Moderation of non-exam assessments: is Comparative Judgement a practical alternative?

Carmen Vidal Rodeiro and Lucy Chambers (Research Division)

Introduction

Many high-stakes qualifications include non-exam assessments (NEAs)[1] that are marked within the centres, by teachers who act as internal assessors. Awarding bodies then apply a moderation process to bring the marking of these assessments to an agreed standard (Joint Council for Qualifications, 2019). During this process, moderators check samples of student work (henceforth portfolios) to ensure that centres have applied the marking criteria correctly. Moderators are usually teachers who have received training in moderation procedures by the awarding bodies. The two main tasks of moderation are to determine whether the rank order of the candidates' portfolio marks within the sample is correct, and to ascertain whether the marks awarded are acceptable or whether adjustments are necessary. Once these tasks are completed, moderators submit their marks for the moderation sample. If the centre marks differ from the moderator's marks beyond a predetermined amount, known as the tolerance level, then adjustments are made to all the centre's marks to align them to the standard (Gill, 2015).

At present, moderation is conducted at centre level; this enables moderators to build up a holistic view of a centre's approach to the course and how they have applied the assessment criteria. However, as work from each centre is usually only viewed by a single moderator, the process is reliant on the moderators applying the same standard across the centres they moderate. This raises challenges for standard maintaining across the whole cohort (currently, standard maintaining across the whole cohort is achieved by the standardisation of moderators and monitoring activities by senior moderators). Given that some NEAs are now moderated remotely, meaning that a central pool of electronic submissions of candidates' portfolios is available, there is the potential to moderate across all centres simultaneously. This means that candidates' portfolios could be allocated across multiple moderators without being bound by the centre. This could help address the maintenance of standards challenge, and thus ensure that the

1 The term NEA, standing for non-exam assessment, is used in this article to cover school-based assessment, internal assessment, or coursework.

marking standard is consistently applied across all centres. The current study sought to explore the use of Comparative Judgement (CJ) as one possible method for achieving this.

CJ is a process where multiple judges compare two (or more) pieces of work, for example pairs of student scripts, and decide which script in each pair is the "better" one (Bramley, 2007; Pollitt, 2012a; 2012b). CJ requires judges to make relative judgements, which are considered to be easier to make than absolute judgements of an individual script against a mark scheme (Pollitt & Crisp, 2004). Analysis of the resulting data places each script on a scale of relative quality and produces an overall rank order of the scripts.

As CJ is designed to create a rank order of scripts (in this case, portfolios), the first moderation task (whether the rank order of the portfolios within the sample is correct) would easily be accomplished via this method. The second task, determining the acceptability of marks, is a little more complex and would entail assigning moderator marks as a result of the CJ analysis (and not directly by a human judge). The CJ produces a measure of quality for each portfolio, the CJ estimate. In order to then apply the usual process of moderation, these CJ estimates need to be converted into moderator marks (i.e., marks that correspond to the particular portfolio after the moderation task). These moderator marks can be compared to the marks given by the teachers within the centres and an adjustment procedure can be carried out if necessary.

This study formed the second part in a strand of research exploring the potential use of CJ for the moderation of non-exam assessments. The first part, a simulation study, explored the theoretical feasibility of using pairwise CJ for moderation (Chambers et al., 2019). The research proved promising, identifying a potential approach of assigning moderator marks to candidates' work using the data from the CJ exercise, and the minimum numbers of judgements and moderation sample size for the CJ to have good reliability.

The current research explored the method further, via an experimental moderation task using portfolios of work. In particular, it investigated its *practical* feasibility. This included aspects such as time taken to moderate, whether CJ can feasibly be used on larger bodies of work (e.g., portfolios) and whether moderators can be confident making CJ judgements on large pieces of candidates' work.

The overarching research question in this study was:

> Is CJ a practically feasible method for moderating non-exam assessments?

The following sub-research questions were also investigated:

- Can moderators view and navigate the portfolios sufficiently to enable them to make the comparative judgements?
- On what basis do moderators make their judgements?
- Are moderators confident making comparative judgements on portfolios?
- How long does it take to make comparative judgements on portfolios?

Method

Portfolios

In this research, the focus was on unit R053 (Sports Leadership) from the Cambridge National in Sport Studies (J813). Students who take this unit build a portfolio of evidence to meet the learning objectives (LOs). This portfolio is centre-assessed, and then moderated by OCR. Centres can choose between three moderation modes: postal, visiting or the OCR Repository (electronic submissions).

For this study, 30 portfolios from the June 2019 session were selected from across the whole grade range and from a variety of centres. The portfolios were drawn from the samples that were submitted to the OCR Repository.

A cover sheet (the unit recording sheet) is attached to each portfolio with a summary of the marks awarded for the task by the teacher and some teacher comments. For the purpose of this research, the cover sheet was not included and all comments were removed, as it was felt that it could exert undue influence on the judging process and could potentially undermine the task. Any identifying information was also removed.

Typically, a candidate's portfolio for this unit consists of multiple documents. In a few cases, these were compiled into a single document for each candidate by the centre; however, for most centres a number of separate documents were submitted for each candidate. As part of this unit, candidate performances of physical activities were assessed. Examples of the evidence required for unit R053 include witness statements and/or filmed/documentary evidence of the physical activities undertaken. For the purposes of the research, portfolios containing videos of performances were excluded and only those portfolios using witness statements were considered, as these formed the vast majority of samples in the OCR Repository. For each portfolio, all documents were stitched together into a single PDF file, which enabled the research to be conducted using the Cambridge Assessment CJ Scaling tool (see below).

