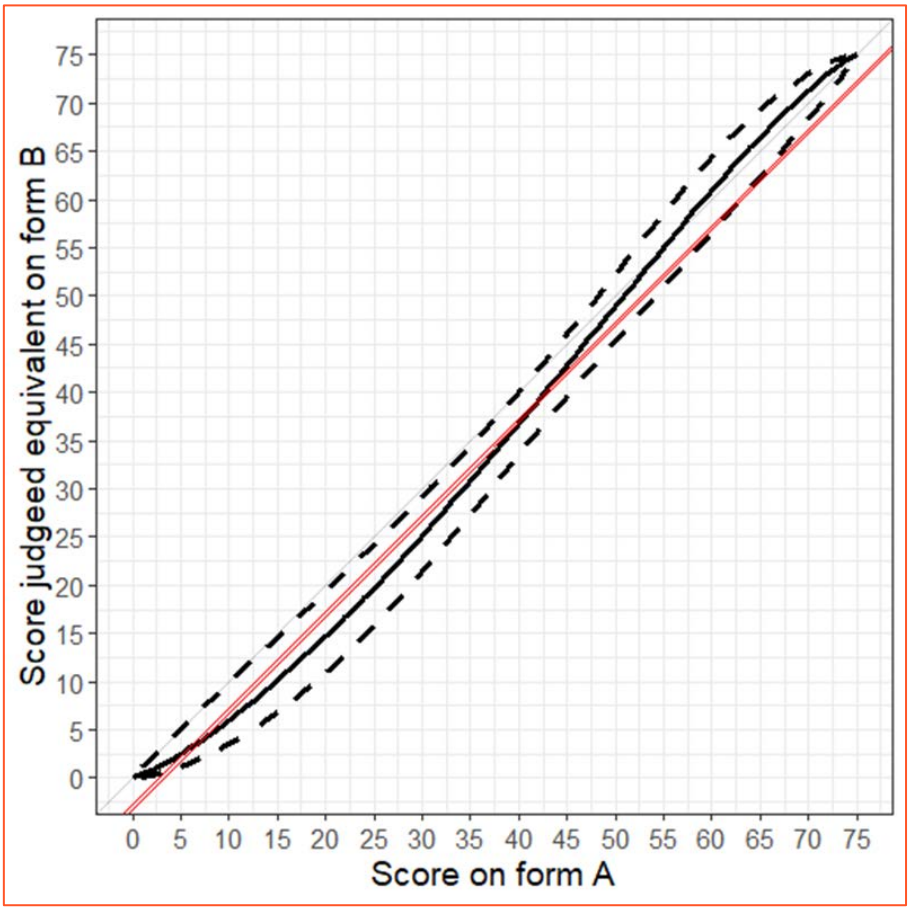

Figure 4: Mapping scores across the full range using simplified pairs. The faint grey line represents a line of equality. The thick (double) red line represents a constant difference of 3 marks between the two test forms.

Finally, for interest, we compare the grade boundaries that were actually set for form B in practice to those that would have been set based on the simplified pairs method alone. The results are shown in Table 3. As can be seen, the estimated grade boundary at grade A was very similar to that set in practice. At grade E the estimated grade boundary is a little lower than that used in practice, however, the estimate confirms that form B was harder than form A at this grade, and the confidence interval overlaps with the grade boundary that was set in practice.

Table 3: Comparing grade boundaries for form B from the simplified pairs method to those used in practice

Grade	Actual form A grade boundary	Actual form B grade boundary	Form B grade boundary from simplified pairs (95% confidence interval)
E	32	30	27.3 (22.9, 30.8)
A	53	53	52.6 (48.7, 56.6)

Adjusting calculations for different designs

So far, we have assumed that each script will be used in exactly one paired comparison. However, the above formulae can still be used if this is not the case. All that is required is to adjust the method of logistic regression. In particular, if the same scripts are used multiple times, the way in which $\boldsymbol{V}(\boldsymbol{\beta})$ is estimated should reflect the non-independence of different pairs. The bullet points below give details of how the logistic regression model should be fitted in difference scenarios.

