Judges' views on pairwise Comparative Judgement and Rank Ordering as alternatives to analytical essay marking

Emma Walland (Research Division)

Background and aim

In exam board settings in England, analytical marking is the typical method used to mark essays. This requires examiners to allocate marks, nested within levels of performance, for different areas of achievement or features of the essay (Meadows & Billington, 2005). However, this method has attracted criticism from the assessment community. Some have argued that relying on narrow and detailed mark schemes is not ideal for subjects such as English language due to the examiner judgement and interpretation involved in assessing extended writing tasks (Meadows & Billington, 2005). Others argue that too much detail in mark schemes could negatively influence teaching and learning, narrowing the focus of teachers and students on what is needed to gain marks (Brooks, 2004; Holmes et al., 2017; Wheadon, Barmby, et al., 2020; Wheadon, de Moira, et al., 2020).

In contrast, holistic methods involve marking a piece of work based on an overall evaluation, rather than viewing features of the text as separate entities. According to Hamp-Lyons (1990), it is "based on the view that there are inherent qualities of written text which are greater than the sum of the text's countable elements and that this quality can be recognized only by carefully selected and trained readers, not by any objectifiable means" (p. 79). Pairwise Comparative Judgement (PCJ) and Rank Ordering (RO) are holistic methods in which examiners make judgements about the overall quality of essays in comparison with others, and the final scores awarded to students are derived from a combination of several judges' inputs. The methods require examiners to choose a better essay between a pair (PCJ) or to sort larger packs of essays into order from best to worst (RO), guided by the assessment objectives.

PCJ and RO have been the focus of much previous research (Holmes et al., 2017; Wheadon, Barmby, et al., 2020), and researchers are exploring their potential applications for exam boards. A main disadvantage is that the scores obtained provide less detail or diagnostic information about students' performances, and how examiners made judgements is less clear. This could be a concern for

stakeholders, such as teachers, who may prefer more detailed information about how scores are allocated in order to inform their teaching or to make informed enquiries about whether to challenge the marks. There are also concerns about a potential increase in cognitive demand placed on examiners using these methods, and whether they function as well for novice examiners.

Previous research in a variety of contexts shows that comparative judgement methods have the potential to produce high reliability and validity (Benton & Gallacher, 2018; Bramley & Vitello, 2019; Heldsinger & Humphry, 2010; Jones & Inglis, 2015; Steedle & Ferrara, 2016; Verhavert et al., 2019). But there is less reported data on how examiners experience the methods, and particularly on how they, and other stakeholders, may feel about them as alternatives to marking (for some examples of work reporting perceptions in various contexts see Jones et al., 2015; Kimbell et al., 2009). In addition, software to allow RO studies to be completed online has only very recently been developed. As such, the present article is the first to report upon examiner experiences of this approach. Understanding examiner experiences is important because examiner experiences are vital for retention, and stakeholder confidence in the methods is important for ensuring trust in the assessment system. In this study, in the context of GCSE English Language, I looked at perceptions of PCJ and RO in terms of:

- how decisions were made, and the marking strategies used
- cognitive demand and ease of use
- enjoyment
- quality of results
- stakeholder response to the methods
- suitability for new examiners.

Method

Participants

Fifteen GCSE English Language examiners with at least three years' examining experience took part in the study in early 2021. I recruited them via email, following the ethical procedures according to the British Educational Research Association (BERA) (2018). The participants were broadly representative of the diverse group of examiners that mark live examination papers in terms of their roles (seniority), teaching experience and previous marking performance ratings. For most participants, it was their first time using the methods. Three had previously used PCJ and four had done paper-based rank ordering or something similar in a school setting.

Procedure

Two separate sets of 150 essays were sampled from the OCR GCSE English Language June 2019 series for use in the PCJ and RO studies respectively. They were non-fiction essays worth 40 marks. The essays used for each comparative judgement approach were different but had the same distribution of scores from traditional analytical marking.

For the PCJ study each essay was included in 20 separate paired comparisons creating a total of 1500 pairs. The participants were each given 100 pairs of essays to judge. For the RO study each essay was included in 8 separate packs of 10 essays that needed to be ranked. This created a total of 120 packs of 10. As such, for this study, each judge was assigned 8 packs of 10 essays.

The participants were given detailed instructions, marking guidance and technical guidance for the software for each task, in writing and during a Microsoft Teams meeting. The tasks were carried out remotely using browser-based CJ software and the order in which they used the methods varied. For PCJ, they were asked to choose which essay of each pair was better and for RO, they were asked to rank packs of 10 essays in order from best to worst. The rankings were to be based on the assessment objectives for the essay, similar to Bramley and Vitello (2019). They were instructed not to re-mark the essays but to use a holistic professional judgement to make decisions. (The specific instructions given to participants are given in the appendix).

After marking with each method, the participants completed questionnaires (developed using SurveyMonkey) about their views and experiences of the methods. The questionnaire was a combination of single item scales and free-text comment boxes. At the end of the experiment, the participants also took part in 30-minute semi-structured interviews (via Microsoft Teams), which were recorded and transcribed.

Analysis

I report the data from the single item scales using descriptive statistics and graphs produced using SAS Enterprise Guide version 7.1. The free-text responses and interview data were analysed in MAXQDA 2020 (VERBI Software), using thematic content analysis (Braun & Clarke, 2006).

Findings and discussion

Participants' views and experiences of the methods were grouped into several themes during analysis. The themes are supported with illustrative quotations from participants and, where applicable, with data from the closed-response items in the questionnaire (five-point Likert-type items).

