How do judges in Comparative Judgement exercises make their judgements?

Tony Leech and Lucy Chambers (Research Division)

Introduction

Comparative judgement (CJ) in the context of assessment is a method in which judges compare a series of two or more candidate scripts directly, to rank them in order of quality. The judgements are intended to be holistic and quick, relying on a judge's internalised sense of what constitutes better performance in their subject. CJ takes account of the psychological fact that it is often considered easier (see for instance Pollitt & Crisp, 2004) to make relative decisions (comparing things to each other) than absolute decisions (comparing things to targets or standards).

There are two main applications of CJ in assessment (Bramley & Oates, 2011). The first is as an alternative to marking. All the judgements of the judges are combined in a statistical model to create a single numerical value for each script representing its perceived quality. The second application is for maintaining standards (the process whereby grade boundaries in an exam are decided such that it is no easier or more difficult for a candidate to get a grade in the current year as in previous years). Here the idea is to use CJ to compare samples of scripts from two different exams that have been marked in the usual way. The mark scales of the two exams can then either be linked via the measures of perceived quality (e.g., Bramley, 2005), or the difference in difficulty between the exams (in marks) can be estimated directly via logistic regression using the "simplified pairs/ranks" method of Benton (2021).

Expert judgement has a role in current (non-CJ) procedures for setting grade boundaries on GCSEs and A levels. It involves comparing exam scripts from the current year to a previous benchmark year. Firstly, statistical analysis of cohort prior attainment data is used to identify suggested grade boundaries on the current test. Secondly, in a judgemental element, candidate responses from the current year around the statistically recommended grade boundaries are compared to those around the same grade boundary in the previous year, and judges are instructed to determine if those of the current year demonstrate the same grade-worthiness (and therefore whether they can endorse the recommended boundary as representing the same standard of performance as previously). Thus, this judgemental element is secondary to statistical methods.

This process has been criticised for using a small number of judgements and relying on judges being able to recognise a candidate script as, for example, embodying the characteristics of "A-grade-ness". For more on current approaches, see Curcin et al. (2019, p. 17).

Standard maintaining using CJ involves judges having to compare packs of two or more candidate scripts, with each pack containing scripts from both the current year and a benchmark year, to decide which candidate responses are better (Bramley, 2007). The judgement is made on the basis of a prompt question e.g., "which script exhibits the best overall performance?" In packs involving pairs of scripts, judges will choose the superior script. In larger packs, e.g., of four or six scripts, judges rank the scripts in order from best to worst. Each judge will see multiple packs, and each script will be seen by multiple judges. A large number of scripts from across the mark range are used in a CJ exercise, unlike the handful of scripts, all around key grade boundaries, which are used in the judgemental element of current standard-maintaining processes. The outcomes of comparisons are processed using statistical models so a more precise determination of the difference in difficulty between the two years' papers can be identified, which could lead to grade boundaries which are more likely to represent this difference in difficulty than when set with current approaches. For more specific details of the methods, see Benton, Cunningham et al. (2020). Using CJ in standard maintaining is presumably harder for judges than using it as an alternative to marking, as they must take into account potential differences in difficulty between the questions set in different years in their judgements.

Two issues for CJ in relation to standard maintaining which are highly relevant are "what processes do judges use to make their decisions?" and "what features do they focus on when making their decisions?". These issues were discussed briefly by Curcin et al. (2019, pp. 87–93) where the authors found that judges in their pilot CJ exercises mainly judged scripts question by question, gave questions with more marks a higher weighting in their overall judgement, used missing responses as a differentiator of quality and based their judgements on mark scheme requirements. However, no judges explicitly suggested that they were re-marking scripts. Subject-specific features of candidate responses were important to judges, while, pleasingly, superficial features were seldom mentioned.

This article extends discussion of these issues by reference to outcomes of a series of OCR/Cambridge Assessment studies exploring the use of CJ for maintaining standards, conducted using in-house CJ software. Our contribution is to focus explicitly on what CJ judges are doing when judging, and what they are attending to in their judgements. We hope thereby to render more explicit some of the assumptions underlying both comparative judgement, and standard maintaining, both in its CJ and current forms. We explored whether judgements were holistic, whether judges were able to take into account differences in difficulty between papers from different years, and what parts of papers or types of questions were attended to the most. This focus is important so we can better understand the validity implications of CJ and their impacts on decisions made using the judgements.