Judges

Six moderators (team leaders for the unit) and the principal moderator were recruited from the pool that moderated the June 2019 series. Although there were seven participants in total, the principal moderator was only included in certain aspects of the research due to availability (see below).

The participants had between 3 and 20 years moderating experience and all but one had marking experience as well. Only one participant, the most experienced in terms of years marking and moderating, had taken part in a CJ exercise before.

Information about the study and full instructions and guidance on how to perform the CJ moderation task were provided at the onset. In order to re-familiarise themselves with the assessment task for unit R053, participants were also given a copy of the assessment task and associated mark scheme.

Research task

The participants were asked to make comparative judgements on pairs of portfolios from unit R053. They were presented with two portfolios at a time (a pack) and they had to decide which was better based on a holistic judgement of the overall quality of the work. In this particular case, the question they were answering was:

> Which portfolio better demonstrates the knowledge, understanding and skills required to be an effective sports leader?

The portfolios were loaded into Cambridge Assessment's comparative judgement online tool (https://cjscaling.cambridgeassessment.org.uk), referred to as the 'CJ Scaling tool' in this article. The portfolio allocation to each pack was random, thus any pair could potentially contain portfolios that were similar in terms of the marks received or portfolios with very different marks. The six team leaders comprised the panel of judges carrying out the CJ task. In total, each of these six judges made 30 paired-comparison judgements and, therefore, each portfolio was judged 12 times. This resulted in some judges seeing the same portfolio more than once.

The principal moderator was not included in the panel of judges due to availability. However, they were able to make 30 additional judgements (with the same allocation of the scripts as one of the other six participants) in a separate judging session. This allowed the principal moderator to carry out the CJ task and experience the CJ tool and, therefore, made possible joining in for the subsequent aspects of the research.

Although the judges were provided with the assessment task and the mark scheme, they were instructed not to re-mark the portfolios. Instead, judges were asked to make a holistic judgement about each portfolio's quality and its overall merit, relative to the other portfolio in the pair.

After the task, judges were invited to complete a short online questionnaire. This gave them the opportunity to provide feedback and enabled the researchers to gather additional information on their judging behaviour. The judges were also asked either to agree to be observed by the researchers (while doing some of the judging using the CJ Scaling tool) or to be interviewed. Five of the judges (four moderators and the principal moderator) were interviewed, while the remaining two were observed while doing the CJ task.

For the observations, one of the researchers observed each judge for approximately 1 hour, while they were making their judgements. The observation was conducted on Microsoft Teams, which allowed the judges to share their screen so that the researcher could see what they were doing at any given point. This was supplemented by a think aloud procedure in which the judges verbalised their thoughts while making their judgements.

The interviews (which took approximately 30 minutes) were also conducted on Microsoft Teams, after the judges completed their judging and had submitted their survey responses.

The observations and the interviews were recorded and automated transcripts generated.

Findings

The analysis comprised the evaluation of four types of data: CJ data, observation data, survey responses and interview data.

CJ data

In total, there were 180 judgements for the 30 portfolios considered in the research (made by the six moderators in the judging panel). This meant that, as two portfolios were seen in each judgement, each portfolio was seen 12 times. This number is slightly lower than typically recommended in CJ studies and by the Chambers et al. (2019) feasibility study but suitable for the purpose of this experimental work.

The data on pairwise judgements was downloaded from the CJ Scaling tool and fitted to the Bradley-Terry model (Bradley & Terry, 1952).

The Scale Separation Reliability or SSR (i.e., the reliability of the CJ) was 0.76, which is slightly lower than the reliability in other CJ studies carried out recently at Cambridge Assessment and elsewhere. For example, Chambers and Cunningham (2022) reported an SSR around 0.80 when they asked judges to rank scripts from a GCSE in Physical Education in an awarding context, and Holmes et al. (2020) found that the reliability of several CJ exercises looking at AS History scripts was between 0.85 and 0.88. A higher number of judgements per portfolio in the current study could potentially have increased the reliability of the CJ.

Judge statistics

Measures of judge fit, such as infit and outfit, were calculated and used to check the quality and consistency of the judging (see, for example, Linacre (2002) for details on these measures). Typically, these measures are examined with a view to assessing whether any judges were misfitting the model to such an extent they might be affecting the estimates of script quality. In some contexts, this might be a reason to exclude their judgements. In this research, however, the focus was on the judges' behaviours and perceptions of the method, so the analysis of the CJ data was not to evaluate the method itself, just to give an indication of how it was performing (in a "live" study there would be more portfolios and more judges). Therefore, no judges were removed on the basis of their fit statistics.

As stated above, in this study, judge fit was determined with regard to how well the judgements agreed with what would be expected given the CJ measures of each portfolio as derived from the Bradley-Terry model (Benton, Cunningham et al., 2020). The judge infit values (see Table 1) were within an acceptable range (between 0.5 and 1.5, as stated by Linacre (2002)), suggesting that the judges were reasonably consistent in their judgements. The one judge outside this range (Judge 4) had very low values of infit and outfit suggesting a surprisingly high level of agreement between their judgements and the rank order of the CJ measures. This is not normally a concern in terms of the quality of measurement.

Furthermore, it is worth noting that these fit statistics are based on relatively small numbers of pairs per judge. The majority of the judges had low outfit (under 0.5), which means that they exhibited more predictable judgement patterns than was expected by the Bradley-Terry model.

Table 1: Judge fit statistics.

