The concurrent validity of Comparative Judgement outcomes compared with marks

Tim Gill (Research Division)

Introduction

In Comparative Judgement (CJ) exercises, examiners are asked to look at a selection of candidate scripts (with marks removed) and order them in terms of which they believe display the best quality. The comparisons can either take the form of ranking of pairs of scripts ("paired CJ" or "PCJ") or of ranking of more than two scripts ("rank ordering" or "RO"). By including scripts from different examination sessions, the results of these exercises can be used to help with maintaining standards.

Results from previous CJ studies have demonstrated that the method appears to be valid and highly reliable in many contexts, including for marking of essays (Steedle & Ferrara, 2016) and standard maintaining (Benton, Leech & Hughes, 2020; Curcin et al., 2019). However, it is not entirely clear why CJ works as well as it does. Proponents of the method argue that it is because of the physical and judgemental processes involved in making comparative judgements. That is, the physical act of placing two scripts next to each other and deciding which is better based on an intuitive, holistic and relative judgement of quality. In particular, they argue that it is the relative aspect of the judgement that is important, because humans are better at making relative than absolute judgements (Laming, 1984). An alternative explanation, proposed by Benton & Gallacher (2018), is that the CJ method works well because CJ exercises capture a lot of individual paired comparison decisions quickly. In their study, they found that the predictive validity of scores derived from a CJ exercise was no better than the predictive validity of pseudo-CJ scores derived from comparing marks. This would suggest that CJ works well because of the number of judgements involved, not because the judgements come from the physical act of putting scripts next to each other and making a holistic relative comparison.

The analysis presented in this article adds to the research on this question by comparing the concurrent validity of the outcomes of CJ paired comparisons with the concurrent validity of outcomes based on the original marks given to scripts.

The focus here is on the validity of the outcomes of individual paired comparisons (the smallest building block within the CJ process), rather than the validity of scores allocated to scripts by a statistical model (such as the Bradley-Terry model) following multiple comparisons. The aim is to discover whether the decisions of a human judge directly comparing two pieces of work have more validity than those based on comparing the marks of two scripts, when these are derived independently and (usually) by different markers. As such, this research provides direct evidence on whether the idea that humans are better at making relative rather than absolute judgements (Laming, 1984) applies in the context of educational assessment when absolute judgements are supported by a mark scheme. Previous research in the context of awarding (Gill & Bramley, 2013) found that examiners were better at making relative judgements of quality than absolute judgements.

Data and methods

For this research, we re-used data from several previous CJ studies undertaken by Cambridge Assessment. All of these were experimental trials of the CJ method, with the aim of determining whether CJ had the potential to be used in standard-maintaining exercises in GCSEs and AS or A level qualifications in England. Each of these CJ studies used exam scripts taken from qualifications offered by the OCR awarding body (either GCSEs or AS levels). In all cases, the method was similar: either five or six examiners were asked to make comparisons of exam scripts (either in pairs or in packs of four) and to order the scripts from best to worst, in terms of the overall quality of the work. In most of the studies, at least some of the paired comparisons involved scripts from the same exam paper, but a version taken in a different exam session and the results of the comparisons were then analysed statistically to give an indication of the relative difficulty of the two papers. In total, there were 20 datasets which were all analysed separately. Details of these are presented in Table 1.

Most of these CJ studies asked examiners to make comparisons between pairs of scripts, but there were three which asked examiners to rank order packs of four scripts instead. For these studies, the rank ordering outcomes were converted into paired comparisons data (i.e., 1st beats 2nd, 1st beats 3rd, 1st beats 4th, 2nd beats 3rd etc.).