What processes do judges use to make their decisions?

The evidence to answer this question came from a CJ study using a GCSE Physical Education (PE) component. In the task, judges were asked to rank packs of four scripts in order from best to worst overall performance. Each pack contained two scripts from the 2018 assessment and two scripts from the 2019 assessment; judges were provided with the question paper and mark scheme for each paper in order to re-familiarise themselves with the papers used in the study. The four scripts appeared in a random order and were labelled A–D. The paper was out of 60 and candidates wrote their responses in a structured answer booklet which also contained the questions. There were a mixture of short-answer and mid-length questions.

Within the CJ software, judges were presented with packs of four scripts and instructed to "rank these in order from best to worst overall performance". Figure 1 shows the judge view of the tool; judges could view each script by clicking on the buttons A–D on the left-hand side. Once the judges were ready to rank a script, they could drag it over to one of boxes 1–4 on the right-hand side, the position they chose indicating their view of its quality, with box 1 indicating the best script and box 4 the worst. Scripts could still be viewed from within these boxes. If the judges changed their mind, they could reorder the scripts by dragging the letter to a different rank position. When judges were satisfied with their rankings they clicked Submit and would be automatically presented with the next pack of four scripts.

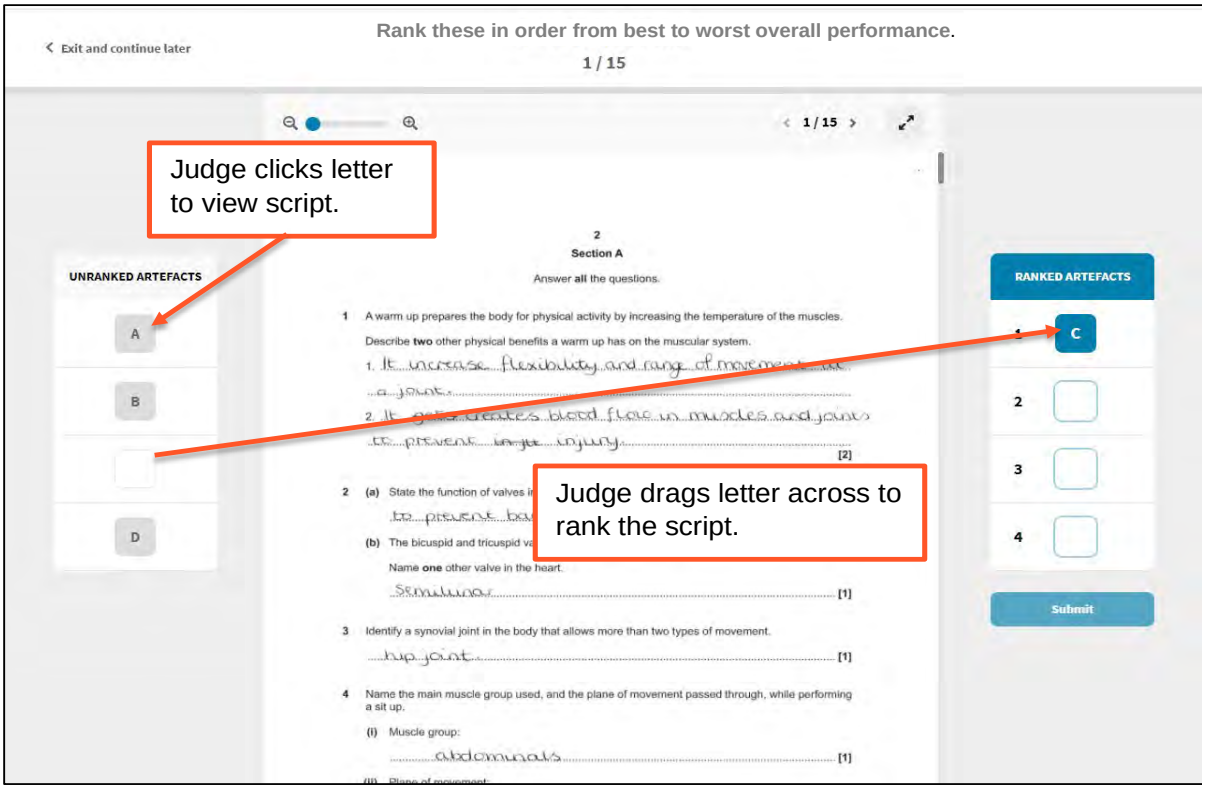

Figure 1: Annotated screenshot of judge's view of CJ software.

