Comparing music recordings using Pairwise Comparative Judgement: Exploring the judge experience

Lucy Chambers, Emma Walland and Jo Ireland (Research Division)

Introduction

Comparative Judgement (CJ) involves judges comparing two or more artefacts (often exam responses or scripts) to decide which is better. Multiple judgements of each artefact are statistically modelled to assign each a relative measure of quality and consequently create a rank order of artefacts. CJ has been widely investigated in educational assessment as an alternative for marking (Pollitt, 2012; Steedle & Ferrara, 2016; Walland, 2022; Wheadon et al., 2020), for standard maintaining (Benton et al., 2022; Curcin et al., 2019), for monitoring comparability (Bramley, 2007; Jones et al., 2016) and, more recently, for moderation (Chambers et al., 2024; Vidal Rodeiro & Chambers, 2022).

In the context of alternatives to marking, Bramley (2022) noted in his editorial for issue 33 of *Research Matters* (which focused on CJ) that the key questions are the reliability and validity of the resulting scores, the feasibility and cost, and transparency from the candidate perspective. This article seeks to add support to the validity argument by addressing the judge perspective. It is important to verify that the judges are able to make appropriate CJ decisions just as "it is necessary to ensure that the judges themselves believe in the validity of what they are doing if stakeholders more widely are to be convinced" (Bramley, 2022, p. 7).

Decisions within a CJ context are considered to be holistic; the judges consider the evidence presented as a whole and make an evaluation. This is as opposed to the more traditional analytic method of marking using a detailed mark scheme. On the surface the CJ task appears simple, but it is actually the result of considering many pieces of interconnecting evidence. Leech and Vitello (2023) proposed three central concepts that "should define holistic judgement in an assessment context" (p. 4). Namely, the ultimate output is singular in nature, the process involves the combination of comprehensive construct-relevant evidence and that the process considers the interconnectedness of the evidence. By evaluating the judge experience, we can establish to what extent these concepts have been fulfilled.

To date, the vast majority of studies have involved written or text-based artefacts. There are a small number of studies using Art or Art and Design portfolios (Mason & Garelli, 2022; Newhouse, 2014; Tarricone & Newhouse, 2016) and one project using voice recordings (RM, 2022). To our knowledge there have been no studies involving wider-ranging artefacts, for example, recordings of music. This study sought to address this gap.

As part of a project exploring alternative ways of marking Non-Examined Assessments (NEA), we investigated using Pairwise Comparative Judgement (PCJ) to assess OCR's GCSE Music portfolios.[1] Previous work has shown that using CJ on larger bodies of NEA work (i.e., larger in size than an exam script) is practically feasible (Vidal Rodeiro & Chambers, 2022) and this study built on this by using portfolios that were primarily auditory in nature.

Previous work has also shown, however, that making comparative judgements can be challenging in certain circumstances. In a synthesis of participant questionnaires from multiple studies exploring CJ in a standard setting context, analysis has highlighted the challenges in making comparisons when the work is very different in nature (Leech & Chambers, 2022). With GCSE Music, certain differences are inherent as candidates will use different instruments, different mediums (e.g., live instrument versus sequencer) and different musical genres. In addition, pieces will be of different technical difficulty. Thus, we were keen to explore what, if any, level of challenge this might raise for the judges.

This article examines the judges' perceptions of using CJ in this context with reference to the *Dimensions of judge decision-making* model (Leech & Chambers, 2022) and makes comparisons with the findings from text-based studies.

Method

In England, OCR's GCSE Music (J536) involves one written paper (examined) and two performance-based components (Non-Examined Assessments). For the current study, we used one of the performance-based components: the integrated portfolio. This consists of a solo performance and a composition to a brief set by the candidate. The portfolios consisted of audio files, musical scores and any other accompanying documentation.

A sample of 150 NEA candidate submissions were selected from the 2019 exam series. The sample was selected using stratified random sampling based on candidate final grade. The original marks awarded by the teachers were removed, as well as any teacher commentary about how they evaluated the work. The candidate work was separated into performance and composition (so that the two elements could be judged separately) and loaded onto a bespoke online marking software. The software was user-friendly and allowed participants to listen to the audio recording (while simultaneously viewing the musical score and any other documents) and record their judgements all in one place.