Judge	Number of pairs judged	Infit	Outfit
1	30	0.57	0.32
2	30	0.60	0.34
3	30	1.04	0.57
4	30	0.27	0.14
5	30	0.68	0.38
6	30	0.53	0.27

CJ measures

The rank order of the portfolios based on the CJ analysis was compared with the rank order based on the final marks awarded during the "live" assessment (i.e., after moderation) in the June 2019 session. This analysis was carried out in order to make sure that the judges' decision-making was similar to that of a centre following the mark scheme accurately and appropriately applying the national standard. A poor correlation would indicate that the decisions were either being made on a different basis or that the method itself was introducing differences.

Figure 1 shows comparisons of candidate marks in the portfolios, as awarded in June 2019, with the CJ measures.

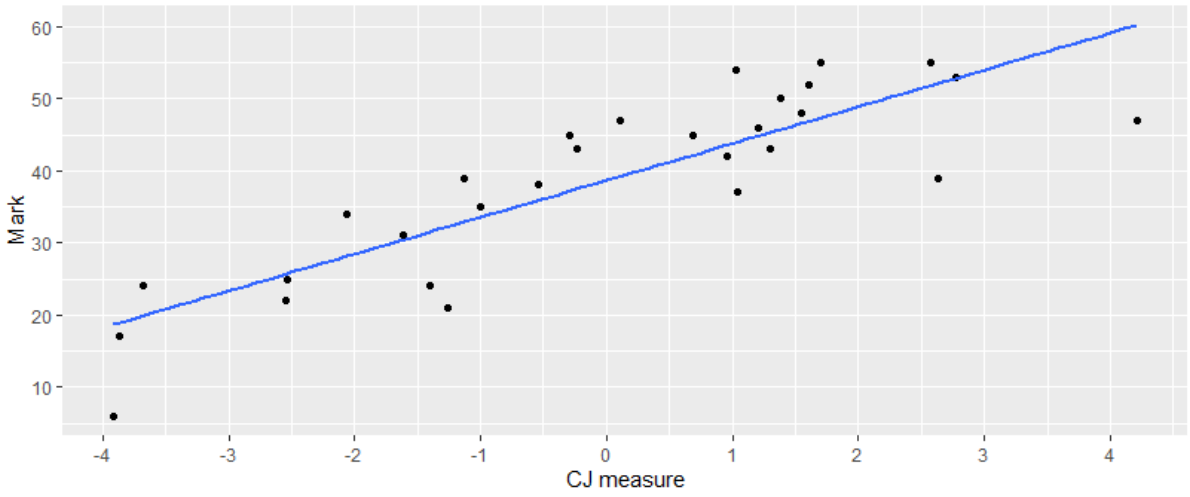

Figure 1: Portfolio marks vs. CJ measures.

The correlation of 0.85 between marks and portfolio CJ measures indicates that the candidate rank orders were similar for marking and CJ judgements.

Time required to complete the task

The estimated total time required to complete the task varied from 5 hours to just under 8 minutes. The estimates are based on the time taken from the start of a judging session to the moment a decision is submitted – we cannot be certain whether active judging was happening throughout all that time. The average time per pack (in minutes and seconds) was 4m 46s and the median time per pack was 2m 23s. Table 2 below shows the time required to complete the task for each of the judges who took part in the study.

Table 2: Time required to complete the task.

Judge	Number of pairs judged	Total time	Median time per pack
1	30	5h 0m	2m 51s
2	30	2h 29m	3m 50s
3	30	2h 27m	3m 52s
4	30	1h 27m	1m 55s
5	30	46m 28s	53s
6	30	7m 56s	8s

Compared to the CJ judgements of exam scripts, the judgements of portfolios were found to take a similar amount of time. For example, Benton et al. (2022, this issue), who summarised the results of 20 CJ studies using exam scripts in the context of awarding, reported that the average time per pair of scripts was around 5 minutes, which is not very different from the average time per pack of two portfolios observed in this study (4m 46s). This can be an indication that CJ is practically feasible for comparing portfolios in terms of time taken.

As shown in Table 2 above, judges varied in the time taken to make judgements. Figure 2 below shows a box plot of time taken in minutes for each judge. Judges 5 and 6 were the quickest (in fact, they were much quicker than the other judges) and Judge 1 was the slowest.

Note that some of the times recorded may be long because of the online observations (for example, Judges 1 and 3 were observed by the researchers while judging two packs). Talking out loud while conducting a task, the presence of an observer and judges having the CJ tool open prior to the start of the observation could all contribute to longer judging times which may account for the outliers.

