Research Matters / 33
A Cambridge University Press & Assessment publication

ISSN: 1755-6031

Journal homepage: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/

How are standard-maintaining activities based on Comparative Judgement affected by mismarking in the script evidence?

Joanna Williamson (Research Division)

To cite this article: Williamson, J. (2022). How are standard-maintaining activities based on Comparative Judgement affected by mismarking in the script evidence? *Research Matters: A Cambridge University Press & Assessment publication, 33,* 80–99.

To link this article: https://www.cambridgeassessment.org.uk/Images/research-matters-33-how-are-standard-maintaining-activities-based-on-comparative-judgement-affected-by-mismarking-in-the-script-evidence.pdf

Abstract:

An important application of Comparative Judgement (CJ) methods is to assist in the maintenance of standards from one series to another in high stakes qualifications, by informing decisions about where to place grade boundaries or cut scores. This article explores the extent to which standard-maintaining activities based on Comparative Judgement would be robust to mismarking in the sample of scripts used for the comparison exercise. While extreme marking errors are unlikely, we know that mismarking can occur in live assessments, and quality of marking can vary. This research investigates how this could affect the outcomes of CJ-based methods, and therefore contributes to better understanding of the risks associated with using CJ-based methods for standard maintaining. The article focuses on the 'simplified pairs' method (Benton et al., 2020), an example of the 'universal method' discussed by Benton (this issue).

How are standard-maintaining activities based on Comparative Judgement affected by mismarking in the script evidence?

Joanna Williamson (Research Division)

Introduction

Providing evidence that can inform awarding is an important application of Comparative Judgement (CJ) methods in high-stakes qualifications. The process of marking scripts is not changed, but CJ methods can assist in the maintenance of standards from one series to another by informing decisions about where to place grade boundaries or cut scores. The research described in this article set out to increase understanding of the risks associated with this use of CJ. Specifically, the research explored how robust the outcomes of CJ-based awarding activities would be to mismarking in the script evidence.

In recent years, Ofqual has investigated various CJ methods for identifying cut scores in standard maintaining, and Curcin et al. (2019) reported the results of a large-scale pilot of several variants. This article focuses on the "simplified pairs" method (Benton et al., 2020), an example of the "universal method" discussed by Benton et al. (2022, this issue). Like other CJ methods, simplified pairs (SP) harnesses the information from paired comparisons in order to put the scores from two different assessments onto a common scale, but it does so without the need to fit a Bradley-Terry model and without the need to include individual scripts in multiple comparisons. Previous research has shown SP to be an efficient method, and comparisons with statistical equating have provided further evidence of the ability of SP to correctly determine the relative difficulty of two assessments, as well as for the ability of judges to account for the difficulty of different assessments in their comparisons (Benton et al., 2020).

In this article we explore the extent to which the SP method would be robust to mismarking in the sample of scripts used for the comparison exercise. In a particularly extreme case (e.g., if every script sampled from one assessment happened to be marked by a particularly harsh examiner, who undermarked by 10 marks), it is clear that the relationship estimated between scores on assessment A and assessment B would reflect this. More realistically, we know that mismarking can occur in live assessments, and quality of marking can vary, and it is therefore desirable to know how CJ-based awarding activities may be affected.

The simplified pairs method

In a typical application of SP for standard maintaining, there are two assessments (form A and form B), and existing grade boundaries or cut scores for form A. The SP method is applied in order to find the scores on form B that represent an equivalent level of performance to the grade boundary scores on form A. In the most straightforward case, we assume a fixed overall difference in difficulty between the two assessments, and the purpose of SP in this context is to find the difference d such that for scores x_A and x_B representing equivalent levels of performance on forms A and B respectively, $x_B = x_A + d$.

In an SP study, judges are asked to compare pairs of scripts, always comparing one form A script with one form B script, and decide which one is superior. Scripts from the extremes of the score distribution are excluded from the judging process, since where candidates have answered everything (or nothing) correctly, there is no basis for judging either to be superior. Scripts are sampled from a sub-range (e.g., those with scores between 20 and 90 per cent of the total available score), and paired for comparison in such a way that pairs include a wide range of score differences – Benton et al. (2020) recommend differences should span at least -20 to +20 per cent of the maximum available score. A typical SP study uses each script only once, to maximise the new information gained from each judgement, and can include several hundred pairs of scripts (Benton et al., 2020, pp. 5–6). This contrasts with typical CJ study designs, which would involve a smaller set of scripts from each assessment, that are then judged multiple times.

The overall difference in difficulty between form A and form B is found via logistic regression analysis of the judges' decisions. For the *i*th pair of scripts judged by judge *j*, the decision is represented by the outcome variable y_{ij}, where $y_{ij} = 0$ if the form A script is judged superior, and $y_{ij} = 1$ if the form B script is judged superior. The difference between the form A script score and form B script score is the independent variable and is notated d_{ij}, so that the modelled relationship is the following:

$$\text{log odds}\,(y_{ij} = 1) = \beta_0 + \beta_1 d_{ij}$$

where β_0 and β_1 are the intercept and slope in the linear relationship between score difference d_{ij} and the log odds[1] of the event $y_{ij} = 1$ (the event that the form B script is judged superior) in the logistic regression model. Since scores on form A and form B are considered equivalent when scripts with those scores have an equal probability of being judged superior, the overall difference d is d_{ij} where $P(y_{ij}=1) = 0.5$. Figure 1 gives a graphical example of this analysis: the blue markers and blue line show the percentage of script pairs at each mark difference where the form B script was judged to be superior to the form A script. The solid red line shows the fitted logistic regression line, and the dotted red lines show its 95 per cent confidence interval. The purple lines show d, the estimated overall difference in difficulty between form B and form A (in this example, 8 marks) and its estimated confidence interval.

1 The log odds or logit of the event $y_{ij}=1$ is $ln\left(\frac{p}{1-p}\right)$, where p is the probability that $y_{ij}=1$.

If the estimated relationship between script mark differences and judgement of superiority is very weak, the slope of the fitted logistic regression will be shallow and – in extreme cases – the SP analysis may result in 'flatlining'. This term describes a result such as that shown in Figure 2, where the dotted red lines representing the upper and/or lower 95 per cent confidence intervals for the logistic regression line fail to intersect the line y=0.5 at all. This indicates "a complete failure of the CJ method" (Benton et al., 2020, p. 8) – the relationship between script marks and judges' CJ decisions is so weak that it is impossible to produce a reliable confidence interval for the estimated difference in difficulty, meaning that the CJ method is unable to produce the evidence sought for awarding. The occurrence of mismarked scripts is a factor that can weaken the estimated relationship between mark differences and judgements of script superiority. It is, therefore, important to investigate quantitatively how robust SP analyses are to changes in the quality of marking in the selected script evidence.