- **Each script used in a single pair.** In this case every pair is independent of every other pair[5], and so ordinary logistic regression can be used.
- **Each script is used in a single pack producing several pairs**. For example, judges might rank packs of 6 scripts (3 from each form) yielding 9 cross-form comparisons[6]. These 9 pairs are clustered within a single pack and are not independent of one another. To address this, we may use any method of logistic regression that adjusts standard errors to account for clustering. In our analyses we have used the function *svyglm* from the R package *survey* (Lumley, 2004).
- **Each script from one form used in multiple pairs but scripts from the other form used only once**. This might occur if one form has an abundant availability of script evidence whereas the other does not. In this instance, we can treat all pairs as being clustered within scripts from the form used multiple times, and again adjust the standard errors accordingly.
- **Scripts from both forms used multiple times**. If scripts from both forms are used in many paired comparisons, we account for the impact of individual scripts using mixed effect (multilevel) logistic modelling. One set of random effects are included in the model to relate to the impact of individual form A scripts, and another set are included relating to the impact of individual form B scripts. Note that, since we never compare scripts within the same form, such a multilevel logistic model is fairly easy to set up. In practice, we have used the function *glmer* from the R package *lme4* (Bates et al, 2015) to fit this kind of model.

Once the logistic regression model has been fitted in the appropriate manner, analysis proceeds using exactly the same set of formulae as described earlier.

[5] If we ignore judge effects. These could theoretically be captured within the logistic regression model (possibly as random effects within a multilevel model) but this is beyond the scope of this report.
[6] Within-form comparisons are discarded.

2. Evaluating simplified pairs by comparison to equating

Having thoroughly described the method of simplified pairs, this section gives details of an experiment designed to evaluate whether the method is actually effective in identifying the relative difficulty of two tests.

In order to make such an evaluation of the simplified pairs method, it is necessary to compare the outcome of this method with an empirical estimate of the difference in the difficulty of the two tests, such as would be provided by statistical equating (Kolen & Brennan, 2004). In other words, we need to compare the difference in difficulty between two assessments as indicated by the simplified pairs methods, with 'known' differences in difficulty that can be derived from statistical equating of actual assessment results. Crucially, however, in order to establish these 'known' differences in difficulty, it is essential that at least one of the following three conditions be satisfied in the study: that some of the students take both of the assessments being equated; that some of the same items are present in both assessments; or, that some students taking either assessment version also take a common anchor assessment.

Using the first of these criteria, the present study will compare the results of the simplified pairs method to the outcome of statistical equating on two assessments that were undertaken by the same set of students. In this way, the study will satisfy one of the conditions needed for generating a 'known' difference. Note that, beyond its direct relevance to simplified pairs, this study will also overcome the limitations of previous studies into the use of CJ in standard maintaining in general (e.g. Bramley, 2005; Curcin et al, 2019; Gill, Bramley, & Black, 2007). These earlier studies evaluated the application of CJ by simply comparing grade boundaries that would be set based on the CJ method to those that were used in practice. While this type of comparison does provide some evidence of the plausibility of the results from the CJ method, it does not - owing to its failure to satisfy any of the three conditions (referred to above) for statistical equating - generate a 'known' difference to which the results of CJ studies can be compared. With this is mind, the present study was purposively designed to allow direct comparison to a 'known' difference. This, in turn, provides the study with the capacity to directly evaluate the accuracy of the simplified pairs CJ method.

Moreover, beyond this aim of evaluating the accuracy of the simplified pairs method, the present study attempts to address a further, separate, concern: that raised by Baird (2007) that "examiners cannot adequately compensate in their judgements of candidates' work for the demands of the question papers" (page 142). In order that the truth of this claim can be investigated, this study has deliberately chosen two assessments that were known in advance by the researchers to have notably different levels of difficulty. Using assessments with a notable disparity in difficulty allows us to observe whether the judges, across the totality of their judgements, are able to account for the demands of the questions in their judgements of corresponding answers. It is important to note here that the difference in difficulty between the papers (while known to the researchers) was not be known to the judges prior to them reading the scripts (the respective marking schemes for the two assessments give no indication of this difference). Consequently, in the case where the judges are indeed able to accurately identify this difference, their being able to do so could not be attributed to their prior knowledge of these disparities. Rather, their accuracy would attest to their ability to judiciously account for differences in difficulty between assessments and would thus provide some evidence that Baird's concern is not always justified.

Choice of assessments

The first step in this study involved selecting appropriate assessments to be used for the analysis. The assessments chosen for this study were from GCSE English Literature exams and required students to write a single essay in response to a prompt. The specific essay questions are provided below:

1) **Assessment 1 (A1)**. Maximum available score of 40. Students were given 45 minutes to answer the following question.

   "In what ways does Priestley make this such a striking and revealing moment in the play? Remember to support your ideas with details from the passage and the rest of the play."

2) **Assessment 2 (A2)**. Maximum available score of 24. Students were given 90 minutes to answer the question below and one other question. Only scores from the question below were used in analysis.

   "How does Orwell make this moment in the novel so shocking? Remember to support your ideas with details from the passage as well as the rest of the novel."

