Using Adaptive Comparative Judgement for assessing GCSE History NEA responses

Research Report

June 2022

Using Adaptive Comparative Judgement for assessing GCSE History NEA responses. Research Report.
Author: Vasile Rotaru, Senior Research Officer Qualifications Wales

Acknowledgements

I would like to express my gratitude to Tom Anderson (Head of Research and Statistics at Qualifications Wales) for his substantial contribution to the concept and design of the research, as well as for his help in critically revising the findings and the subsequent drafts of this report.

I would like to thank two members of the QW Research and Advisory Group - Dr. Joshua McGrain (Associate Professor in Educational Assessment at Oxford) and Anne Pinot de Moira (Honorary Norham Fellow at University of Oxford, Department of Education) - for comments and suggestions that greatly improved the report.

I am most grateful to Mark Hogan for his help with my questions on the statistical aspects.

My thanks also go to the teachers who participated in this research for their willingness to contribute and for their open discussions.

I would like to thank RM education for providing free access to the RM Compare™ software.

The acknowledgements do not imply endorsement of the current study and its findings.

Published in May 2022 by Qualifications Wales

To cite this publication: Rotaru, V. (2022) Using Adaptive Comparative Judgement for assessing GCSE History NEA responses. Qualifications Wale

Table of contents

1. Introduction

Comparative Judgement (CJ) is a method in which a group of judges repeatedly compare stimuli drawn from the same set. Usually, the stimuli are presented in pairs; for example, comparing pairs of essays written by the learners from the same class. For each pair, judges would choose a winner based on any 'quantitative or qualitative attribute about which we can think "more" or "less" for each specimen' (Thurstone, 1927, p.273), which loosely could also mean "better", "greater", "more appealing" etc.

Research undertaken by Qualifications Wales aimed to explore the opinions of History teachers about CJ as an alternative method to traditional marking and to investigate their decision making when judging. Therefore, this study is primarily qualitative although the CJ results could add to the existing findings on the reliability and validity of CJ.

2. Research background[1]

The CJ method is based on the law of comparative judgement developed by the psychologist Louis Thurstone (Thurstone, 1927). He assumed that each time a person compares a pair of stimuli, each stimulus gets placed on a scale or continuum by a 'discriminal process'. This is a 'kind of process in us by which we react differently' to stimuli and categorise them according to degrees of possessing a certain attribute. By placing the stimulus on the continuum, we assign to it an instant value. This is not an exact value, but by repeatedly comparing a stimulus we get a set of values which form a normal distribution. The mean value of the distribution is the approximate scale value of the stimulus (Thurstone, 1927). The values of all stimuli are on an interval scale where the rank order of the stimuli and the differences between scores are taken to be meaningful. This value is used to rank order a stimulus in relation to other stimuli.

The current research focuses on the use in educational assessment of a form of CJ, in which a group of judges is asked to compare pairs of objects drawn from the same set. Alternative approaches to using CJ or using comparisons exist as well. For example, judges can be asked to compare and place a set of responses into a rank order, instead of comparing the pairs (see Bramley, 2005). CJ can also be used to design a two-stage assessment in which a group of judges initially calibrate a set of responses, and afterwards selected responses from this set are used as anchor points on a scale, against which markers compare the responses (McGrane et al., 2018).

[1] In the Winter/Spring of 2022, the Frontiers in Education Journal, and the Research Matters (a publication of the Cambridge Assessment Network) published a number of articles on CJ which are relevant to this study. As these articles have been published after this report was finalised, they are not included in this report's literature overview and findings comparisons.

CJ implementation requires that judges compare enough pairs to determine with sufficient reliability the values of the stimuli (Thurstone, 1927). The pairs could be selected at random or by using an adaptive approach. Adaptive Comparative Judgement (ACJ), which is used in this study, does not pair objects that are further apart (that is an object which almost certainly has a greater parameter value than the other), as a decision on such a pair does not add useful information for the ranking purpose. As most of the time we do not have any information about an object's value, the pairs are initially picked at random. Once the judging process provides some information about the value of the objects, an algorithm guides the selection of pairs. With each round of this adaptive selection, better object separation occurs to the extent that reliable results can be achieved with a relatively reduced number of comparisons (see Pollitt, 2012), and in about ten rounds earlier than is the case in the non-adaptive (purely random) selection of pairs (Kimbell, 2021).

According to Marshal et al. (2020) the use of CJ in educational assessment has been increasingly researched. CJ has been explored, piloted and used in assessment (both on formative and summative tasks) and in maintaining standards when awarding qualification grades.

Findings suggest that for certain types of assessments, CJ has some potential as an alternative to conventional marking and external moderation. Tarricone and Newhouse (2016, p.13) even argued that CJ is a 'highly reliable alternative to traditional analytical marking' which may improve the validity and reliability of traditional high-stakes examinations.

Regarding validity, in practical craft skills (wood and metal-based decorative processing techniques) it has been shown that the use of CJ allows learners more freedom of creation (by not requiring them to follow rigid assessment criteria) and enhances the capacity of assessment criteria to change in response to how students align their work with evidence of capability (Seery et al., 2012). Likewise, it has been shown that in mathematics the CJ assessment approach encourages the design of less structured and more problem-based tasks compared to the typical mark scheme-based tasks (Jones and Inglis, 2015).

A factor that has made CJ gain interest were the findings that it is more reliable than traditional marking (Kimbell, 2021). For example, an overview by Bramley (2015, p.6) found that nine publications describing studies which had 9 to 15 comparisons per script, indicated reliability levels that ranged from 0.86 to 0.97, with one exception of 0.73. These studies included such subjects as Maths, Chemistry, Design, and different forms of writing.

However, the studies are not conclusive about the potential benefits of CJ. For example, the validity of CJ when assessing different competences was questioned,

because a single holistic judgement could not encapsulate 'the uneven nature of candidates' performances' (Barkaoui, 2011 cited in Whitehouse and Pollitt, 2012). Similarly, although CJ assessment usually has high reliability indices, it has been pointed out that adaptivity tends to overestimate the reliability coefficient (Bramley 2015). It has been suggested that this problem can be remedied by fine-tuning the selection algorithm, while also ensuring that around 15 judgement rounds are conducted (Kimbell, 2021).

Most of the existing research evaluates the merit of CJ assessment using the results of judging (e.g. correlation/reliability coefficients, average judging time). A few studies explore judges' opinions of the method or their approach to judging. In some of them researchers surveyed judges (Jones and Inglis, 2015; Marshal et al., 2020; Pollitt, 2012), in others they explored comments that judges left about the scripts or their judgements (Daal, 2016; Jones et al., 2015; Whitehouse, 2013). To our knowledge, only two studies (Coertjens et al., 2021 and Newhouse, 2014) have explored qualitatively the opinions of judges. Not all of the judges in these studies would conventionally assess on a regular basis (or at all) responses similar to those judged.

However, if CJ is to be used in assessment, we consider that a more in-depth investigation of the opinions of assessors is needed to examine how CJ would fit into a given assessment paradigm and how it would address the constraints that assessors face.

Because the use of CJ implies a change of the assessment approach, a useful aspect to explore relates to the markers' openness to accept CJ and the factors that could help, or conversely hinder, a buy-in. Closely related to this is also the need to explore what type of ongoing support judges might require when using CJ in assessment. Another aspect worth exploring is whether the CJ decision-making process allows judges to pay attention and assess the features that are assessed through conventional marking. Markers' perceptions of the method could be helpful in exploring these aspects.

In addition, it would be useful to investigate if CJ could be a viable, or even a better, assessment tool in terms of manageability, validity and reliability than conventional marking for summative assessment of extended writing responses. The general qualifications system in Wales relies on constructed responses, including extended response items (some of which are marked by teachers internally), as opposed to using multiple-choice or closed response questions. Such an approach is often thought to ensure greater validity of the assessment (see, for example, Murphy and Yancey, 2009; Kimbell, 2021). However, in the case of extended writing, research suggests that the assessment can be affected by a relatively low marking accuracy. For example, Rhead et al. (2018, p.4) have shown that the probability of receiving the

'definitive grade'[2] varies considerably across qualifications in England. The marking consistency was the lowest for exams that contained extended writing answers: 0.52 for English language and literature qualifications, with history having the second lowest consistency. On the other hand, the reliability of marking for mathematics was 0.92. In another study, the reliability of marking AS History source-based exam writing questions was in the range of 0.52 to 0.62 (Holmes et al., 2018, p.17).

It is likely that teacher-led assessment, as contrasted with external marking, has a further effect on marking accuracy because of the additional aspects that teachers might face or consider when marking (e.g. biases developed from teachers knowing the students, or because qualifications outcomes are used in school accountability processes) (see for example, Kellaghan et al., 2019, p.263-264).

There has been research that showed that CJ could be a valid method for GCSE or A level writing tasks (Bramley and Vitello, 2019; Steedle and Ferrara, 2016; Whitehouse and Pollitt, 2012; Pollitt and Crisp, 2004)[3]. For example, Whitehouse (2013) showed that judges assessing AS level Geography scripts used the mark scheme and assessment objectives (AOs) when judging, even if not provided with the mark scheme. These studies have used actual or modified exam item responses, but in all of them the scripts were relatively short. However, extended writing response items could have at least three handwritten pages. For example, GCE AS/A level History response should be between 3,000 and 4,000 words; the KS4 National/Foundation Welsh Baccalaureate Individual Project should be presented in written form of 1,000 – 2,000 words.

Lengthier scripts could make judgements more difficult, which could affect the efficiency of CJ as more content has to be considered to make a judgement. For example, Jones et al. (2015) assessed an entire GCSE Mathematics paper using CJ, with the length of some scripts going up to 47 pages. They found that scripts' length (and, in our opinion, probably the nature of these responses) made it more difficult to form a global judgement and rendered judging difficult, stressful, and less efficient.

[2] In the cited study the term 'definitive' refers to the mark given by a panel of senior examiners at item level for each seeding response.

[3] To note, some studies used extended responses to explore the use of CJ in maintaining standards (see examples cited in Steedle and Ferrara, 2016).

3. Research aims

The aims of this study were to:

- explore teachers' opinions about using CJ to mark extended writing responses (defined as responses that have at least three handwritten pages) including information about how the judgements were made;
- collect information relevant to the validity, reliability and manageability (e.g. time needed for judging compared to marking) of using CJ for the summative assessment of extended answers.

4. Method

4.1. Judges

Twenty-two history teachers who had marked either GCSE History or A level History were recruited to take part in a CJ workshop in January 2020 via an invitation letter sent to all schools in Wales. None of the participants knew about CJ before the workshop. The workshop took place in Cardiff and most of the teachers were from south and south-west Wales. However, as can be seen in Table 1, participants had a good range of marking experience (GCSE or A level) and represented the main centre types in Wales.

Table 1: Teachers' marking experience and type of school

School type	Years of experience				
	1-2	**3-4**	**5-6**	**7+**	**Total**
Maintained	-	1	1	12	14
Independent	2	2	-	1	5
Further Education College	1	1	1		3
Total	3	4	2	13	22

4.2. Scripts

Participants judged 23 GCSE History Non-Examination Assessment (NEA) part (a) responses that had been conventionally marked in 2019.

NEA is a compulsory component of GCSE History. It complements the external examinations and it is worth 40 marks, which is 20% of the GCSE History qualification. The completed NEA comprises two pieces of written work which are marked separately. The part (a) task asks candidates to analyse and evaluate historical sources addressing the chosen topic area and it is worth 14 marks. The part (b) task asks candidates to discuss an issue of historical debate that arises from the chosen topic area and is worth 26 marks.

Each centre selects the topic area that will be covered when completing the NEA unit. The part (a) responses are assessed across three AOs:

- AO1 - Demonstrate knowledge and understanding of the key features and characteristics of the periods studied (2 marks)
- AO2 - Explain and analyse historical events and periods studied using second-order historical concepts (4 marks)
- AO3 - Analyse, evaluate and use sources to make substantiated judgements, in the context of the historical events studied (8 marks).

The NEA tasks are accompanied by a pack of documents containing up to 25 sources (photographs, cartoons, speeches, etc.). The pack includes contemporary and later sources, which outline at least two interpretations of the topic. Candidates use the source pack as the basis for their 8-to-10-week research, which is conducted under limited supervision. The part (a) NEA task then requires students to select four sources and analyse how useful and reliable they are in explaining a certain aspect of the studied topic. Learners must complete the entire controlled assessment within five hours of formal supervised time, and part (a) should take up to two hours to complete. Candidates can handwrite or type the responses. It is advised that part (a) responses should be about 1,000 words, but this is not a mandatory limit.

Once part (a) response is completed, teachers mark it and internally moderate the marking, and then the awarding body selects a certain number of responses at random to be externally moderated.

For the CJ workshop, we asked three participants – from a comprehensive English-medium school, a bilingual (English/Welsh-medium) school and an independent school – to submit responses from an entire class via a secure website. The responses from two schools addressed the question of how useful and reliable the selected sources were in showing John F. Kennedy's foreign policy. The responses from the other school examined how useful and reliable the sources were in showing how difficult life had been for the inhabitants of the East End of London in the second part of the 19th century.