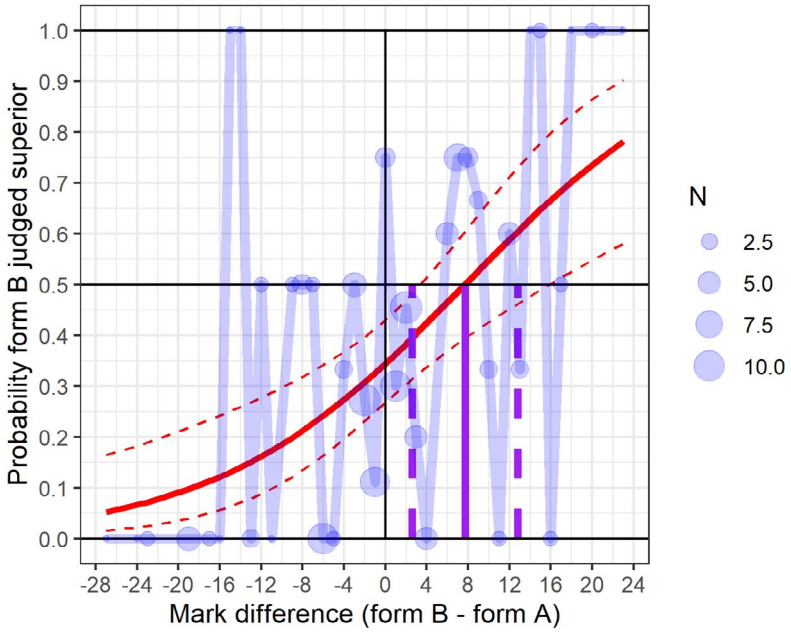

Figure 1: Example of a successful simplified pairs analysis.

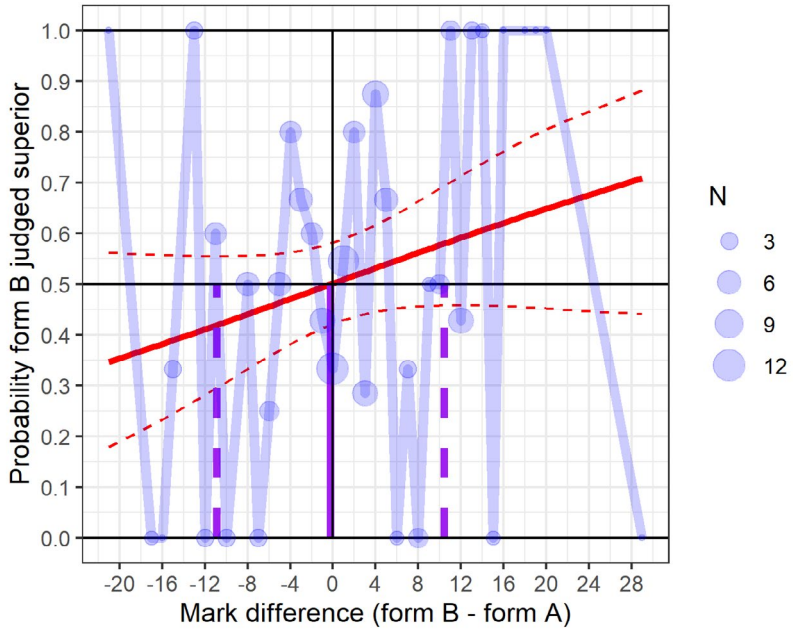

Figure 2: Example of a flatlining simplified pairs analysis.

Research overview

The overarching research question was addressed via three specific sub-questions, to explore robustness against mismarking in slightly different scenarios:

1. What is the impact on SP outcomes of large, one-off marking errors in the script evidence?
2. How many moderately sized marking errors can occur in the script evidence before SP analyses fail?
3. What is the impact on SP outcomes of a degradation in marking quality?

The first two questions were addressed using simulations based on data from previous SP studies, while the final question was addressed by simulating a large number of SP studies from scratch. All data simulation and analysis was carried out in R (R Core Team, 2021).

Impact of single large marking errors

The first set of simulations explored the impact on SP analyses of single large marking errors in the script evidence – such as could be introduced by a transcription error on a script (e.g., recording 13 as 31). These simulations were based on data from three real-life SP studies comparing different versions (forms) of various GCSE and AS level components.

To simulate a large one-off marking error in one of these SP studies, the mark difference for a single pair of scripts was manually altered (without changing the judge's decision) before re-running the SP analysis. To investigate the range of outcomes that such an error could cause, this was repeated, in turn, for every paired judgement in the dataset. For each SP study, we investigated four variants

of large errors, so each of the original SP studies therefore resulted in 4*n* simulated SP studies, where *n* was the number of pairs in the original study. The four types of large error were generated by altering the mark difference of the "marking error" script pair to one of the following values:

1. The largest positive mark difference between paired scripts in the study.
2. The largest negative mark difference between paired scripts in the study.
3. 70 per cent of the component maximum mark.
4. -70 per cent of the component maximum mark.

Figure 3 shows the distributions of estimated mark differences *d* for one of the original SP studies, under the four simulation conditions. The estimated difference between form B and form A in the original study (i.e., before deliberately introducing error) was -3.38 marks, and this value is shown by the vertical dotted line in each panel. The largest positive mark difference (form B script–form A script) between paired scripts in the original study was 15 marks, the largest negative mark difference was -15 marks, and the component maximum mark was 50 marks. A script pair selected as the "marking error" pair therefore had its mark difference altered to 15 marks, -15 marks, 35 marks and -35 marks in the four simulation conditions respectively. It is worth noting that the "error" introduced could therefore change the direction as well as the magnitude of the actual mark difference for the pair. It is clear from Figure 3 that the estimated mark differences from the simulated studies were all close to the originally estimated mark difference. While the shape of the distribution differed according to which particular large error was simulated, in all cases the estimated differences were very close to the originally estimated difference d in absolute terms. Although the values appear spread out along the x-axis, the scale is very fine-grained, and all estimates from the simulated studies were within a fifth of a mark of the originally estimated value for d.