Is faster better?

In comparative judgement, a core feature of the method is that the judgements are intended to be quick to facilitate the large number of comparisons that are needed to produce sufficiently valid and reliable results. However, some of the participants expressed concern over the speed at which they were making judgements. They worried about potentially making judgements too quickly, being too influenced by the first paragraphs of the essays or overlooking the finer details. For example, Participant 6 noted that speed in both methods could lead to mistakes, saying:

> It's not necessarily a good thing to be quick. I marked one exam where the Chief Examiners deliberately made everybody go slowly on one question because there were so many mistakes on it.

Similarly, Participant 8 felt that speed of PCJ resulted in overlooking details. They said:

> I've always had markers on my team over the years where they just will go too fast and I always have to slow them down and say, 'look at the detail. Look at why this one is better for this reason. Look at the vocabulary'. Sometimes you do have to get into the detail of a script, don't you, in order to assess it? And I did worry how much it would throw it if you had certain markers like that.

Furthermore, Participant 10 felt a sense of fear about the speed of PCJ, saying:

> It was possible to reach a conclusion sometimes after reading the first paragraph or two [for PCJ] ... very few of them did I have to read the whole script which on the one hand works completely counter to how I've always worked as an examiner ... I think there's a real fear at some point because you sort of have this sense of, 'am I going through these too quickly?' So, I'd stop periodically and go through those again and I'd spend a few more minutes but still come to the same judgment.

Similar feedback from participants was also found by Jones et al. (2015), in the context of mathematics assessment, where examiners felt that skimming the work and not carefully examining each response was unprofessional. These findings suggest that examiners would need reassurance and encouragement that assessing in a quicker way can lead to equally, and hopefully more, reliable and accurate results. Examiners and other stakeholders would need to be made more aware of the benefits of gathering multiple judgements about each essay which compensates for the loss of time each individual examiner spends on each decision.

In contrast, some participants did not find the methods speedy at all. They noted finding it difficult and time consuming to make decisions and had to read the essays several times or use some form of analytical marking criteria (either the mark scheme or their own marking scale) to inform their judgements. Such strategies have been found in previous comparative judgement research too (Bramley, 2007).

Confidence with holistic judgements

Resorting to analytical marking strategies is not ideal as it undermines the intended holistic nature of the methods and their ability to capture judgements efficiently. However, this seems an understandable response to the uncertainty and lack of confidence that some participants had in holistic marking. Participants had varying views about using holistic marking or relying on their gut instincts. While some appreciated having more freedom to use their professional judgement, others felt uncomfortable with this. An example of a positive view from Participant 10 was, "It was a liberating experience to use gut-reaction and

professional judgement, rather than becoming bogged down in an overly complex mark scheme".

In contrast, others felt it was subjective, they lacked confidence in their decisions and believed that stakeholders may not accept it. Participant 8 said, for example, "I found it hard, and I found it hard that feeling of not being certain after I'd done it either". Similarly, Participant 2 said, "I feel the method [PCJ] would be very successful if all examiners were confident in marking holistically. It can be difficult getting into that mindset, if you have spent many years marking in the traditional way". Finally, Participant 12 noted, "if you tried to explain that to a parent whose child has just done an essay, 'well it was a gut feeling', I don't think it would go down too well".

Making direct comparisons among essays in a holistic way is a departure from the usual analytical method examiners are used to. Therefore, using these methods will require a period of adjustment from both examiners and other stakeholders. It is likely that comfort and confidence with holistic marking would increase over time with more training and practice.

Quality of results

As RO and PCJ are quite different from traditional analytical marking, I was interested in exploring participants' views on the quality of results they perceived that the methods would produce. The participants were also asked for their opinions of how other stakeholders (such as teachers, parents and other examiners) might view this. A limitation of this data is that it is based on expert opinion, rather than gathering views directly from other stakeholders. However, they provide a good indication of possible reactions, and all participants did have teaching experience to draw upon.

The results from the questionnaire (as shown in Figure 1) showed that most participants were fairly or very confident in the quality of results produced by the methods. PCJ had the highest proportion of positive responses.

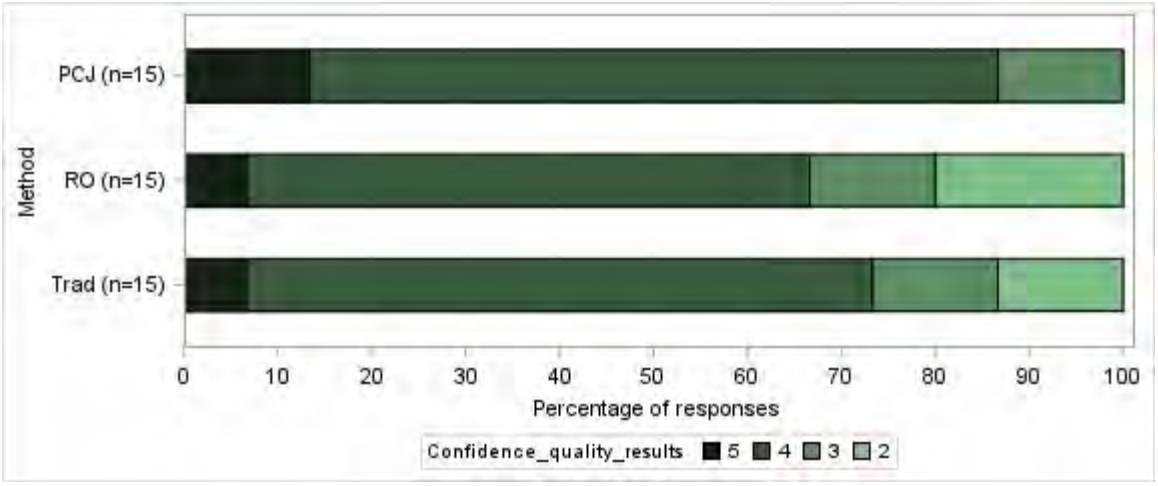

Figure 1: Participants' responses about their confidence in the quality of the results produced by the methods on a scale from 1 to 5. Darker shading represents more positive responses (increased confidence). 5 was "very confident", 4 was "fairly confident", 3 was "not sure", 2 was "fairly unconfident" and 1 was "very unconfident".