This study (for more on its method see Chambers and Cunningham, 2022) included an observational element. Ten judges were observed via online meeting software for 30–40 minutes while engaged in the CJ standard-maintaining task. During this observation, judges were asked to "think aloud" while they were judging, allowing the researchers to gain an understanding of their approach to the task and their decision-making process. This section details the behaviours drawn from the observations concerning the overarching CJ method employed by the judges i.e., how they approached the task. All quotes from the observations are written verbatim.

The 10 judges differed in how they approached the task and the key features evident in their behaviour are recorded in Table 1. Since the observation was a "snapshot" of their judging, presence or absence (rather than a count) of each feature was recorded. It is possible that the behaviour exhibited during the observation did not reflect the rest of the judging, however given the candid comments made by the judges, the authors believe it is unlikely to have been fundamentally different.

Table 1: Judge behaviour as witnessed in the observation.

Judge	Looked at 2018 and 2019 scripts as two groups	Dragged scripts to rank position as went along	Evaluated each question	Looked at mark scheme multiple times	Re-marked (Tallied up marks)	Made comparative references to other scripts	Returned to previous scripts
1	✓	✓	✓	✓	✓	✓	
2	✓	✓	✓	✓		✓	
3			✓		✓	✓	
4	✓	✓	✓			✓	✓
5	✓		✓	✓		✓	✓
6		✓	✓	✓		✓	✓
7			✓	✓	✓		
8	✓	✓	✓	✓		✓	✓
9	✓	✓	✓	✓		✓	✓
10						✓	✓

The judges developed a preferred method of viewing the scripts within a pack. Six of the judges chose to look at both scripts from one year before moving on to the other year's scripts. Interestingly, judge 4, who was observed at the start of their judging, did their first pack in the order presented by the CJ software but by the second pack they judged the two years separately. Ease of comparison and the use of the mark scheme may have exerted some influence over the judges' preference for judging by year:

> Now this is where we get in difficulty, because this now goes into the next question paper. They don't actually follow on. And so, I'm actually going to go back and look in C instead rather than jump around. And it's not C either so I'm going to open B or D.
>
> What I've actually been doing and to start off with it. Uh, there are two papers and completely different ones. One's [20]18 and one's [20]19 … what I've been doing is, I've been opening up the scripts or the candidates' responses and I've been checking to see which two pair up and so when I, when I, basically open up the mark scheme, see, it's easier to cross reference rather than having to change the mark scheme all the time.

Three judges selected the scripts in the order presented by the CJ software. One judge (10) picked a different starting script in each pack; this was because "otherwise I find you end up with all the A's being number ones" [ranked top].

While making their judgements, six of the judges dragged each script across to a rank position as they finished looking at each script. The first script was generally put in position two or three and the positions reordered as further scripts were attended to e.g., "Will put that in at number 2 for now", "And so I'm now going to look at, and I'm going to assume it's fairly high, so I'm going to move A across into sort of second position to start with and we'll see how we go." If the script was particularly weak or strong then some judges would move it straight into the top or bottom positions, e.g:

> For my starting process I sort of put which ever I start with either two or three generally unless it's a boss, boss work and you might stick it up, you think that's going to be the best or, or absolutely the worst. I'll put it at two or three.
>
> So, I like that. I like that is a, is a, is a good start and so I'm going to put that up, up at the top at the minute. It's a strong paper. I would you know. I would categorise that in there in the top, top third for sure.

The remaining judges moved the scripts into the rank position once they had looked at all the scripts, e.g:

> Yeah so, do you know what. I'm gonna pop C down there, B down there, pop A down there and I'm going to stick D down there. So yeah, I think that is the right order.