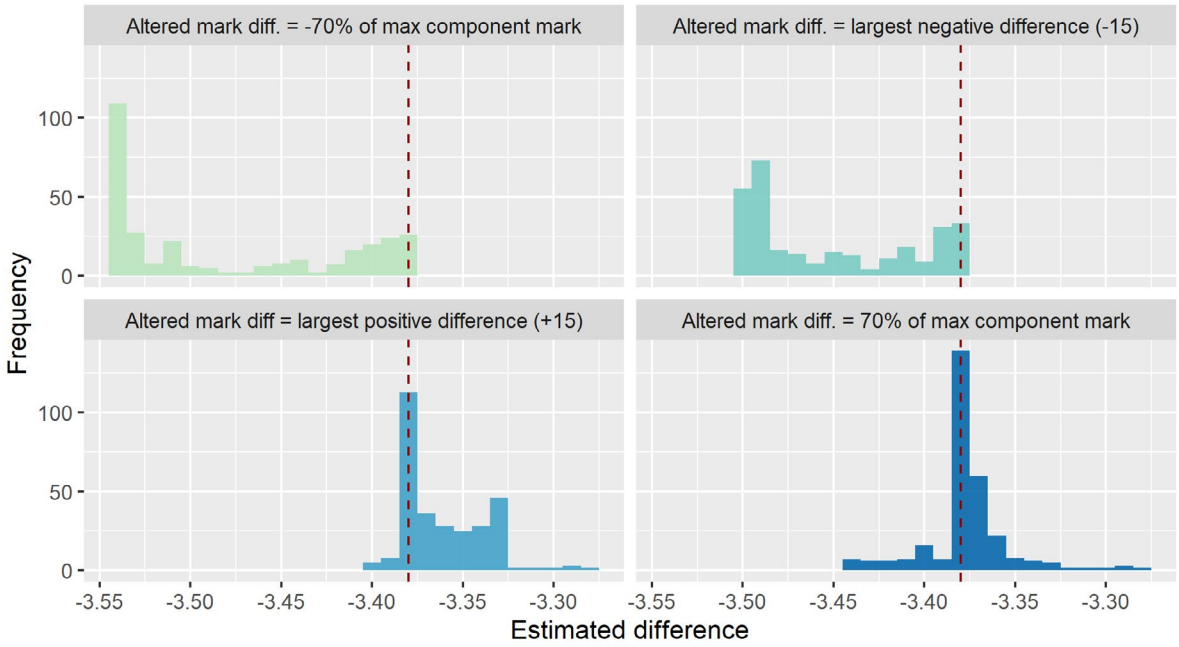

Figure 3: Estimated difficulty differences from simulating one-off large marking errors, Assessment 2 (reference line shows the original estimated *d*, before simulation of marking error).

For all four of the original SP studies, the estimated difficulty differences and associated standard errors changed little when a single large marking error was simulated. Table 1 summarises the range of outcomes from the simulated SP studies, in comparison with the original SP study results. In all cases, the estimated difference d was very close to the estimated difference from the original study (i.e., before simulating a large marking error), and the standard errors of estimates increased only moderately.

Table 1: Summary of single large marking error simulations in comparison with original studies.

Component	Max. mark	Pairs (*n*)	Original study *d* (SE)	Min *d* (SE) from simulated studies	Median *d* (SE) from simulated studies	Max *d* (SE) from simulated studies
Assessment 1 (English Language)	80	292	1.38 (1.14)	1.25 (1.10)	1.43 (1.16)	1.72 (1.36)
Assessment 2 (Maths)	50	300	-3.38 (0.49)	-3.54 (0.48)	-3.39 (0.49)	-3.28 (0.54)
Assessment 3 (Sociology)	75	289	-2.61 (1.33)	-3.04 (1.29)	-2.65 (1.35)	-2.43 (1.53)

How many marking errors can occur before SP fails?

The second set of simulations made use of data from the same three real-life SP studies (Table 1), but this time simulated the occurrence of multiple moderately large marking errors. The purpose of these simulations was to explore how many such errors could occur before the SP method broke down.

For each original SP study, the simulations were carried out as follows:

1. Randomly select *n* pairs from the original study.
2. Add a fixed "marking error" *e* to the observed mark difference for each of these pairs[2].
3. Re-run the SP analysis.
4. Retain/calculate:
    a. whether the analysis flatlined or not
    b. whether the 95 per cent confidence interval for the estimated overall difference d includes the value estimated in the original study (pre-error d)
    c. the difference between the estimated d and the value estimated in the original study (pre-error d).

These steps were carried out for two values of "marking error" *e*, equal to 10 per cent of component maximum mark, and -10 per cent of component maximum, and 1000 studies were simulated for each combination of conditions. For each *n* investigated, 6000 simulations were therefore carried out (3 original studies x 2 values of "marking error" x 1000 repetitions). The simulations were carried out at values of *n* from 10 up to 150. To give some context to the "marking errors" in this set of simulations, the value of 10 per cent of component maximum mark was chosen as a marking error that would be moderately large but of the magnitude that could occur in real life assessment scenarios. In the case studies presented by Ofqual (2014, pp. 31–32), for example, which analyse mark changes following enquiries about results for Geography A level and French A level, 1 per cent of mark changes made were of a magnitude of 10 per cent of the total raw marks, or larger.

As in the simulation of single large marking errors, the results showed that the SP studies were robust. Figure 4 shows the proportion of simulated studies for which the 95 per cent confidence interval for d contained the original (pre-error) estimate, according to number of marking errors introduced. The proportion only fell below 1 once the number of pairs of scripts containing marking error was large: around 50 pairs (out of 300) for Assessment 2, and only after 75 pairs for the other two studies.

Figure 5 shows how the estimated overall differences d deviated from the original (pre-error) estimates as more marking errors were introduced. The mean size of these deviations (expressed as percentages of component maximum) increased linearly, and at a moderate rate: for simulations adding marking errors to 50 pairs of scripts, the average deviation from original d was up to 2 per cent of the component maximum mark. The size of the deviations in d increased at a

2 This method (adding "error" to pairs of scripts selected on the basis of their original marks) results in a set of script pairings with a different distribution of mark differences than if scripts were selected on the basis of observed marks that already included large marking errors. Most obviously, the added "error" may cause mark differences to fall outside the original range of mark differences. The method used here should produce similar or worse outcomes (i.e., overestimate rather than underestimate risk).

higher rate when the sign of the marking errors introduced matched the sign of the original difference d. For Assessment 1, for example, the originally estimated overall difference was positive (1.38 marks), and the mean size of deviations in d increased faster for marking errors of +10 per cent than for marking errors of -10 per cent. The results show that, across all cases studied, at least 25 script pairs would need to contain such a marking error in order to alter the estimate by at least 1 per cent of the maximum.

None of the simulated SP studies resulted in flatlining.