In total, 67 responses were submitted, of which 34 were handwritten. Responses from one school were all typed and a small number of the responses from the two other schools were also typed. Most of the typed responses were between 1,500 –2,000 words long; a much smaller proportion had over 2,000 words; and only a few had around 1,200 words. Almost all these responses were three or four pages. The length of the handwritten responses varied from five to ten pages. Almost half of them had eight pages, and about a quarter had six pages.

The range of marks available for part (a) response is 0 to 14. The distribution of the marks in the submitted sample varied from 7 to 14 with a median of 11, so did not represent the full mark range available on the task. Across the three schools, the average mark ranged from 10.7 to 12.1.

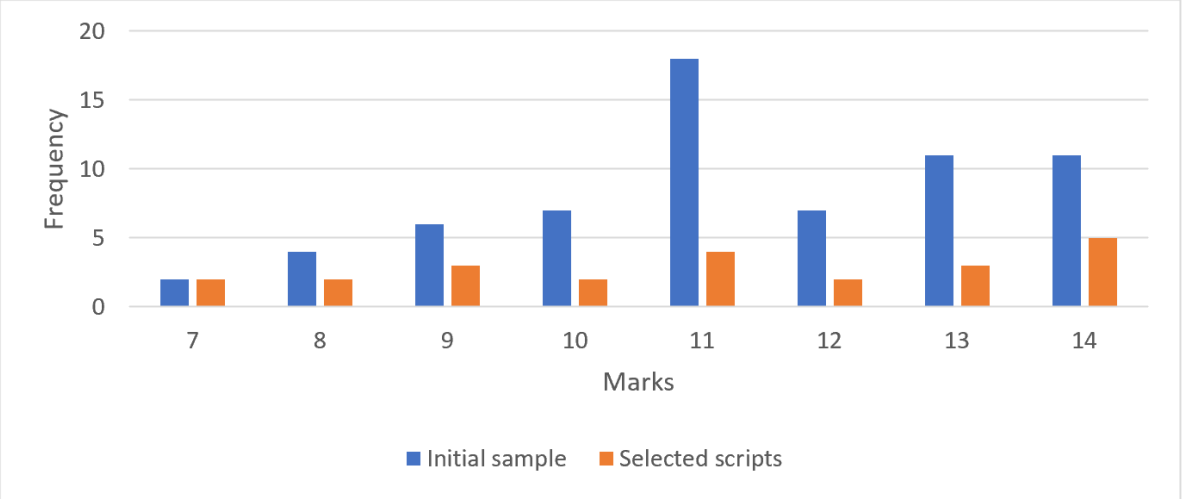
To reach an acceptable number of judgements for each response, only 23 responses, hereinafter scripts, were selected for the CJ exercise, of which 11 were handwritten (Table 2).

Table 2: Initial sample of responses and selected scripts

	Initial sample	Selected scripts
Total number	67	23
Of which, handwritten	34	11
Median mark	11	11
Mean mark	11.4	11.04
Range of marks	7-14	7-14
Standard deviation of marks	1.93	2.36

We selected the scripts at random but aimed to have a representative sample across each mark group within each class. We also aimed for responses with the lowest marks to be over-represented, to account for a small number at the lowest mark values. The mark distribution of the selected scripts is slightly different from the initial sample, although the median mark was the same as in the initial sample (11), while the average was close to the initial sample – 11.04 with a higher standard deviation of 2.36 (attributable to the oversampling of lower scoring scripts). The typed scripts' length varied from 1,147 to 1,810 words, and the handwritten scripts had a similar distribution of the number of pages to the initial sample.

Figure 1: Distribution of marks initial and selected samples



All scripts were anonymised, and the existing annotations and marks were removed to ensure judges' opinions were not affected by them.

4.3. Procedure

Participants attended a one-day workshop. During the first part of the workshop, they learned about CJ and conducted a trial judging session in which they compared the handwriting legibility of the selected responses. Not only did this act as a training exercise but it also provided information allowing us to assess whether handwriting introduced bias into judgement of quality (see section 5). In total they conducted 20 judgements rounds, achieving a reliability coefficient of 0.82.

The judging of responses was conducted in the second part of the workshop and lasted for about four hours. The question that judges were asked to consider was formulated similarly to the question in Pollitt and Crisp (2004): *Which script represents a pupil who is better at History?*

We had informed participants about the topics of the responses and sent them the relevant sources before the workshop. On the day, we provided them with the mark scheme and instructed that they could use it as a guide when deciding the winner but that this was not mandatory. Participants were instructed to make the judgements at a convenient pace and to not worry about needing to complete a certain number of judgements or worry about being behind others in completing the task. We instructed them to make the judgement whenever they felt they were ready to do so.

Participants were divided into two equal groups using stratified randomisation to ensure a similar distribution of teachers with the same years of marking experience.

One group (hereinafter Group F) was asked to judge the responses and provide feedback or comments on each response. We instructed that the comments could be short sentences or even a group of words that would describe the response and that the comments did not need to be similar to the annotating made during marking. Participants could see and add to their comments when encountering a script, the next time, but they could not see the comments made by other judges. The other group (hereinafter Group NF) judged the same responses but did not have to provide any feedback/comments.

Judgements were conducted using the RM Compare™ software. It presents pairs of students' responses online via an internet browser and the judge selects either the left or right response by clicking the relevant assigned key on the keyboard.
In the first judgement round, the software selects the pairs at random. In the next three rounds, it pairs the responses using the Swiss Tournament method. Under this method, in the first out of the three rounds, the winners and losers are paired respectively with winners and losers, and in the next two rounds, the method pairs the scripts that have the same number of wins. After the four rounds are completed, the software pairs the scripts using an algorithm. The algorithm selects the scripts that have the closest quality parameter values, which are estimated anew in each round, based on the judgements made by all judges (for a more detailed description see Rangel-Smith and Lynch, 2018). Therefore, the more rounds that are conducted, the harder it should be to make the judgements, as the paired scripts should have increasingly closer values. The selection of scripts based on the results from previous judgements makes the judgements adaptive, hence the method is called Adaptive CJ (Bramley and Vitello, 2019). In our case, the responses were also chained, i.e. during the adaptive phase, the software selected one response from the previous judgement to be part of the next pair.

Participants did not have any noticeable difficulties navigating the CJ software and got used to it quickly. The only inconvenience mentioned by many was the need to scroll up the page to press the 'next page' button. Participants thought that having this button positioned at the bottom of the page would make the interface more user-friendly.

4.4. Post-workshop interviews

After the workshop, we conducted semi-structured interviews with each participant. Out of 22 participants, we interviewed 10 during the first week after the workshop, and the rest in the following two weeks. Twenty participants were interviewed face-to-face and two were interviewed over the phone.

At the beginning of the interview, participants were asked to make two judgements while thinking aloud about how they choose the winning response. The think aloud

method asks participants to verbalise their thinking while solving a task. In this study it was used to triangulate what participants recalled about conducting judgements with observing the aspects they considered during the interview judgements. For example, the aspects mentioned when evaluating scripts during the think aloud part were corroborated with the responses during the interview. For this exercise, six scripts that were ranked above the average in the CJ exercise and had similar parameter values were selected. The intention was to select responses that would not be very easy to compare, with the hope of eliciting more comments from participants during the think aloud session.

After the think aloud session, participants were asked questions about their judging experience during the workshop and their views on CJ. When considering the usefulness of CJ, participants were asked to imagine an assessment system in which teachers would judge responses from other schools, as this avoids biases related to knowing the learners. The participants were also told to imagine that the whole exercise would not take more time than they spend on marking NEA responses. This instruction was given to avoid participants elaborating on this aspect, for which they would not have enough information to make an accurate judgement when considering the costs and benefits of the method.

5. Comparative Judgement results

This section briefly presents and discusses the quantitative results of the judgement sessions. Even though the focus of the study was to explore teachers' opinions, comparative judgement results will help put those opinions in context and allow comparisons with similar studies.

5.1 Number of judgements

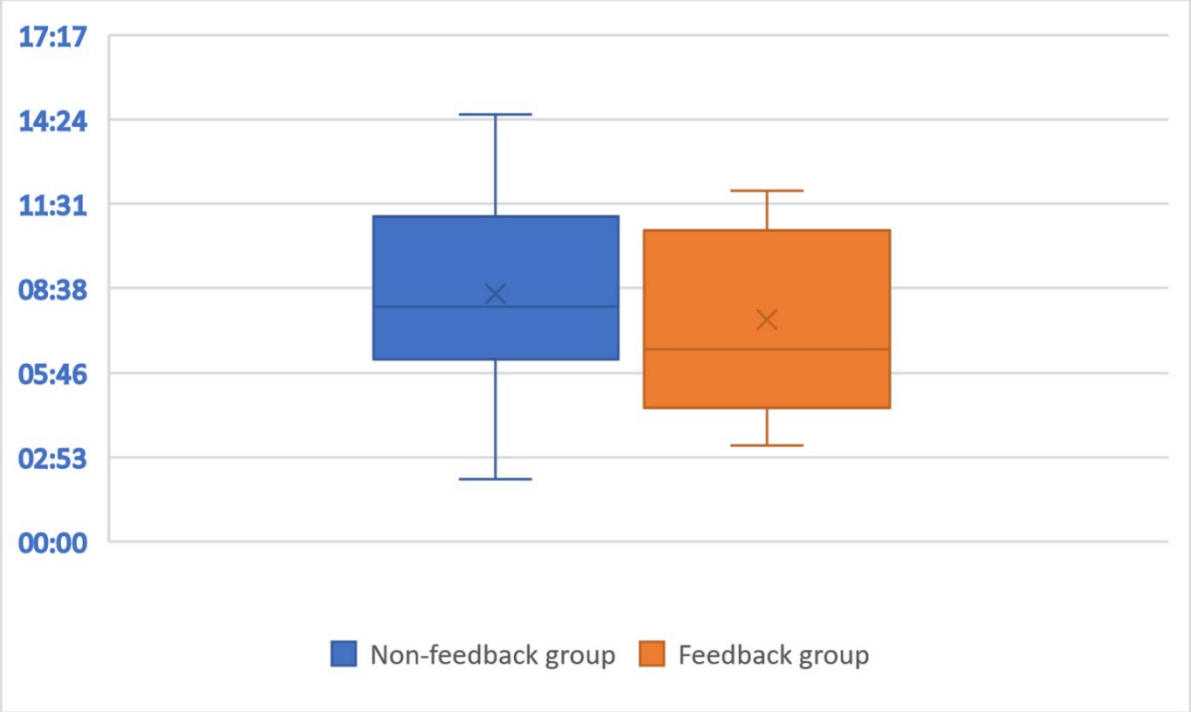
At the time of the study, the RM Compare™ software limited the number of judgements to 21 per judge. Most participants conducted 21 judgements and stopped, as the system would not allow them to make any more judgements; two participants made 17 judgements; one participant made 20 judgements; and another was somehow able to bypass the imposed limitation and make 28 judgements. There were 230 judgements made in each group (460 judgements in total); therefore, each script was judged on average 10 times. The range of judgements per script was 17 to 25 in Group NF, and 18 to 24 in Group F.

5.2 Judging times

The mean time for one judgement was 7 min 34 sec in Group F and 8 min 05 sec in Group NF (or 8 min 52 sec if excluding the times of a judge whose average judging

time was 2 min 8 sec[4]). These means are shown as crosses in Figure 2 below. The median judging time for Group F was 6 min 34 sec and 8 min 01 sec for Group NF. The average of judge averages was respectively 7 min 35 sec and 8 min 27 sec (including the 2 minute 8 second outlier). Figure 2 shows that there was quite a large overlap between the average time for each judge in both groups. The data shows that participants in Group NF were overall slower and had a wider spread of judging times than those in Group F, which confirms the subjective perceptions of most participants that providing feedback helped speed up the judgement process (see section 6.4.4).

Figure 2: Average judging time for each judge (minutes)



The time for one judgement is measured as the time elapsed between generating the pair allocation and recording the decision. To avoid the recording of idle time, we encouraged participants to log off from their accounts when taking a break (the software would only automatically log off participants after about 24 minutes of inactivity). But some participants would still occasionally enter a conversation (as far as we can tell not related to the judgement they had to make) with adjacent colleagues without logging off. This means that the mean judging time was probably lower than stated above.

[4] Despite their speed, the participant who spent an average of two minutes judging each pair was consistent with the other judges. Infit and reliability for all judges and scripts is discussed in greater detail later.