[1] Currently such portfolios are marked by teachers and then moderated by Awarding Organisation trained assessment specialists. For details of the process see Gill (2015).

Fifteen participants were recruited to take part in the study. They were drawn from the pool of OCR assessment specialists for GCSE Music and, as such, they were familiar with the material and assessment objectives. They were a mixture of current and retired teachers.

Each participant judged 80 pairs of performances and 80 pairs of compositions, in the order of their choosing. The pairs were determined and allocated using a randomly generated design such that each candidate's work was included in 16 comparisons. The same design was used for both the performances and compositions. Participants were instructed to choose which of each pair better demonstrated the construct of interest:

- For performances: Which student performed with better technical control, expression and interpretation (accounting for difficulty)?
- For compositions: Which student demonstrated the highest level of successful compositional skills?

Previous research (Leech & Chambers, 2022; Vidal Rodeiro & Chambers, 2022; Walland, 2022) reported that participants sometimes found it challenging to make holistic judgements and sometimes resorted to analytical marking. Thus, in this study, we enhanced the training and made specific efforts to address potential discomfort with the method. This involved familiarisation, practice, and small group online training meetings where we discussed the judgements and provided strategies to assist with decision-making. Some participants raised queries about how the method would work in practice; we asked participants to try to concentrate on the exercise and not think about the logistics. In order to mimic the support of a traditional Team Leader,[2] we supported the participants throughout the judging and offered individual online meetings to discuss any further queries.

The participants completed their judgements at their own pace, working towards a final deadline. We designed and distributed an online post-judging questionnaire where we collected participants' views and experiences of the method. Topics included likes and dislikes with the method, ease of shifting from marking to CJ, any challenging comparisons, confidence in decision-making and whether the participants found themselves re-marking or using the mark scheme.

Frequencies of responses to selected closed questions are reported alongside the question. The open-ended comments were analysed and grouped into themes that spanned across the questionnaire (i.e., the themes did not directly correspond to specific questions) – firstly, according to the *Dimensions of judge decision-making* model (Leech & Chambers, 2023) and then into other data derived themes.

When reporting results, representative comments (rather than all) are presented to capture the full breadth of opinions. Obvious typographical errors were corrected to aid readability. Px denotes the participant number.

[2] A Team Leader will guide and co-ordinate a team of assistant examiners to ensure they are all marking to the same standard.

Findings

We started by asking the participants how easy they found the shift from traditional analytical marking to PCJ. Overall the participants found the shift to be straightforward saying it was “a simpler task!” (P2), that “judging quality rather than analysing criteria felt quite natural” (P7) and that it “was not looking to fit a piece into a box – just establish if it was better or worse than a second piece” (P5). Three participants were neutral and only one reported finding the shift difficult, saying that:

> “It was challenging to change to the comparisons but once I had done a few learners’ work I felt more at home with it. It was a different way of addressing assessment and I did enjoy it by the end of the work” (P6)

Considering it was the first time that participants had encountered the PCJ approach, their reaction was promising.

We now look in more detail at decision-making and any challenges experienced by the participants. In order to frame the participants’ perceptions of the method, we drew on the *Dimensions of judge decision-making* model (Leech & Chambers, 2023). This model (Figure 1) highlights that a judge’s CJ decision-making is related to: their individual approach, the structure and features of the question paper, the way that the candidates have answered items and the unique comparative requirements of the CJ task. The arrows in the model illustrate that these dimensions impact and interplay with one another. Using this model allows us to interrogate whether judges are making appropriate decisions and therefore creating valid outputs. Table 1 summarises the judges’ decision-making features found in this study. That a number of construct-relevant features are present in each dimension supports the second concept of holistic judgement specified earlier (Leech & Vitello, 2023). The sections that follow report the findings from the current study for each of the points in Table 1 in turn and, where relevant, provide reflections on how these findings compare to those from past CJ studies that involved text-based artefacts.