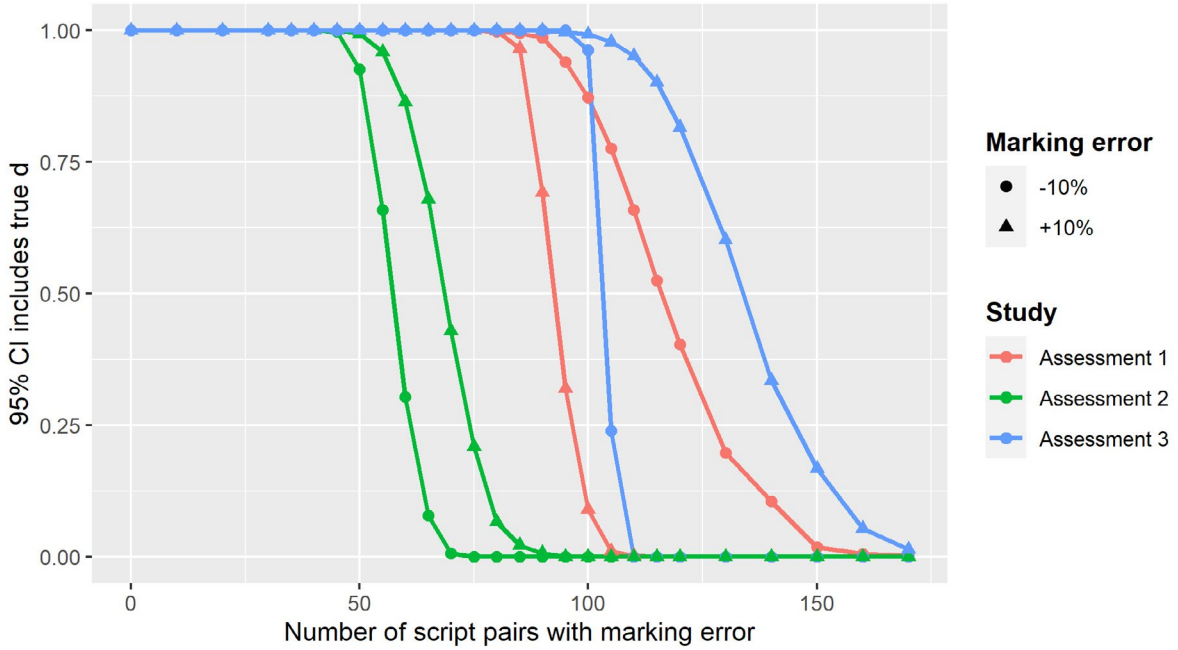

Figure 4: Proportion of simulated SP studies on target, by number of script pairs containing marking error.

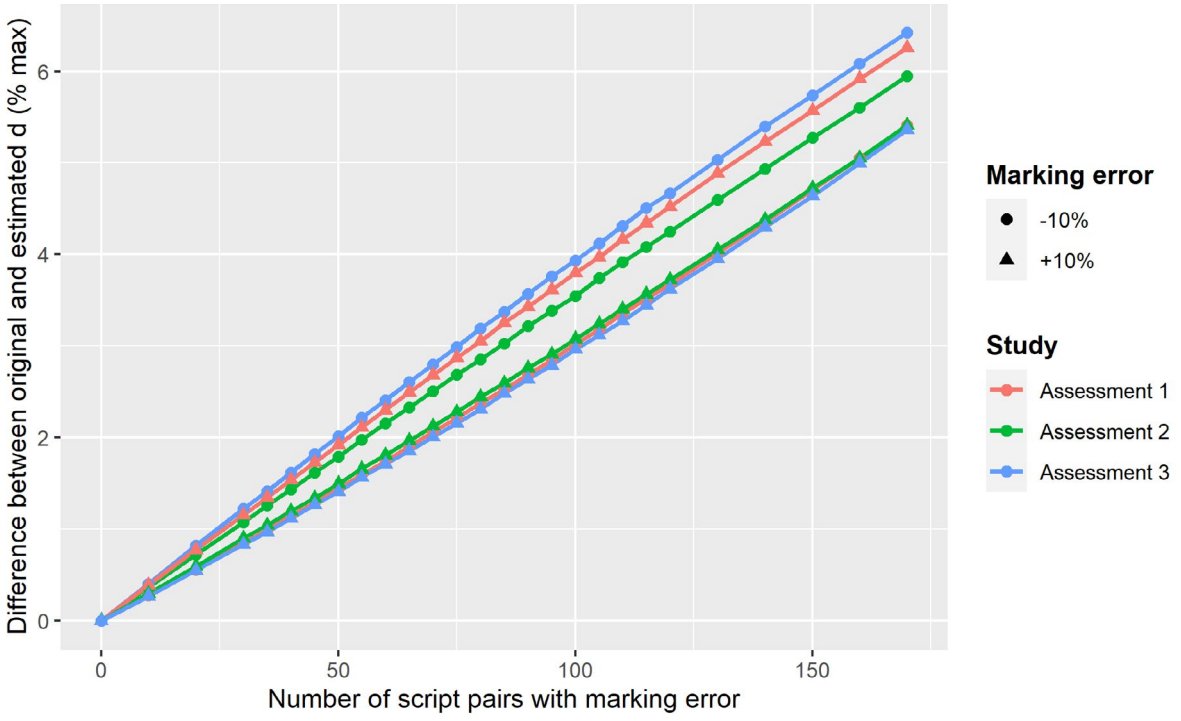

Figure 5: Mean absolute difference between original and estimated *d*, as a percentage of maximum mark.

The impact of progressively degrading marking quality

The third and final research question was addressed by simulating a large number of SP studies from scratch. The purpose of these simulations was to investigate the impact on SP results of progressively degrading quality of marking. These simulations differed from the earlier simulations by focusing on the overall relationship between awarded marks and script quality, rather than on single large marking errors or a fixed number of over- or under-marked scripts. The simulated SP data therefore needed to contain plausible data on mark differences, and simulated comparative judgements for these mark differences, and we needed to simulate how the relationship between mark differences and judgements would vary if marking quality decreased.

In the section below, we first explain the model relating marks and CJ measures, and how this relationship varies with marking quality. We then describe how the relationship between mark differences and CJ judgements can be expected to vary as marking quality varies, which is the foundation for the simulations. Finally, we explain how specific values for the key parameters were chosen.

Simulating SP study data

Throughout this section, we label all marks as x_i and all true CJ measures as θ_i. The CJ measures θ_i are the holistic measures of script quality that would result from analysing the outcomes of paired script comparisons using a Bradley-Terry model (Bradley & Terry, 1952). By "true" CJ measures, we mean the CJ measures if they were measured without error (i.e., with an extremely large number of comparisons for each script). The CJ measures are on a logit scale, which means that the difference between two scripts' measures $(\theta_j - \theta_i)$ is equal to the log of the odds of script *j* being judged higher quality than script *i* in any single paired comparison. For the time being we ignore differences in difficulty between different versions of assessments that may be included in a CJ exercise.

Following the approach in Benton and Elliott (2016) and Bramley and Gill (2010) we assume that over the range of interest[3], the relationship between marks and CJ measures can be summarised in the form:

$$\theta_i = \beta x_i + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$. There are two parts to the relationship between marks and measures:

1. First is "σ" (the standard deviation of the normally distributed residuals), which expresses the extent to which scripts with the same mark may have different "true" CJ measures. This might be because marking and CJ in fact measure slightly different constructs – so that even if scripts were marked perfectly and even if we included each script in a huge number of pairwise comparisons, we still wouldn't achieve a perfect correlation between marks

3 As previously noted, SP studies – like other CJ studies – exclude scripts from the extremes of the mark distribution, where the linear regression relationship would be affected by the floor and ceiling effects of the fixed total mark range.

and measures. It might also be a result of marking error. Higher levels of marking error will result in a larger value of σ.