To compare the concurrent validity of CJ decisions with decisions based on the marks we needed the original marks given to the scripts and a measure of concurrent validity. Each CJ dataset contained the centre and candidate numbers of each candidate included in the paired comparisons, the original mark given to each script by the original examiner in the live exam session and the outcome of the paired comparison (i.e., which script was judged to be better). Candidates were matched (using centre and candidate numbers) to their marks achieved on other component(s) in the same qualification. These marks were used as the measure of concurrent attainment. Where all candidates within a study took more than one other component in the same qualification, marks were summed and the total used.

Some of the previous CJ studies only included paired comparisons between scripts from the same exam paper taken in different sessions, while others also included some comparisons between scripts from the same paper taken in the same session. For these latter studies, the datasets were split, so that the comparisons of scripts from the same exam session were analysed separately from the comparisons of scripts from different exam sessions. For example, we created three different sets of data for component AS level Geography Paper 1: comparisons between scripts from June 2018 and June 2019; June 2018 only comparisons; and June 2019 only comparisons.

In each dataset, the scripts were labelled as being either from the version 1 ("v1") paper or from the version 2 ("v2") paper. Every paired comparison included one v1 script and one v2 script. For the analysis of paired comparisons of scripts from different exam sessions, the scripts from the earlier session were designated as v1 and scripts from the later session as v2. For the analysis of paired comparisons of scripts from the same session, we needed to decide arbitrarily which of each pair of scripts would be the v1 script and which would be the v2 script. This was done by sorting each pair by the centre and candidate number and choosing the first script as the v1 script.

Table 1: Details of CJ study datasets used in the analysis.

Qualification and subject	Paper(s)	v1 exam session	v2 exam session	Pairs (PCJ) or Rank Order (RO)?	No. of judges	No. of scripts	No. of comparisons
AS Geography	Paper 1	June 18	June 19	RO	6	400	400
AS Geography	Paper 1 June 18	June 18	June 18	RO	6	200	100
AS Geography	Paper 1 June 19	June 19	June 19	RO	6	200	100
AS Geography	Paper 2	June 18	June 19	RO	6	400	400
AS Geography	Paper 2 June 18	June 18	June 18	RO	6	200	100
AS Geography	Paper 2 June 19	June 19	June 19	RO	6	200	100
AS Sociology	Paper 1	June 18	June 19	Pairs	22	140	1337
AS Sociology	Paper 2	June 18	June 19	Pairs	5	569	289
GCSE Eng Lang	Paper 1 PCJ	June 19	Nov 19	Pairs	14	124	517
GCSE Eng Lang	Paper 1 June 19 PCJ	June 19	June 19	Pairs	14	57	210
GCSE Eng Lang	Paper 1 Nov 19 PCJ	Nov 19	Nov 19	Pairs	14	70	303
GCSE Eng Lang	Paper 1 RO	June 19	Nov 19	RO	9	141	772
GCSE Eng Lang	Paper 1 RO June 19	June 19	June 19	RO	9	70	193
GCSE Eng Lang	Paper 1 RO Nov 19	Nov 19	Nov 19	RO	9	70	176
GCSE Eng Lang	Paper 1 SP	June 19	Nov 19	Pairs	5	570	285
GCSE Eng Lang	Paper 2 PCJ	June 19	Nov 19	Pairs	15	129	555
GCSE Eng Lang	Paper 2 PCJ June 19	June 19	June 19	Pairs	15	57	235
GCSE Eng Lang	Paper 2 PCJ Nov 19	Nov 19	Nov 19	Pairs	15	72	371
GCSE Maths	Paper 1	June 19	June 19	Pairs	6	600	300
GCSE Eng Lit	Paper 1 / Paper 2	June 16	June 16	Pairs	6	572	286

Table 1 includes three different datasets for GCSE English Language Paper 1. This is because they were taken from a Cambridge Assessment research project investigating which method of paired comparative judgement (PCJ), rank ordering

(RO) or simplified pairs (SP)[1] was most helpful for identifying grade boundaries (see Benton et al., 2022, this issue). Therefore, three different CJ exercises were undertaken. For GCSE Maths, the v1 and v2 sessions were the same because this study involved splitting the June 2019 paper into two halves and making comparisons between scripts from each half (see Benton, Leech & Hughes, 2020). Similarly, for the GCSE English Literature exercise, the v1 and v2 sessions were the same since comparisons were made between different papers in the same session (see Benton, Cunningham, Hughes & Leech, 2020).