When viewing a pack, one judge (10) skimmed through the scripts, dipping into certain questions to evaluate them more fully. The remaining nine judges evaluated every, or nearly every, question of each script in turn. Where scripts differed significantly in quality, one would expect that such a full evaluation would not be necessary – it is possible that with a pack of four (as opposed to pairs) the quality of multiple other scripts is unknown and so the judges felt more comfortable evaluating each script more fully. The presence of the observer may have also caused them to be more thorough.

Hand in hand with the evaluation of each question was the frequent use of the mark scheme. The judges were given the mark scheme for familiarisation with the explicit instructions "Please do not use the mark scheme to re-mark the scripts; the mark scheme is available only so you can be clear about the constructs being assessed". Despite this, seven of the judges actively referred to the mark scheme while going through the scripts. Two of the three that did not were clearly very familiar with the mark scheme and possibly did not need to refer to it. Example comments:

> I'm just going through by question, by question. I'm looking at the handwriting, but I'm just going through in terms of the, the knowledge really and comparing it with the, with the mark scheme.
>
> Good, good aortic valve and is, is accepted I believe just let me just double check that with the mark scheme.
>
> Let's look at the mark scheme very quickly.

Three judges were observed to be fully re-marking the scripts and totting up the total marks the candidate would have received. Comments made during the observation included:

> I've got a pen and paper as well because I use that quite a lot for just making notes where they've actually got marks. So, it is a bit like actually marking it. When I know it says don't mark it, but...
>
> Well, I mean, I'll be honest. I mean, I did sort of tally up what I thought was

> worth a mark to compare them in on the certainly the first 10 I did. And if that messes up, at least you know, then you can use that information.
>
> Just two marks for that. Scroll down, 5, 10, 15, 20, 25, 30, 34. So it's B first...

In a CJ exercise, one would expect the judges to make comparative references to the other scripts in the pack. All judges, except one, did so. This judge (7), treated the exercise purely as a re-marking one, totting up marks and making the final judgement purely on marks attained. Examples of comparative comments:

> I don't think it's as good as the first one.
>
> That's good, it's nowhere near as bad as the last one.
>
> C is definitely the worst.
>
> The other one was much better in comparison to this one – that, particularly in the early questions.
>
> I would not put this in the same category as the other student. I would put this lower so that one would go at the top.
>
> True, so this this kid's already better than the previous one. So, in terms of ranking, that would, that D will be better than A.
>
> There's some good examples across the top three, I think. Obviously, A is possibly slightly better in terms of overall holistic, but it's very close.

Related to this, six of the judges revisited previous scripts when deciding on their final rank order. This was generally to confirm their choice or help decide between two scripts.

> Just wanna check it against B though cos even though...
>
> Just let me check on the bottom two. I'm happy with D and A. When I look at the first page just to make my overall judgement, we've got...
>
> At the moment I think I've got my first and last. Second and third very difficult. I look at the six marker...

In summary, the judges varied in their approach to the task: three judges re-marked the scripts, one marked purely holistically and the others used a mixture of both approaches. Many judges relied heavily on the mark scheme; it could be that the nature of the paper (many short answers) encouraged this. What is reassuring is that most of the judges were actively comparing the scripts against each other, which is the purpose of comparative judgement. This suggests that the issue of concern is in making holistic judgements rather than the comparative nature of the exercise.

The level of re-marking and in-depth evaluation of each question suggests that judgements were only partially, if at all, holistic. Moreover, judges made frequent reference to the mark scheme. In other words, the judges appeared to be engaging in activity that had similarities to marking. However, given that these

studies were about CJ in the standard-maintaining context, we had intended them to be engaging in processes more like those undertaken in the judgemental element of current awarding procedures – i.e., making holistic, whole-script assessments of quality. That the judges were not judging in the way we had intended them to has implications for the validity of CJ outcomes. In the PE study we observed the judges so we know how they approached their judgements but typically we would not.