2. Second is the coefficient "β", which expresses the strength of the association between marks and the decisions made by judges. Even if σ =0 (meaning that CJ and marking measure the same construct, and there is no marking error) it is likely that individual judges' decisions will not correspond perfectly to the marks that were awarded. However, the higher β is, the stronger the association. The CJ measures (θ_i) are constrained to have a mean of zero and the unit size (the logit) is directly related to judges' discrimination between scripts: a difference between two scripts of zero logits means that the scripts are equally likely to be judged superior (i.e., the probability of script *j* being judged superior is 0.5), and a difference of 1 logit between scripts means that the higher-rated script is judged superior with a probability of just over 0.7. When the coefficient β is higher, the same level of discrimination (e.g., a 1 logit difference) is associated with a smaller mark difference than when the coefficient β is lower. Alternatively, seen from the perspective of marks, a higher value of β means that the same mark difference between scripts corresponds to a higher probability of the higher-rated script being judged superior than when β is lower. Assuming a fixed level of reliability for CJ itself, then lower marking reliability would result in a lower value for β.

The logistic model describing CJ judgements tells us that for true CJ measures θ_i, the probability of script *j* being judged superior to script *i* is:

$$P(j \text{ beats } i) = \frac{\exp(\theta_j - \theta_i)}{1 + \exp(\theta_j - \theta_i)}$$

Via transformation and substitution (shown step by step in the Technical Appendix), we can re-express the likelihood of a script *j* "win" in terms of the mark difference between the scripts compared, and the two parameters β and σ reflecting marking quality. This means that the slope of the logistic regression linking mark differences and the probability of judges deciding script *j* is superior to script *i* (for brevity, written "GLM slope" from here on, for Generalised Linear Model slope) is given by:

$$GLM\ slope = \frac{1.7\beta}{\sqrt{1.7^2 + 2\sigma^2}}$$

Once a plausible value for the GLM slope is chosen, this value, together with a suitable set of mark differences (consistent with the methods used to sample pairs of scripts for an SP study) is sufficient to simulate a dataset of SP judgements.

Choosing values for β and σ

To simulate the SP studies, we estimated values of β and σ using data published in the appendices of Curcin et al. (2019). Using data from 20 pairwise comparison

studies[4] we used linear regression to estimate the relationship between marks (as a percentage of the total) and CJ measures of the holistic quality of papers. Across all 20 linear regressions, the median coefficient for β was 0.13 and the median value of σ (the standard deviation of estimated residual variance in the regression) was 1.3. Using these values with the GLM slope formula above, the expected slope of the logistic regression between mark difference (as a percentage of maximum mark) and judges' decisions would be the following:

$$GLM\ slope = \frac{1.7 * 0.13}{\sqrt{1.7^2 + 2(1.3)^2}} = 0.09$$

For the purposes of simulating a realistic SP study, 0.09 is therefore a reasonable value for the GLM slope of simulated data. The focus of this research, however, was on the extent to which the SP method would be robust to decreases in marking quality. Higher levels of marking error will result in higher values for σ and lower values for β, and hence smaller slope values.

In general, then, the simulations explored slope values lower than 0.09. In order to link slope values to a (quantified) degradation in marking quality, we calculated the values of σ and β (and hence, slope) that would correspond to specific decreases in marking reliability for a given SP study. This was done via substituting in marks x_i^* with added marking error, in the following way:

$$x_i^* = \rho x_i + \epsilon_i \sqrt{\left(1 - \rho^2\right)}$$

where var(ϵ_i)=400, and ρ represents the level of marking degradation – so that if the original marks x_i are perfectly reliable, then x_i^* would have marking reliability of ρ^2. The variance of ϵ_i in these simulated error-affected marks is set at 400 because a typical CJ study includes scripts with marks between 20 and 90 per cent of the available total, and roughly evenly spread (as reflected by the simulation steps in the next section). If the script marks are evenly spread between 20 and 90 per cent, their variance will be approximately 400[5].

Now, $\theta_i = \beta x_i + \varepsilon_i = \beta\rho x_i^* + \left(\beta\epsilon_i\sqrt{\left(1-\rho^2\right)} + \varepsilon_i\right)$, so we can use new values of beta and sigma to calculate the likely slope of the GLM, using $\beta^* = \rho\beta$ and $\sigma^{*2} = \sigma^2 + 400\beta^2\ (1 - \rho^2)$.

Simulation steps

We simulated a large number of SP studies from scratch. Varying levels of marking quality degradation were simulated via varying the GLM slope linking

4 Data from the rank ordering, "pinpointing" paired comparisons, and teacher paired comparison studies were not included. The rank ordering studies were analysed as pairs (and this may not be accurate), while the "pinpointing" and PCJ with teachers do not reflect Cambridge University Press & Assessment's normal practice.

5 The variance of a single-variable uniform distribution between values min and max is $\frac{1}{12}\ (max - min)^2$, see https://reference.wolfram.com/language/ref/UniformDistribution.html

mark differences and probability of script 2 "win". As shown above, this slope is dependent on both marking reliability and the strength of the relationship between marks and CJ measures.

The steps carried out were the following:

1. Simulate data from an SP "study" comparing two assessments (form A and form B) with 300 pairs of scripts, on a 0–100 mark scale:
    a. Simulate 300 script 1 marks from form A, sampled uniformly between 20 and 90 marks.
    b. Simulate 300 script 2 marks from form B, in the same way as for the script 1 marks.
    c. Pair "scripts" from form A and form B and calculate the mark difference (script 2–script 1). Scripts were paired so that the mark differences were approximately normally distributed around zero and 90 per cent of mark differences lay between -30 and 30 marks. The maximum mark differences ranged from ~-60 to ~60.
    d. Simulate a paired comparison decision for each pair of scripts by random draw from a binomial distribution, with the probability of success (script 2 "win") for each judgement being given by the logistic function of g*(mark difference - d), where g is the GLM slope and d is the overall difficulty difference (in marks) between form A and form B.
2. Analyse the simulated SP data using logistic regression.
3. Retain/calculate:
    a. estimated difficulty difference in marks (d)
    b. 95 per cent confidence intervals for d
    c. whether the estimated slope flatlined or not.

A simulated study was recorded as flatlining whenever either boundary of the 95 per cent confidence interval for the predicted probability of a script 2 "win" failed to intersect the line y=0.5 within the study's range of mark differences. This would occur, for example, if all lower bounds of the 95 per cent confidence intervals were lower than 0.5, or all upper bounds of the intervals were above 0.5, for the study's range of mark differences.

The simulation steps were carried out for two levels of true mark difference between form A and form B (d=0 and d=10), and for slope values ranging from 0.01 to 0.09, with 5000 "studies" simulated per condition. The entire set of simulations was then repeated for a simulated study size of 150 pairs of scripts, to give a sense of the impact on smaller SP studies. A true mark difference of 10 marks (i.e., 10 per cent of the mark range) between the two assessments compared is a fairly large difference, and the purpose of simulating at d=10 was to explore outcomes for a difference at the upper end of normal variation.