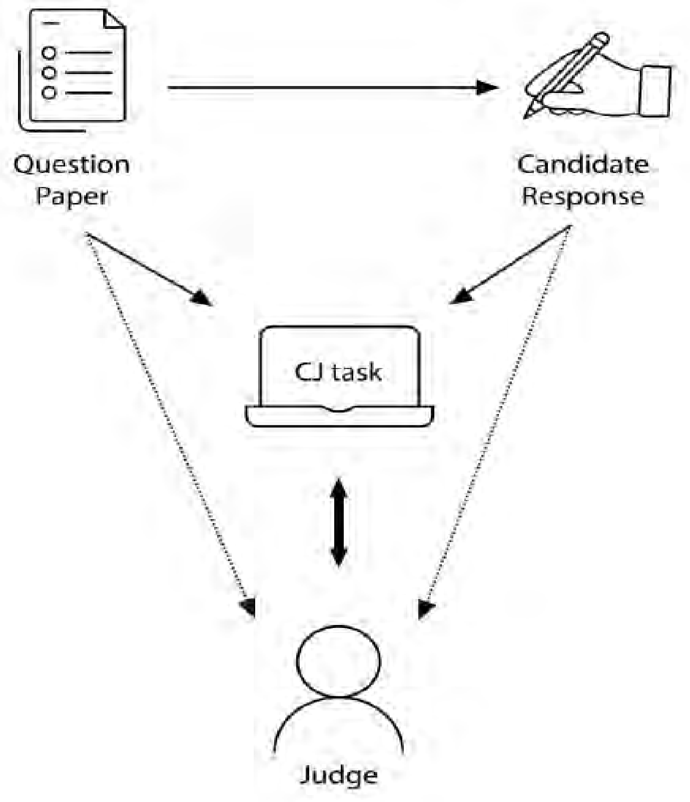

Figure 1: Dimensions of judge decision-making

Table 1: Summary of decision-making features identified in the current study by dimension

Judge-centred dimension	Question paper features dimension	Candidate response and CJ task dimensions
• Ability to make holistic judgements • Confidence • Understanding the process • Cognitive load • Judge bias	• Performance versus composition • Many pieces of information (e.g., score and recording)	• Instrument, genre/style, medium (sequencing versus live) • Piece difficulty • Balance of different response elements • Closeness in quality

Judge-centred dimension

The first dimension of the model we will examine is the judge-centred aspect. One of the key features within this dimension is whether judges were actually able to make holistic PCJ decisions – a central tenet in ensuring the validity of the method. Whether or not participants showed any marking behaviours while conducting PCJ may be an indication of this. We found that the majority of participants reported that they never or rarely engaged in these behaviours (see Table 2). This is in line with the instructions they were given during training, which emphasised that the participants should try to avoid marking the work. Nonetheless, some participants did note that:

> "I would have found it easier on several occasions to award individual marks for technical skills, expression and difficulty, and then come up with a final total to make a judgement" (P12)

> "Although I found this quite easy, I did struggle with not giving pieces a mark. I had to keep mentally referring back to the old mark scheme as there is no real guidance for marking in this way" (P15)

Table 2: Participants' self-reported engagement in marking behaviour

	Never	Rarely	Sometimes	Frequently
During the PCJ part of the study, how often did you find yourself re-marking students' work (i.e., awarding marks in the traditional way)?	7	7	1	0
During the PCJ part of the study, how often did you need to refer back to the traditional mark scheme in order to make decisions?	8	3	1	3

When making holistic decisions we expect judges to draw on their experience and their knowledge of what "good" looks like and acknowledge that this may vary across judges. In the assessment context this will inevitably include knowledge of the assessment objectives and expected standards, thus reference to this would be expected. However, if these judgements become mechanistic (e.g., marking) and breach the third interconnected aspect of holistic judgement (Leech & Vitello, 2023) then the judgement is no longer holistic, which is a threat to validity. It is encouraging that marking behaviours were infrequent. In addition, the presence of some marking behaviour is not unprecedented as previous CJ studies have found that judges sometimes re-marked the work explicitly using the mark scheme or their knowledge of it (Leech & Chambers, 2022; Vidal Rodeiro & Chambers, 2022; Walland, 2022).