There were several reasons why the above assessments were chosen for analysis. Firstly, a large sample of students completed both these assessments[7] thus allowing us to complete formal statistical equating as a comparator to the simplified pairs method.

Secondly, as is indicated from the similarity of text of the questions, the two assessment were both designed to measure the same assessment objectives. This was further confirmed by reviewing the mark schemes. These indicated that, for both of the two assessments, equal weight should be given to the following two assessment objectives (AOs):

1) Respond to texts critically and imaginatively; select and evaluate relevant textual detail to illustrate and support interpretations.
2) Explain how language, structure and form contribute to writers' presentations of ideas, themes and settings.

The identical weighting of assessment objectives made it easier for the judges to make a clean comparison that is based solely on the relative quality of the two essays. That is, the comparison process is not confounded by the judges having to account for the two assessments having different assessment objectives.

The final reason for selecting these particular assessments relates to the aim outlined above: the notable difference in their level of difficulty. This will be illustrated later.

Collecting paired comparison data

Once the assessments had been chosen, 6 experienced markers were recruited from OCR to act as judges of the exam scripts for the simplified pairs method. These judges were each issued with 50 pairs of scripts, with each pair containing one script from assessment A1 and one from A2. The judges were then asked to indicate "which is the better essay in response to the question?". To help them make judgements, each judge was provided with the original mark schemes from the examinations. Judgements were made on-screen using the Cambridge Assessment Comparative Judgement Tool (https://cjscaling.cambridgeassessment.org.uk/). No

[7] On separate days within the same exam session.

marks or other annotations were visible to the judges on any of the scripts. As such, judges' decisions were based purely on the content of students' responses.

By design, the pairs shown to judges were compiled such that, across the whole set of pairs, there was a wide range in mark differences between the A1 and A2 scripts. In other words, there were some pairs composed of scripts with identical (or very similar) marks as a percentage of the number available, some pairs containing scripts with radically different marks (e.g. a 40% difference between the two scripts), and other pairs with mark differences somewhere in between. Ensuring this wide range of mark differences within the pairs was of vital importance: only by including judgements from across a wide range of mark differences do we allow ourselves scope to accurately identify the specific mark difference where the two scripts are deemed to be of equal quality. Details of the scripts that were included in the pairs are displayed below in Table 4.

Table 4: Descriptive information on scripts from each paper included in the simplified pairs study

	Assessment – A1	**Assessment – A2**
Number of scripts	300	300
Mean score	23.49 (out of a possible 40)	13.47 (out of a possible 24)
Standard Deviation	6.66	5.09
Mean score (as % of max)	58.7	56.1
SD of score (as % of max)	16.7	21.2
Minimum score included	7	4
Maximum score included	36	22

Once each judge had made their decisions on of their 50 allocated pairs, these decisions were analysed using the method described in section 1. The results of this analysis were subsequently compared to the 'known' difference generated by the statistical equating of the actual scores given to each assessment. This comparison subsequently provided us with a means of observing the accuracy of the results of the simplified pairs method.

It is important to note here that only scripts marked by team leaders were included in the study. This criterion was implemented because, as would be expected, team leaders are generally the most experienced and competent markers. Consequently, their marking is, on average, likely to be more reliable than that of the wider pool of examiners. Ensuring this maximum reliability in the marks of the included scripts was done to help maximise the reliability of the study with respect to its central aim. Note that, for consistency, equating was also conducted using only scripts marked by team leaders.

Another element of interest in this study relates to understanding the relationship between judge speed (i.e., how long does a judge take between first being presented with the scripts and making a final judgement) and the accuracy of their judgement. Accordingly, the time taken by each judge to make each decision was recorded for all their judgements.

Results

Overall difference in difficulty

Table 5: Results from the statistical equating of the actual scores of pupils taking the A1 and A2 assessments. All pupils who took either test and those who took both tests are shown separately.

	All Pupils		Pupils taking both tests	
	A1	**A2**	**A1**	**A2**
Number of pupils	1638	1429	145	145
Mean score (as % of total)	66.28	56.89	65.17	55.89
SD (as % of total)	13.06	17.13	12.57	16.49
Difference in means	-9.38		-9.28	
SE of difference in means	0.56		1.18	
Confidence interval for difference in means	[-10.47, -8.29]		[-11.59, -6.97]	

To begin with, we show the actual empirical difference in the difficulty of the two assessments. Table 5 shows the results both for all pupils who took either one of the assessments and for those pupils who took both assessments. Starting with those pupils that took both assessments (the right-hand side of Table 5), we can see that A2 was substantially harder than A1. For the same set of pupils, the difference in mean scores was around 9 per cent of the maximum number of marks available. In other words, assessment A2 is about 9 per cent harder than assessment A1 overall.