5.3 Reliability and misfit values

The Bradley-Terry statistical model behind the RM Compare™ software calculates the reliability and the infit values as well as other statistics pertinent to CJ. The judges in Group F achieved a reliability coefficient of 0.88 (+/-0.02). The judges in Group NF achieved a lower 0.74 (+/-0.02) reliability. Although the reliability indices for both groups were high, they were not as high as in other studies. For comparison, in a study judging geography essays with a length of no more than two pages (probably handwritten) and which could be given a maximum of 15 marks, the reliability was 0.97 after 12.5 rounds (roughly 6.2 judgements per script) and more than 0.9 after only seven rounds (Whitehouse and Pollitt, 2012). Another study compared responses to one question from the writing section of a GCSE English Higher Tier. Responses ranged from one to several pages and could be given a maximum of 40 marks. After about 14 judgements per script, a reliability of 0.97 was reached (Bramley and Vitello, 2019). It is possible that the lower reliability in our study was caused by participants having to judge responses from two different tasks. The effect of having to judge writing responses based on different tasks on reliability should be further explored.

It is also possible that the reliability coefficients in our study are lower because they are not inflated. Rangel-Smith and Lynch (2018), following Bramley's (2015) criticism that adaptivity inflates the reliability coefficients, explored the issue and suggested ways to fine-tune the pairing algorithm to address the issue. RM has implemented the recommendations and it is now suggested that 15 rounds of judgements should remove the reliability inflation (Kimbell, 2021). In our study, 20 judgements' rounds were conducted.

The statistical model also calculates the differences between the expected and the observed outcomes in a given judgement for each judge and script. These differences can be standardised in two ways and are called infit and outfit values, and are based on Rasch's logistic model which is equivalent with the model formulated by Thurstone (Pollitt 2012). Infit values are used to identify judges that were not consistent with the rest of the group, or to identify scripts that made judges less consistent when judging them, which is what Pollitt (2012) calls misfitting. As a convention, Pollitt (2012) indicated that infit values within two standard deviations of the average are considered acceptable, and those that are greater than two standard deviations are considered significant differences or misfitting. Judges whose infit values go beyond the acceptable limits take more unexpected decisions than the rest of the group. This could be interpreted as them having a different understanding of what good quality work is (Bramley, 2007 p. 260). However, when such interpretations are made, it should be kept in mind that a misfit statistic is a relative measure as it quantifies the extent that the decisions made by a judge differ from the

decisions that other judges in the group made. As such, misfitting does not necessarily indicate poor judging (Pollitt, 2012).

The average infit value of the judges in Group F was 1.03 with a standard deviation (SD) 0.345. One participant had a higher than acceptable infit value – 1.87. This judge left very short and general comments, to the point that they might not have been useful to help with judging[5]. This contrasts with what Daal et al. (2016) showed in their study where the one misfitting judge (higher infit values) had left more arguments per comment than other judges. Perhaps, the cause of misfitting is not the length of the comments, but the extent to which comments can usefully inform current or future judgements with extremes being correlated with more unexpected decisions. The infit values of the remaining judges ranged from 0.57 to 1.14.

The average infit value of the scripts in Group F was 0.99 with an SD of 0.3. All scripts had acceptable infit ranging from 0.59 to 1.54, except for one script that had a low infit of 0.08 and was ranked last, i.e. 23/23 (ranked 6/23 in Group NF). This script was given 12 marks. Please note that this is the only script that judges made no comment about. After obtaining the permission of the teacher who provided this script, we shared it with participants along with two other scripts that were also further away in the resulting ranking compared to mark ranking. We told teachers that the three scripts had some more pronounced differences in ranking and asked them to investigate if there was anything that would make them assess these scripts differently during CJ than when marking. Three participants replied, and the level of detail was not sufficient to gain any insights into the possible explanation for the misfitting. Judges' average infit value in Group NF was 1 (SD 0.22). One judge had a higher than acceptable misfit value of 1.47. The infit values of the remaining judges ranged from 0.76 to 1.42. The judge with the average judging time of 2 minutes and 08 seconds had an acceptable infit of 0.92.

The average infit value of the scripts for Group NF was 0.96 with a SD of 0.265. All scripts had acceptable infit values, ranging from 0.64 to 1.38, except for one script that had a low infit value of 0.08 and was ranked last (ranked 15/23 in Group F). This script was awarded 12 marks and is from a different school than the script with the low infit in Group F. The judges in Group F commented that this script exhibited good knowledge and contextual description but was weak on evaluation and analysis. This inconsistency could partially explain the difference of opinions on this script.

Both judges whose infit values were unacceptably high had more than seven years of marking experience. Their misfit values do not necessarily indicate that they did a

[5] Participants made 204 comments, that is, they made on average 9 comments per response, which means that judges left approximately one comment per judgement they made. Of the total number of comments, 20 were made in addition to the initial ones. Comments ranged from two words to 64 words.

worse job than the others, but that they had a different view about the quality of the responses. An interesting piece of data around the misfitting judges is that participants in this study would typically manage to conduct two to four judgements during the first four rounds, that is before the adaptivity and chaining would start. The judge with the unacceptable misfitting value in Group F conducted 11 judgements out of 21 during the first four rounds. On the other hand, the judge with the highest misfit value in Group NF conducted 5 out of 20 judgements in the first four rounds. Moreover, a judge from the same group conducted 11 out of 20 judgements in the first four rounds (similar to the highest misfitting judge in Group F) but had an acceptable infit of 0.9. Given this, the number of judgements done before adaptivity and chaining has a complex relationship with misfit.

Although some CJ studies (e.g. Newhouse, 2014; McMahon and Jones, 2015) use the two-standard deviation rule suggested by Pollitt, in Rasch analysis other rules are also used to identify misfitting values. For example, another approach is to consider the values that are outside of 0.70-1.30 range as misfitting (see for example, Hodge and Morgan, 2017). If this rule is used, then in group NF there would be two judges with infit values greater than 1.30 and two scripts below the 0.7 cut-off value. In Group F there would be one judge with the infit of 0.64 and two judges with the infit greater than 1.3. There would also be three scripts with infit values below, and three scripts with infit values above the cut-off points.

There was no correlation between the ranking of perceived legibility of the handwritten responses and their CJ rank, which suggests that the legibility of the handwriting did not affect the validity of the script comparisons.

The correlation between the rank order based on marks awarded to the scripts and the rank order based on CJ for Group F was 0.5 ($p=0.014$) and 0.73 ($p<0.01$) for Group NF. The correlation test performed was a two-tailed Spearman's rank correlation. To exclude ties for the mark-based ranking, in each mark group we ranked the highest response with the same number of marks that was ranked highest in the CJ, and so on. When ties are included, the correlation coefficients are much lower and not statistically significantly different from zero: for Group F the correlation coefficient was 0.25 ($p=0.24$) and for Group NF it was 0.31 ($p=0.15$). We computed the 95% confidence level of the Spearman's rank correlation coefficients by bootstrapping using RVAideMemoire package in R. For Group F the CI was from -0.18 to 0.64 and for Group NF CI was from -0.18 to 0.71.

In the Whitehouse and Pollitt (2012) study mentioned above, the achieved correlation with traditional marking was 0.63 after about six judgements per script. In the Bramley and Vitello (2019) study, the Pearson correlation of CJ results with the principal examiners' marks was 0.74 (Bramley and Vitello, 2019). In a study involving

shorter AS level History exam answers, the correlation varied from 0.84 to 0.86 (Holmes et al., 2018, p.22).

In this study the correlation test results are not conclusive. On the one hand, the correlation coefficients when ties are included suggest a weak relationship between the rank orders from CJ and marking. On the other hand, there is a large uncertainty around these coefficients indicating that correlation could be as high as 0.71 and as low as -0.18. This uncertainty is likely caused by the small sample sizes, the clustering of the scripts in a small range of marks and the large number of tied ranks in the mark scale.

A factor impacting the validity could be the length of the scripts. It is possible that the lengthier the scripts, the more difficult it is to achieve a higher correlation, assuming that more information should be evaluated and kept in mind when comparing the scripts holistically, in contrast to analytical marking where marks are given for different aspects and then combined into a final mark. Having to judge scripts on different topics could have impacted the validity as well. It is also possible that unlike annotating, which specifically targets assessment objectives, making open, unguided and brief comments made those comments diverge from the actual marking criteria which then gets exacerbated once the number of judgements increases. Therefore, further exploration on the validity of judging extended writing responses is needed.

The correlation coefficients (when excluding and including ties) were lower in Group F. Judges in both groups stated that they used the mark scheme and the assessment objectives to guide their decisions. As we discuss further, the only exception to this could have been the participants who never marked GCSE NEA or exams. When we divided participants into groups, we aimed that they were broadly representative of the type of school and years of experience. By chance, all teachers who had never marked GCSE were allocated to Group F. This could be a plausible explanation as to why the convergent validity was lower for this group. Another explanation for this discrepancy could be the fact that commenting affects the validity of judgements. Comments made earlier were sometimes used in decision making, and this could be another explanation.

6. Participants' experience of CJ and the aspects they considered when judging

Overall, almost all participants mentioned that they had enjoyed the exercise. Participants contrasted CJ with conventional marking and believed that CJ was a more engaging experience. As we will show in the next section, they thought marking was cognitively more difficult. In addition, the more structured approach

makes marking feel at times 'boring', 'tedious', and 'tick-boxy'. Conversely, CJ was considered more engaging because of the holistic approach to assessing. Moreover, participants thought that CJ would encourage learners to write more creative and less formulaic responses, which would make assessing responses even more engaging, although we have not explored whether this would make judging more challenging.

> *The more traditional form of marking does feel even more prosaic than it did before now, somehow.*
>
> *I was expecting myself to get a bit bored. Click, click, click. But actually,... once I got into the swing of it, I was okay.*

Participants thought they got used to the new assessment approach quickly and that they got better at it as they made more judgements. They felt that the accuracy of their decisions remained the same throughout the day. However, some participants mentioned that the background noise in the room might have affected their concentration.

Participants felt that it was more difficult to judge by the end of the workshop. They considered this to be 'a natural slowdown' which would happen during marking as well, and which was not worse than when marking. In this regard, participants mentioned that it was necessary to make sure judges take breaks, which were considered even more important given that responses are on screen.

6.1. Approaches to reading responses

Based on observations during the think aloud sessions and on participants' recollections during interviews, it was established that judges took several approaches in familiarising themselves with the two responses before making their judgement. Some participants read one response and then read the second one. Another group of participants would read a bit of each response consecutively. Most of these judges would usually read the analysis and evaluation of one or two sources from each response, although a few would read one page. This latter approach was used as a way of quickly finding the place where they had left off in the previous response. A third group of judges would first 'get a little bit of a feel' of at least one response before reading the response(s) in full. For example, they would read the introduction or the first page, go to the conclusion and then read the entire response.

When reading, the participants would sometimes use skim reading as a technique. In this case, they would select only some elements from each response, then decide which one was better. Most of the time, skimming consisted of reading more

attentively the introduction, the conclusion and one or two sources, while only scanning the rest of the response. This approach to skimming seems to assume a well-structured response and may not be as useful for less structured responses.

Another approach to skimming was to scan-read the entire response while selecting some content for more detailed reading. Participants said they use this latter approach when marking, although they felt that during marking, they needed to be more thorough because they needed to annotate. This type of skimming seems to rely on identifying the 'key' or 'buzz' words and then zooming into the text around those words.

When explaining their approach to skimming, many participants could not describe it as it was something they would do instinctively.

We do not have exact data to investigate whether different approaches to reading the responses had any effect on the quality of judgements. The way judges read the responses could be further explored to establish whether there is one most efficient way of reading the responses, regardless of personal preferences.

6.2. The need to read responses in full

A decision that participants had to make was whether they could make the judgement after reading or skimming only a part of a response.

A small proportion of participants suggested that to make a fair decision, a judge needs to read both responses fully. These participants said that they would keep reading or, at least, skimming the responses in full, even if it was clear that one response was better than the other. They felt that the quality of a response can change, therefore it would not be fair to make a judgement after familiarising themselves with only a part of it.

> *I could never ever just read a little bit and mark off that 'cause I don't think it's fair.*

There was some evidence that at least some of these participants would read the entire response out of habit, but with more judging experience they could change their approach.

> *P: I have to read it all. But I'm 99.9% sure that B will be better.*
> *I: So, why do you think you need to read it all?*
> *P: Because you need to just make sure. I … if I'm honest what I'll do now is how I would mark.*

But the vast majority of participants (including the two with unacceptable misfit) agreed that it was unnecessary to read or skim a response in its entirety when it was clear which one was better.

In such cases, these participants would either stop reading the response or would selectively read some additional elements (e.g. conclusion or another source). The decision to have this additional check seemed to depend not only on the strength of their conviction, but also on the preference of each judge.

> *I would at this stage go to the judgement at the end just to see the quality of it before, and I'll see if there's a need to go back to the rest of the answer.*

Participants who considered that it was not always necessary to read each response completely said that they do not take the same approach when marking because they felt assigning a mark was different from deciding which response was better.

The conviction that one response is better than the other can be formed quickly. For example, during think aloud sessions, participants were already able to make a decision after only having read the analysis of one or two sources in the second response.