To generate the measures of concurrent validity, the following process was undertaken for each dataset:

- For every paired comparison, a variable (called "v2CJsuperior") was created and was given a value of 1 if the v2 script was judged superior, and 0 otherwise.
- A variable (called "v2marksuperior") was created and was given a value of 1 if the V2 script was given a higher mark by the original marking, and 0 otherwise. For studies where the v1 and v2 were from different exam sessions, marks were converted to Uniform Mark Scale (UMS) marks so that they were directly comparable[2]. For the two studies (GCSE English Literature and GCSE Maths) where the papers being compared were from the same exam session, all candidates took both papers (or half papers in the case of GCSE Maths) being compared. This meant it was possible to use statistical equating (using the equipercentile method) to find the equivalent marks on v2 for each mark on v1.
- For the candidates in each CJ exercise, the total marks achieved in the other component(s) in the same specification in the same session were found ("concurrent marks"). For studies where the v1 and v2 scripts were from different exam sessions (and therefore the concurrent marks were also from different exam sessions), the marks were converted to UMS so that they were directly comparable. These variables were called "v1concurrentmark" and "v2concurrentmark".
- Pearson correlation coefficients[3] were calculated between both "v2CJsuperior" and "v2marksuperior" and the differences in candidate mark on the concurrent assessment(s) (v2concurrentmark-v1concurrentmark).

[1] The Simplified Pairs method of CJ enables the mapping of marks between different tests without the need to estimate values on a common scale by fitting a statistical model (such as the Bradley-Terry model) to the experts' judgements. See Benton, Cunningham et al. for a more detailed description of this method (2020).

[2] UMS marks are on a common scale, so that they can be directly compared between exam series (see https://ocr.org.uk/students/getting-your-results/calculating-your-grade/). If we had not done this it would mean that, if the two exams differed in difficulty, it would not be possible to say which script was judged to be superior according to the raw marks. As it happens, the differences in difficulty were all very small, meaning that there were very few instances of the order of pairs of marks changing after converting to UMS.

[3] With one binary variable and one continuous variable this is equivalent to a point biserial correlation.

- A multiple logistic regression was undertaken of "v2CJsuperior" on the two concurrent marks. The pseudo R-squared value was recorded[4], as a measure of the model fit.
- A multiple logistic regression was undertaken of "v2marksuperior" on the two concurrent marks, and the pseudo R-squared value recorded.

By comparing the correlation coefficients and the pseudo R-squared values, it was possible to determine whether the individual decisions based on marks had higher concurrent validity than those derived using CJ. The correlation coefficients indicate the strength of the relationship between wider candidate ability (as measured by the marks on assessments taken concurrently) and which candidate was judged to be better by either the paired comparison or the marks. As the value of v2concurrentmark-v1concurrentmark increases we would also expect the likelihood of the v2 script winning to increase.

The purpose of undertaking the logistic regressions was to allow for the possibility that the UMS had not completely controlled for difficulty. The pseudo-R square measure can be thought of as an indication of how well the outcome (which script was better according to either CJ or marks) was predicted by the independent variables (marks on concurrent components). A higher pseudo-R square value for the prediction of the CJ outcome would be an indication of better concurrent validity for the CJ outcome than for the marks outcome.

As shown in Table 1, most of the data came from CJ exercises which were comparing scripts from different exam sessions (hence the need for two separate concurrent marks in the above description). However, there were several datasets where all the data came from a single session, so that the concurrent marks were directly comparable. For these, it was only necessary to calculate and compare the correlation coefficients.