Results

As GLM slope value decreased, that is, the simulated relationship between mark difference and judges' decisions weakened, the proportion of simulated SP studies that flatlined increased (Figure 6). The size of confidence intervals for the estimated d increased (Figure 7) along with the variability of estimates, although estimates for d remained on target until the very lowest slope values (Figure 8). In comparison with the full SP studies using 300 pairs, outcomes deteriorated sooner when the number of pairs per simulated SP study was reduced to 150. Outcomes were better for the SP studies with no overall difficulty difference (d=0) than for those with an overall difference of 10 marks.

At a slope value of 0.09 (the GLM slope estimated from the median values for β and σ in the Ofqual studies), the simulated SP studies were successful: none flatlined, and the difficulty difference was estimated with confidence intervals comfortably smaller than 10 marks for 300-pair studies, and smaller than 15 marks for 150-pair studies. The "worst" values[6] in the Ofqual studies reported by Curcin et al. (2019) were σ =2 and β =0.09, which produced an estimated GLM slope of 0.046. The simulated SP study outcomes for a slope of this magnitude were slightly worse: Figure 6 shows that flatlining occurred for such studies with a non-zero difficulty difference, and for the 150-pair studies; and Figure 7 shows that 95 per cent confidence intervals for the estimated difficulty difference had a median size of around 10 marks, for the "best case" condition of no overall difficulty difference and *n*=300 pairs.

6 The study producing these values was AS Psychology specification 2, paper 1, year 1.

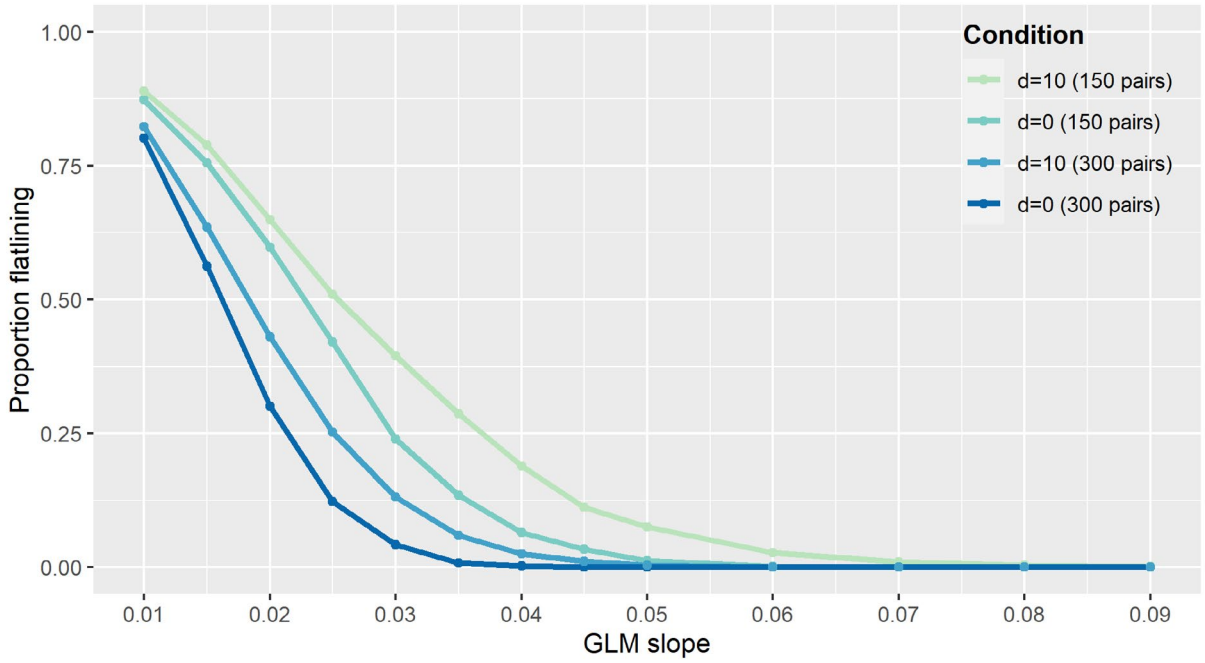

Figure 6: Proportion of SP studies flatlining.

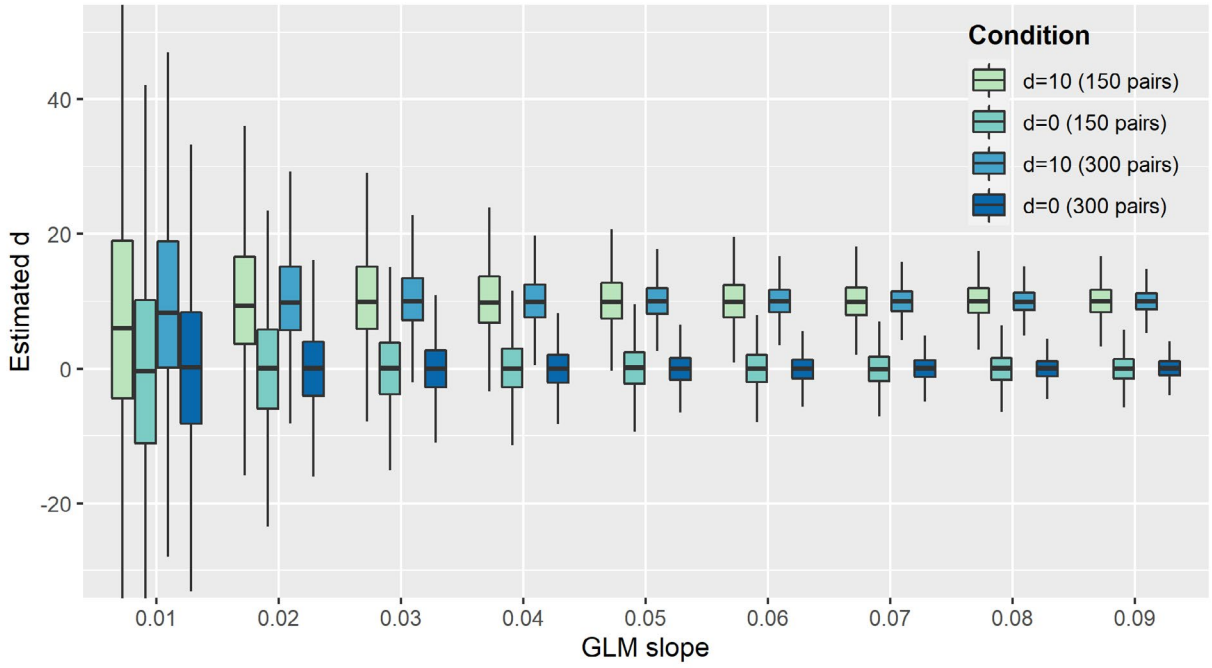

Figure 7: Distributions of confidence interval sizes (outliers not plotted; y-axis cropped).