> *Yeah, he's written eight sides, but even just read the first page and a half, you can tell this is better because he interacts with the source.*
>
> *Two pages in, this one is significantly better than this one. I don't even need to read the rest, because I can see that. And it's by the end of each paragraph you can see that they're doing that higher order thinking in terms of the analysis.*
>
> *You get an idea from the first two sources. I'd say 95%, 99% of the time source three and four will be the same as source one and two.*

The important question is whether the incomplete reading of the responses affects the quality of the judgements. Almost all participants were confident about the quality of their decisions. This was mainly because they considered their marking experience had helped them and because they felt comparing was easier than marking, which requires having to 'pinpoint actual marks'. The two standard deviation misfit statistics presented earlier and the correlation of CJ results with the marks seem to show that the incomplete reading, did not have a notably negative effect. However, more investigation of this phenomenon would be needed, especially considering that more misfitting was identified when applying the 0.7-1.3 rule-of-thumb.

Judging chained responses was a particular case of incomplete reading. All participants said that they would not read or look at the response that was shown the second time in a row unless the responses were of a similar quality, thereby making it harder to make a final decision. In this case, participants would return to the chained response to selectively check parts of it.

Participants liked chaining as they thought it helped them make quicker judgements. They also felt that chaining did not make their judgement biased.

The CJ software would always place the chained response first. For each group, a statistical test[6] was performed to determine the strength of the statistical evidence that there was a bias in choosing the first response over the second response.

We removed the decisions from the first five rounds: in the first four rounds the chaining does not exist and in the fifth it is not yet evident. In the remaining rounds the first (chained) response in Group F was selected 84 out of 173 times or in proportion of 0.49 (99% CI: 0.388-0.584; P – 0.704) and in Group NF 97 out of 173 times or in proportion of 0.56 (99% CI: 0.463- 0.658; P – 0.110). In neither case was the p-value statistically significant at the 1% significance level. Therefore, there was no strong statistical evidence in this study that the chaining led to a bias in selecting the chained object more often (although the sample size may limit the analysis).

Pollitt and Crisp (2004), who also implemented chaining, mentioned that no bias had been found in any studies (most probably referring to educational field) implementing this approach. However, a study using paired comparison of objects related to economic and environmental goods showed that when the scripts are encountered more times (thereby offering the possibility of remembering them) the independent assumption of CJ was open to question (Brown and Peterson, 2009, p. 82). This issue needs to be further researched.

6.3. Aspects considered when making the decisions

Participants familiar with the mark scheme stated that they used it to decide which response was better. These participants would consider the same aspects that they would when marking. For example, participants would check whether learners answered the questions of utility and reliability of the examined sources; they would assess learners' understanding of the context; and the style of their exposition. However, because they knew the mark scheme well, they did not have to look at it during judgements.

[6] Test of binomial proportion assuming a normal approximation for which the assumptions ($np>5$ and $nq>5$) were satisfied.

> *You're still doing the same thing, you're still reading it and you're still applying the mark scheme, so it's still the same process.*
>
> *I didn't think we needed the marking scheme because we all knew what was entailed in the essay.*

At least three of the participants out of the five who did not have any experience with marking GCSE responses did not use the mark scheme and judged the responses against their understanding of a well-written response. The infit values of these five participants were within the acceptable range.

> *[The colleague who sat near me] and me were both from colleges where we don't have that mark scheme; we didn't know that mark scheme. We had a quick look over it before we started [judging], but we went with our gut and what we deemed to be a good historian, and that would be someone who can back up source material with their knowledge.*

Table 3 below presents the aspects that were included in the comments made by participants in Group F with additional data for judges who had never marked GCSEs. Participants from both groups also mentioned these aspects during interviews and some of them are further explored in the subsequent sub-sections.

Table 3: Aspects mentioned in comments by Group F

Aspect	Total number of comments	Percentage of total comments in which the aspect is mentioned	Number of comments made by judges who marked only A levels	Percentage of comments in which the aspect is mentioned by judges who marked only A levels
Evaluation	82	40	28	34
Knowledge/contextual knowledge	79	39	41	49
Analysis	53	26	30	36
Conclusion	52	26	18	22
Reliability/utility	37	18	27	33
Source content	21	10	6	7
(Non-)Mechanistic/formulaic answer	18	9	10	12
Introduction	9	4	2	2
Structure	5	3	3	4
Reference to Assessment Objectives (AOs)	2	1	-	-
Spelling	2	1	1	1
General comments (e.g. attempts to answer the question; valid, but generic; not 100% focused on the question set)	40	20	15	18

All of the aspects mentioned in the comments, except those referring to the mechanistic writing, introduction and structure are reflected in the mark scheme. Each judge (except one who left one/two words comments) referred to various aspects in their comments and did not seem to focus only on some aspects. It is likely that they mentioned the aspects that were most noticeable in the script they commented on.

As can be seen from Table 3, judges who had never marked GCSEs, and had only marked A level papers, were overall in line with the whole group regarding the aspects they mentioned in their comments. This can be expected considering that A

level History NEA mark scheme emphasises the analysis and evaluation criteria as well. The language used in the comments suggests that these judges were, to a certain degree, guided by this mark scheme. For example, these judges mentioned whether the scripts reached or had a supported judgement, an aspect that is emphasised in the A level specification. For comparison, only two out of the remaining six judges mentioned judgement in their comments. These five judges also paid more attention to the knowledge aspect which again seems to be mentioned and weighted more in the A level specification. These observations could explain the low convergent validity (when ties were excluded) for this group, and also confirm that markers would rely on mark schemes when judging.

It is not clear what weighting was placed on the aspects mentioned in the comments. However, it is reasonable to assume that they played at least some role. From the table above and the comments made during the interviews, it is obvious that participants followed the aspects included in the mark scheme, which suggests that CJ could be a valid method of assessing extended writing responses, but this should be confirmed quantitatively by research with more robust design.

Even if participants used the mark scheme and the AOs as reference points, they felt that decision making during CJ was different from marking. When comparing, participants took a more holistic approach to assessing the responses because keeping the atomised approach that they would use during marking would not be efficient or reasonable.

Participants did not consider that using this holistic approach negatively affected their decisions. They also mentioned that holistic assessment should not be difficult to adopt, because, to a certain degree, it is already used when marking some subjects. A level History and Welsh Baccalaureate assessments were given as examples of such an approach. However, for some, this novel way of looking at the responses felt ‘daunting’ at the beginning, but they were able to quickly adapt.

To arrive at their decisions, participants used different approaches. For example, some participants answered the question, ‘What do I think of this essay?’. This overall impression then was used to decide which response was better. Similar to findings in Marshal et al. (2020), other participants would assign a band to responses or, in the case of two participants, an approximative mark.

Regardless of their approach, it seems that many participants were unaware of all the aspects that affected their decisions. Participants often mentioned in this regard their reliance on gut feeling. They found it difficult to describe how gut feeling worked or what it was. When describing it they would use words like ‘being in the zone’ and ‘having a feeling for the work’.

Participants were sure that this intuition about a response comes with experience of assessing responses. Therefore, they thought that newly qualified teachers might have more difficulties in being consistent if not guided by the mark scheme.

6.3.1. Evaluation and analysis of the reliability and usefulness of the sources

During the interviews, participants mentioned the analysis and evaluation of reliability and usefulness of the historical sources as the most important aspects when judging the responses. Reliability and usefulness were mentioned in 37% of the comments, while evaluation and analysis were mentioned in about two-thirds of the comments (40% and 26% respectively). This is roughly reflective of the mark scheme in which analysis and evaluation are mentioned as part of AO3, with analysis also being mentioned under AO2.

Participants stated that when considering how well learners evaluated and analysed the utility and reliability of the sources, they used the same approach they would follow during marking. For example, when considering how well the learner analysed the reliability of a source, the participants would assess whether they considered bias, elements of exaggeration, and other factors that might affect reliability (e.g. propaganda or purpose of the source).

> *The bulk of the mark then would be looking at the origin, seeing who wrote it and why they wrote it, how it came about that their conclusions come out the way they do. The date - that's in the origin as well. Is the date significant to why they wrote it? So, that all, that would come in there and then you've got the purpose then. Who is the audience? Why are they writing it? What are they trying to get across?*

6.3.2. Knowledge

Part (a) responses can be awarded only two marks for knowledge. These two marks also include aspects of quality of the written communication. Participants felt that during marking these two marks are usually easily given and, to gain them, the learner is not expected to have a great level of knowledge.

> *So, when I'm reading the first paragraph, I normally expect to see some sort of knowledge about the question. Now in the mark scheme, there's not many marks for knowledge, so that knowledge is absolutely fine.*

> *…with [the mark scheme] being in my head, I know they don't have to have too much knowledge.*

When marking, participants would therefore pay more attention to aspects relevant to AO2 and AO3. The importance of knowledge, however, seems to be greater during CJ, and especially when participants encounter responses that display similar levels of historical skill. In conventional marking, the knowledge element of both responses would have been given two marks (barring the absence of obvious errors). However, during CJ, a response gains an advantage if the judges consider it to show more sophisticated or more detailed knowledge.

Therefore, it seems that CJ could place a greater emphasis on contextual knowledge used to support the evaluation. In fact, participants from Group NF mentioned knowledge or awareness of context in about 40% of their comments, which is similar to the proportion of the comments mentioned under evaluation. However, it is possible that during marking knowledge is also considered beyond AO1 as arguably learners can not analyse and evaluate the sources without showing a grasp of contextual knowledge.

6.3.3. Complexity of the analysis

Participants noted that the exhibition of broader historical skill was another important aspect considered during CJ. They regarded it as a further demonstration of the quality of a response in addition to the examination of the reliability and usefulness of sources. This skill consisted of taking the arguments about a source's reliability and usefulness and developing them further by, for example, explaining whether the arguments could be limited, or by comparing sources. Another aspect of this skill was the ability to integrate knowledge into evaluation and analysis which would also result in less mechanistic writing. A well-developed response would also explain the results of the analysis, highlight the insights, and consider the wider context, which is also important for the basic analysis of reliability and usefulness.

> *This one's got quite a nice bit of sophistication the way the candidate weaves in. She analyses the source and then she contextualises it with what she knows and then she brings it back to the question and explains why it's useful. She has got a nice critical tone in it as well. So, she's still critiquing the evidence. It's nice and sophisticated.*

6.3.4. Mechanistic writing style

The avoidance of a mechanistic writing style was another aspect that participants considered during judgements. Although it seemed difficult to define concisely what a mechanistic writing style was, it mainly related to an exposition that aimed to tick the boxes in the mark scheme. For example, a mechanistic response would have each paragraph addressing a criterion of the mark scheme and/or would overemphasise

certain key words (e.g. reliable, useful). Because of this formulaic approach, the response does not feel like an integrated work and lacks 'flair'.

Participants perceived that, when marking, such responses are likely to get the same marks as the more sophisticated answers, because they tick the marking criteria. However, during judging they were at a disadvantage when compared to responses that instead of 'going through the motions' overcame this mechanistic framework and showed more thinking.

> *The other one is fine, but it is clearly, you know, they're talking about reliability but it's not really doing anything because it is doing it because they've been told to do it.*

> *So they've both done Source A3, so they've said about utility. It's... a more sophisticated style of writing as well in A; it seems far more mechanistic in B. They're, sort of you know, the way they're introducing things is repetitive and very much following a set structure, rather than flowing.*

> *[B is] more aware of the importance of the key words than A. But A is ... it's, [...] warmer. It's more human, you know. I'd rather have a chat with this student than this student.*

6.3.5. The role of the structural elements

Apart from content requirements, at least some participants considered the structural elements of a response when judging. For some of them, a good structure that consists of 'an introduction, a conclusion and four paragraphs' was an indication that the response should be broadly in line with the requirements for a good essay. This does not contradict participants' point on valuing non-mechanistic writing to the degree that this concept is applied to the content of the response and not the structure. Even a non-mechanistic response needs to be organised and flow from one structural element to another.

> *My initial processes were just to skim and scan and look at the actual structure.*

The introduction was considered helpful for getting an understanding of what the learner will cover or address in the paper. When marking, teachers would not explicitly consider giving points for the introduction because a response does not strictly 'have to have an introduction', but also because an introduction would usually contain generic information that would not count towards the candidate's marks. Nevertheless, learners might lose marks if the introduction is missing or weak, as it could create a negative impression of the response.

During CJ, participants used the introduction as a deciding criterion when two responses were of similar quality. Participants in Group F mentioned the introduction in about 5% of the comments. However, references to the introduction were general statements, such as 'clear focus in the introduction'; 'generic introduction'.

The conclusion is an important structural element, and a weak conclusion or lack of it would lower the scores when marking. During CJ, the conclusion was used to create the overall impression about the response; as a tiebreaker for similar quality responses; and as an aspect that is considered in more detail when skim reading or not reading an answer fully. The participants mentioned the conclusion in more than a quarter of their comments.

6.3.6. The importance of the key words

Participants used key words to make judgements about the quality of the response and to identify the text that they would read more carefully when skim reading. Overall, certain key words would create a more positive impression about the response.