The self-report of the participants suggested, that for the most part, the decisions were valid. In fact, the participants noted that the exercise made them reflect on the essential features of effective performance and composition. Participant 14 noted that "it made one think harder about the fundamental principles of composing and performing to assess why one piece was better/worse than the other".

Another related feature is the judges' level of confidence in making their PCJ decisions. When asked directly, most participants reported that they were confident or very confident (see Table 3).

Table 3: Participants' self-reported confidence in their PCJ decision-making

	Very confident	**Confident**	**Neither**	**Not confident**	**Not at all confident**
How confident were you about your PCJ decisions?	2	10	3	0	0

Their comments also emphasised their confidence, for example, Participant 13 noted that "I'd say 90 per cent of the time very confident. There were just a few where I doubted my judgement" and Participant 11 reported that "generally I felt confident in the choice I made because I felt it was clear in the majority of cases". Participants cited their experience, previous marking and moderating, and their ability to play many instruments as contributing reasons for their confidence. Participant 7 gave a succinct reason for their confidence: "because it was a straight comparison of quality and musicianship". Other reasons stemmed from there often being a clear difference in quality between the pieces, and the knowledge that they were not solely responsible for the candidate's final mark. Participant 11 summed this up:

> "Some pieces were very easy to compare as the standard was so vastly different. Some were harder but I took comfort in the fact that I wasn't the only person marking the candidate so it didn't all fall on my shoulders" (P11)

One participant, who rated their confidence as "neither", reported that they found it "very difficult to compare. We are not used to doing this. We mark/moderate individuals but don't compare" (P13). This suggests that unfamiliarity may have played a part in their level of confidence.

It is also possible that the research context affected confidence levels. In fact, two participants alluded to this as increasing their confidence:

> "Actually, I felt very little pressure in doing this marking, I guess because it is a research project using 'old' candidate work. When marking/moderating 'live' work, one is much more conscious that what you do has a direct effect upon an individual's/centre's results" (P2)
>
> "The process has been enjoyable but I felt under no pressure of time" (P5)

This aligns with findings from the text-based standard maintaining studies cited in Leech and Chambers (2022): judges involved in live (exam session) trials of the methods found judging more challenging than those in pilot studies. Nonetheless, the high levels of confidence in PCJ found in the current study reflect those from text-based research (Vidal Rodeiro & Chambers, 2022).

Another judge-centred feature was judges' understanding of the process. Several participants wanted more information on the method – evidence on how it would work in practice (e.g., who would make judgements) and what the outcomes would be (e.g., how would final marks be derived, what feedback could be given to schools). These points go beyond the current article's focus on judges' experiences of making judgements but have the potential to affect judges' wider confidence in the method.

In terms of the decision process, several concerns were expressed, for example, feeling bad for the losing candidate, having to make a judgement when the participants felt the pieces were of the same standard, or not seeing the benefits of the method:

> "A lot were very easy, but sometimes I liked both or thought both were not so good. Sometimes there was a very good performance and an exceptional performance and I felt bad saying the exceptional one was better, when the very good one would have been the best in many other pairings" (P8)

> "Very often there was a distinct difference between the two pieces being listened to. I was just a little uncomfortable marking one piece as being better than another when they were of the same standard (especially at the top end)" (P15)

> "I'm not sure what the gains would be or what would be achieved beyond the traditional methods unless comparisons were made between pieces of a similar type. Even then how would you compare a rock singer with a more classically trained singer" (P14)

> "I found it straightforward to shift but I'm not confident about the results it will produce, even when all the moderators' decisions are put together, some decisions could have gone either way. I think the top and bottom candidates will be in the right place but I'm not sure about all the ones somewhere in the middle" (P9)

Related to understanding the process, some judges commented on the method itself:

> "I can see the benefits of the PCJ method and I believe that if moderators are trained to complete this approach it would be successful. I would think that moderators would listen to more pieces of music which again would be a good thing" (P6)

> "This method is very subjective" (P8)

> "It just felt a bit random to me. It didn't seem like I was rewarding the candidate's work" (P9)

In previous research on text-based studies there has often been one or two judges who did not favour CJ as a method (Vidal Rodeiro & Chambers, 2022; Walland, 2022), so it is not surprising that some caution about how the method worked was expressed by some of the current participants.