Figure 2 shows that participants were fairly positive about stakeholder reactions to RO, but less sure for PCJ. The findings in the following themes help to explain these results.

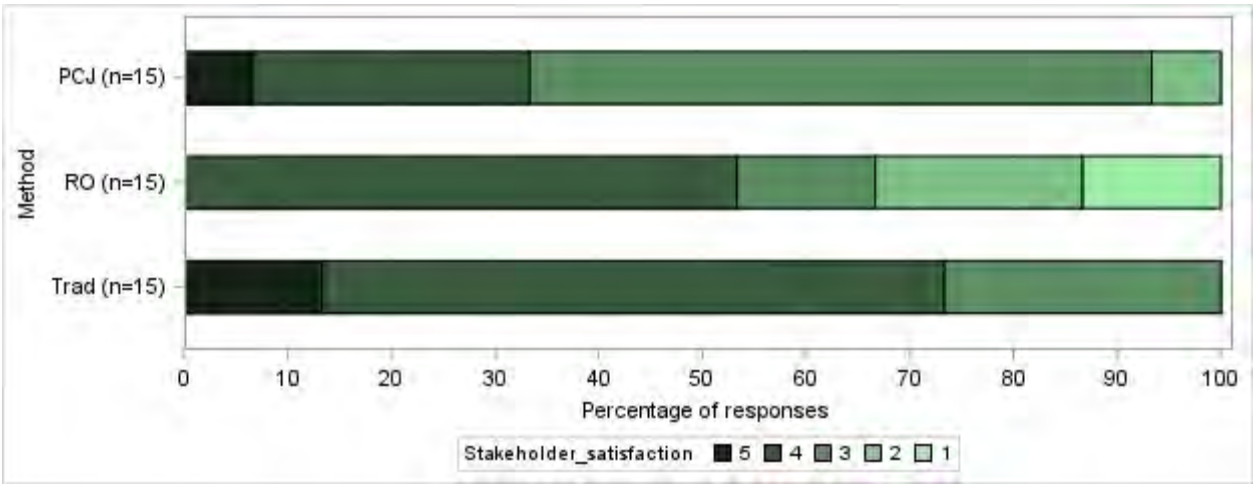

Figure 2: Participants' responses about their opinions on stakeholder satisfaction with the results produced by each of the methods on a scale from 1 to 5. Darker shading represents more positive responses (greater satisfaction). 5 was "very satisfied", 4 was "fairly satisfied", 3 was "not sure", 2 was "fairly dissatisfied" and 1 was "very dissatisfied".

The benefits of multiple marking

One driver behind positive views of the methods was multiple marking, by which I mean the fact that CJ scores are derived from the decisions of several examiners. In contrast, in traditional marking, the vast majority of essays are marked by only a single marker. Participants saw the formation of a consensus view among examiners as a highly positive feature that stakeholders would appreciate, and they felt it would help with the subjectivity possible in a subject like English.

For example, Participant 12 said about RO, "It is reassuring to know that other examiners are marking the same scripts, so there is support and my individual decision is not the ultimate one". Similarly, Participant 9 noted, about PCJ, "For a subject like English Language, a group judgement would result in a less subjective response". Similar feedback was raised by participants in a study by Kimbell et al. (2009), albeit in a different context (design and technology e-portfolio assessment).

There was only one negative comment about multiple marking. For PCJ, one participant wondered whether examiners might be less careful if the responsibility for marking was shared. Participant 11 said, "I wonder if this sense of security and the anonymity of judgement might result in less careful choices". This highlights the importance of ensuring accountability in marking, whichever methods are used.

Individual versus comparative approaches

Comparative judgement methods differ from traditional marking methods because, rather than marking each essay individually, they are considered in a pair or group and in direct comparison with each other. There were mixed views from participants about this mechanism. Some participants felt it was a positive feature that would lead to more accurate and reliable results, and they enjoyed comparing essays with each other. One advantage of comparative methods is that the results are not influenced by examiners' individual leniencies or severities, as they are not making absolute judgements.

In contrast, others felt that the best method would be one that considered each essay on its own. They preferred an approach that was more closely tied to a marking scheme where each essay could be judged on its own merits and felt that stakeholders would prefer this too. Furthermore, they noted that how comparative methods translate into grades may be more difficult for stakeholders to understand (also noted by Steedle & Ferrara, 2016). For PCJ, some participants felt that the approach was too subjective and dependent on which essays were in each pair.

Examples of positive views about making comparisons included Participant 15, who said, "I would definitely say that the paired marking makes you more consistent. Because you're constantly thinking about how you've [made judgements]". Similarly, Participant 10 noted:

> My difficulty when I've been an examiner for many years has been the ability to show consistency over large groups of scripts … in the past, there has been a sense when moving from script to script of thinking back to one paper, say, ten scripts ago, and wondering if I had marked it too generously or too severely. I came out of Rank Ordering with a reasonable amount of confidence (and not too much difficulty) that my ranks were accurate.