The key words that participants paid attention to could be categorised into three types: subject words, assessment criteria words, and transitional/analysis words.

Subject words are specific to the topic of the response. Examples of these words are 'propaganda' and 'communism'. Assessment criteria words, such as 'reliable', 'biased' and 'useful' are relevant to AOs. Both these groups of words were used to identify the passages that needed more attention during skim reading.

The third type of key words, such as 'however', 'this suggests' flag the analytical elements of the response. For example, when learners use a 'this suggests' type of sentence, it would show judges that those learners were 'reading into the source' and thinking about what the source means. These words are also important during marking, although they are not necessarily reflected in the mark scheme.

> *It talks about useful, useful, useful. Evaluative useful, highly useful, also useful, there we go. … Yeah, okay, so it's wordy but it's doing what it says on the tin.*

> *[…] they've used the word [reliable]; have they actually assessed it? It may well be biased … yeah, so it's very, very vague, […] not really properly examined.*

> *I'm looking for a 'however', which is, which is what, you know, when a student looks at the balance, and there's not a 'however' in this at all.*

I haven't found a lot of [key words] in A, so it makes me think that candidate B is the, I don't know, the better one, you could say.

As soon as I see those words, I'm focusing on those sentences obviously to check the reliability comments and the usefulness comments and how they're applied to the answers.

6.4. Challenging aspects of judging

Most of the time participants did not have any difficulties deciding which response of the two was better. However, in a small proportion of judgements, participants found they needed to consider more aspects of the responses before deciding. This happened partially because of natural tiredness, but also because the responses presented certain difficulties which are discussed in the following subsections.

6.4.1. Judging responses of similar value

Participants felt that in the second half of the judging session, the paired responses were of a more similar value, which is in line with how the CJ algorithm would pair the responses. These responses needed more careful consideration and were more difficult to decide regarding relative quality. Of these responses, participants thought that it was particularly difficult to discriminate between two weaker responses because they 'all say the same thing' or 'sound the same'. At the same time, strong responses would always have something that would help scorers discriminate more easily between them.

Having to compare and decide a winner between responses of similar value was felt 'a bit artificial' because during marking, both responses would be given the same marks. During CJ, participants adopted various strategies to make these decisions.

Some participants would re-examine certain elements of a response, such as the text around the key words, the conclusion or other 'triggers' of the assessment objectives.

When two are quite similar to each other, then I'd be thinking, "What am I actually looking for?" And I think ... and in those instances, I was looking at if they'd understood the bias of the source.

I was focusing on to see is one more slightly sophisticated than the other, or have they included the audience and the other one hasn't?

Other participants did not focus on one aspect but tried to find any subtle differences between the two responses.

So, it could have been just something that perhaps in one there was ... it could have been something as simple as spelling, punctuation and grammar. As we mark this as well because AO1, it's about coherence, quality of written communication and knowledge understanding.

Many times, participants would rely on their gut feeling or intuition to decide between two similar responses. Most participants were open about it, while some were more reserved in recognising it.

Participants were instructed that they could choose a response at random when not sure which response is better (e.g. by flipping a coin), but we are not aware if any of the participants did so. Those participants who were asked about this approach during interviews mentioned that they preferred to decide themselves than to leave it to chance. Theoretically, this could be a problem as the model would assume a random decision when the scripts have the same value on the latent attribute. Not deciding at random could create a bias depending on the aspect that is used to decide the winner. In our study, very few such instances were mentioned.

6.4.2. Judging responses on different topics

Some participants found it difficult to judge responses on different topics, or thought that some teachers might find it difficult to judge such responses.

So, I, I constantly felt like I was probably not going to ever be 100% accurate [judging] those. Whereas if they were all 19th century social conditions, or I think that's the one that we did for the other one, I, I feel like I would've been far more accurate if I was constantly doing the same one.

Participants felt that knowing more about one topic than the other can affect the accuracy of the judgement, as judges might not be sure of the quality of the response on the less familiar topic. This might make them assess the response on the unfamiliar topics either more harshly or more leniently (allowing for the benefit of the doubt).

So, it could be problematical if the person marking doesn't actually teach that area, doesn't know much about it. So how do you know what they're saying about the conditions of the East End? How do you know that that's factually correct?

[Before the workshop] I looked at the source and on the background, but that was maybe a little bit challenging. [...]I think it would be really useful if the person is teaching that particular topic.

Teachers might also enjoy one topic more than the other, and this can create an unconscious bias towards favouring responses on the preferred topic. As one participant explained, in such cases, judges can have a good feel about a response, but this might be for the wrong reasons.

Despite these difficulties, participants still considered that CJ could work well even when judging responses on different topics. It was felt that having responses on different topics could even make judges focus more on the skill of the learner as opposed to getting bogged down in the content.

However, these judgements would be more cognitively demanding, require more time and probably be less reliable. To make judging responses on different topics more accurate, it was suggested that teachers obtain some background knowledge on the unfamiliar topic before judging.

To make CJ feasible, it would be then required that the NEA topics are restricted to a reasonable number[7]. Participants overall were not against this idea. Some stated that they would not mind if schools studied the same topic, as they felt this would level the playing field and make the assessment results fairer. However, other participants disagreed with this because it would advantage the schools that have more experience with the topic. These participants supported only the idea of a limited number of topics, as this would allow schools to choose a topic closer to their expertise and learners' needs.

> *I think if everybody was doing something similar, you could monitor in terms of the support and the access to the sources that they should be finding independently easily.*

> *I would be concerned about those pupils who've taken A level because they've got a natural passion for something and may want to do a topic that isn't necessarily within the syllabus. They could then be penalised for that.*

Overall, participants agreed that it would be more difficult and less efficient to compare responses on different topics. They believed that it would be even more difficult to compare part (b) responses as those are more complex, and there was an even greater need to consider the historiography. However, more research would be needed to evaluate these opinions quantitatively.

[7] Another approach would be to ask teachers to compare the work of their learners mixed with a smaller proportion (e.g. 20%) of pairs using scripts from other schools. The scripts from other schools act as anchors and their estimated values are used to calibrate the values of all the scripts involved in CJ (see Wheadon, C., et al., 2020a). The drawback of this approach compared to the approach used in this study is that it does not remove some of the teacher bias and that it probably does not have a similar professional development impact.

As with the similar value responses, it was felt that weaker responses on different subjects were harder to judge because there is less evidence of higher order skills, and historical knowledge becomes more important to make the judgement. However, the judges would not be able to appreciate it because the topic was not familiar to them.

6.4.3. Judging handwritten responses

Unsurprisingly, participants mentioned that it was much easier to read typed responses than the handwritten ones, especially those with less legible handwriting. Having to read scanned handwritten responses on screen made the reading even more difficult. But this did not seem to be an issue that had a great effect on judging, as a large majority of participants did not feel any inconvenience in the process of making judgements.

Having to read handwritten responses did not seem to affect the opinion about those responses. As mentioned earlier, we asked participants to rank the legibility of the handwritten responses. The ranking of response legibility did not correlate with the CJ rank order in any of the two groups.

But this does not mean that teachers are not paying attention to the quality of handwriting at all. During the interviews, for example, teachers commented on handwriting and characterised it as 'neat' or 'immature'. Sometimes teachers referred to the learner using the personal pronoun 'she', inferring the gender from the neatness of the handwriting.

> *Let's start with this one. And I do remember this one [from the workshop]. 'Cause I remember saying, "Must be a girl, you can tell." Be interesting you could do a girl or boy handwriting and I bet we'd get it right.*

Overall, participants mentioned that judging typed responses would be preferable. However, they doubted that all schools could ensure access to computers. Some also worried whether learners could type their responses quickly enough in the context of a time limited assessment.

> *We've got a massive problem with access to computers for students that don't do IT, DT and subjects that need computers.*

6.4.4. Making comments

As mentioned earlier, half of the judges were asked to provide feedback or comment on the responses.

Participants made 204 comments in total, making on average 9 comments per response and 18.5 comments per judge, which means that judges left approximately one comment per judgement they made. Of the total number of comments, 20 were made in addition to the initial ones. Comments ranged from two words to 64 words. Shorter comments were more general and probably less useful in contrast to more detailed comments, as can be seen from the below comments made on three different responses:

> Participant 1: *Weak introduction*
>
> Participant 2: *Evaluates sources*
>
> Participant 3: *More generalised introduction, although focused. Analyses quotes and provides historical detail which is relevant. Analysis of usefulness is a little formulaic e.g. content then origin. Some simplistic points e.g Daily Mail might be biased as it was not American. Conclusion is on the question.*

Participants said that the comments they made were meant to help them judge and were not intended as feedback to learners, except one participant who addressed their comments in the form of feedback for the learner. For example:

> *When making analysis of the reliability or utility of sources explain what this means for your judgements. Consider the purpose of the sources you investigate.*

Participants felt that making comments was easier than annotating during conventional marking. This is because conventional annotating is more detailed, while for CJ, the participants left more holistic and high-level comments.

Although some participants considered annotating during marking helpful, others did not feel it was useful to them and felt that it made them spend more time with marking. In contrast, participants from both groups overwhelmingly preferred to make comments because they thought commenting would help decide when comparing the responses, especially when encountering the response the next time[8].

[8] We should note, though, that this would not be the case if CJ is implemented at a larger scale, as the possibility of encountering the same response could be lower.

Indeed, participants who commented often used their comments as a tiebreaker when looking at the response a second time and when the other script was of similar quality. Sometimes they primarily based their decision on the comments alone, quickly scanning the script for confirmation (or vice versa). Some participants felt that, sometimes, only reading the comments would be enough to make the judgement.

> *If I'd written on it 'an excellent paper', [and], if the other paper then was appalling, very poor then I wouldn't read it again.*

> *Once you've seen [a response], you could just look at the comments [made about the response] to be reminded about [it]. Then you didn't have to read it again unless the B example was quite similar, and then you had to look at them again and then decide.*

> *So, I was looking at the comments to double-check, and sometimes I had comments on both of them, and I just, I'd just go back to quickly scan just to make sure that my comments were correct. And so, from the comment bank then, I, well, comments, I could decide which one was the best.*

Participants believed strongly that making comments sped up the judgements. Even teachers who did not make comments shared this opinion, as they could observe that the colleagues from the other group were not behind them with the number of judgements accomplished. This subjective perception was confirmed by the average judgement time in each group. Making comments was thought to be especially useful for long responses, for example, part (b) responses, which can be as long as 18 pages.

Some participants (especially in Group NF) worried that existing comments would prevent judges from properly reading a script when it is encountered the next time, thereby affecting their accuracy. Conversely, some participants from Group F felt that reading comments made their decisions more accurate. However, there was a concern that judges might think a comment made previously referred to the response they were currently reviewing, whereas it was in fact a comment referring to a different response.

> *You might see the comments and say, "Okay, that's the one," but then it might be the other one.*

These observations suggest that making comments can add to the variability of the judgements in at least two ways. Firstly, making comments can affect the judgement process and lead to different outcomes compared to when comments are not made.

It is possible, for example, that aspects that are included in the comments trigger an availability bias when the final decision is made, thus distorting the holistic judgement of the exercise. Secondly, relying on previous comments could affect the independence assumption of the statistical model used to analyse the data. Considering that judges in Group F felt that commenting was helpful, and it did not slow down the judgements, it seems that commenting could help judge longer written responses more quickly. However, this advantage probably does not exist when CJ is applied on a large scale, in which case the probability of encountering previous scripts is very low. Nevertheless, the effects of commenting should be further explored for when a small pool of responses needs to be judged.

Aside from being useful during CJ, comments could have a formative value if passed to the teacher, (e.g. to improve future coursework). However, for this to happen, comments would need to be more specific and detailed, rather than comprise a few generally descriptive words. Comments could also be used to examine the misfitting scripts or when examining appeals against CJ results.

Participants suggested that commenting could be further sped up by including ready-to-use comments in the software (e.g. 'detailed knowledge'). Some felt the need to annotate a response and suggested that the software could make available short codes that would resemble annotation, e.g. R for repetition. Participants also felt that it would be useful to be able to highlight text (e.g. use distinct colours for different AOs) and if possible, integrate comments in the body of the paper.

6.5. The CJ question

The question asked of judges can present some challenges if it is not clear. The majority thought that the question asked - *Which script represents a pupil who is better at History?* - was in line with what they would look at during marking.

> *The one who I think would be better at history would be the one who'd get the most marks for reliability, usefulness. So, it does ... it did work.*

Some participants however thought that it might be interpreted differently as referring to the inclusion of other aspects than those considered during marking. They suggested that the question should be worded so that it is closer to the question the learners answer in their essays, e.g. 'Which learner is better at looking at the usefulness and reliability of sources?'.

7. The perceived benefits of CJ

7.1. Effects on the assessment

7.1.1. Effects on the validity and creativity of the responses

Research studies (see Wheadon et al., 2020a) have found that the use of mark schemes for writing, although intended to increase reliability of assessing, could also negatively affect validity and learners' creativity, as they start to target the mark scheme in their responses.