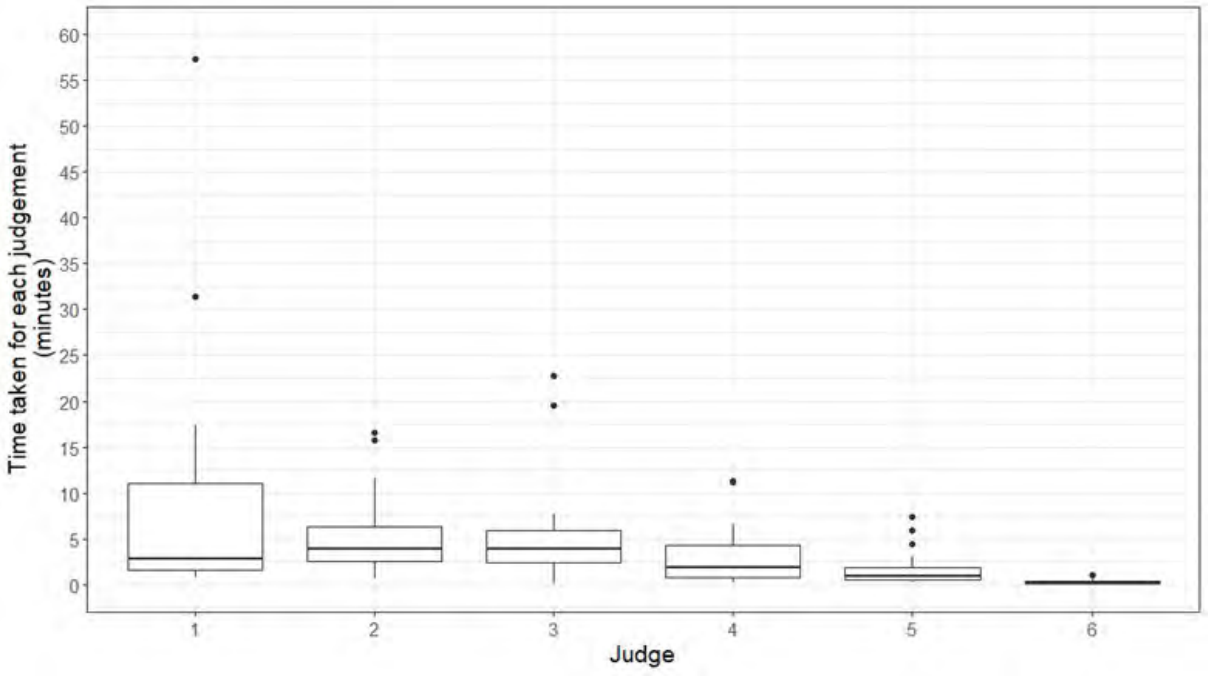

Figure 2: Time per judgement (minutes), by judge.

Observations

Two judges (Judge 1 and Judge 3) were observed, for between 30 and 45 minutes each, while they were making their judgements. Thematic analysis of the recordings of the observations provided evidence of the way the judges carried out the CJ task (e.g., their approach to the task, what they paid attention to, their use of the CJ tool, the navigation through the portfolios, etc.).

This section of the article starts by presenting behaviours drawn from the observations concerning the way the judges approached the CJ task (i.e., the CJ method in general). It then describes some of the difficulties the judges encountered while doing their judging (with either the task or the CJ Scaling tool). All quotes from the observations are written verbatim.

Note that it is possible that the behaviour exhibited during the observation did not reflect the rest of the judging. However, although judging while observed might have taken a bit longer than the rest of the judging, the general method employed to make the decisions about which portfolio would "win" the comparison is unlikely to have been fundamentally different.

The observations showed that the judges differed in how they approached the task and that they used different methods when viewing the portfolios within each pack.

For each learning objective, Judge 3 looked at each portfolio in turn, stating what they were looking for in the work to assign a specific mark band and mentioning what they were finding or what it was missing. For example:

> It's not strong as I would be looking for mark band three, so I would say that that final one was mark band two.

> But we need to see links, the information, the descriptions, is good for mark band one, but then when we get to mark bands two and three it's links […] and definitely now it is in mark band three.

Although the mark scheme was not mentioned directly by Judge 3, it was clear that they were very familiar with it and made frequent use of it. Comments made during the observation were:

> So, for me, for the first learning objective, those are both in mark band three.

> The one on the left has mark band three work for every learning objective. The one on the right has not.

As shown above, Judge 3 seemed to have been following their normal way of working when carrying out moderation under the traditional procedures, rather than following the instructions of the research study and making a holistic judgement about the quality of the portfolios. Similar behaviour was evident in the observations of a recent CJ study set in an awarding context (Leech & Chambers, 2022, this issue).

Judge 1, however, worked through both portfolios at the same time, dipping into certain learning objectives to evaluate them more fully. They did not necessarily go through a whole learning objective for one portfolio before moving on to the next. In fact, the judge was scrolling down both portfolios simultaneously while looking at the different learning objectives. Furthermore, they did not refer to the mark scheme or appear to use it when doing the judging (they were also not looking for specific key words). Their approach seemed to be more holistic, and in line with the instructions given to carry out the CJ task. For example:

> Immediately I'm starting to like the one on the left-hand side because there is more detail in it.

> At the moment the left-hand side one is winning in my mind.

Judge 1 was actively comparing extracts of the portfolios against each other, which is within the purpose of CJ, while Judge 3 seemed to compare each of the portfolios with what they were expecting to see. Some comments reflecting these behaviours are given below:

> Judge 1:

> But there's a lot more detail on the left-hand side.

> You've got knowledge of activities on the left-hand side and they straight away give you an example […] which is what you don't really get on the right-hand side.

> Judge 3:

> I can see for this first sample of work there's a thorough risk assessment. They've identified lots and lots of different hazards or risks.

> I can't see examples. They've not come out and said an example of a manager is. But I can see again that they've talked about [...].

Overall, the observations showed that one of the judges was using their knowledge of the mark scheme and their moderation techniques to carry out the CJ task, while the other judge used a more holistic approach.

During the observations, there were some concerns raised by the judges. The concerns related to potential malpractice, presentation of the work, and IT issues with the CJ Scaling tool. Some of these issues, however, could also be encountered during traditional moderation and they were not inherent to the CJ Scaling tool or the CJ task.

In terms of presentation of the portfolios, both judges made comments about the amount of text (they much preferred pieces of work with diagrams or bullet points). In addition, Judge 1, the holistic judge, made comments about quality of scanning. Both features could be concerning if these construct-irrelevant features influence the judges' decision-making.

The third learning objective for the Cambridge National unit considered in this research is usually assessed via a witness statement and, when moderating, the judge needs to rely on the information the teachers are providing after witnessing a practice session. One of the judges mentioned that, in the traditional moderation process, they would have looked at several witness statements either in the repository or in the physical work, to make sure they were different from each other. However, in the CJ task, they would have to assume that the witness statement has been written specifically for that particular learner. This was slightly concerning for the judge, and they suggested there could be malpractice going unnoticed if witness statements were not individualised.