It should be noted that judges in the PE study and the other studies discussed in this article, were all experienced markers of the papers they were judging. Many, but by no means all, were also involved in the judgemental element of the current standard-maintaining process. While these two tasks are conceptually dissimilar, they are often undertaken by the same people (often, the most experienced markers are selected as standard-maintaining judges), and so the same approach was taken for the CJ method. However, this raises a potential problem, which exists implicitly in current standard-maintaining processes but which we have highlighted explicitly here. To solve this problem, if examiners without experience of the judgemental element of current standard-maintaining processes are used in CJ exercises, they will need to set aside their marking experience and apply a new, more holistic technique. (This is also true for examiners who take part in the judgemental element of current standard maintaining – who must apply different techniques at different times – so this issue is not unique to CJ). What is apparent from this study is that support is needed. We recommend that judges have training on making holistic CJ decisions involving practice, feedback and discussion.

What features do judges focus on when making their decisions?

In this section of the article we broaden out the question of what judges attended to by exploring their answers to survey questions. Online surveys were all administered on completion of the studies to which they related. Each survey took around 10–15 minutes for the judges to complete. Most of the surveys related to multiple, parallel CJ exercises (sometimes judges took part in two exercises as part of the same study). The surveys covered various subjects and levels (GCSE English Language, AS level Geography and Sociology, A level English Literature and Psychology, Cambridge Technical in Digital Media, and Cambridge Nationals in Child Development, Enterprise & Marketing, and Information Technologies) across different CJ approaches[1] (see Benton et al., 2022, this issue, for details). Results from the PE study discussed above and in Chambers and Cunningham (2022) are also included where appropriate. In total, 108 judges took part in the surveys. The surveys were subject-specific and covered more ground than is discussed here, e.g., they included issues relating to the specific setup of the particular studies, the time taken to judge etc. However, for the topics discussed here, the questions were similar enough across the surveys to allow us to make useful comparisons.

1 Some of the studies were pilot studies and some were part of live operational standard-maintaining activities, in which grade boundaries were set.

From the responses, we developed a model of the different dimensions which underpin a judge's decision-making, as shown in Figure 2. We see that a judge's CJ decision-making is related to: a) their individual approach; b) the way that the question paper is constructed, such as how many short-answer questions it has, etc.; c) the way that the candidates have answered items; and d) the unique, comparative requirements of the CJ task.

The thick arrow between the judge and CJ task reflects the fact that all the judgemental work here was carried out within the context of a comparative task; the arrow is two-way to reflect the fact that the judges nonetheless interpreted CJ requirements slightly differently. The solid arrows indicate elements that invariably impact one another, while the dotted arrows highlight that, though the main influence of question paper and candidate factors comes through the task, factors like the structure of the paper or whether context-irrelevant features were judged are not unique to CJ.

These different elements interplay with one another – for example, candidate responses are naturally conditioned by the requirements of the question paper, while the fact that a CJ task is different from normal marking tasks (which embody only the three outer elements) highlights the importance of the two-way judge–task relationship here. In other words, how do judges individually interpret the requirement to make comparative, holistic judgements? In what follows, we have used judge survey responses to highlight these four broad areas. Factors relating to each dimension are summarised in Table 2, and we explore each in turn.

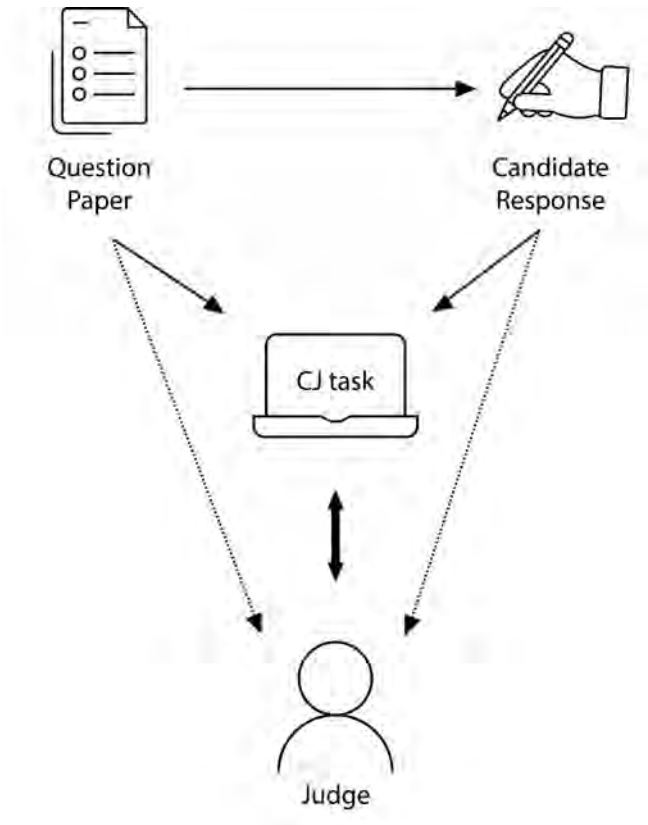

Figure 2: Dimensions of judge decision-making.