Participants believed that using CJ in marking would eventually steer learners to be more creative and to focus on skills and understanding rather than on the 'ability to jump through hoops'. This is because CJ requires a holistic approach to assessing, and the value of a response is defined by the relative quality of other responses. They believed that while this steer would motivate the high-end learners to write better answers, the less able learners would not be disadvantaged, because writing responses should not be more difficult than it currently is. It was suggested that this shift would offer a better preparation for further progression in terms of promoting creativity when approaching their assessments.

> *Students would perhaps be better rewarded for their individuality and their flair than perhaps they are if they're marked against the, just against the mark scheme. Because if they're marking just against the mark scheme, and they haven't done something that the mark scheme specifies they need to do, it might be as brilliant as you like but they're still not going to get the marks for it, whereas this allows you to reward that flair and individuality perhaps more.*

Participants felt that CJ would also encourage a teaching style that focuses more on the skills than on the elements that need to be included in their responses. The emphasis of correct knowledge of historical events by some participants when judging does not necessarily contradict this, as it could be considered as being integral to skill. Some participants thought that if CJ is used in marking, they would rather focus on exemplar answers (including A level and undergraduate papers) and teach historical skills from them. Some suggested that they could even use CJ as a formative tool for the learners by asking them to compare different responses.

7.1.2. Fairness

Participants considered that teachers marking their own learners were prone to a degree of subjectivity. This is because teachers' preconceived ideas about their learners' abilities makes objective marking difficult. They also pointed out that it is

likely that teachers become 'emotionally invested' in their own learners and therefore are more lenient when marking them.

> *[CJ] would take some subjectivity away because there is, I think, a ... whether we like it or not, there's always a bias towards our learners. And you always try and be as, I would be honest, be as tolerant as possible.*

Some participants also believed that teachers' approaches to marking might not be consistent across schools. For example, participants felt that there was a reluctance among some teachers to give zero or even one mark for a separate AO because this could be perceived as 'mean' or not supportive of the learner's effort. Conversely, some teachers would be more reluctant to give full marks than the others. These predispositions might unfairly discriminate against learners in some schools, especially when a limited range of marks is available, as it is the case for part (a) response.

Participants thought that assessing NEA responses using CJ would make assessment conducted in centres more objective than using conventional marking, as it would remove some biases of conventional marking. For example, they thought that comparing two responses removes the leniency or harshness bias that some markers might display when marking. The accuracy of the assessment would also increase because many more teachers would consider the same response.

> *I mark it once, I take it to moderation, it might be moderated, it might not. So, that script goes in. Done. Never seen by another colleague.*

Regarding bias, it is worth mentioning that there was a possibility that if only a few of all the judges make the judgements in the first rounds, they could bias the adaptivity process if they were prejudiced in any way (e.g. they all placed the same disproportionate weighting on a certain aspect when judging responses, e.g. the neatness of handwriting). In our study, all judges started judging at the same time. However, out of 46 judgements in the first four rounds in Group NF, about 57% of judgements were made by three judges and 67% by four judges (out of 12 judges in total). In Group F, out of the same number of judgements and judges, 50% of judgements were made by three judges and 61% of judgements by four judges. We cannot tell if this biased the results in our study. However, the disproportionate number of judgements made by each judge in the initial rounds also suggests that it would be very difficult to ensure all judges would conduct similar numbers of initial judgements in a large-scale application of CJ if no restrictions are in place. One solution would be to design the sorting algorithm so that it does not allow the more active judges to proceed beyond the first rounds until all judges have participated.

Finally, several participants mentioned that CJ could reduce the possibility of cheating. They mentioned that it was possible that teachers could give more marks to all responses and then ask learners whose work was selected for external moderation to correct or add to their responses. Participants did not say they knew this had happened, but the mere possibility of this being possible seemed to be of concern to them. Although CJ would not entirely prevent this, it was believed that it could make it less likely, because teachers would not feel they need to defend a mark in external scrutiny. Besides, uploading the responses shortly after they are completed creates fewer possibilities for inappropriate reworking.

7.1.3. The ease of decision making

Participants considered that comparing responses was easier than assigning a specific mark to each response.

> *When you first start out, marking an assessment, it is really hard and I, I don't think in teacher training, you ever really get an idea of how to mark. But I think everybody can judge. So, I think for newer teachers and for most people, you are able to judge the difference between something that's better or worse.*

> *If I had a new starter, I'd rather them compare these rather than try to put a specific mark on something. I think that's far harder.*

Participants mentioned that the ease of comparing is especially helpful for subjects such as history, which they considered to be highly subjective and therefore difficult to mark accurately. This perception of subjectivity echoes the findings that questions which are aimed at the exploration of sources are the most difficult to mark in History exams because of the possibilities of such varied responses and interpretations of the sources (Holmes et al., 2018, p.12).

> *You could have three colleagues in a school that all mark a set and then come together. They've taught it the same way; they've marked it the same way. And they sit down to moderate, and they just don't agree.*

> *But then you've still got to try and work it out against the mark scheme as to what mark you're actually going to give it. And you're deliberating, "Oh, is that 11 or is that 12? I don't know." Whereas this, I think I find it easier to say, "No, that is better than that one."*

The difficulty of achieving accurate marking places an additional psychological burden on teachers because they feel that given the narrow range of marks available, it is more likely that a one-mark error could make a difference for learners. To reduce this psychological pressure, participants said that sometimes

they had to consider a response extensively or even, as one participant put it, 'agonise' over the correct marks.

> *Because that's where I find marking coursework stressful, is when you're in that dilemma, "Do I give it this? Do I give it that? What mark am I going to give it?"*

Participants felt that CJ would remove this psychological pressure, not only because comparing is easier but also because many more teachers would assess the same response, which reduces the likelihood of a mistake. This should hopefully result in each script getting 'its right place and right mark'.

Participants mentioned that aside from deliberating how many marks a response is worth, they would also compare different responses. This helps them calibrate their marking and ensures that marking is fair overall. Sometimes this means that they must go back to already marked papers to check that the awarded marks feel right.

> *…you ask a lot of people in moderation, they're doing that anyway, it's almost like they're marking it twice in a way.*

> *You know it's all subjective. It takes time. You kind of second guess yourself for lots. You go back and re-read and, "Should I give that a couple more?"*

The tendency to compare responses has been ascertained in other studies as well (e.g. Daal et al., 2016) and our participants thought it was even more prominent when teachers mark their own learners, because teachers might want to ensure that the marks reflect the rank order that they have assigned (if only mentally) to their learners.

CJ was considered more efficient, not only because it already addressed the need to compare the responses, but also because it removed the need to go back and re-check previous decisions.

A downside of CJ is the need to make many comparisons which can feel daunting.

> *That's right, if you get a counter at the top that says 500, and then it marks down and you're thinking "Oh I've worked hard" and you look at it and go "I've got 394 left", I think that might be a little bit stressful.*

7.1.4. Convenience

Participants had mixed views on how convenient it would be to assess responses using CJ compared to conventional marking.

Some participants mentioned that because they wanted to ensure that marks are consistent across responses, they preferred to mark them during a short period. This pressure does not exist in the case of CJ, so assessing can be conducted over longer periods. This is a significant advantage, as CJ can be conducted whenever the teacher has a timeslot to do this, providing that controls exist so that teachers conducting judgements at different times do not distort the results of CJ.

However, some participants thought it might be more difficult to assess responses during school breaks or lunch time as they currently do when they mark. They believed that interruptions, which are likely in a school environment, would probably make them start reading the response from the beginning. This does not happen during marking, as teachers could use the annotations to remind them about the response quality.

> *And the school environment does, I suppose, disturb you quite a lot from being able to sit there for two hours on something, or an hour on something.* So, I think *[CJ] would be something that most people would probably do outside of school rather than in the school, and then I don't know whether they would be so keen on that because teachers can be quite precious about time outside of a classroom ...*

Participants had mixed views on the advantages of using a computer to assess responses. Some participants considered that the preference to mark paper responses is mainly a matter of habit. Once teachers get used to the CJ software interface, they would find it easier and quicker to assess the responses on screen than marking responses on paper. Participants said that onscreen marking gives them quicker access to the responses. In addition, having a touchscreen laptop or tablet makes the process even easier than when using a mouse.

> *That's a bit of a weird one. I would have said in the past I did not like onscreen marking until I came to the judgement session, and then I was surprised how much I didn't mind it. I thought that I wouldn't be able, I thought I'd miss things on a screen that I wouldn't necessarily miss if I had it written down in front of me, and I was physically doing it. I didn't really feel like that ...*

However, marking paper responses has its own advantages. For example, teachers can take the papers away with them easily and mark whenever it is convenient. These opportunities are more limited in the case of CJ as it requires a computer/tablet with

a reliable internet connection. However, the upside is that teachers will no longer have to worry about the security of the responses.

> *I think you're very much tied then to... you've got to be sat at your desk. Whereas with a hard copy you could do it on the train, you could do it anywhere.*

7.1.5. Administrative burden

Participants suggested that CJ could ease the administrative work that is currently required in relation to the NEA responses.

When learners complete their responses, teachers collect a signed form from learners that confirms that the work they have submitted is their own. Teachers store the completed papers in a secure place until the marks are confirmed. Because the learners can undertake the tasks at any time during the duration of the course, the storage can take months. Some teachers mentioned that this can cause a certain amount of anxiety as they would worry about the safekeeping of the papers.

Once a sample of papers is selected for the external moderation, teachers need to mail them to WJEC. Some participants mentioned that this could create additional pressure if the submission deadline is around a bank holiday.

Participants felt that CJ would make this administrative process simpler because the papers would be uploaded to a dedicated site.

> *I think the processes of paper marking [...] is just so labour intensive and there's postage and worrying about scripts getting lost and those sorts of things. Where I think if everything's uploaded to a system there's a lot more peace of mind that it's kept safe.*

However, this benefit assumes that learners type their responses. Handwritten responses must be scanned, and this would offset the advantage because it was felt that scanning could take even more time than is currently needed for the administrative work.

We should note though that the administrative and marking convenience are not only advantages of CJ, and this could be achieved with onscreen conventional marking as well.

7.2. Effect on professional development

Participants mentioned that receiving feedback on the quality of the responses their learners produce was an important avenue towards improving their teaching. The feedback they receive comes from internal or external moderation. Some participants found moderators' reports useful and worried that if CJ is implemented, teachers would no longer receive that feedback. On the other hand, participants mentioned that the moderation exists to ensure consistency of marking. So, any feedback coming from moderation covers only a limited part of the aspects that might be relevant for their professional development.

Moreover, some participants believed that the information received during moderation is not always helpful or reliable. They felt that groupthink affected internal moderation, while external moderation reports were not always specific, which limited how helpful they could be. The participants, for example, mentioned that the external moderation report would often contain only general phrases, such as the 'marking was fine' or the marking was 'too harsh or too lenient'. The usefulness of this report is further limited if the same moderator reviews the marking system of schools over the years.

> *Well, it's not great at the moment, because at the moment you get a moderator's report, and it tells you if they agreed with your marking, if they thought you were out, but it doesn't actually give you any detail. So, they may tell you that you've been too generous, but it doesn't tell you how generous or was it at the bottom end or the top end. So, it's a bit of an enigma. You send this work off, somebody says, "Yes, it's okay," and then you just rinse and repeat the following year.*

In contrast, participants considered that judging other schools' responses was a great professional development opportunity because it allows teachers to see how other schools approach the NEA task and have a better understanding of how standards compare across centres.

> *It was eye opening of how could we ... how could I set out the coursework for my own students and ... you know, like sharing good practice really.*

> *From seeing some of those [responses], particularly the Jack the Ripper ones that I teach, I'm like, "Oh yeah, I'm definitely on the right lines. I'm telling the kids the right things."*

> *... in our school we're in a little bit of a bubble. So, whilst I think my NEAs may be really good, I could look at somebody who is in another school in Cardiff and*

mine could be average. So, whilst I could scale internally, I'm not really scaling to the national picture. Whereas I think [CJ] would really help.

8. Implementing CJ as an alternative to conventional marking

Participants felt that CJ would be easy to conduct by experienced and inexperienced teachers. A large majority of participants considered that CJ was better suited for assessing NEA part (a) responses than marking. If CJ would not take more or much more time than conventional marking, these participants said that they would prefer to assess using CJ.

This section explores the aspects that participants mentioned would need consideration before CJ could be implemented.

8.1. Involving and informing teachers

Despite CJ appearing to be a great tool for assessment, participants cautioned that teachers would resist its implementation if they feel it is imposed on them. Participants also mentioned that teachers might be reluctant to change something they already know, and are used to, for something new. The resistance to change was also invoked by a few participants as the reason they would still prefer conventional marking.

I'd go for traditional but then I'm at the end of my career so, that's what I've been used to all along. If it was, if I was at the beginning of my career, I'd probably go for that. I'd go for the opposite way.

But from my own perspective I've been using traditional [marking] for over 20 years and I think I've got a natural confidence marking that. So, I'd be inclined to say traditional.