There were some IT difficulties during the observations. In particular, one of the judges found it quite difficult to have just one script on the screen and to adjust the size of the text (e.g., enlarging it to make it easier to read). The system's response was quite slow and took the judge over 5 minutes to set up the screen and font size the way they wanted.

Other IT difficulties were related to the amount of time it took to load the portfolios, and to moving (scrolling) through the students' work. Examples of these issues, encountered by one of the observed judges, are given below:

> These are quite big documents. They've often got colour photographs, so I know they take a while to load up.

> It is difficult to navigate because it flicks very easily between the sections, I can't quite get to the bottom of the page. It won't let me move it up or down. And the little scroll is bringing everything connected. I can't move it so I can't actually very easily see the bottom of the pages [...] little scroll is too sensitive.

Survey and interviews

On completion of the CJ task, judges filled in a short survey, which contained a mix of Likert scale and open response questions. Analysis of the survey data provided insights about how the judges approached the CJ task, the usefulness of the CJ Scaling tool for the task, and what features of the portfolios they attended to. The interviews were designed to further explore the findings of the survey; thus, the video recordings were analysed along similar themes. Interview findings have been interwoven into the survey findings.

Use of the CJ tool and navigation

The judges used a variety of devices to carry out the CJ task: three used a laptop, two a desktop and two a MacBook. Generally, the judges found the screen size suitable for the task, although some noted that they had to zoom in on certain PDFs when the candidates' handwriting/font size was small. This zooming made the task less efficient as it took longer and involved additional mouse clicks. Judges reported that some portfolios took longer to load than others, particularly those with images, and that sometimes there was a time lag when scrolling down through the portfolios. These aspects were reported to interrupt the flow and caused some frustration.

Figure 3 below shows the judges' responses to further questions about their experience with the CJ Scaling tool. None of the judges strongly disagreed with any of the statements.

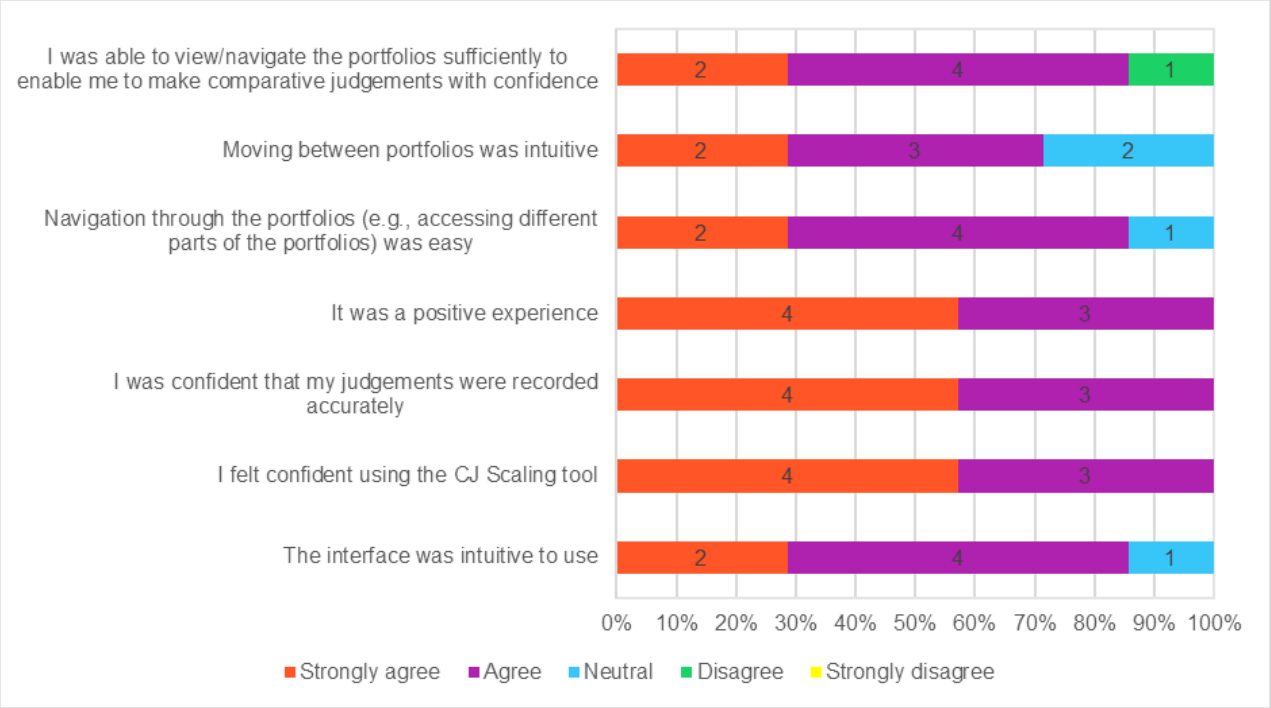

Figure 3: Judges' experiences with the CJ Scaling tool.

Responses were, in general, quite positive. All judges felt confident using the tool, were confident that their judgements were recorded accurately and found using

the tool to be a positive experience. When asked about the interface in the CJ Scaling tool, six of the judges agreed that it was intuitive to use.

In terms of viewing or navigating the portfolios, judges were mostly positive. Six judges found navigation through the portfolios to be easy and five found moving between portfolios to be intuitive. Six judges agreed that they were able to view and navigate the portfolios sufficiently well to enable them to make comparative judgements with confidence; only one judge disagreed with this last statement and explained that they could not download one portfolio properly and therefore could not view the entire document.