Table 5 shows a similar difference in means if we base analysis on all pupils (the left-hand side of Table 5). In other words, these are students who completed one of the assessments (and were marked by a team leader) but not necessarily the other. This suggests that the pupils taking both tests are fully representative of the wider body of pupils who took either one of them. As a consequence of this, we were able to use the data from the wider body of pupils in the equipercentile equating method, which is outlined later. This was clearly advantageous given that the equipercentile method requires as large a volume of data as possible, and that being able to use around 1,500 pupils from each exam is a substantial improvement upon having to rely upon the 145 pupils who took both of them.

Simplified pairs results

Having determined using empirical equating that A2 is harder than A1 by about 9 per cent, we next generate an estimate of the difference in difficulty using the data from the simplified pairs study. This analysis is shown in Figure 1.

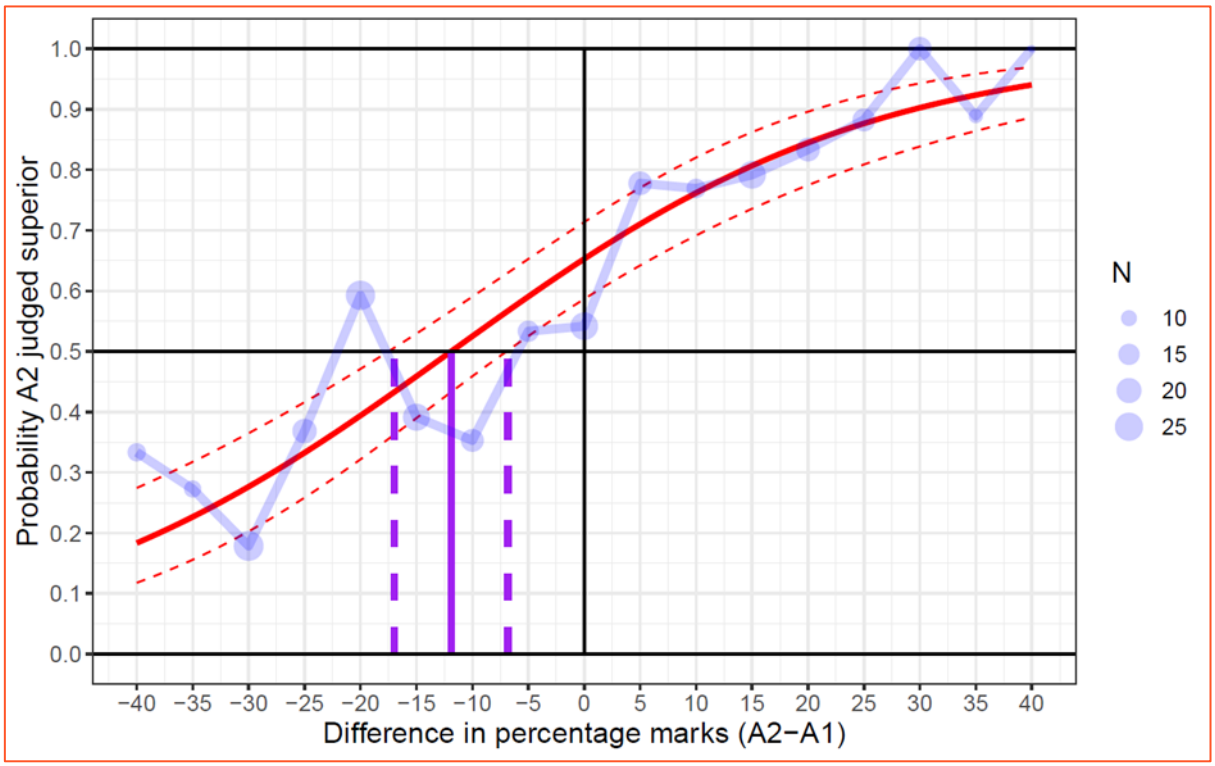

Figure 5: Graphical depiction of the results of using simplified pairs to gauge the relative difficulty of two assessment versions.

Figure 5 depicts the relationship between the difference in percentage mark awarded to two scripts in a pair, and the probability of the A2 script being judged superior. Accordingly, the blue points on the graph represent the proportion of pairs with a given difference in percentage marks between the A2 and A1 script where the A2 script was selected as being superior. It is worth noting at this stage the use of percentage difference (as opposed to absolute difference) in marks between the two scripts. This is to account for the difference in total available scores of on the two assessments (A1 is scored out of a possible mark of 40; A2 out of a possible mark of 24); a difference which – because it means that a single mark is not of the same worth on both scripts - renders any comparison based on absolute marks difficult to interpret. Also note that, to improve readability on this chart (but not for the analysis more generally) percentage differences in marks represented by the blue dots have been rounded to the nearest 5 per cent. Larger points depict mark differences where a larger number of pairs were included in the study.