In terms of cognitive load, the pieces of music were often quite long, and some participants struggled with remembering the first piece after listening to the second, for example, Participant 13 said that it was "too long after listening to both examples to remember the first one sufficiently". Participants reported that they sometimes took notes as a memory aid. Participant 1 noted that "listening to music takes time! Notes needed to be taken in order to remember back to piece 1". Also related to note taking, Participant 3 commented that "it became rather dull in places as a lack of marking / note writing to help lead to a conclusion led to a lack of brain power / interest at times".

The cognitive load needed to complete the activity has been discussed in other text-based studies (Vidal Rodeiro & Chambers, 2022; Walland, 2022). Interestingly, the challenge noted here, of recalling the first artefact, was not apparent in the text-based tasks, as a quick view or skim of the first text-based script would be enough for the judge to recall the content – with music, there is an absence of such cues.

Some participants mentioned judge bias as a feature:

> "I found judging drummers very hard with other performances and I wonder if I was harsher there on the drummers" (P11)
>
> "No real dislikes – sometimes a close call was hard to make. Possible scope for bias by the assessor against work in certain genres, meaning that the wrong piece is preferred...?" (P7)
>
> "In a real situation I feel judgement could be clouded at times when hearing something new or refreshing i.e., a steel pan after listening to 3 or 4 vocal pieces in a row" (P3)

This is an interesting finding since judge bias has not been previously raised by participants in text-based studies.

Overall, the judges felt able to and were confident in making judgements. However, similarly to text-based studies (Leech & Chambers, 2022; Vidal Rodeiro & Chambers, 2022), the participants did experience challenges in making the judgements due to the interplay with other dimensions. The next sections discuss the other dimensions.

Question paper features dimension

For GCSE Music NEA, there is no question paper as such. However, candidates produce a recorded performance and performed composition, so we can think of these as essentially two items, weighted equally. The participants found that compositions appeared to present more problems than performances. Participant 6 noted that "the performances were more straight forward". Other participants also reported this and added additional detail about the interaction with medium and cognitive challenge:

> "I found performances easier than compositions to compare especially if it was a live composition played well compared to a computer export" (P11)

> "Composition required a consideration of the whole piece more so than performances" (P7)

There were often many pieces of information, for example, cover sheets, musical scores and the candidate recording. The interface of the software was designed to be user-friendly, however navigating through this work and viewing it clearly was sometimes a challenge for judges. One participant noted that "some candidates had about 60 pages of score" (P1). Another noted that:

> "I would have liked to be able to jump between documents. There were numerous occasions where I would have liked to have jumped to a cover sheet, which was the final document, but I had to scroll through page after page of score to get to it. Also, a zoom function would have helped at times" (P12)

A related issue has previously been found with text portfolios, where participants experienced some difficulties when scrolling through many pages of work due to time lags (Vidal Rodeiro & Chambers, 2022) and difficulty making decisions due to the layout of portfolios.

We found that the features apparent for this dimension were quite different to text-based judgements, due in part to the absence of a question paper containing discrete items. For text-based tasks, the features mentioned by judges were: number of short items, the presence of longer questions involving evaluation or explanation and the focus on more discriminating items over others (Leech & Chambers, 2022).

Candidate response and CJ task features dimension

The candidate response features mentioned by participants included elements such as instrument, genre and style, difficulty of piece and medium. We found discussion of these features to be inextricably bound with discussion of the PCJ task. Comments centred around balancing the different response features when making comparisons between the candidates.[3] As a result, we discuss both dimensions together.

Participants reported that for the most part the decisions were straightforward, and that "most of the time there were few problems differentiating pieces" (P14). However, when the pieces were very different in some way – for example, "perhaps one was technically accurate but emotionless, another full of expression but out of tune" (P9), or "a difficult piece played badly with an easier piece played really well" (P10) – then comparison could be more challenging. Interestingly, participants

[3] This may be in part due to the nature of the survey question. In this study we asked a question about whether they found any comparisons challenging rather than an explicit question on how the participants made their decisions.

differed in what they found challenging. Table 4 highlights some of the response elements and the differing views.