Participants believed that teachers should be informed about CJ, and consequently involved in conducting judgements, so that they could get a feel for it (e.g. as a formative or professional development tool). Participants believed that most teachers would favour CJ once they try and understand it.

Participants suggested that some of their colleagues would be reluctant to move from marking on paper to onscreen marking. Some participants gave examples of markers who they knew refused to mark when the awarding body asked them to do so onscreen. Having this gentle introduction to CJ would help such teachers overcome their reluctance or even fear of having to work with technology.

Another aspect that participants suggested would make teachers reluctant to accept CJ was the uncertainty over the final results, and the concern over whether their learners would receive the grade they deserve. Participants mentioned that the current marking system gives teachers more certainty about this.

> *I know pretty much that the mark that I give them is almost a 100% the final mark that they'll get. So that's the main thing. It's the unknown, it is.*

> *I think people will think, "Well, how do I know the people making the judgements on our centre are viewing it as we would?" I think people would just be very unsure.*

> *It would be scary as a teacher to think that, "I don't know where that mark's going to be now when it comes back," because, you know, I bank on, "Right, this is this way," do whatever, for their coursework.*

The concern about the final marks also seemed to arise from the fear of teachers judging responses from schools that have different performance levels compared to their own school. Teachers from higher attaining schools were concerned that their learners would be marked down, as judges would not consider additional aspects that add value to a response. On the other hand, participants from lower attaining schools worried that their learners' work would look poor when compared to schools with learners coming from more advantaged neighbourhoods.

> *I feel a little bit sort of uneasy that somebody else is marking my coursework, if their coursework is awful; that'll be one disadvantage.*

> *… because they're from really good areas, those pupils naturally can write better than some of the pupils I teach [who] find reading difficult. So if you're going to have an answer from one of the more prestigious schools and you're going to have an answer from one of my pupils […] so my pupil will always be put down in comparison to that and I would feel that that's really unfair because although they might be getting to the skills, their fluency might not be as good and I'd feel that that would be a bit of a disadvantage for them.*

Participants thought teachers might have similar concerns in relation to borderline C grade learners who they felt could be currently 'pushed' to get a C.

We believe these concerns could be addressed by properly informing teachers about how results can be arrived at when using CJ. Emphasising the safeguards of CJ to prevent arbitrary decisions (such as misfit), more teachers looking at the same response, the possibility of having a separate process of determining grades and grade boundaries, could also help.

8.2. CJ guidance for teachers and learners

Participants indicated that teachers would need clear guidance on how to choose the better response. They believed that the current mark scheme includes the relevant aspects, therefore, they suggested that it could be used as a starting point in drafting the guidance. However, it was perceived that the new guidance should be more holistic and offer more 'free rein' and ensure that learners have a chance to write more creative responses.

> *So, you need to have some kind of reference points for the examiner to look to see what to look for. This needs to be as scaffolding without limiting creativity.*

Participants also suggested that the aspects included in the guidance should be described at a more general level and should avoid too much detail. They noted the A level History mark scheme for NEA as an example of such an approach. In their opinion the guidance should comprise bands of descriptors, for example, what is good, very good and excellent. They also felt that this is how the current mark scheme is used in any case.

> *The mark schemes are not really that much of, of a guide, and sometimes it's more of a, of a gut instinct or the fit of that level. Even when marking. You know, you hear stuff in conferences where there's lots of dispute in terms of tolerance with responses because the mark schemes, it's very, they're just wordy. And it could be, just be a change of one word that then flips you from getting a level, a Band 5 to a Band 6. So, I think if we were going down the comparative judgement route, I think the mark schemes would need to change so that they're a bit more holistic ...*

Apart from guidance for teachers, participants felt that there would be a need for a framework that would outline the expectations or success criteria to learners. Participants thought that using guidance for teachers could be difficult for learners to understand, as is the case with the current mark scheme. Therefore, they opted for having a separate document in clear language for learners.

> *Whereas, you know, we can't help but go, "Ladies and gentlemen, here's the mark scheme, what do you understand by that?" And they don't, and then it's this uphill battle.*

CJ can be used to create a scale of calibrated exemplars which can then be used when assessing responses (McGrane, J. A, 2018). This could be another way of maintaining consistency, although it would then be a different approach to assessing responses than the one discussed in this study.

8.3. Training

Participants felt that working with the CJ software was easy and some short practice exercises would be enough for teachers to get used to it. Some suggested that a short activity, like the one they conducted at the beginning of the workshop, when they compared the legibility of writing, would be enough for teachers to learn the technical aspects of navigating the software and making judgements.

Participants believed more training would be needed to teach judges how to correctly make the decisions. However, they thought that judging was relatively easy, and that there was no need for substantial training, or at least no need for training that is any more onerous than the training teachers already receive to become markers.

As part of their training, teachers would need to familiarise themselves with how CJ works (including the random order of responses in the pairs when not chained) and with the judgement guidance. This theoretical part should be then followed by a session of CJ conducted under supervision, or in groups, which would offer the opportunity to receive feedback on the judgements made. This stage should then be followed by a session of CJ using responses with known values.

Some participants thought that CJ training could be held within schools, whereas others suggested that it could take place online. It was also suggested that involving teachers as trainers would be desirable.

> *I would say teachers giving training to other teachers would also be good because we tend to trust each other as professionals ...*

Participants felt that informing judges about the quality of their decisions (e.g. how far away from the consensus they were, or flagging responses that were difficult to judge) would further help teachers to calibrate their judging skills. The training could be complemented with additional insights into effective judgement once this information was available.

> *Because I looked over my note as well and I thought ... and after ... yes, because several times whichever I chose for B appeared in A. That happened several times, and I was thinking, "Oh, why's that happening? Is B better than A?" Or vice versa.*

8.4. The length of responses

Participants considered judging long responses to be more difficult than shorter ones. They felt that long answers make retaining the information more challenging and that skimming could become less efficient because important aspects could be missed.

Short responses are also easier to decide on because they create the sensation of quicker progress. On the other hand, exceedingly long responses could be judged more harshly as they can generate a negative bias against them.

As mentioned earlier, the specification indicates that part (a) responses contain about 1,000 words. But this is only a recommendation and is not mandatory. Therefore, many responses are much longer than the recommended length. Participants believed that a maximum number or range of words should be enforced, and most of them believed that a range of 1,000-1,500 words would be acceptable for part (a). This would increase the manageability of the CJ process with little risk to the validity of the assessment.

8.5. Response types

Participants had mixed views on whether judgements could become more, or less, difficult depending on what is assessed. Some participants felt that part (a) responses were straightforward to judge but part (b), which is an interpretation essay, would be cognitively more demanding. On the other hand, some participants felt that part (b) responses would be easier to compare because they are more complex, which means that it would be easier to discriminate between responses.

Similarly, participants thought that CJ could be appropriate to mark A level History NEA as the assessment approach in A level is more holistic than for GCSEs. The only downside to this suggestion is that A level responses are long (4,000 words) and are written on various topics, which would make comparisons more difficult.

It is expected that responses of different complexities could influence the ease of judging. However, it is not clear how this would manifest and whether the demand would differ from marking.

8.6. Time needed

As mentioned earlier, participants' opinions of CJ as a marking tool assumed that it would not require more, or much more, time than marking does. Under this assumption, most participants considered CJ to be preferable to traditional marking.

An important question for the successful implementation of CJ then is whether the assumption of a similar amount of time is correct. Two variables have an important effect on the time needed to complete CJ: average judgement time and the number of judgements required.

The median time for making a judgement in Group NF was approximately eight minutes. If we assume that the responses for the part (a) task would be capped at 1,500 words and that they will be all typed, we might hope that on average a judgement should take about five minutes.

Table 4 illustrates the approximative time needed for various tasks under current marking arrangements, based on a class of 22 learners. Most of the participants said that they would mark three to four responses an hour. If we take this as an assumed average marking time, it means that a teacher would need about six hours to mark 22 responses. Participants also mentioned that it might take up to one day to complete the paperwork for one class, upload the marks, and send the selected papers for external moderation. It could take a further half day to one day to internally moderate the marking. Thus, the total time needed currently is about 18-22 hours.

Table 4: Approximative time spent on traditional assessment based on 22 learners

Task	Marking responses	Administrative tasks	Moderation tasks	Total
Time (hours)	6	8	4-8	18-22

CJ would not require internal or external moderation and, providing that the responses are typed, CJ would require an insignificant amount of time for the administrative tasks related to processing the answers. Considering this, we would assume that to achieve the same time efficiency, the actual judging should take no more than about 17 hours for a class of 22 learners (thus allowing 1-5 hours for administrative and reviewing tasks). This means that to achieve similar time efficiency, we would need to limit the number of judgements per response to about 9-10. This is similar to the number of judgements conducted in this study.

As we mentioned earlier, the reliability in Group NF was 0.74 (we are focusing on this group because we consider not providing feedback as the probable approach to CJ when used in a real assessment context, and also because they had a lower reliability which should ensure a more conservative approach to the discussion that follows). Bramley (2015) indicated that adaptability inflates the reliability coefficients and Bramley and Vitello (2019) suggested 16 judgements per object were needed to

achieve good reliability[9]. As we noted earlier, inflation should not take place with the software we used, as its algorithm was optimised to reduce the inflation. As Kimbell (2021) indicates, Bramley agreed that the changes made by RM to its software would mean that 15 judgement rounds (i.e. 7.5 judgements per script) would remove the risk of reliability inflation. Given this, we would assume that the 0.74 reliability achieved by Group NF was not affected by inflation. If we take this as a base value, would it be acceptable if conventional marking were replaced?

Essay-type questions elicit the lowest levels of conventional marking reliability because unlike, for example, a mathematical test, it is very difficult to precisely determine how to allocate each mark included in the scheme (Kellaghan et al. 2019 p. 167). As mentioned earlier, a study conducted by Rhead et al. (2018, p.4) has shown that the probability of receiving a 'definitive grade' for exams in England was lowest for extended responses for English language and literature (0.52) and history had the second lowest consistency marking. Holmes et al. (2018, p.17) found that the reliability of marking AS History source-based exam writing questions from three different papers was in the range of 0.52 to 0.62.

We do not know whether the reliability of the conventional marking for these responses examined in this research is similar to what others have found. On the one hand, we can assume it might be lower, given that the responses were lengthier; on the other hand, it might be higher if we consider the external moderation. If we assume that the reliability of NEA GCSE history assessment is around 0.6, then 10 judgements per response would probably offer a rank order that is at least as reliable as conventional marking. However, additional research to establish this would be needed before a more definitive conclusion about this aspect of the CJ feasibility can be made.

Overall, we believe that for extended writing it is likely that CJ reliability is at least as good as that of the conventional marking, as it is generally accepted that assessors are better at relative than absolute judgements. In a study, Gill and Bramley (2013) asked a group of history teachers to determine the grades (A, B or C) of a set of History exam papers (probably A level) using relative and absolute judgements. Most scripts in this study were around the A/B boundary which made the decisions more difficult (scripts ranged from 57 marks up to 75 and the A/B grade boundary was at 68 marks while the B/C boundary at 61 marks). Although judges had approximately similar levels of confidence about their accuracy when using both approaches, they displayed less variability and were much more accurate when using relative judgement (66.1% of papers assigned the original grade as a result of comparative judgement versus 41.7% when making an absolute judgement).

[9] In another simulation study conducted by Crompvoets, E. A. et al. (2020), it was found that one needs on average 20 comparisons to reach a reliability of .80.

If conventional marking reliability is higher than what can be achieved after 10 judgements per script, a possible way of dealing with this would be to test if guiding the initial selection of pairs by providing additional information about the value of responses would increase the reliability of CJ while keeping the number of judgements low (see more Bramley, 2015). As Bramley explains, overestimating of reliability indices during ACJ happens because the objects' values are not known beforehand. Therefore, all parameters are estimated concurrently as the data from the judgement are fed into the model. This produces 'spurious separation' of the objects. Introducing some information about the value of the objects will reduce the risk of overinflation. In our case, this could be some estimates provided by teachers based on the knowledge they have about the learners' work. However, this idea needs to be further investigated to check whether the initial values are not biased for various reasons, which could in turn create problems with the adaptive stage of the process.

9. Study limitations

The focus of the study was the qualitative analysis of the judgement process and teacher opinions of the method. The study is based on a convenience sample that ensured a reasonable representation of teachers. We cannot verify whether participants' views on CJ could be applied to the entire teacher population. Nonetheless, most of the opinions expressed regarding CJ are in line with what studies collecting feedback from judges have found (e.g. Pollitt 2012), although there are few studies of such sort. Therefore, additional research to confirm the findings and conclusions reached in the current study would be recommended.