Before going any further, it can be seen immediately that judges tended to infer for themselves that A2 was harder than A1. For example, where both scripts had been awarded a similar percentage of marks (e.g. between -5 and +5 per cent of one another), more often than not judges selected the A2 script as being superior. Thus, the results from simplified pairs agree with those from empirical analysis with regard to which of the two assessments was more difficult.

The solid red line running through Figure 5 represents the outcome of a logistic regression analysis. The x-axis value that corresponds to the point at which this line crosses 0.5 probability

constitutes the percentage mark difference whereby both scripts have an equal chance of being judged superior. In this case, the line crosses the 0.5 probability mark at a percentage mark difference of -11.9 per cent. In other words, the simplified pairs method estimated that A2 was 11.9 per cent harder than A1; that is, an A2 script had an even chance of being judged superior to an A1 script even if it had been awarded 11.9 per cent fewer marks. A 95 per cent confidence interval for this value (the dashed vertical lines) indicate that the results from simplified pairs analysis are consistent with any actual difference in difficulty between -6.8 and -17.0 per cent.

With regard to the study's primary objective - the accuracy of the results generated by the simplified pairs method - the most noteworthy result of this comparison is that both simplified pairs and equating identified A2 to be significantly harder than A1 (9.3 and 11.9 per cent more difficult for the statistical equating and simplified pairs methods, respectively). Beyond these similar mean differences, the agreement between the methods is further reflected in the clear overlap between the confidence intervals for the mean difference indicated by the simplified pairs methods and the 'known' mean difference emanating from the statistical equating method. This tells us that the real difference in difficulty between the tests fits nicely within the range of possible differences in marks indicated by the simplified pairs method. Clearly, this agreement between the simplified pairs method and the 'known' /difference constitutes strong evidence in support of the capacity of the simplified pairs method to generate accurate results.

Judge fit

Before generating more detailed results on the relative difficulty of the two tests across the score range, we first review the data on the fit of the 6 judges involved in the study. Statistics on the decisions made by each of the judges are shown in Table 6.

Table 6: Displays figures which show judge fit and speed for each of the 6 judges.

Judge	Number of Pairs judges	Proportion of times selected A2	INFIT	OUTFIT	Point biserial Correlation between difference in percentage marks and selecting A2	Median time per judgement
1	50	0.58	1.03	0.96	0.46	3.63
2	50	0.56	0.81	0.74	0.62	1.62
3	50	0.68	1.15	1.22	0.33	1.53
4	50	0.58	0.90	0.82	0.60	4.04
5	50	0.52	1.11	1.10	0.42	2.53
6	50	0.62	1.06	1.00	0.44	8.60

A key point from Table 6 lies in the general agreement between the judges in relation to the proportion of times they selected assessment A2 to be of higher quality. Indeed, for the random selection of 50 pairs assigned to each of them, all 6 judges were more likely to select the script from A2 as being superior than the script from A1. This indicates some agreement across all the judges that A2 was a harder assessment.

With respect to the point biserial correlations, there was found to be no statistically significant differences between the values for any of the judges. This is important because it indicates a high level of consistency between the way in which judges handled the CJ task.

This agreement between the judges is further illustrated in graphical form in Figure 6. The most evident feature of this chart is the similar trajectories for the line of best fit representing each of the judges: all of the lines move from a lower probability on the left of the chart to a higher probability on the right of the chart. This tells us that there is general agreement among the judges concerning their probabilities of judging the A2 script to be superior across all the percentage differences shown in the chart.

Beyond this general trend, the most significant finding indicated here relates to all the judges having deemed the A2 assessment question to be harder than A1. This is reflected – as with Figure 5, above - in the observation that the points on the x-axis that align with each line of best fit crossing 0.5 probability, all lie to the left of 0. This, in effect, tells us that all the judges deemed the answers to be of equal quality when A2 had been awarded somewhere between 5 per cent and 25 per cent fewer marks, thus clearly implying that the A2 question was harder.