The CJ tool and judgements

When asked whether or not the use of the CJ tool might have impacted the quality of the judgements, six judges reported that it had not. Below are some of their comments:

> The two pieces of work I was comparing each time, were usually easy to see which one was better. Some were more similar and so required much more scrutiny.

> It was easy (and very quick) to make a decision/judgement where the two pieces of work were very different. When similar, I had to spend much more time looking for key identifiers in each LO [learning objective] and MB [mark band] to be able to find differences. I was still able to moderate to the standard, but time was spent unequally on different pieces of work / pairs.

However, one judge reported that the use of the tool had compromised the quality of their judging stating that "it was difficult at times to give an overall comparison rather than as we do usually and give marks for each learning outcome". This judge elaborated on this during the interview, saying that the centres gave marks by learning objectives and so felt that they should be moderated this way too. This judge also expressed that they were not at all comfortable providing a whole portfolio holistic judgement.

The judges' experiences mirror other CJ findings (Leech & Chambers, 2022, this issue) in that judgements were harder when the work was similar in standard and that some assessors find the move to making holistic judgements challenging.

Making holistic judgements

Despite some concerns having been raised, all of the judges reported the process of making holistic judgements of the portfolios to be somewhat or very straightforward. Comments included:

> Once I became accustomed to the process it became easy.

> In most cases, a clear comparison was noticeable. Seeing pieces of work multiple times helped to get to know the work too. Some were closer in quality and needed more thought and scrutiny.

> It was straightforward but just different from the process I am used to.

When the judges were asked how confident they felt making a holistic judgement of each portfolio, four judges were very confident and two were somewhat confident. These judges attributed their confidence to previous experience of marking that unit, clarity about the task and the fact that the portfolios were viewed side by side. One judge was not sure about their confidence level, reporting that they wanted to judge by learning outcome.

Features on which judgements were based

Judges were asked to detail the main portfolio features on which they made judgements. As one would expect, answer detail, use of examples, correct terminology and relationship to the mark scheme were key features. Some of the judges' responses are shown below:

> Information contained within the answers to achieve marks in some mark bands.

> Detailed descriptions [...] supported with relevant examples.

> I looked for inclusion of detail with examples, and appropriate terminology being included. Inclusion of progressions in lesson plans. I looked for key information in witness statements.

> I made a table which contained key info from each LO [learning objective] and each MB [mark band]. I looked for key areas to be covered for the bottom and top MBs. [...]. I identified areas such as detailed or basic, some or good range. Looked for links, evaluations and key improvements.

Time taken

Regarding the time taken to judge each pair of portfolios, at the onset of the project the researchers estimated that each judgement could take around 10 minutes. This estimate was based on previous research on comparative judgement of exam scripts (e.g., Benton, Leech, et al., 2020), as there was no research available looking at the use of CJ with portfolios.

As part of the survey, judges were asked if 10 minutes was an appropriate estimate of the time taken to make a CJ judgement. Two judges thought that 10 minutes was not enough, four judges agreed that 10 minutes was about right, and one judge thought 10 minutes was too long. This is an interesting finding when we compare it to the actual time taken; it appears the judges generally felt that the judging took far longer than it did. In the interview, some judges elaborated on time taken and made comments around the following themes:

> This method was quicker than traditional moderation.

> The time taken to carry out the CJ task was about right because the judges were experienced moderators. It was suggested that the task would have taken longer if the judges were new or inexperienced moderators.

10 minutes was about right at the start, but felt they became quicker as they did more judging[2].

Comparison of moderation methods

Judges were asked to compare traditional and CJ moderation methods in terms of whether they were easier/harder to do, whether they were more or less cognitively demanding and more or less enjoyable. Figure 4 shows that, in terms of sentiment, judges were split. However, they tended to be consistent across all three questions.

Additional explanations offered by the judges overlapped across the three questions so are summarised by sentiment below.

Positive sentiment

- Making comparisons on the tool was easier than on paper.
- Comparisons were easier particularly where the work was very different in quality.
- Ease of scrolling down the page.
- Not having to justify the decision.
- Not having to scrutinise how to mark each learning objective.
- It was a good way to get a feel for the work.

Neutral sentiment

- Both methods were comparable – when moderating a centre's work there are often a variety of portfolios.
- Only checking the rank order.

Negative sentiment

- Easier to work with paper copies.
- Preference for looking at work in relation to centre marks and per learning objective.
- Difficult to judge work which was similar.

2 This perception may be based on the judges' increased familiarity with the CJ Scaling tool or task and/or on the fact that some portfolios were seen more than once. There did not appear to be any noticeable patterns in the timing data to support this perception.