Table 4: Differing participant views (quotations) with respect to candidate response features

Response feature	Perspective – easy	Perspective – neutral	Perspective – challenging
Difficulty of musical piece	It was easy to compare performances where the difficulty level was different. It was much harder to compare performances which were very similar in standard (P8)	Overall, regardless of the instrument, there were several performances that were difficult to determine which was better and sometimes it was the difficulty of the piece that was the decider (P3)	The biggest challenge for me was comparing pieces with widely different difficulties. There were easy pieces that were played fluently and with style, compared with significantly harder performances that had hesitations, etc. (P12)
Instruments	… I found it okay to compare performances on different instruments (P8)		Difficult when marking completely different instruments i.e., Piano versus Indian Raga vocal line (P3)
Genres/styles	I actually found it quite straightforward to compare a range of different genres. The quality of a great composition or performance shone through regardless of the genre (P7)	I think it is always hard to mark things that one is less familiar with such as classical Indian music or sequencing (P4)	It was sometimes difficult when marking the same instrument which were similar in credit but of different styles i.e., a Big Band drummer playing live versus a Grade 8 Rock drummer (P3)
Medium (sequencing versus live)			Sequencing against “live” instrument was difficult. … (P5)

Interestingly, these features seemed to have more impact on participant comments than some of the features to be assessed as set out in the mark scheme (e.g., for performance: technical control and fluency and expression and interpretation; and for composition: sense of style, a range of musical elements, composition techniques, stylistic and structural conventions). This could be evidence of these response features getting in the way or perhaps evidence of the participants judging holistically.

Several participants reported that it was challenging to judge between two candidates whose work was very similar in quality: Participant 11 reported that "very occasionally I wanted to say it was a tie as I really felt both pieces were the same standard", Participant 15 cited instances "where the same mark would have been awarded to both in the usual mark scheme" and went on to report that "when work was of an identical standard there was no option to show this - you still had to choose which one was better". This is a challenge that has been seen across previous CJ studies. Judges often struggle in this scenario as it goes against their many years of training and their wish to do right by the candidate. In the training as part of the current study, we tried to reassure the participants and explained that the method, with multiple judgements, would ensure the appropriate outcome for the candidate. The fact that participants worried about this issue despite the training suggests that further reassurance and evidence needs to be provided to judges (and other stakeholders).

The participants reported a number of strategies for dealing with the challenge of comparing work of similar quality:

> "In most cases one candidate's work seemed clearly better than the other. When this was not the case I made my best judgement and trusted that the system would work" (P4)
>
> "Where there were close calls, it was back to basics – who was the most accurate and the most musical and which piece was delivered the most successfully given the challenge of the repertoire" (P7)
>
> "With some less able musicians it was sometimes a case of which one was worse rather than better and working it out that way" (P5)
>
> "Another challenging performance was a Rap artist whose performance was stylish and professional versus an alto sax performance. I found myself taking other things into account opting for the sax as this candidate would have had to learn how to play the instrument and follow the music over a longer period of time" (P3)

This last comment shows how other, potentially unintended, factors might be used where judgements are difficult. Some participants' comments showed their awareness of the need to know the criteria to be used even when making comparative judgements:

> "It is easier to compare 2 pieces rather than trying to fit them into a level category. You still need to know/understand the criteria on which you are judging the pieces" (P5)

The features described in this dimension were again often different from those found in text-based studies. For text-based studies, candidate response features were centred around response consistency, depth of responses, clarity/structure, spiky profiles[4] and omitted questions, and use of examples, facts and statistics (Leech & Chambers, 2022; Vidal Rodeiro & Chambers, 2022; Walland, 2022). For Music, as there was only one non-text task in each condition, some of these responses were not present (e.g., use of examples, facts and statistics) or were presented differently (e.g., an imbalanced performance instead of a spiky profile).

For GCSE Music there are many variables (e.g., instrument, medium, difficulty and genres, etc.) and it appears that it is the interactions between these features and the many permutations and combinations that prove challenging.