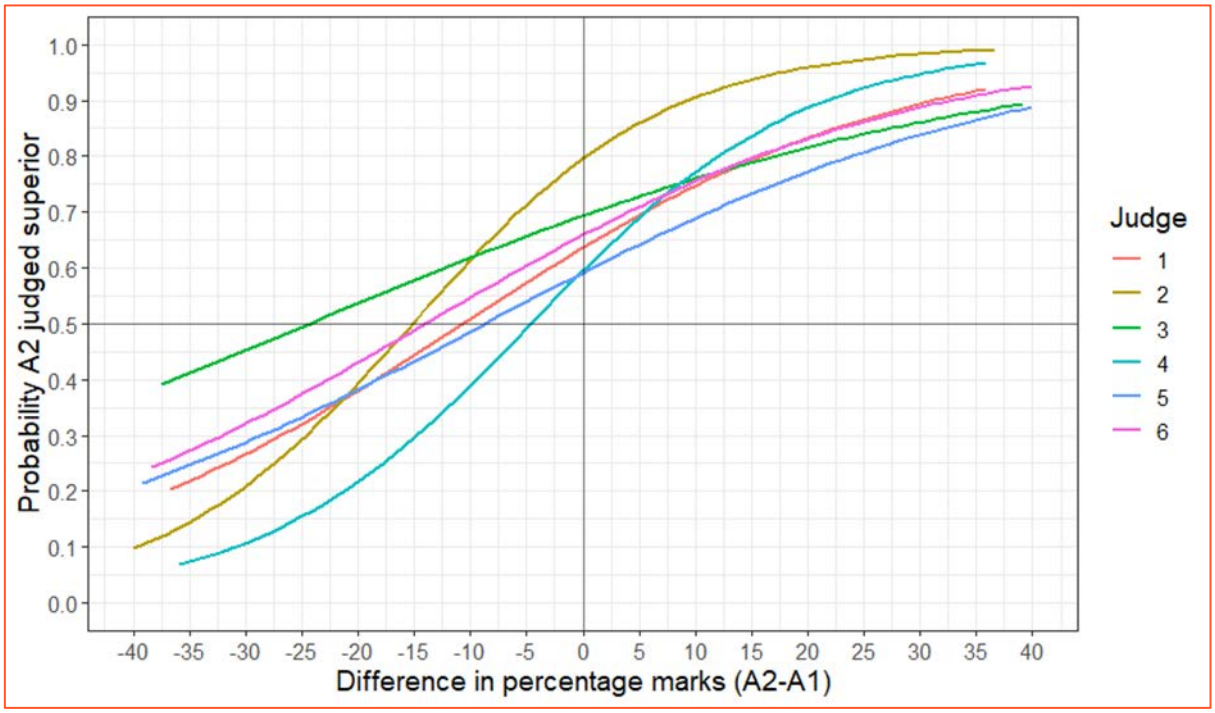

Figure 6: Graphical depiction of the relationship between difference in percentage marks and the likelihood of selecting A2 as being of superior quality for each judge.

Beyond this agreement between the judges, a second key finding displayed in Table 6 is the lack of any relationship between median time per judgement and the point biserial correlation. This is perhaps best reflected in the observation that the highest median time per judgement (8.6 minutes for judge 6) corresponds to only the fourth highest point biserial value. Similarly, further illustrating this lack of relationship, the two quickest median judgement times (1.53 and 1.62 minutes for judges 3 and 2 respectively) correspond to the lowest and highest point biserial values respectively. As such, these observations suggest that the speed at which the judges made their decisions had no meaningful relationship with their overall accuracy.

Equating across the score range

Both equating and simplified pairs methods were used to estimate the mapping from of scores from A1 to A2 across the available range. For empirical equating, this was done using the equipercentile method. Because equipercentile equating requires a large amount of data, data from all students was used for this analysis (as opposed to just those who had both essays marked by a team leader). As outlined above, the descriptive analysis found that there was no notable evidence of any systematic difference in these groups of students, meaning that we could use the larger data set without it having any substantive impact on the conclusions. For simplified pairs, this was done using the method described in section 1.