In contrast, examples of negative views included:

> I think most stakeholders would expect a student's essay to be marked in detail and would lack confidence in this method. Personally, if I wrote an essay under exam conditions, I would expect it to be marked and scored against the specification as an individual piece of work (Participant 5).

> There would be some of my students who would be motivated by this marking method, other students would be intimidated or disheartened by being directly compared with others for a decision to be made (Participant 7).

> There is still no mark scheme, which I think is important to give clarity to students, teachers, parents about how to improve and what to aim for to achieve a level (Participant 4).

> I disliked the absolute nature of [PCJ]. It sometimes felt as if you were doing a disservice to a good student, simply because they were up against a marginally better one, and you were unable to reward them for their achievements. Similarly, you were unable to reward the achievements of weaker students as they were inevitably not chosen. I missed the satisfaction of the finer points of assessment and the awarding of a final score (Participant 9).

The methods were seen to have less transparency as they were less closely tied to a mark scheme and did not leave details of how examiners made judgements, a point also raised in previous literature (Bramley, 2007; Holmes et al., 2017; Steedle & Ferrara, 2016). Relevant here, as noted by Aloisi (2020), is the notion that stakeholders do not like black boxes in marking and they desire 'explainability' in addition to reliability and validity.

Simpler versus detailed marking criteria

The two methods used far simpler written judging criteria – a summary of the assessment objectives – compared with the original analytical mark scheme, which is long and detailed, indicating what needs to be achieved for each level. There were mixed views about whether simpler or more complex marking criteria are better. Some participants enjoyed not having to interpret complex and ambiguous terminology in mark schemes, such as phrases like "deliberately adapted" versus "confidently adapted the form of the text", which could be interpreted differently by different examiners (see, for example, Nadas et al., 2021). Similar themes were raised by participants in regard to assessing design and technology e-portfolios, where it was noted that PCJ could be seen as fairer due to the holistic nature of marking, as the existing marking criteria can be too limiting (Kimbell et al., 2009). Some participants in the current study also appreciated having more freedom to use their professional judgement. They also felt that stakeholders would prefer simpler marking criteria as it would enable teachers and students to better understand what is being assessed. Simplified marking methods were also felt to be useful in encouraging new examiners and new teachers to mark. For example:

> This methodology [PCJ] brought back the sense of being able to enjoy a student's work, rather than the highly mechanised use of rigid marking criteria and in-depth analysis of the response (Participant 10).

> [PCJ] was a more joyful process, not being hamstrung by constant reference to statements of the mark scheme, being able to enjoy the development of trains of thought uninterrupted (Participant 1).

In contrast, some participants preferred a more detailed mark scheme and also felt that stakeholders might prefer it too, due to the reasons given in the examples below:

> Although it is time consuming, marking against a detailed mark scheme and assigning a level and choosing a mark means that each answer is viewed in much more detail. I feel looking closely at the SPaG [spelling, punctuation and grammar] elements of the mark scheme require the essay to be marked (Participant 7).

> A brief mark scheme [as in RO], as opposed to a more detailed one, may implant in their mind a sense that major strengths and weaknesses of their children's work are being overlooked and that perhaps the proper level of rigour is being inconsistently applied … Parents, teachers and other stakeholders may view this as overly simplistic and a watering-down of grades (Participant 10).

What informed judgements?

Previous research on holistic marking methods has suggested that, in comparative judgement exercises, examiners may be more influenced by construct-irrelevant features, such as handwriting and essay length (e.g., Meadows & Billington, 2005). This is a concern worth exploring, although Benton and Gallacher (2018) found evidence that essay length was not a particular concern for PCJ in comparison to other methods. In the current study, I analysed the features that participants felt had influenced their judgements, particularly when judgements were difficult. A limitation with this data is that it is self-reported, but it does provide an indication of what they thought they were attending to.

Encouragingly, I found that most participants reported making decisions in line with the constructs being assessed as per the assessment objectives. Some participants also mentioned more abstract constructs such as "flair", and how some students showed originality, imagination and creativity. Some also noted how the choice of topic could influence the quality of the work, for example, choosing a more ambitious topic and supporting it with facts and statistics, rather than relying on personal experience. Assessing some of these more complex constructs could arguably be better facilitated by a more holistic marking process (see also Jones & Inglis, 2015).

Only one construct-irrelevant feature was noted by two participants, and one noted using it more so than the other. This was graphology (or handwriting). While this could have negatively influenced the quality of their judgements, this could likely be prevented in a live setting through training, support from the team leader, and through monitoring and quality control processes.

To annotate or not to annotate?

In this context, annotation refers to practices like underlining spelling or grammar errors in an essay or highlighting where the student has met part of an assessment objective. Summative comments are a few sentences produced after a mark has been allocated to explain the mark. They are usually produced as part of the traditional analytical marking process. Previous research has found that

annotations could provide cognitive support for examiners while marking, support communication between markers and their team leaders, and help examiners explain their marking rationales to others (Crisp & Johnson, 2007; Johnson & Nadas, 2009).

In the current study, these were omitted as they would reduce the efficiency of the methods (see also Jones et al., 2015). While participants agreed that omitting them made the assessment process far less time-consuming, they also had negative views about this, which should be considered if the methods were to be implemented. Table 1 highlights the contrasting views from participants.

Table 1: Views of participants about annotating versus not annotating.