Although the main focus of this research was on the validity of the outcomes of individual paired comparisons, a further analysis was undertaken to compare the concurrent validity of the CJ "measure" (see below for an explanation of the term "measure") with the concurrent validity of UMS marks. If the concurrent validity of CJ is substantially improved by using the measure instead of the outcomes of the individual paired comparisons, then this will be a further indication that it is the way in which CJ incorporates the many judgements that makes the method successful. For this analysis we just used data from the studies where each script was involved in multiple comparisons (AS level Sociology Paper 1, GCSE English Language Paper 1 PCJ and RO, and GCSE English Language Paper 2). For these studies, the paired comparison data was analysed using the Bradley-Terry model (Bradley & Terry, 1952). This generated a measure of quality for each script, based on the number of times each script was judged superior across the multiple comparisons it was included in. Pearson correlation coefficients were calculated between the measure and the UMS marks on the concurrent component, and these were compared with correlations between UMS marks on the component of interest and the UMS marks on the concurrent component.

[4] Proc Logistic in SAS software reports the Cox & Snell (1989) calculation of R-squared.

Results

Table 2 presents the results of the correlations and the pseudo R-squared values for each dataset. For further details about the logistic regression (including the regression equation and some example output from one dataset), see the Appendix.

Table 2: Correlation coefficients and pseudo R-squared values for CJ study datasets.

Paper	Corr between concurrent marks and CJ outcome	Corr between concurrent marks and marks outcome	Pseudo R-square for CJ outcome	Pseudo R-square for marks outcome	Decision with higher concurrent validity
AS Geography Paper 1	0.37	0.38	0.14	0.15	Marks-based
AS Geography Paper 1 June 18	0.41	0.44	n/a	n/a	Marks-based
AS Geography Paper 1 June 19	0.36	0.48	n/a	n/a	Marks-based
AS Geography Paper 2	0.34	0.27	0.12	0.08	CJ-based
AS Geography Paper 2 June 18	0.37	0.33	n/a	n/a	CJ-based
AS Geography Paper 2 June 19	0.47	0.20	n/a	n/a	CJ-based
AS Sociology Paper 1	0.52	0.58	0.28	0.35	Marks-based
AS Sociology Paper 2	0.22	0.39	0.07	0.16	Marks-based
GCSE Eng Lang Paper 1 PCJ	0.57	0.66	0.33	0.44	Marks-based
GCSE Eng Lang Paper 1 PCJ June 19	0.63	0.74	n/a	n/a	Marks-based
GCSE Eng Lang Paper 1 PCJ Nov 19	0.48	0.60	n/a	n/a	Marks-based
GCSE Eng Lang Paper 1 RO	0.37	0.47	0.14	0.23	Marks-based
GCSE Eng Lang Paper 1 RO June 19	0.41	0.47	n/a	n/a	Marks-based
GCSE Eng Lang Paper 1 RO Nov 19	0.33	0.50	n/a	n/a	Marks-based
GCSE Eng Lang Paper 1 SP	0.37	0.40	0.16	0.18	Marks-based
GCSE Eng Lang Paper 2 PCJ	0.51	0.61	0.25	0.38	Marks-based
GCSE Eng Lang Paper 2 PCJ June 19	0.50	0.54	n/a	n/a	Marks-based
GCSE Eng Lang Paper 2 PCJ Nov 19	0.58	0.63	n/a	n/a	Marks-based
GCSE Maths Paper 1	0.56	0.59	n/a	n/a	Marks-based
GCSE Eng Lit Paper 1 / Paper 2	0.45	0.38	n/a	n/a	CJ-based

The “n/a” in the table indicates CJ exercises where all the data came from the same session and so it was not necessary to run a logistic regression model. The final column in the table indicates which decision (CJ-based or mark-based) had higher concurrent validity, according to the results of the correlations and the pseudo-R squares.