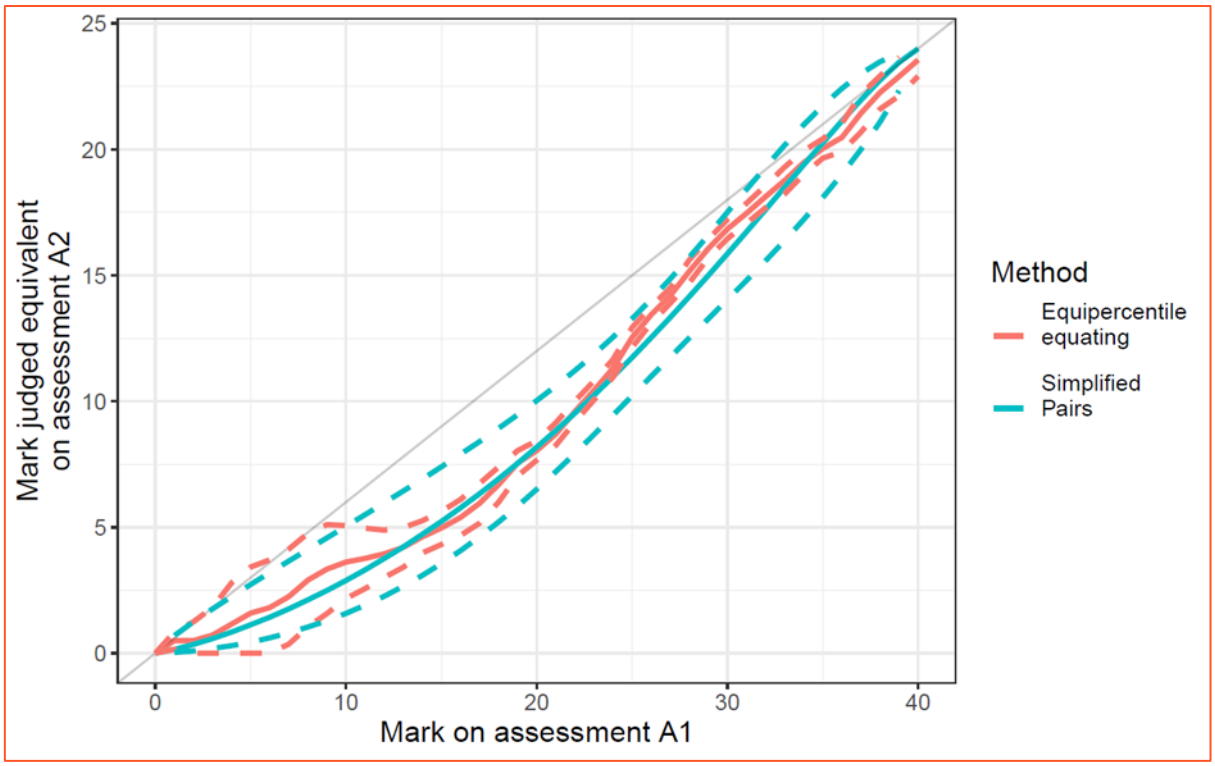

Figure 7: Graphical depiction of the results of the equating across the score range.

Figure 7 shows us the equivalent mark on A2 for any given mark on A1 for both the equipercentile equating and the simplified pairs method. As is displayed in the graph's key, the solid blue line represents the simplified pairs method and the solid red line represents the equipercentile equating. The dashed lines represent the corresponding confidence intervals for each of the methods.

For reference a straight diagonal line of proportional equivalence is provided in Figure 7. This broadly reflects the equivalencies we might expect from looking at the bands within the mark schemes. For example, for A1 the mark scheme indicated that a mark of at least 35 out of 40 (87.5 per cent) should be awarded for answers in the top band of marks that demonstrate "sophisticated critical perception in response to and interpretation of text", "cogent and precise evaluation of well-selected detail from the text" and "sensitive understanding of the significance and effects of writers' choices of language, structure and form". Exactly the same descriptors were used for the top band in assessment A2 with the expectation that such answers should be awarded at least 22 out of 24 (91.6 per cent).

The results from empirical equating again confirm that A2 tends to be harder than A1. For example, a mark of 20 out of 40 (50 per cent of the maximum) on A1 equates to a score of 8.2

on A2 (34 per cent). This difference in difficulty is particularly evident towards the middle of the score range.

The result of most interest from this chart is the close agreement between mapping identified using simplified pairs and the results of equipercentile equating. Across most of the score range the two lines themselves are very close, and, even where some differences are visible (for example at a score of 30 on A1), the “known” mapping of equivalent scores from equating is within the confidence interval for the simplified pairs method. This finding concerning the similar trajectories of the two methods clearly constitutes further evidence in support of the capacity of the simplified pairs method to produce results with a level of accuracy which is somewhere close to that of the statistical equating methods – even for two tests with differing levels of difficulty.

Conclusion

This primary objective of this study was to evaluate the accuracy of the results generated by Benton’s (2019) simplified pairs method by comparing them to a ‘known’ difference generated by statistical equating methods. The second objective was to address a concern raised by Baird (2008) relating to the supposed inability of examiners to account for varying levels of demand associated with different exam papers when they are deciding where to place grade boundaries.