Table 2: Dimensions of decision-making and relevant factors.

Judge-centred dimension	Question paper features dimension	Candidate response features dimension	CJ task dimension
• Ability to make holistic judgements • Confidence • Understanding the process and where their judgements fit	• Structure of the paper e.g., short answer versus longer response • Existence or otherwise of key discriminator questions	• Missing responses • Spiky profiles • Supporting examples and evidence • Clarity/structure • Construct-irrelevant features e.g., handwriting	• Balance of different response elements • Balance of answers from different years • Closeness in script quality within a pack

Judge-centred dimension

One of the major judge-centred dimensions of decision-making relates to whether they found it straightforward to make a holistic judgement. It can be assumed that a judgement would be less straightforward if it required the bringing together of complex material in unsystematic ways. Of course, this dimension is not independent of the requirements of the paper or task, or candidate responses, as we discuss later. Across the various surveys discussed here, including the PE study, judges generally responded that it was at least "somewhat" straightforward to make their judgements – with many describing the process as "entirely" straightforward. Whether "entirely" or "somewhat" was the modal value differed across the surveys, but there did not appear to be any consistent pattern in this. One PE judge noted that, while it took a while to get into the process, "once a few scripts were marked it was pretty straightforward".

However, judges who took part in CJ judging used to inform OCR's live grade boundary setting in autumn 2020 generally found the process more challenging than those who judged in pilot studies. 9 out of 17 judges in the live context described the task as at least "somewhat" straightforward – with the other eight either neutral or critical. These more critical judges highlighted various task and candidate response factors as making their judgements more challenging, as discussed below.

Though not mentioned by judges, it is possible that the fact that the judgements informed real grade boundary setting led to judges believing they needed to do the best possible job on every judgement in order not to do a disservice to candidates. It is worth noting, however, that the fact that judgements were not necessarily experienced as straightforward by all judges does not mean they were not providing useful information. Perhaps judges did not appreciate the fact that their individual judgements alone did not decide students' results, but rather that they were statistically combined with other judges' judgements. Ensuring that judges understand the context of their judgements is therefore important.

Overall, answers reveal that while most judges found the task they were being asked to do straightforward enough, inhibitors include the context of the task and the challenge of weighing up papers where the candidates had answered differently well on different parts. This highlights the interplay of the different dimensions. For example:

> This was quite difficult and time consuming, ultimately it did slow down the process because you don't want to disadvantage the students and so the mark scheme has to be applied accurately judging the subject knowledge of the content for that individual.

While overall, the surveys suggested that CJ is straightforward for most judges, in many surveys there was at least one judge who just found the process challenging – perhaps because it was very different from processes like marking. It does not appear that there are obvious characteristics distinguishing these individuals from others, so perhaps this is purely a case of individual preference.

Question paper features dimension

Features relating to the design of the question paper are also central to decision-making. A selection of comments from the PE study are illustrative of the range of judge views of many of these issues across the studies:

> So many questions on the paper, mostly very short answers. Difficult to avoid totting up correct/incorrect answers.

> It was difficult to not 're-mark' as a lot of 1-mark questions and also ignore the fact that I knew one paper was slightly harder than the other so had slightly lower grade boundaries.

> Because the scripts were from 2 different exams, I felt that the best way of comparing them was by the number of questions they got correct. However, it wasn't a comfortable decision as the 2 exam papers may not have been of the same difficulty.