Not annotating	Annotating
Avoids distraction, allowing more focus and appreciation of each essay.	Helps some markers stay on track while marking.
Speeds up the marking.	Is more time consuming.
Accurate marking can take place without annotations.	Annotations can help some examiners make more accurate judgements.
Annotations are not necessary for teachers.	Annotations can be beneficial for teachers to see how marks were allocated.
Some see it as unnecessary and meaningless.	Is satisfying for some examiners, for example, giving them a chance to share feedback.

An example of a positive view about not having to make a summative comment was from Participant 9, who said:

> It was quite nice not having to put the summative comment on because I always found that I was just sort of like scrabbling for something from the mark scheme just to justify the match. To me that seemed a little bit meaningless. Actually, if you want annotation, just look at the mark. If this is the mark you've got, then look at the mark scheme to see the justification.

In contrast, Participants 4 and 6 raised some perceived benefits of annotating, saying respectively:

> I also think that some form of annotation is important, as it reassures parents, teachers and students that the script has been marked thoroughly. Also, it shows them where the standard was reached in the script.

> [It was] much less satisfying [not annotating] in that I couldn't say what I really thought about each piece of work. No piece of writing is wholly good or bad and in a good piece, we usually underline a few errors and in a poor piece, we try to give credit for something. This is often done with annotations or in the comments.

In a study by Kimbell et al. (2009), judges also raised concerns about the lack of formative feedback to schools, in the context of assessing design and

technology e-portfolios. Participants' differing experiences about annotations and summative comments indicate the individuals might perceive and benefit from them in different ways, which was also found in Crisp and Johnson (2007). Due to the mixed views, further research is needed to understand how teachers and examiners perceive and use annotations and summative comments in this context. Annotation could be a useful communicative and training tool, although previous research found that it did not have a dramatic effect on marker reliability (Crisp & Johnson, 2007).

How easy were the methods to use?

Apart from some technical problems with the RO task (due to the large pack size), the participants found the software for all methods very straightforward, simple and easy to use. It is helpful to confirm that the software was not a cause of any frustration or discontent with the methods for the most part. Regarding ease of use of each method compared with traditional marking, most respondents reported that the new methods were a little easier or much easier to use, as shown in Figure 3. PCJ in particular was reported as the easiest to use. This was expected as the task appears simpler for participants than applying a complex marking scheme (Benton & Gallacher, 2018).

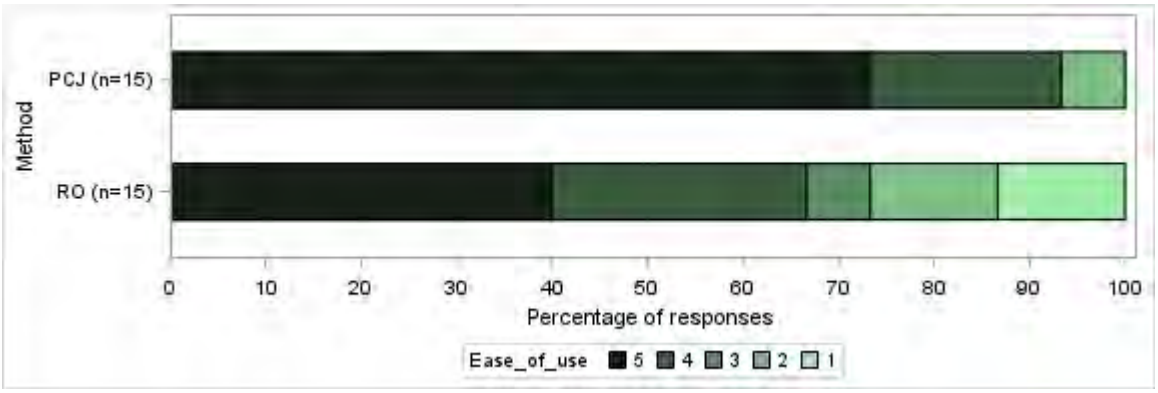

Figure 3: Participants' responses about the ease of use of each of the methods in comparison with analytical marking on a scale from 1 to 5. Darker shading represents more positive responses (easier to use). 5 was "much easier to use", 4 was "a little easier to use", 3 was "much the same", 2 was "a little harder to use" and 1 was "much harder to use".

Regarding cognitive demand, the data shows that many participants were either unsure or found the new methods less cognitively demanding (Figure 4). For about 50 per cent of participants, PCJ and RO were less demanding than traditional marking. About 20 per cent of participants, however, found RO much more cognitively demanding.

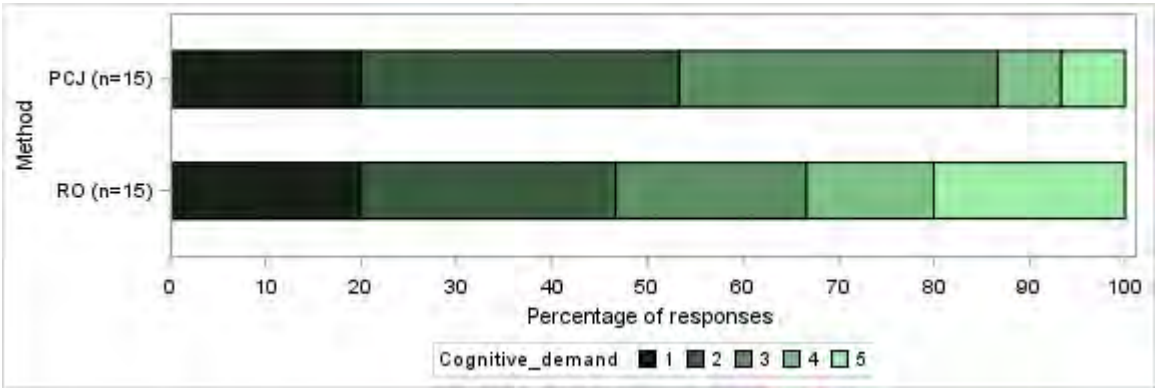

Figure 4: Participants' responses about the cognitive demand of each of the methods on a scale from 1 to 5. Darker shading represents more positive responses (less cognitively demanding). 5 was "much more cognitively demanding", 4 was "a little more cognitively demanding", 3 was "much the same", 2 was "a little less cognitively demanding" and 1 was "much less cognitively demanding".