With respect to the first of these objectives, the main finding to emerge from this study concerns the agreement between the results of the simplified pairs method and those of the statistical equating: both methods found assessment A2 to be notably harder than assessment A1. This finding clearly provides strong support for the potential accuracy of the simplified pairs method in determining the relative difficulty of two papers. The uncovering of such accuracy is likely - given the attendant time efficiency benefits of simplified pairs - to be particularly encouraging for those with an interest in enhancing the CJ process. As such, in line with this suggestion, we hope that this finding will serve as a catalyst for future research that seeks to replicate, and further develop where possible, the findings uncovered herein. Such research may constitute a key step on the path towards simplified pairs being established as a widely adopted means of operationalising CJ processes within awarding.

Relating to the second aim of the study, it appears that the judges were indeed able to account in their decisions for varying levels of demand associated with different assessment papers. This conclusion was rendered even more robust by the fact – as outlined above – that the judges had no prior knowledge of differences in difficulty between the two papers. As such, the judges’ accurately determining that A2 was a more difficult assessment clearly reflects their ability to reliably distinguish between answers that display differential levels of quality. Such a finding would, then, appear to lend support to the notion that there is indeed some value in retaining a prominent role for expert judges in the positioning of grades boundaries.

It may be evident from our results that expert judges were not able to match the levels of precision that can be attained through statistical equating. That is, the confidence intervals for simplified pairs are usually somewhat wider than those from equating. For this reason, where formal statistical equating is possible, it is likely to provide a superior source of evidence to expert judgements of script quality such as those based on CJ. In addition, formal equating provides the most relevant measure of the relative difficulty of tests in terms of how well pupils performed on each of them. However, given the difficulty of meeting the requirements for formal statistical equating in many practical situations, it is encouraging that a method based on expert judgement such as simplified pairs may provide a workable alternative form of evidence.

As mentioned above, the data required for formal statistical equating, by which we mean equating supported by common items or pupils across different test forms, cannot be obtained in many practical situations. However, other forms of statistical evidence may be available. For example, in England, setting grade boundaries often makes use of the "Similar Cohort Adage" described by Newton (2011) above saying "if the cohort hasn't changed much, then don't expect the pass rate to change much either". This adage is converted into a more formal statistical methodology by the comparable outcomes approach to awarding high stakes qualifications in England described by Benton & Sutch (2014).

Statistical evidence such as that provided by the comparable outcomes approach to awarding provides a healthy dose of common sense into the way we award grade boundaries. However, such an approach only accounts for evidence about pupil ability in previous exams and not for evidence about the quality of work they have produced (Benton & Bramley, 2015). Simplified pairs provides a way for this important second source of evidence to influence decisions about grade boundaries. Rather than replacing statistical sources of evidence, we would envisage it supplementing them. For instance, results from simplified pairs could be used to verify whether grade boundaries set using comparable outcomes are plausible or whether they have failed to identify substantial improvements in the quality of work produced by students. Ensuring that any changes in the quality of work are recognised and may help maintain public confidence in examination standards.

References

Baird, J.-A. (2007). Alternative conceptions of comparability. In P. E. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/487054/2007-comparability-exam-standards-f-chapter4.pdf.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Benton, T. (2019). *Maintaining standards using paired comparisons (but without Bradley or Terry).* Cambridge Assessment Internal Research Report.

Benton, T. & Bramley, T. (2015). *The use of evidence in setting and maintaining standards in GCSEs and A levels*. Cambridge: Cambridge Assessment.

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

Bramley, T. (2005). A Rank-ordering Method for Equating Tests by Expert Judgement. *Journal of Applied Measurement*, 6(2), 202–223.

Bramley, T. (2020). Comparing small-sample equating with Angoff judgement for linking cut-scores on two tests. *Research Matters: A Cambridge Assessment Publication,* 29, 23–27.

Bramley, T., & Gill, T. (2010). Evaluating the rank-ordering method for standard maintaining. *Research Papers in Education*, 25(3), 293-317.

Curcin, M., Howard, E., Sully, K., & Black, B. (2019). *Improving awarding: 2018/2019 pilots*. Ofqual report Ofqual/19/6575. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/851778/Improving_awarding_-_FINAL196575.pdf.

Gill, T., Bramley, T. & Black, B. (2007). *An investigation of standard maintaining in GCSE English using a rank-ordering method*. Paper presented at the British Educational Research Association Conference, 5–8 September in London, UK. https://www.cambridgeassessment.org.uk/Images/109760-an-investigation-of-standard-maintaining-in-gcse-english-using-a-rank-ordering-method.pdf.

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

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement*, 46(3), 330-343.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1), 1-19

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

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 26(5), 541-562.

Wright, B.D., & Masters, G.N. (1990). Computation of OUTFIT and INFIT Statistics. *Rasch Measurement Transactions*, 3(4), 84-85. https://www.rasch.org/rmt/rmt34e.htm.