Figures 1 and 2 illustrate the relationships visually for two of the datasets (GCSE English Language Paper 1 PCJ, with a relatively high correlation and pseudo-R squared, and AS level Geography Paper 2, with a relatively low correlation and pseudo-R squared). The figures compare the range of mark differences in the concurrent attainments (v2concurrentmark-v1concurrentmark) by whether the V2 script was judged superior and by the judgement type (CJ or marks).

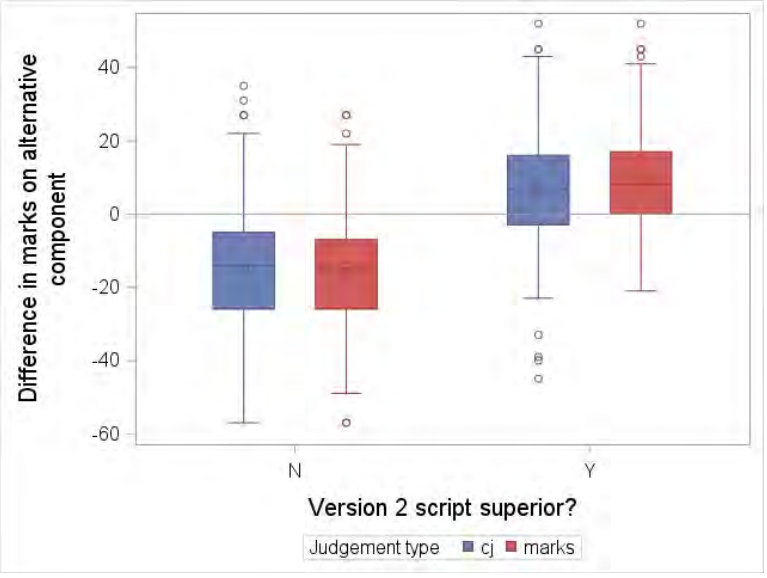

Figure 1: Distribution of v2concurrentmark- v1concurrentmark by superiority of v2 script and by judgement type (GCSE English Language, Paper 1, PCJ).

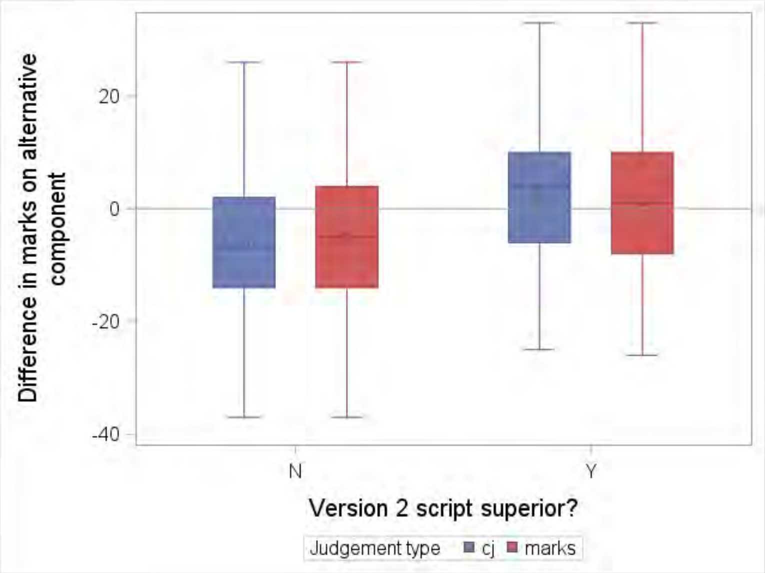

Figure 2: Distribution of v2concurrentmark–v1concurrentmark by superiority of V2 script and by judgement type (AS level Geography, Paper 2).

For example, Figure 1 shows that for V2 scripts judged to be superior according to CJ, the average difference in marks on concurrent components was around 10 marks. In contrast, when the V2 script was judged to be inferior, the average difference was around -15 marks. Figure 2 shows a much smaller difference in the average mark differences, being around 2 marks for V2 judged superior and around -5 when V2 was judged inferior.