An open question about how judges made their judgements was asked in many surveys. In some cases this was asked explicitly in relation to script features they were looking for, but not all. Answers varied substantially across the surveys, though there was no obvious pattern by subject or CJ pack size. Instead, different judges within the same survey seemed to have looked at different things, revealing the interplay of these different dimensions. For example, in the digital media exercise, two judges described how they used the mark scheme, three used "key discriminator questions" (one judge defining these specifically as those worth the most marks) and two counted up the marks. Some judges used more than one of these techniques. Half of the PE judges wrote "Number of correct answers" or equivalent as the first part of their response. For two respondents this was their complete response. This reflects the observations where re-marking was evident.

Question paper structure impacts decision-making; re-marking was at least somewhat more common in papers with a greater number of short-answer items. There is a relationship between task type and response, with judges tending to

agree more with the idea that they focused on certain question types (implicitly because they viewed other question types as weaker discriminators) if the paper either contained more structured questions or came from a vocational qualification (or both). Possible explanations for this could include that in these papers it is harder to discriminate between candidates on shorter-answer questions, or that it is harder to avoid just re-marking them and totting up the scores. It could also be that higher-tariff, more extended responses are designed to test higher-order skills, and therefore that this question type was appropriately more likely to be chosen as a discriminator.

Judges were asked about the extent to which they agreed with a statement that some types of questions were better discriminators of script quality than other types of questions. Judges tended to suggest that there were certain types of questions that mattered more than others. For instance, in the survey relating to the Cambridge National in Information Technologies, five out of eight judges agreed with this statement, with only one disagreeing and two neutral. In Enterprise & Marketing, three judges "entirely" agreed with this statement, and four "somewhat" agreed, with only two neutral and no-one disagreeing. The agreeing judges suggested that "evaluative questions" and "questions that require more depth" are good discriminators, while multiple-choice questions are not.

These views were shared by some judges in other surveys. In the PE study, all but one judge reported that they entirely or somewhat agreed with the statement. The better discriminators in this study were reported to be the longer questions, especially those requiring examples, evaluation, description, or explanation. In particular, the 6-mark question was cited as it "... requires a full response which combines different parts of their learning". These question types were seen as better discriminators as they allow candidates to demonstrate their knowledge and whether they fully understand a topic.

So, did the judges actually focus on certain question types more than others? In many cases the answer seems to be yes. For the Cambridge Technical in Digital Media, three out of five judges agreed that they did focus on certain question types. Those that elaborated noted the importance of essays and long-answer questions to their decisions. The same was roughly true for the Cambridge Nationals exercises as well, with judges highlighting the importance of longer questions and calculations. However, for AS Sociology, 13 of 19 judges said they looked at the whole script, with a minority highlighting the importance of the longer essays at the end of the paper as tiebreakers. AS Geography saw a more neutral response, with equal numbers of judges agreeing and disagreeing. All but two PE respondents agreed that they focused on certain question types more than others when making their judgements. The other two were neutral, with one citing that they focused on "... just the number of right answers and therefore the marks". This reiterates the fact that question paper and candidate responses are nonetheless interpreted differently by different judges.

Candidate response features dimension

The responses of the individual candidates were a major element in the decision-making of judges. For example, in AS level Geography, a considerable number of approaches were highlighted. These include, from one single judge, "Consistency across all questions, clarity and structure of longer answers, use of supporting evidence, understanding of geographical concepts, the ability to evaluate and use of geographical terminology". Other responses complemented this judge's focus on these features, with other judges referring to "depth of geographical explanation", the use of "place examples" and specific geographical terminology and the number of correct short answers as well as quality of longer essays.

In other surveys, elements cited as making it more difficult to perform a holistic judgement included missing answers to questions, the poor expression of some candidates, and where scripts exhibited "spiky profiles" – in other words, where candidates answered some questions well and others poorly. PE respondents reported they focused on a number of other elements, for example, clarity/command of the written language, handwriting, spelling, overall impression of the script, short answer-questions and not repeating the question in the answer. For example:

> Different years meant you had to look for terminology in the shorter questions but not the same amount of shorter questions. Easier to find terminology in shorter questions. Lots of comparison of the 6 mark question as it's on every paper.