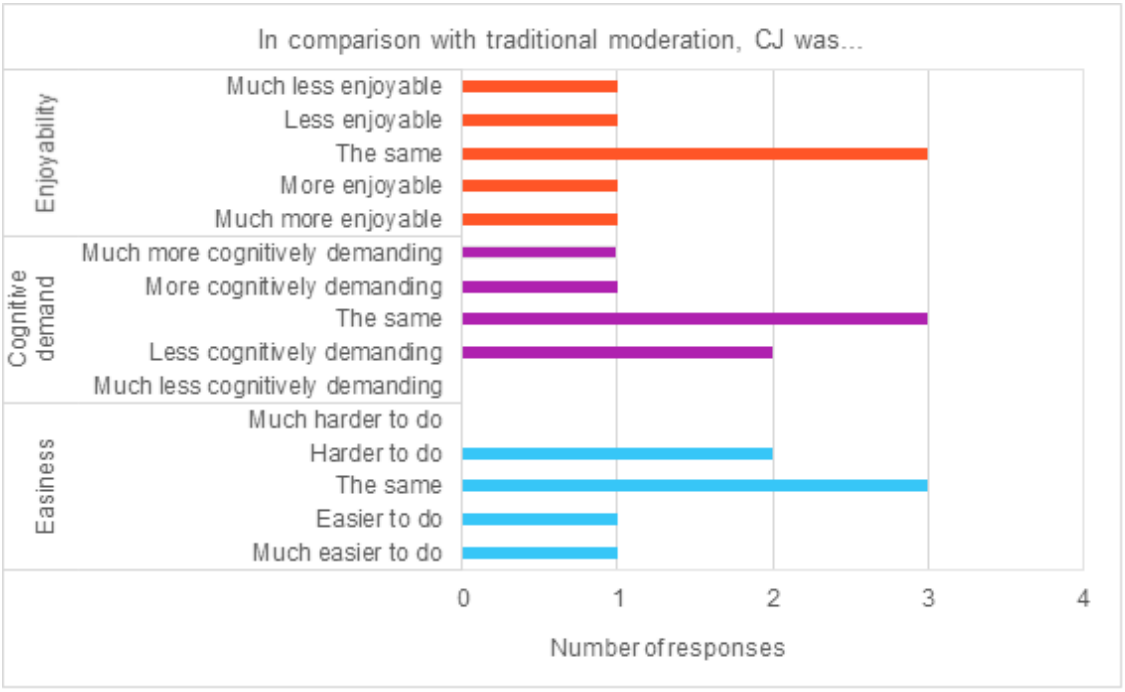

Figure 4: Responses to the prompt "In comparison with traditional moderation, Comparative Judgement was...".

Some judges struggled with the difference between using CJ for moderation and the traditional moderation task. In particular, the lack of centre marks and the fact that they were not asked to verify the marks seemed to be an issue for some judges, as shown in the quote below.

> The unit relies on a teacher completed witness statement for one LO [learning objective]. When moderating we check that these are different and unique for each learner in the cohort – this cannot be checked when completing CJ. So to carry out moderation for this unit to the standard we currently work to, the LO would need to be changed or the type of evidence submitted.

There appeared to be a difference in opinion as to whether using CJ for moderation was more or less like marking. One judge noted that "It was a good way to get a feel for the work and moderate it. Traditionally, the temptation is to re-mark the work – particularly if the centre mark is very different to the moderator's mark" while another noted "we are moderating their marks so need the centre marks, this felt like I was marking the work". This judge elaborated that, without the marks, they were not establishing whether they agreed or disagreed with someone.

In the interview we were keen to hear the judges' views about viewing the portfolios digitally; this was in an attempt to establish whether any concerns they shared about the task were due to not having the materials in physical form or to the on-screen CJ method. We found that judges' opinions were mixed, and in fact one judge reported that they liked to have a mix of mediums in their allocation. Some judges preferred to lay the scripts out to view them, with one explaining that it meant they could revisit them and another stating that they could then

compare them to the standardisation scripts. Others felt fine about viewing portfolios digitally with two judges admitting that they were getting more used to it.

The judges liked the single PDF file provided for the CJ task, noting that it was far better than the repository where they often had to open multiple files. They also appreciated that pages were in the correct order, and that they were all the right way up and they did not have to rotate them.

Conclusions

The overarching aim of the study was to establish whether CJ could be used as a feasible alternative for moderating NEAs. The conclusions are presented with reference to the initial research questions.

Is CJ a practically feasible method for moderating NEAs?

This study, in conjunction with the previous simulation study (Chambers et al., 2019), provided evidence that CJ is a feasible method for moderation and one that should be explored further. The judges were able to perform the task, make decisions with confidence, and the indicative statistical analysis looked promising.

There are a couple of practical considerations that should be borne in mind if the method is taken forward. Firstly, candidates' work would need to be submitted to an online repository to be able to moderate all centres in the same way (e.g., visiting and postal moderation would not be available). Secondly, NEA moderation samples (i.e., portfolios) vary substantially in both their inherent formats and structure. For example, formats can include standard document types (Word, Excel, PowerPoint), artwork (pictures and sculptures), videoed performance and computer code. Centres also vary in how they submit the work, ranging from a clearly labelled and organised submission to a single structureless folder containing everything a candidate has produced; this tends to be influenced to some extent by the qualification/unit/task. The CJ Scaling tool used in the study requires a single PDF file for each artefact. Thus, for the current study we used a unit (R053) where the portfolio could be readily presented in this form. This had a simple structure, easy formats to work with and centre submissions were relatively well organised. If the method was to be utilised, then consideration would need to be given to the software used.

Can moderators view and navigate the portfolios sufficiently to enable them to make the Comparative Judgements?

Overall, the judges were able to view and navigate the portfolios easily and found using the CJ Scaling tool to be a positive experience. A few issues were reported concerning time taken for certain portfolios to load, time lags when scrolling or where a centre had organised the submission in a non-standard way making the evidence harder to find. While these issues are independent of the CJ method and are largely a result of local internet connection and centre submissions, they are features that should be borne in mind if the method is taken forward.

On what basis do moderators make their judgements?

Judges reported that they made their decisions based on features such as answer detail, use of examples, correct terminology and relationship to the mark scheme. These are all appropriate.

However, during the observations, the judges made comments about context-irrelevant features (e.g., amount of text, tabulation, quality of scanning). It could be concerning if these features were to influence the judgement process. It was also clear that some judges were essentially trying to re-mark the portfolios.