In Figure 1, the red boxes are slightly further apart than the blue boxes, indicating a stronger relationship between the marks-based decision and the mark difference than between the CJ-based decision and the mark difference. This implies that the marks-based decision had higher concurrent validity. In contrast, the blue boxes were further apart than the red boxes in Figure 2, implying that the CJ-based decision had higher concurrent validity.

Table 2 shows that for 16 out of the 20 data sets analysed, marks-based decisions had higher concurrent validity than CJ-based decisions. All but one of the pseudo R-squared values was higher for marks than for CJ. The only exception was AS level Geography, Paper 2, which had an R-squared of 0.12 for the CJ outcome model, compared with 0.08 for the marks model. For the 12 datasets which only included comparisons within the same session (and therefore with no logistic regression undertaken), there were only three occasions where the correlation coefficient was higher for the CJ outcome than for the marks outcome. These were for component AS level Geography, Paper 2 (both the 2018 only and the 2019 only datasets) and for the comparison between GCSE English Literature, Papers 1 and 2.

The AS level Geography, Paper 2 study used rank ordering, but otherwise the results showed no evidence of any different pattern for rank ordering studies compared with paired comparison studies.

Table 3 shows the correlation coefficients between the script measures (generated using the Bradley-Terry model) and the UMS marks on the concurrent component. It also shows the correlations between UMS marks on the component of interest and UMS marks on the concurrent component.

Table 3: Comparison of correlation coefficients of script measures and UMS with concurrent component UMS.

Component	No. of scripts	Corr between script measure and concurrent UMS	Corr between UMS and concurrent UMS
AS Sociology Paper 1	139	0.67	0.66
GCSE Eng Lang Paper 1 PCJ	124	0.77	0.84
GCSE Eng Lang Paper 1 RO	137	0.68	0.81
GCSE Eng Lang Paper 2 PCJ	129	0.71	0.77

These results mainly follow the pattern seen in Table 2, with higher correlations for marks-based outcomes (UMS) than for CJ-based outcomes (script measure). The only exception to this was for AS level Sociology, where the correlation between the script measure and concurrent component UMS was very slightly higher. This contrasts with the results from Table 2, where the correlation between the CJ outcome and concurrent component UMS (0.52) was lower than between the marks-based outcome and concurrent component UMS (0.58).

Having seen that individual decisions based on marks had higher concurrent validity than those based on CJ (Table 2), we had hoped that the additional analysis in Table 3 would illustrate how this is overcome by the way CJ incorporates many judgements. This effect was visible in only one of the four studies. Specifically, we found that for AS Sociology Paper 1, although the concurrent validity of individual CJ decisions was lower than that of marks-based decisions (Table 2), the concurrent validity of CJ estimated measures was higher than that of the original marks. However, the expected effect was not visible in the other papers. Our expectations may have been confounded elsewhere because, although the CJ validity benefits from combining many judgements, the concurrent validity from marks also increased, for a different reason – namely that, analysing it in this way used the marks awarded to scripts, not just which of a pair is higher.

To think of this another way, it is clear that our earlier analysis provided a straightforward like with like comparison. Individual choices between two scripts based on judges' opinions were compared to individual choices based on marks. However, in this additional analysis we are comparing scores on one scale based upon multiple pairwise comparisons of each script (and different numbers of these for different components) to scores on an entirely different scale based on detailed marking. As such, meaningful interpretation is much harder.

It should be remembered that, in this section, we only have results from a relatively small number of studies, each of which only incorporates a fairly small number of scripts. As such it is important that we do not overinterpret these particular findings.