Participants noted that the cognitive demand increased when the essays were similar in standard. Some also found it more cognitively demanding in general because they had to consider two essays at once, in terms of the marking criteria and both assessment objectives together. For example, Participant 1 said, "This process necessitates holding many different aspects of two responses in your head at once and is therefore more mentally tiring".

For RO, some participants found it more difficult for the following reasons:

- they had no marking tool for support with difficult decisions
- they found ranking 10 essays at once to be challenging
- they had difficulties with the software (due to the large pack size)
- they had to hold a lot of information in their heads at once
- they had to use many different skills at once
- some had to re-read essays several times
- there were no annotations to guide them and keep them on track.

For example, Participant 4 said:

> Having to judge ten scripts in one go was intense, there is a lot of information to process at once … Initially I would be quite alert to the differences, but as it progressed to the seventh script and beyond my mind started to lose track a little bit of where I would be putting the script.

Some of these concerns could be minimised by reducing the pack sizes and/or making the software more user-friendly for larger pack sizes. Previous research by Black (2008) suggested that using three scripts per pack, "Thurstone Triples", might still be cognitively meaningful (but less cognitively demanding) as well as more efficient than pairs. Further research on this would be worthwhile. However, in theory, the larger the pack sizes, the greater the value of the information conferred about each essay from the rankings in each pack and, as such, the larger the gain in efficiency.

Another area for further research concerns which RO strategies are the easiest to use and the most efficient, a point also made by Bramley (2007). In the current study, participants reported trying various different strategies to achieve the final rank order. For example, some read through all essays first and then ranked them, while others read and ranked them one by one. Some skim-read all of them to look for obviously good or poor ones to use as benchmarks. Others used marking criteria to assign a mark to each essay before placing them in order.

How enjoyable were the methods?

Marking a live series takes place in a pressurised and somewhat stressful environment, and participants' enjoyment of the methods is an important consideration from an examiner retention perspective. Previous literature has suggested that holistic methods could be more enjoyable for some examiners (Brooks, 2004). Similarly, I found that some participants enjoyed a more holistic approach, while a few enjoyed the detail that analytical marking brings. Their enjoyment may also have been influenced by how easy or difficult they found the methods to be, as discussed in the previous theme.

As shown in Figure 5, PCJ was the most enjoyable compared with traditional marking. The data for RO was mixed, although more respondents gave the lowest two ratings. A limitation of these findings is that enjoyment may have been inflated by the relative lack of pressure in the experimental (rather than live) marking setting, and the novelty of the methods. On the other hand, a new method that examiners have less experience with could negatively affect their enjoyment. It should also be remembered that this data is from a fairly small sample of examiners.

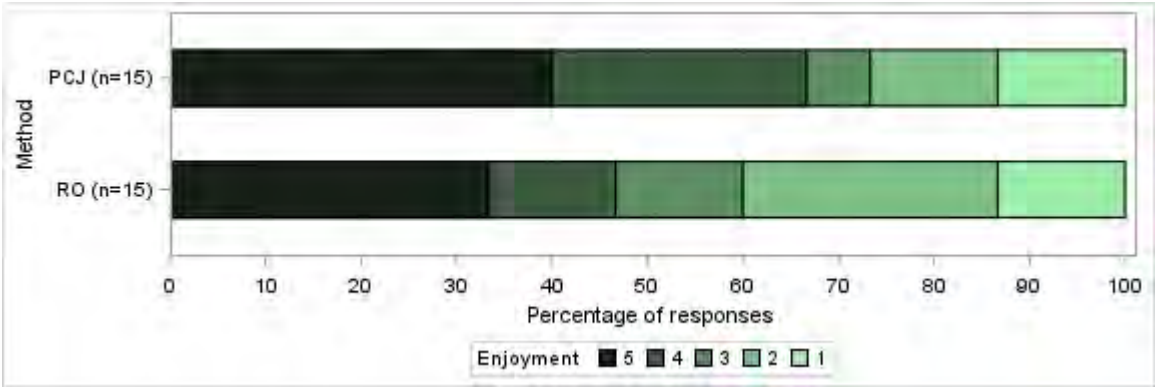

Figure 5: Participants' responses about their enjoyment of each of the methods in comparison with analytical marking on a scale from 1 to 5. Darker shading represents more positive responses (more enjoyment). 5 was "much more enjoyable", 4 was "a little more enjoyable", 3 was "much the same", 2 was "a little less enjoyable" and 1 was "much less enjoyable".

Various factors appeared to have influenced their enjoyment of the methods, some of which were mentioned in previous themes. For PCJ, participants reported enjoying it because it was easier and less time-consuming. They also noted that not being tied to a mark scheme enabled them to enjoy the students' work more. Some found it less stressful due to the lower cognitive demand. Similarly for RO,

participants enjoyed reading students' work without a rigid mark scheme, and for some, marking in packs allowed them to see the variety of responses more clearly. For example, Participant 1 said, "Part of the enjoyment comes from the variety of responses on a single topic, which becomes more acute when assessing a number of responses together".