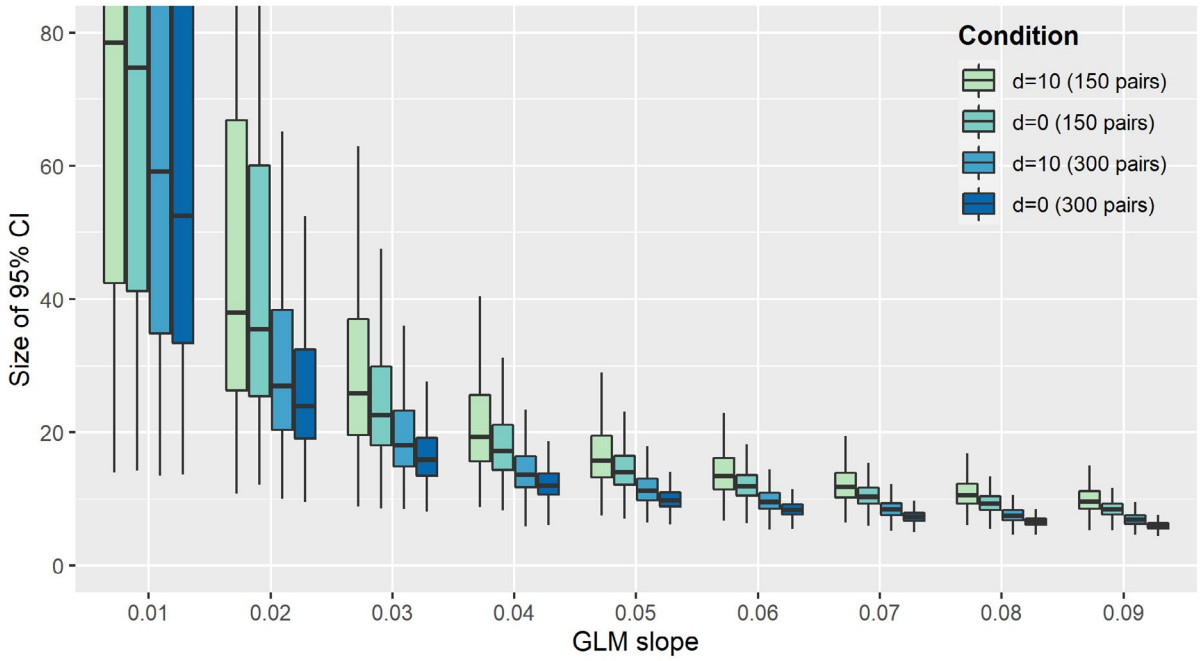

Figure 8: Distributions of estimated overall differences (outliers not plotted).

To consider the impact of specific levels of marking degradation, we simulated reductions in marking quality from the starting point of these "worst" values from the Ofqual studies (σ =2 and β =0.09, producing estimated GLM slope 0.046). Since the assessments in the Ofqual studies represent a selection of typical actual GCSE and AS level assessments (not chosen to be in any way extreme), this is a reasonable starting point to consider. Table 2 shows the estimated GLM slopes corresponding to increasing levels of marking degradation from this starting point, and the corresponding percentages of studies that flatlined at each level. Table 3 shows the median confidence interval sizes at each level of marking degradation.

For ρ =0.9, a modest degradation in marking that would result in a slope value of 0.04 (and corresponds to reliability of 0.81, if the original marking reliability is assumed to have been perfect), less than 1 per cent of 300-pair studies flatlined when the difficulty difference was zero, and 2.4 per cent flatlined when the difficulty difference was 10 marks (Table 2). The widths of the 95 per cent confidence intervals for d were about 12 marks and 14 marks respectively (Table 3). When the number of pairs per simulated SP study was reduced to 150, however, the same levels of marking degradation resulted in much more problematic outcomes: 7.3 per cent of studies flatlined when the difficulty difference was zero, and almost 20 per cent when the difference was 10 marks (Table 2). The median confidence interval sizes, meanwhile, were around 17 and 19 marks respectively (Table 3).

For higher levels of marking degradation, the results of the simulated SP studies deteriorated further. At marking degradation of ρ =0.775 (corresponding to reliability of 0.60, if original marking assumed perfect) the estimated GLM slope was 0.032. Of the 300-pair SP studies simulated with this slope, 1.8 per cent and 9.5 per cent flatlined (for d=0 and d=10 respectively), and median confidence interval sizes were around 15 and 17 marks. In the simulated 150-pair studies, the proportions flatlining were 18.4 per cent (d=0) and 32.9 per cent (d=10), and the median confidence interval sizes were around 21 and 24 marks.

Table 2: Flatlining in simulated SP studies, by condition (*n*=5000 studies per condition).

Marking degradation (ρ)	Revised slope	Percentage of studies that flatlined			
		300-pair studies		150-pair studies	
		d = 0	d = 10	d = 0	d = 10
1	0.046	0.04	0.50	2.42	9.76
0.975	0.045	0.00	1.10	3.02	11.68
0.95	0.043	0.02	1.40	4.76	14.64
0.925	0.041	0.06	1.78	5.24	16.90
0.9	0.040	0.06	2.38	7.26	19.94
0.875	0.038	0.34	3.20	9.06	22.90
0.85	0.037	0.48	4.26	11.32	23.62
0.825	0.035	1.14	5.78	13.66	27.06
0.8	0.034	1.34	7.90	16.54	30.26
0.775	0.032	1.78	9.48	18.38	32.90
0.75	0.031	2.64	11.36	23.02	36.90
0.725	0.030	4.32	13.40	25.16	40.96
0.7	0.028	6.02	16.00	30.88	43.24
0.65	0.026	10.22	22.34	38.42	50.10
0.6	0.024	15.78	30.32	45.64	54.56
0.55	0.021	24.24	38.02	54.42	61.02
0.5	0.019	34.02	46.50	62.04	68.32

Table 3: Confidence interval sizes for d in simulated SP studies, by condition (*n*=5000 studies per condition).

Marking degradation (ρ)	Revised slope	Median size of 95% CI for d			
		300-pair studies		150-pair studies	
		d = 0	d = 10	d = 0	d = 10
1	0.046	10.46	11.98	14.96	16.81
0.975	0.045	10.81	12.49	15.46	17.56
0.95	0.043	11.23	12.97	16.08	18.19
0.925	0.041	11.79	13.33	16.63	18.84
0.9	0.040	12.13	13.92	17.33	19.91
0.875	0.038	12.59	14.34	18.10	20.51
0.85	0.037	13.08	14.85	18.91	20.93
0.825	0.035	13.65	15.70	19.43	21.98
0.8	0.034	14.15	16.11	20.45	22.92
0.775	0.032	14.83	16.69	21.10	23.72
0.75	0.031	15.37	17.53	22.12	24.84
0.725	0.030	16.06	18.12	23.01	26.20
0.7	0.028	16.68	18.92	24.10	27.06
0.65	0.026	18.29	20.76	26.53	29.78
0.6	0.024	19.86	22.67	29.12	32.07
0.55	0.021	22.33	25.13	32.80	35.41
0.5	0.019	25.03	28.19	36.62	40.41

Conclusions

The research has two main conclusions. The first is that the SP method appears robust to single large marking errors, and to fairly large marking errors in quite high proportions of sampled scripts. The simulations of one-off large marking errors indicated that the estimated overall difficulty difference was affected only slightly, with numerical values very close to the originally estimated value, and only slightly increased standard errors. The simulations of multiple marking errors with a magnitude of 10 per cent of the component maximum mark, meanwhile, showed that the SP method failed only when large numbers of sampled scripts were affected – starting at around 50 out of 300 pairs. Similarly, it would take the occurrence of such marking errors in at least 25 out of 300 pairs to alter the estimated difference in difficulty between two tests by even 1 per cent of the maximum. These results are both reassuring and encouraging – the SP analyses proved robust even in the face of unusually large and unusually numerous errors in the script evidence, increasing confidence that the outcomes of SP analyses can be used to support maintenance of standards.