Conclusion

The main conclusion from this analysis is that the concurrent validity of the decision based on marks was generally higher than the concurrent validity of the CJ decision. Two possible reasons for this finding suggest themselves: firstly, CJ decisions reward different skills to marks (and ones that are less related to marks on other components). This may be because of the different processes involved. In CJ, the judges make holistic and relative judgements of quality, without direct reference to a mark scheme. In contrast, in live marking, the total mark is an absolute judgement of quality based on the summation of marks given for responses to individual items, with direct reference to the mark scheme. An alternative explanation is that individual CJ decisions are of lower quality than decisions based on marks. In other words, judges are less able to make reliable judgements of the relative qualities of scripts when using the quick holistic approach required of comparative judgements.

This finding adds further evidence in favour of the contention in Benton & Gallacher (2018) that it is not the physical process of making intuitive, holistic and relative judgements of quality that makes CJ successful, but rather that it is able to capture many individual paired comparison decisions quickly.

The results here contrast with a previous study evaluating examiners' holistic judgements of script quality (Gill & Bramley, 2013), which found that examiners were better at making relative judgements of quality than absolute judgements. The results of the current research suggest that the absolute judgements (i.e., marks) were better than the relative judgements (CJ). This difference may be because in practice marking also involves some form of relative judgement, versus a fixed mark scheme. This differs from the context of the previous study (Gill & Bramley, 2013) where the absolute judgements were made without access to the mark scheme and therefore dependent only on the judges' own idea of what grades should look like.

This research was opportunistic, in that it used already available datasets. Further research which is designed to answer a specific research question would be worthwhile. For example, it would be interesting to investigate which of CJ decisions or marks-based decisions in one component is a better predictor of CJ decisions in a related component. If CJ decisions are better then this would suggest that they are indeed rewarding different skills to marks.

References

Benton, T., & Gallacher, T. (2018). Is comparative judgement just a quick form of multiple marking? *Research Matters: A Cambridge Assessment publication, 26*, 22–28.

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 and Assessment publication, 33*, 10–30

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 Assessment.

Benton, T., Leech, T., & Hughes, S. (2020). Does comparative judgement of scripts provide an effective means of maintaining standards in mathematics? Cambridge Assessment Research Report. Cambridge Assessment.

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.

Cox, D. R., & Snell, E. J. (1989). *The Analysis of Binary Data*, (2nd ed.). Chapman and Hall.

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

Gill, T., & Bramley, T. (2013). How accurate are examiners' holistic judgements of script quality? *Assessment in Education: Principles, Policy & Practice, 20*(3), 308–324. https://doi.org/10.1080/0969594X.2013.779229

Laming, D. (1984). The relativity of 'absolute' judgements. *British Journal of Mathematical and Statistical Psychology*, *37*(2), 152–183. https://doi.org/10.1111/j.2044-8317.1984.tb00798.x

Steedle J. T., & Ferrara, S. (2016). Evaluating Comparative Judgment as an Approach to Essay Scoring. *Applied Measurement in Education, 29*(3), 211–223. https://doi.org/10.1080/08957347.2016.1171769

Appendix – details of logistic regression

Logistic regression equation:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 v1concurrentmark_i + \beta_2 v2concurrentmark_i$$

Where p_i is the probability that in comparison "i" the version 2 script was judged superior, v1concurrentmark and v2concurrentmark are the independent variables and β_1 and β_2 are the regression coefficients.

Table A1: Example output from logistic regression (AS level Geography Paper 1, dependent variable = CJ-based decision)

	Parameter estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	0.0771	0.5441	0.0201	0.8874
v1concurrentmark	-0.0652	0.0114	32.6160	<0.0001
v2concurrentmark	0.0679	0.0122	31.1551	<0.0001

Table A2: Example output from logistic regression (AS level Geography Paper 1, dependent variable = marks-based decision)

	Parameter estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	0.7557	0.5456	1.9182	0.1661
v1concurrentmark	-0.0800	0.0118	46.1882	<0.0001
v2concurrentmark	0.0549	0.0119	21.2937	<0.0001