Fairness

Moving beyond the dimensions model, another related theme that came up in the responses was that of fairness. One participant stated that "it just doesn't seem very fair, the two being compared are so different, e.g., a big band composition on Sibelius compared to a garage band piece, or a film music composition compared to a piano piece" (P9) and "it would be fairer to compare similar instruments where possible" (P9).

Another concern was that candidates would not receive a fair grade, as Participant 3 noted:

> "I didn't enjoy this method. It felt less personal and less hands on with a lack of professional opinion. I felt that in some cases, there wasn't a need for expertise or musicianship to be able to determine 'which was better' and that the candidates would not receive a fair and considered grade." (P3)

In contrast, some participants saw the inherent fairness in the method itself due to multiple judgements, for example, "it felt fairer that the marks would be based on lots of people's opinions" (P4) and "I guess the more times a candidate's work is viewed by different assessors, the more chance there'll be of establishing a true and fair assessment" (P2). Fairness was also cited in comparison to the current moderation process:

> "It appears to be a fairer system of marking. Although it is still subjective, the fact that a number of people would mark the same pieces should make for a better consensus. It would no longer be the school's opinion versus the (single) moderator's opinion" (P5)

> "A range of markers look at work from a range of centres, so one marker is not responsible for marking all the work of one centre – this provides a balance of opinion" (P7)

> "I think centres would welcome the idea that the work is marked multiple times to establish a clear overview of the relative standard of the work" (P7)

The current finding regarding the benefit to fairness of multiple judges evaluating

[4] A spiky profile is where candidates answer some questions well and others poorly.

one candidate’s work echoes views reported in a text-based GCSE English Language CJ study (Walland, 2022).

Conclusion

This study sought to investigate, from the judge perspective, the use of PCJ with auditory-based artefacts. In this study we used GCSE Music, and as such it is important to note that the findings as detailed relate to music recordings and will not necessarily apply to all auditory-based artefacts. This study used a small sample of participants (n=15) and one component of GCSE Music. It involved only self-report data; observational research could add richness and support to the findings. As such, these factors should be borne in mind if generalising the findings more broadly.

The use of auditory-based artefacts, in particular music files (as in this study), is an under researched context for CJ. At the start of this article, we noted two aspects of the judge experience that are necessary to support the validity of the CJ method. Namely, whether judges are able to make appropriate CJ decisions and whether they believe in the validity of what they are doing.

The enhanced training, familiarisation activities and support we gave participants appeared to have proved effective. We saw that for the most part judges were able to make appropriate decisions, there was little evidence of participants re-marking or attending to construct-irrelevant features, and the judgements involved the balancing of different response elements. This also suggests that the second and third aspects of holistic judgement, as defined by Leech and Vitello (2023) (that comprehensive relevant evidence is used and interconnectedness is considered), were met. (Note that Leech and Vitello’s first criterion of holistic judgement is also met, since the participants provided a singular judgement for each pair of performances or compositions.)

In terms of whether the judges believed in the validity of what they were doing, the findings were mixed. Participants could see the benefits of having multiple judgements of each candidate’s work and there was also some evidence that participants were revisiting the fundamental principles of composing and performing. Some participants, however, appeared unconvinced by the method. Sometimes it was a lack of understanding or belief in the process – this was particularly for work they considered to be of the same standard. Further training, experience and provision of evidence could help alleviate this.

A key concern related to candidate work that was very different in some way – for example utilising different genres or instruments, or when the piece difficulty varied. This is harder to address. Leech and Chambers (2022) noted that in the CJ context “there is no immediately clear way to determine which paper of a pair or pack is the superior if each is better in a different way” (p. 45). They discussed the tension between the way current exam papers are set up (i.e., to be marked) and holistic CJ judgement which relies on a judge’s conception of what constitutes better performance. Leech and Vitello (2023) described this as an “informal rubric” where judges determine which features to prioritise. They “contend this informal rubric should be made more formal by the provision of more explicit guidance, and the comparison simplified by, if at all possible, ensuring the similarity of form between different artefacts” (p. 18).