The second conclusion is that the SP method is more vulnerable to a general degradation of marking quality. The final set of simulations showed how SP analyses became problematic when the relationship between marks and CJ measures weakened – from whatever cause. The simulations showed that a non-extreme degradation in marking quality, from the starting point of values seen in published CJ studies, could result in failure of analysis (flatlining) and/or very wide confidence intervals around estimated differences. Importantly, the simulations showed that the deterioration in outcomes occurred much sooner for smaller studies (n=150 pairs), and when the actual overall difference between assessments was non-zero. Reducing the sample size in operational SP studies would, therefore, represent a substantial increase in risk to the success of the SP analysis and its ability to provide useful information for standard maintaining. In practical terms, SP analyses for a reduced sample size such as n=150 pairs have a much higher likelihood of failure than SP analyses for a full study of n=300 pairs, which would more than offset the advantages associated with choosing to run a smaller study. The actionable recommendation from this finding, therefore, is to avoid reducing sample sizes in operational SP studies for standard maintaining.

References

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

Benton, T., & Elliott, G. (2016). The reliability of setting grade boundaries using comparative judgement. *Research Papers in Education*, *31*(3), 352–376. https://doi.org/10.1080/02671522.2015.1027723

Benton, T., Gill, T., Hughes, S., & Leech, T. (2022). A summary of OCR's pilots of the use of Comparative Judgement in setting grade boundaries. *Research Matters: A Cambridge University Press & Assessment publication, 33*, 10–30.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345. https://doi.org/10.2307/2334029

Bramley, T., & Gill, T. (2010). Evaluating the rank ordering method for standard maintaining. *Research Papers in Education, 25*(3), 293–317. https://doi.org/10.1080/02671522.2010.498147

Camilli, G. (1994). Origin of the Scaling Constant *d* = 1.7 in Item Response Theory. *Journal of Educational Statistics*, *19*(3), 293–295. https://doi.org/10.3102/10769986019003293

Curcin, M., Howard, E., Sully, K., & Black, B. (2019). *Improving awarding: 2018/2019 pilots* (Ofqual/19/6575). Ofqual. https://www.gov.uk/government/publications/improving-awarding-20182019-pilots

Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error*, Technical Report No. 15 (Office of Naval Research Contract No. 25140, NR-342-022). Stanford University: Applied Mathematics and Statistics Laboratory.

Ofqual. (2014). *Review of quality of marking in exams in A Levels, GCSEs and other academic qualifications: Final report*. Ofqual. https://www.gov.uk/government/publications/quality-of-marking-in-gcses-and-a-levels

R Core Team. (2021). *R: A language and environment for statistical computing*. https://www.R-project.org/

Technical Appendix: simulating SP studies

This appendix shows the derivation of the equation that expresses the slope of the logistic regression linking mark differences and judges' comparative judgements in terms of the two parameters β and σ reflecting marking quality.

As stated in the main article, we assume that over the range of interest, the relationship between marks and CJ measures can be summarised in the form:

$$\theta_i = \beta x_i + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$.

Representing the true CJ measures as θ_i, we know that the probability of script *j* being judged superior to script *i* is:

$$P(j \text{ beats } i) = \frac{\exp(\theta_j - \theta_i)}{1 + \exp(\theta_j - \theta_i)}$$

At this point, we can usefully approximate this logistic model using the probit link function. This relies on a transformation constant of 1.7 as recommended by Haley (1952) and described in Camilli (1994). Having made this approximation, we can use:

$$P(j \text{ beats } i) = \Phi\left(\frac{\theta_j - \theta_i}{1.7}\right)$$

where Φ is the cumulative distribution function for the standard normal distribution.

Combining this equation above with the equation describing the relationship between marks and CJ measures, we get the following:

$$P(j \text{ beats } i) = \Phi\left(\frac{\beta\left(x_j - x_i\right) + \varepsilon_j - \varepsilon_i}{1.7}\right)$$

We can define $\epsilon_{ji} = \varepsilon_j - \varepsilon_i$, and since the difference of two independent normally distributed variables also follows a normal distribution, we know that $\epsilon_{ji} \sim N(0, 2\sigma^2)$.

Next, we think about the nature of the probit function. What it explicitly does is calculate the following: $\Phi(y) = P(z \leq y)$, where $z \sim N(0, 1)$.

Thus,

$$\Phi\left(\frac{\beta\left(x_j - x_i\right) + \epsilon_{ji}}{1.7}\right) = P\left(z \leq \frac{\beta\left(x_j - x_i\right) + \epsilon_{ji}}{1.7}\right) = P\left(\left(1.7z - \epsilon_{ji}\right) \leq \beta\left(x_j - x_i\right)\right)$$

By the properties of normal distributions, we know that $(1.7z - \epsilon_{ji}) \sim N(0, 1.7^2 + 2\sigma^2)$. By realising that by dividing $(1.7z - \epsilon_{ji})$ by $\sqrt{1.7^2 + 2\sigma^2}$ gets us back to a variable with a standard normal distribution we can see that:

$$P(j \text{ beats } i) = \Phi\left(\frac{\beta\left(x_j - x_i\right) + \epsilon_{ji}}{1.7}\right) = \Phi\left(\frac{\beta\left(x_j - x_i\right)}{\sqrt{1.7^2 + 2\sigma^2}}\right)$$

Finally, by reversing the approximation between the logistic and normal distributions we saw to begin with (i.e., multiplying the numerator of the subject of the function by 1.7), we can say:

$$P(j \text{ beats } i) = \frac{\exp\left(\frac{1.7\beta\left(x_j - x_i\right)}{\sqrt{1.7^2 + 2\sigma^2}}\right)}{1 + \exp\left(\frac{1.7\beta\left(x_j - x_i\right)}{\sqrt{1.7^2 + 2\sigma^2}}\right)}$$

This means that the slope of the logistic regression linking mark differences and the probability of judges deciding script j is superior to script i is given by:

$$GLM \text{ slope} = \frac{1.7\beta}{\sqrt{1.7^2 + 2\sigma^2}}$$