These findings have implications for the validity of the method and would need to be addressed, for example, via discussion about holistic decision-making and training.

Are moderators confident making Comparative Judgements on portfolios?

This was a key question and the research showed that judges were confident about their decisions. However, some judges did struggle with the holistic nature of the task, finding it difficult to "let go" of their current moderation practices and switch to holistic judgements. It is recommended that before any study or judging the judges meet with a trainer or facilitator so that a full explanation of the method is provided and there is an opportunity to ask questions. This should be followed by training and practice.

How long does it take to make Comparative Judgements on portfolios?

In terms of the time taken to make CJ judgements on portfolios, the outcomes of this study show that CJ is practically feasible.

When compared to traditional moderation, the judges felt that the CJ method was quicker, which may be explained by the CJ method only focusing on one aspect of moderation, the rank order. The second aspect, moderator marks, would be calculated by statistical analysis using data from the CJ exercise and not awarded by the moderator. Furthermore, the judges said that 10 minutes was an appropriate estimate of the time taken to make a CJ judgement, particularly at the start, but judging could be quicker with increased familiarity with the CJ Scaling tool or task and/or due to the fact that some portfolios were seen more than once.

Compared to the CJ judgements of exam scripts, the judgements of portfolios were found to take a similar amount of time (Benton et al., 2022, this issue). This may be explained by the judges being used to scanning and dipping into portfolios when moderating – this behaviour is congruent with making holistic judgements. Examiners (i.e., exam markers), however, are used to performing a detailed evaluation of each question and may continue to do this even when asked to make holistic decisions (Leech & Chambers, 2022, this issue).

Recommendations and further research

There are a number of specific recommendations that can be drawn from this study:

- As portfolios vary substantially in both their inherent formats and structure, if CJ were going to be used for moderation, consideration would need to be given to: the way portfolios are organised and submitted by the centre; the type of artefacts submitted (e.g., pictures, videos, documents) and how the software would present them.
- In order for context-irrelevant features (e.g., amount of text, tabulation, quality of scanning) not to influence the judgements, discussion of such issues should be covered in moderator training and candidates/centres should consider the format and the presentation of the materials.
- Before any study that would use CJ for moderation, the judges should meet with trainers or facilitators so that a full explanation of the method is provided (e.g., discussions about holistic decision-making) and that judges have an opportunity to ask questions. This should be followed by training, which should incorporate practice and feedback on the task.
- Thought needs to be given to a number of procedural elements: how plagiarism can be identified (including identifying "blanket" witness statements), how to check centre internal standardisation (e.g., consistency of witness statements) and how to provide support to centres (e.g., reporting; feedback).
- Concern about new centres (judges mentioned that there is a difference between experienced and new centres, with new centres needing more support) is a valid issue and enhanced training and support (not only support to carry out a CJ task, but also general support on moderation in general) should be given to new centres. For example, centres could be assigned a moderator who could provide a guidance role at key points throughout the year.

Further research should investigate the feasibility of carrying out a full end-to-end moderation task. In particular, further studies should investigate: 1) the best approach to assign moderator marks to the portfolios based on the results of the CJ analysis (e.g., following one of the methods outlined in Chambers et al. (2019) or exploring alternative methods such as linear equating); and 2) how to adjust centre marks if necessary.

References

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

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

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

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

Bramley, T. (2007). Paired comparison methods. In P. Newton, J. A. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246–300). Qualifications and Curriculum Authority. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/487059/2007-comparability-exam-standards-i-chapter7.pdf

Chambers, L., & Cunningham, E. (2022). Exploring the validity of comparative judgement – do judges attend to construct-irrelevant features? *Frontiers in Education, 6*. https://doi.org/10.3389/feduc.2022.802392

Chambers, L., Vitello, S., & Vidal Rodeiro, C. (2019, 13-16 November). *Moderation of non-exam assessments: a novel approach using comparative judgement* [Paper presentation]. 20th annual AEA-Europe conference, Lisbon, Portugal. https://www.cambridgeassessment.org.uk/Images/563137-moderation-of-non-exam-assessments-a-novel-approach-using-comparative-judgement.pdf

Gill, T. (2015). The moderation of coursework and controlled assessment: A summary. *Research Matters: A Cambridge Assessment publication*, *19*, 26–31.

Holmes, S., Black, B., & Morin, C. (2020). *Marking reliability studies 2017. Rank ordering versus marking – which is more reliable?* Office of Qualifications and Examinations Regulation. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/859250/Marking_reliability_-_FINAL64494.pdf

Joint Council for Qualifications. (2019). *Instructions for conducting coursework 2019-2020*. Joint Council for Qualifications.

Leech, T., & Chambers, L. (2022). How do judges in Comparative Judgement exercises make their judgements? *Research Matters: A Cambridge University Press and Assessment publication, 33, 31–47.*

Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions*, *16*(2), 878. https://www.rasch.org/rmt/rmt162f.htm

Pollitt, A. (2012a). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *19*(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354

Pollitt, A. (2012b). Comparative judgement for assessment. *International Journal of Technology and Design Education*, *22*(2), 157–170. https://doi.org/10.1007/s10798-011-9189-x

Pollitt, A., & Crisp, V. (2004, September). *Could Comparative Judgements of Script Quality Replace Traditional Marking and Improve the Validity of Exam Questions?* [Paper presentation]. British Educational Research Association Annual Conference, Manchester, UK. https://www.cambridgeassessment.org.uk/Images/109724-could-comparative-judgements-of-script-quality-replace-traditional-marking-and-improve-the-validity-of-exam-questions-.pdf