The current study was not designed to collect data for robust statistical analysis. Therefore, the statistical findings of this study should be viewed critically. For example, it is possible that the small pool of scripts made judges see some of them more than once in a relatively short interval. This could have affected the assumption of independence which, in turn, would make it more difficult to interpret the statistical coefficients and indices. The use of comments in one of the groups could have further exacerbated this, as comments made previously were used when the script was judged again. Commenting could also have impacted on the holistic nature of the judgements. Finally, assuming that lengthier responses could be more difficult to judge consistently, probably a greater number of judgements per script were needed to observe the dynamic of reliability indices.

10. Conclusions and further research

The key aim of this study was to explore teachers' opinions of using CJ to assess NEA GCSE History part (a) extended written responses and identify the aspects they consider when judging.

Participants believed that CJ had several advantages over conventional marking. The four prominent advantages were:

- the relative ease of making comparative judgements;
- the mitigation of certain biases in assessments conducted by a learner's teacher;
- the potential for professional development; and
- the potential to reduce the undesirable behaviours linked to targeting the mark scheme.

The method would also provide more assessment information (e.g. estimates of reliability based on the whole entry), which is not currently available for internally marked NEA. Overall, participants would be happy to use CJ, providing it does not take more or much more time than conventional marking.

This study shows that CJ is likely to be a valid alternative to marking and moderating NEA GCSE History part (a) responses. In agreement with Whitehouse (2013)[10], the findings show that teachers, overall, applied the same criteria that they would apply during marking. While Whitehouse (2013) found that judges omitted some of the mark scheme criteria, we did not find evidence of this, as all the aspects included in AOs were mentioned. However, in line with the mark scheme weighting, the aspects that carried fewer marks (e.g. quality of written communication) were mentioned less, which might suggest that they carried less importance during judging.

Participants were not entirely aware of the decision-making process. The findings suggest that some participants could have applied more frugal rules when making decisions (e.g. the participant who managed to judge much quicker than the rest) while others could have been inclined to use more complex, analytical type of approaches (e.g. judges assigning a mark to each script). It is likely that judges would not consider all mark scheme aspects in every judgement. This may especially occur when judges believe that they can make the decision based on skim-reading or on incomplete reading of the scripts. In such cases, it is possible that judges compare the two scripts by looking at one or two aspects that carry more weight. If there is a clear difference between the two scripts, they would decide without reading further or in more detail, although it is not necessarily always the case. If there is no clear difference, judges would consider other aspects, while also reading the scripts more thoroughly. This approach could explain why judges mentioned only some aspects in

[10] Daal et al. (2016) had similar findings in a university setting.

their comments or during thinking aloud sessions and why the more important aspects were mentioned more often.

In this regard, it would be useful to explore how the selective consideration of assessment criteria can impact the validity of the judgements, and whether a set of rules or simple heuristic can be devised for comparing scripts that would increase validity and efficiency.

Similar to Whitehouse (2013), we found that participants relied on aspects that were not included in the mark scheme. In our study, participants mentioned the so-called mechanistic writing style. This aspect was used as a tiebreaker, but also had a role in creating an overall impression of a script. This could impact the validity of the judgements compared to marking, however an interesting question to investigate would be the degree to which these aspects have an impact during conventional marking as well.

There was some evidence that participants could have weighted some aspects of the mark scheme slightly differently. This was considered overwhelmingly to be an advantage of CJ, as it allowed participants to assess and weigh up more appropriately the aspects that they considered needed to be credited. This poses interesting questions about how the holistic approach is applied by each judge. Judges may be influenced slightly differently by different aspects when making the decision which creates variation. Further variation can be created by judges not reading the scripts in full, or by judges encountering more unusual scripts, such as scripts that are longer, or on less familiar or different topics. It is not clear how all these considerations impact the validity. Therefore, more robust research is needed to further investigate this aspect.

Even if CJ is as valid as marking, validity doubts would probably remain given that the holistic decision process does not have the transparency that conventional analytical marking has. Therefore, we consider that when using CJ, assessors would still need guidance akin to the current mark scheme, although the aspects to be considered will need to be formulated to reflect the more holistic nature of the decision making. They may also need training to maintain a common understanding of the assessment construct, and consistency in considering various relevant aspects of a response.

CJ as a method has been shown to be reliable, but its reliability depends on the number of judgements made for each script. As such, reliability is closely related to feasibility. Some studies have shown that a reliable CJ assessment would require no more time than traditional marking (e.g. Kimbell, 2012; Steedle and Ferrara, 2016, but see Wheadon et al., 2020b as an example of opposite findings).

Our findings do not rule out the possibility of CJ being a feasible approach when assessing part (a) NEA History and probably similar types of extended answers. However, this would depend on what levels of reliability (and therefore the number of judgements required to reach those levels) would be acceptable when using CJ for a specific subject or even assessment task. Even if we take the reliability indices obtained in this research at face value, we do not know how they compare with the reliability of conventional marking because we do not have data on the reliability of the latter.

This research suggests that to maintain the same amount of time for assessment, teachers would need to conduct between 9 to11 judgements per script. Increasing the number of judgements with the aim of achieving higher reliability would make the implementation of CJ for these responses less feasible, unless the process is further optimised. For example, teachers could assign each response a provisional grade that would increase the information gained following the judgements, increase the reliability of CJ, and not be a substantial burden. Therefore, this option could be further explored if higher reliability is desired.

Apart from the number of judgements needed, several other aspects (out of those explored in this study) must be addressed to make CJ workable. Firstly, in order to reduce the time needed to process responses, learners should type their answers. Secondly, the length of responses should be capped to a maximum of 1,500 words. This would not only reduce the time needed for judging, but also may improve the quality of judgements. Finally, because judging responses on different topics was considered to be more challenging and is probably less reliable and efficient, we may consider limiting the number of topic areas that one judge sees to two, if not even one. However, further research should be conducted to explore how multiple topics impact the results of judgements and how to address the potential problems.

Chaining the responses, that is, seeing one response two times in a row, is another factor that speeds up the judgements. The law that Thurstone formulated, and the statistical model, assumes that judgements are independent, and a decision made in one judgement does not affect the decision in another judgement. We did not find strong statistical evidence that the chained response was selected more often as the winner, but we can't rule out the effect. This issue needs further exploration because chaining is important in ensuring that CJ is feasible for the extended writing responses.

As an aside, consideration might be given to using CJ as a formative assessment tool for extended writing of such levels of complexity. As noted above, some participants mentioned that they would have liked to use it for peer feedback in their classes. When used like this, CJ would give instant feedback to each learner about their performance that can be updated by using CJ repeatedly over the course of a

subject. In addition, when used for peer feedback, CJ has been shown to improve significantly the achievement of learners that were initially ranked lower (Seery et al., 2019). If used at the national level, CJ could offer teachers information about standards across schools and facilitate an exchange of professional experience. This seems especially useful if more autonomy is granted to schools in relation to the curriculum content. It is worth mentioning that the Independent Review of Curriculum and Assessment Arrangements in Wales by Professor Donaldson (2015, p.97), which served as a basis for the current curriculum reform, mentions that sharing good practice among schools is one of the ingredients for the successful implementation of a good curriculum. Employing CJ for formative assessment would also be easier given that this does not face the same pressures as the high-stakes summative assessment.

An interesting avenue of further investigation would be to examine the assessment of responses that have a higher level of complexity than those used in this study, such as part (b) responses and A level History, or responses given for other subjects.

If CJ is to be considered as an alternative to the conventional marking, further work is needed to investigate how this method could be used in a high-stakes assessment context and consider aspects that go beyond the actual assessing; for example, explore such aspects as the right of learners to appeal, and the organisational aspects of implementing CJ.

References

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice,* 18(3), 279-293.

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgement. *Journal of Applied Measurement,* 6(2), 202–223.

Bramley, T. (2007). Paired comparison methods. In P. E. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds*.), Techniques for monitoring the comparability of examination standards*. (pp. 246-294). London: Qualifications and Curriculum Authority.

Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgment*. Cambridge Assessment Research Report. Cambridge, 36.

Bramley, T. and Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 26(1), 43-58.

Brown, T. & Peterson, G. (2009). *An enquiry into the method of paired comparison: reliability, scaling, and Thurstone's law of comparative judgment*. Gen Tech. Rep. RMRS-GTR-216WWW. Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Research Station, 98 pages.

Coertjens, L., Lesterhuis, M., De Winter, B. Y., Goossens, M., De Maeyer, S. & Michels, N. (2021). Improving Self-Reflection Assessment Practices: Comparative Judgment as an Alternative to Rubrics, *Teaching and Learning in Medicine*, 33(5), 525-535.

Crompvoets, E. A. V., Béguin, A. A., & Sijtsma, K. (2020). Adaptive pairwise comparison for educational measurement. *Journal of Educational and Behavioral Statistics*, 45, 316-338.

Daal, T. V., Lesterhuis, M., Coertjens, L., Donche, V., & Maeyer, S. D. (2016). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice*, 26(1), 59-74.

Donaldson, G. (2015). *Successful Futures: Independent Review of Curriculum and Assessment Arrangements in Wales*, WG23258.

Gill, T. & Bramley, T. (2013). How accurate are examiners' holistic judgements of script quality? *Assessment in Education: Principles, Policy & Practice*, 20(3), 308-324.

Hodge, K.J., & Morgan, G.B. (2017). Stability of infit and outfit compared to simulated estimates in applied setting. *Journal of Applied Measurement*, 18(4), 383-392.

Holmes, S., Black, B., & Morin, C., (2018). *Marking reliability studies 2017: Rank ordering versus marking – Which is more reliable?* Ofqual publication. 33p.

Jones, I. & Iglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educ Stud Math*, 89, 337–355.

Jones, I., Swan, M. & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *Int J of Sci and Math Educ*, 13, 151–177.

Kellaghan, T. & Greaney, V. (2019). *The Reliability of Public Examinations, Public Examinations Examined*. Washington, DC: World Bank. © World Bank. https://openknowledge.worldbank.org/handle/10986/32352 License: CC BY 3.0 IGO.

Kimbell, R. (2012). Evolving project e-scape for national assessment. *International Journal of Technology and Design Education*, 22, 135–155.

Kimbell, R. (2021). Examining the reliability of adaptive comparative judgement (ACJ) as an assessment tool in educational settings. *International Journal of Technology and Design Education*., 1-15.

Marshall, N., Shaw, K., Hunter, J. & Jones, I. (2020). Assessment by comparative judgement: an application to Secondary Statistics and English in New Zealand. *NZ J Educ Stud,* 55, 49–71.

McGrane, J. A., Humphry, S. M., & Heldsinger, S. A. (2018). Applying a Thurstonian, two-stage method in the standardized assessment of writing. *Applied Measurement in Education*, 31(4), 297–311.

McMahon, S. & Jones, I. (2015) A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice,* 22(3), 368-389.

Murphy S., & Yancey K.B., (2009). Construct and Consequence: Validity in Writing Assessment, in *Handbook of Research on Writing: History, Society, School, Individual, Text*, Routledge, p. 450-464.

Newhouse, C. P. (2014). Using digital representations of practical production work for summative assessment, Assessment in Education: Principles, Policy & Practice, 21:2, 205-220.

Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19(3), 281-300.

Pollitt, A. & Crisp, V. (2004, September). *Could comparative judgements of script quality replace traditional marking and improve the validity of exam questions?* Paper presented at the British Educational Research Association Annual Conference, UMIST, Manchester.

Rangel-Smith, C. & Lynch, D. (2018, 18th – 21st June). *Addressing the issue of bias in the measurement of reliability in the method of adaptive comparative judgment*. Conference presentation at 36th International Pupils' Attitudes Towards Technology Conference Athlone Institute of Technology, Co. Westmeath, Ireland.

Rhead, S., Black, B., Pinot de Moira, A., (2018). *Marking Consistency Metrics*. Ofqual report, 48 pages.

Seery, N., Canty, D., & Phelan, P. (2012) The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *International Journal of Technology and Design Education*, (22), 205–226.

Steedle, J.T. & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring, *Applied Measurement in Education*, 29(3), 211-223.

Tarricone, P. & Newhouse, C. (2016). Using comparative judgement and online technologies in the assessment and measurement of creative performance and capability, *International Journal of Educational Technology in Higher Education*,13-16.

Thurstone, L.L. (1927). A Law of Comparative Judgment. *Psychology Review*, 34, 273-286 https://brocku.ca/MeadProject/Thurstone/Thurstone_1927f.html

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.

Wheadon, C., Barmby, P., Christodoulou, D. & Henderson, B. (2020 a). A comparative judgement approach to the large-scale assessment of primary writing in England, *Assessment in Education: Principles, Policy & Practice*, 27(1), 46-64.

Wheadon, C., de Moira, A. P., & Christodoulou, D. (2020 b). The classification accuracy and consistency of comparative judgement of writing compared to rubric-based teacher assessment. SocArXiv https://doi.org/10.31235/osf.io/vzus4

Whitehouse, C. & Pollitt, A. (2012). *Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment*, AQA Centre for Education Research and Policy, Manchester.

Whitehouse, C. (2013). *Testing the validity of judgements about geography essays using the adaptive comparative judgement method*, Centre for Education Research and Policy AQA Education.

ISBN 978-1-912608-19-5