In this study we did provide some guidance, however for the participants it was the first time they had used this method, and it was unsurprising that some challenges remained. It is recommended that similar guidance and training should accompany further CJ studies so that judges feel confident in making independent holistic decisions and are clear which elements would be considered construct irrelevant in any context. This would help ensure the validity of assessment outcomes.

In terms of simplifying the comparison, for GCSE Music, pairing similar artefacts would be practically unfeasible. Even if, for example, pieces were paired on one factor such as instrument, the genre can be vastly different. However, further research utilising observation-based methods could be used to render the methods by which judges resolve this challenge explicit. In parallel, specific research into the effects of instrument and genre on CJ outcomes could also be conducted to explore whether any bias exists.

What these challenges show is the complexity of making CJ decisions – far from an instant decision, a holistic judgement is the “consequence of the aggregation of a series of micro-judgements, each of which might be quite different for each judge making them” (Leech & Vitello, 2023, p. 13). The level of challenge can be further increased when an element of “difference” is added. All artefacts involving some level of candidate choice, whether text or auditory-based, will create challenges for CJ as difference will be inherent. This difference could be for example, topic in History or choice of sport in PE. Music raises this level of challenge further in that so many elements interplay with each other. It is possible that there could be a “difference ceiling” – a point in certain contexts where the artefacts are just too different to be compared validly using CJ, and other methods such as analytic marking or “levels-only” marking would be more suitable (for information on “levels-only” marking see Walland and Benton, 2023).

Some of the challenges the participants experienced were more practical in nature, for example, the cognitive load in remembering the first artefact or ease of viewing any documents while listening to the recording. These factors are unlikely to be restricted to music recordings and could apply to other portfolios containing audio recordings. Care should be given with respect to the length of any recordings. If portfolios containing large quantities of evidence are to be used alongside audio recordings, then it is necessary to consider which pieces of evidence should be included. Clear design and user experience testing of any software are vital.

It is important to note that, for the most part, participants found the shift to PCJ straightforward and felt confident making the judgements. What was particularly clear from this study was that, in general, participants were open to new ideas and ways of working and welcomed the opportunity to be involved in the research.

References

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

Bramley, T. (2007). Paired comparison methods. In P. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246–300). QCA.

Bramley, T. (2022). Editorial – the CJ landscape. *Research Matters: A Cambridge University Press & Assessment publication, 33*, 5–8.

Chambers, L., Vitello, S., & Vidal Rodeiro, C. (2024). Moderation of non-exam assessments: A novel approach using comparative judgement. *Assessment in Education: Principles, Policy & Practice, 31*(1), 32–55

Curcin, M., Howard, E., Sully, K., & Black, B. (2019). Improving awarding: 2018/2019 pilots. *Research Report Ofqual 19/6575. Research and Analysis*.

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

Jones, I., Wheadon, C., Humphries, S., & Inglis, M. (2016). Fifty years of A-level mathematics: have standards changed? *British Educational Research Journal, 42*(4), 543–560.

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

Leech, T., & Vitello, S. (2023). What is a holistic judgement, anyway? *Research Papers in Education*, 1–23.

Mason, K., & Garelli, L. (2022). *Assessment of art and design courses using comparative judgement in Mexico and England.* Paper presented at the annual conference of AEA-Europe, Dublin, November 2022.

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. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education, 22*(2), 157–170.

RM. (2022). *Using Adaptive Comparative Judgement as a reliable way to assess oracy at scale*.

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. P. (2016). A study of the use of pairwise comparison in the context of social online moderation. *The Australian Educational Researcher, 43*, 273–288.

Vidal Rodeiro, C., & Chambers, L. (2022). Moderation of non-exam assessments: Is Comparative Judgement a practical alternative? *Research Matters: A Cambridge University Press & Assessment publication, 33,* 100–119.

Walland, E. (2022). Judges' views on pairwise Comparative Judgement and Rank Ordering as alternatives to analytical essay marking. *Research Matters: A Cambridge University Press & Assessment publication, 33*, 48–67.

Walland, E., & Benton, T. (2023). Multiple marking methods as alternatives to single marker analytical essay marking: Exploring pairwise comparative judgement, rank ordering and levels-only [Manuscript submitted for publication].

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