In contrast, there were also factors that made the methods less enjoyable. For RO and PCJ, some participants did not enjoy that they were less able to reward students individually. Some also noted the tedium and boredom of the tasks due to their simplicity and the repetition of essays. For example, Participant 9 said:

> The fact that you were constantly being presented with the same responses, albeit in different pairings, also took away some of the enjoyment, particularly towards the end of the [PCJ] exercise. It then felt like a real treat to read an essay that I hadn't seen before.

This disadvantage of comparative approaches was also noted by Bramley (2007) and Holmes et al. (2017). Overall, it is encouraging that participants generally enjoyed the methods.

The data about ease of use, cognitive demand and enjoyment can be used to compare individual participants' views of PCJ and RO, by inferring from their comparisons with analytical marking. This data adds extra insights to the previous analyses. Table 2 shows that a majority of participants found PCJ less cognitively demanding than RO. For ease of use, participants either found the two methods to be similarly easier to use, or found PCJ easier than RO. For enjoyment, a majority found PCJ more enjoyable than RO but there were three participants who found them equally less enjoyable. Overall, the perceptions of PCJ appear to be more positive than RO in these three areas.

Table 2: Participants' views about the ease of use, cognitive demand and enjoyment of RO and PCJ, inferred from their comparisons with analytical marking.

	Number who were more positive about PCJ	Number who were equally positive about PCJ and RO	Number who were more positive about RO	Number who were neutral about both PCJ and RO	Number who were equally negative about PCJ and RO
Ease of use	7	6	1	0	1
Cognitive demand	7	4	3	1	0
Enjoyment	6	3	3	0	3

Novice examiners

Previous research has suggested that holistic methods may work better with experienced examiners with similar training backgrounds, as they share a common

view of what a good essay entails (Meadows & Billington, 2005). In the current study, there were mixed views about how the methods would work for new or less experienced examiners.

Some participants felt that they would work well as they are less complex, and most examiners would have teaching experience and knowledge of what makes good writing to draw upon. It was also noted that new examiners may not have the "baggage" of the existing system and may have a more flexible attitude towards adopting novel methods. Some participants noted that the collective element of marking would put less pressure on new examiners and the simpler methods could attract and retain markers. Regarding RO, some felt that exposure to more essays at once would be useful for new examiners to see the range of standards. For example:

> The fact that other examiners would be marking the same scripts allows for collective responsibility and puts less pressure on new examiners as they know that their decisions will not determine a whole selection of scripts (Participant 4).

On the other hand, some participants reported struggling with the methods even though they had years of examining experience. This was a particular concern for RO, due to the number of essays to assess at once, and they felt it could be overwhelming for new examiners. For example:

> I am an experienced examiner, so I think a new examiner might find it quite daunting comparing scripts. He or she would need clear guidance and criteria about what makes one script better than another script (Participant 4).

> This is the paper I've marked for longer than any other, and I was definitely drawing on my experience … and without that experience, I'm not sure how I'd have coped … it would have been more of a guessing game, which is not what you want (Participant 8).

Thus, the findings indicate that while PCJ may be an attractive option for new examiners, RO with 10 essays per pack may be quite challenging. While some of the participants in this study were fairly new to examining, they all had at least three years' experience. Including brand-new examiners in future research would provide us with additional insights.

Conclusion

In this paper, I explored the perceptions and experiences of examiners using PCJ and RO for GCSE English Language essays. The findings help to both broaden and deepen our understanding of how PCJ and RO are perceived as alternatives to analytical marking. It is important that any methods used for marking in high-stakes settings are reliable, valid and fair but are also well received by the assessment community.

The participants in the study expressed a range of, often divergent, views about

their experiences with PCJ and RO. This indicates that, were any of these methods to be introduced as alternatives to marking, there is likely to be a wide range of responses by stakeholders. Overall, there was some positivity about RO and PCJ but also some hesitation and concerns.

The main benefit of multiple marking, as in PCJ and RO, is that the final score captures a consensus among professional examiners (Brooks, 2004; Holmes et al., 2017). Other positives of the methods include the simpler nature of the marking criteria, the potential to improve the consistency of marking, the ease of use of the methods and software (for the most part), and the enjoyment of comparing essays with one another. However, these views were not unanimously shared and if the methods were to be introduced in live marking, examiners would need supportive training and reassurance with data that the methods produce fair, valid and reliable results. For example, one drawback (mentioned by one of the participants) is the potential lack of individual accountability for CJ decisions. Although quick and careless work can be monitored to some extent by analysis of judgement time and fit statistics (e.g., Benton et al., 2020, p. 22.), providing a transparent audit trail that can be used to understand how judges made their decisions is much more difficult than with analytical marking.

Participants expressed both positive and negative views about annotation, and the concerns raised are important to consider were the methods to be implemented as alternatives to analytical marking. While some found them beneficial for marking and teaching, others felt them to be an unnecessary hindrance. Further research and reflection is needed to inform an approach to annotations and summative comments for PCJ and RO methods going forward. In settings where written feedback is needed, PCJ and RO could be more challenging to implement (Jones et al., 2015).

Finally, any change to practices which examiners have been following for many years are likely to take time to adjust to and become comfortable with. However, the factors raised in this research can help advise tweaks to the methods, as well as informing a training, communication and support strategy if the methods were to be implemented.

Limitations

The main limitation of this study is the potential lack of ecological validity. We cannot be sure what influence the experimental setting had on their experiences and views. However, the examiners were instructed to mark as they would in a live series, and they were paid for their participation. The quality of results and interview responses suggest that they completed the tasks seriously and conscientiously.

Another limitation is that the findings are based on self-report data. Observational studies can complement the findings, especially when looking at aspects like how judgements were made. Expert opinion was used to give an indication of the potential and perceived impact on other stakeholders, but ideally consulting other stakeholders directly would be useful in evaluating the methods.

In terms of the generalisability of the findings, another limitation is that it is not known the extent to which the examiners' views and experiences are linked to GCSE English Language essays. Since many essays are marked in similar ways with analytical mark schemes it seems likely the findings would be applicable to other subjects that use essays as assessment tools, however, further research would be useful to compare and contrast views in different contexts.

References

Aloisi, C. (2020, October 12-13). *'Explainability' of machine learning algorithms and implications for reviews of marking and appeals.* Cambridge Assessment Education Assessment Research Seminar, Online.

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

Black, B. (2008). *Using an adapted rank-ordering method to investigate January versus June awarding standards.* Fourth Biennial EARLI/Northumbria Assessment Conference, Berlin, Germany. https://www.cambridgeassessment.org.uk/Images/109767-using-an-adapted-rank-ordering-method-to-investigate-january-versus-june-awarding-standards.pdf

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* (Vol. 246, p. 294). 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

Bramley, T., & 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. https://doi.org/10.1080/0969594X.2017.1418734

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

British Educational Research Association (BERA). (2018). *Ethical Guidelines for Educational Research* (4th ed.) https://www.bera.ac.uk/researchers-resources/publications/ethical-guidelines-for-educational-research-2018

Brooks, V. (2004). Double marking revisited. *British Journal of Educational Studies, 52*(1), 29–46. https://doi.org/10.1111/j.1467-8527.2004.00253.x

Crisp, V., & Johnson, M. (2007). The use of annotations in examination marking: opening a window into markers' minds. *British Educational Research Journal, 33*(6), 943–961. https://doi.org/10.1080/01411920701657066

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom*, (pp. 69–87).

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher, 37*(2), 1–19. https://doi.org/10.1007/BF03216919

Holmes, S., Black, B., & Morin, C. (2017). *Marking reliability studies 2017: Rank*

ordering versus marking – which is more reliable? https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/859250/Marking_reliability_-_FINAL64494.pdf

Johnson, M., & Nadas, R. (2009). Marginalised behaviour: digital annotations, spatial encoding and the implications for reading comprehension. *Learning, media and technology,* 34(4), 323–336. https://doi.org/10.1080/17439880903338606

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educational Studies in Mathematics, 89*(3), 337–355. https://doi.org/10.1007/s10649-015-9607-1

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education, 13*(1), 151–177. https://doi.org/10.1007/s10763-013-9497-6

Kimbell, R., Wheeler, T., Stables, K., Sheppard, T., Martin, F., Davies, D., Pollitt, A., & Whitehouse, G. (2009). *e-scape portfolio assessment: phase 3 report.* Technology Education Research Unit, Goldsmiths, UL. https://www.teachertoolkit.co.uk/wp-content/uploads/2014/08/e-scape_phase3_report.pdf

Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability.* https://filestore.aqa.org.uk/content/research/CERP_RP_MM_01052005.pdf

Nadas, R., Suto, I., & Grayson, R. (2021). Analyse, evaluate, review, synthesise, and argue: Why teacher-assessors' interpretations of command words matter. *Educational Research, 63*(3), 357–377. https://doi.org/10.1080/00131881.2021.1956987

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

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. https://doi.org/10.1080/0969594X.2019.1602027

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. https://doi.org/10.1080/0969594X.2019.1700212

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

Appendix

Excerpts from instructions to judges about how to make their judgements.

Pairwise Comparative Judgement

- You will be presented with a pair of essays side by side (100 pairs in total).
- The question you are answering is: **Which essay demonstrates better performance on the constructs being assessed?**
- To record your decision, click the 'Choose' button above the essay you believe wins the comparison. You cannot edit your decision once you have pressed the 'Choose' button.

Rank ordering

In this approach, you will be presented with packs of 10 essays ... and your task is to put them in order from best to worst. What constitutes better or worse performance should be guided by the constructs being assessed (as described in the Assessment Objectives).

Guidance for ranking the scripts:

- Your judgements should be holistic and intuitive. **Do not re-mark** the essays to come to a decision. Read each essay, think about which ones are better or worse and put them in order.
- Gut reaction/instinct is fine – you do **not** need to provide any explanation or justification for your decisions. The fact that, in your opinion, essay A is better than essay B, which is better than essay C etc. is enough.
- Try not to dwell on your decisions for too long. Previous exercises suggest that the packs may take approximately 40 minutes on average. Some may be quicker and some may take more time.
- You may not need to read all essays as thoroughly as you usually would. It may be clear that some are better than the others even from a quick skim-read.
- No tied ranks are allowed. Even if you feel that some of the scripts are very similar or the same in their performance, you will need to put them in order.
- There is not a right answer! The 'right' answer is the one you determine by making a holistic judgement of each script's quality.
- If the script is faint and difficult to read, please make the best decision you can and let me know about the issue.
- How you rank a candidate who has, in your view, done well on some parts and poorly on others against another candidate who demonstrates a consistent performance is up to you – the crucial thing is you make a holistic determination of the quality of the essay.