In the PE study, part of the focus was on whether judges focused on construct-irrelevant features – that is, features that are not part of the mark scheme. The study found that judgements did not appear to be influenced by spelling, punctuation and grammar or by the visual appearance of the responses (e.g., crossings out, writing outside the designated area and text insertions). Missing responses rather than zero-mark answers and hard-to-read candidate handwriting were shown to have a negative influence on judgements (see Chambers and Cunningham, 2022).

The issue of judges potentially focusing on certain questions and question types brings out an interesting tension between the issue of holistic judgements having to take into account many different skills at once, and judges attending to certain parts of the paper more than others. Benton, Leech, et al., in a 2020 paper on the use of CJ in mathematics standard maintaining, discuss this tension and its implication for validity – though they note that these tensions are not limited to CJ.

> The hypothetical situation where a script which had overall received fewer marks but was judged superior due to the judge preferring its writer's answers to problem-solving questions, for example, raises certain questions about comparative judgement-informed standard maintaining processes. (p. 15)

They go on to report that some might argue the opposite, "that it is a good thing that judges concentrate on certain, better-discriminating, questions, if these can be seen as identifying the characteristics of the superior mathematician more

efficiently" (p. 15). What is of paramount importance is that both the scripts and the benchmark sessions used are representative. If the judges give most weight to a particular question and this question results in unusual performance in one year, then this has implications for the standard. Likewise, if scripts chosen are not representative of others on that mark, particularly with respect to the discriminating questions, then this again has implications (see Bramley, 2010). We need to avoid a situation where the scripts are marked against one set of criteria (a mark scheme) and the grade boundaries are set using a potentially different set of criteria (holistic CJ judgements focusing on unrepresentative questions/features/scripts).

It might be argued that in this case (and similar cases such as English language) that the judges are effectively highlighting the fact that, within the mark schemes for these subjects, a wide variety of skills are required to gain high marks, and thus that a holistic judgement must take into account a high number of different skills all at once. There may be a tension here between the idea of a whole-paper holistic judgement and the concept of "key discriminator" questions or skills, particularly in subjects based on extended-response items.

CJ task dimension

Finally, the comparative nature of the task was a new dimension for judges that they had to take account of. This includes various factors making comparisons challenging, including the fact that papers from two different years will feature candidates answering different questions and the need to make a judgement when papers were very close in quality. For example, as one of the PE respondents relates:

> Most of the scripts were fairly easy to 'pigeonhole' and put in rank order. However, when scripts were very close, it was difficult to make a decision as to which was the best. Also, I found it difficult to compare the scripts from the 2 different exam papers.

Judges in the PE study were asked how straightforward it was to make judgements of packs containing two scripts from each of two different years. Three respondents found this "not very straightforward"; the others were either neutral (1), or found it "somewhat straightforward" (3) or "entirely straightforward" (3). In other surveys, judges were asked about whether they thought papers from the different years were of similar or different levels of demand, and in most cases were able to make a determination. While it was a new experience for judges used to marking to compare scripts from different years in a CJ context, this comparison is required in the judgemental element of current standard-maintaining procedures too.

Moreover, judges in other surveys highlighted that "balancing" different tasks, performed to different degrees of quality by different candidates, was a challenge – particularly in English language papers with reading and writing sections where candidates may have done well on one section but not the other. Other specific issues such as tight time requirements for judging and the "high degree of subjectivity" in judgements were also highlighted.

Respondents were asked whether they were faced with any situations where one of the scripts they were judging was better in one sense, but another script was better in a different sense (and both senses were significant for determining which script was better overall, meaning that situations like this were difficult to judge). GCSE English Language judges frequently saw these cases, with many citing scripts where one student had done better at writing tasks and the other at reading, and others mentioning missing answers to particular questions (i.e. that the student with a missing answer had done better on the questions they did answer than the student who had answered all of them).

Similar issues were evident for judges of many of the other papers, reflecting the fact that different parts of papers may test different, equally important skills. Eight of the respondents in the PE study encountered this situation. Two respondents cited the balance between the number of correct low and higher tariff questions, e.g., "One script had better short answers, while another script had a few better extended answers. I made my judgement by totting up the right answers as well as using a bit of gut instinct". One respondent "... used the MS [mark scheme] to help with ... responses and compared the [highest tariff] question" and another "Reviewed the scripts – gave each a 'grade' e.g., high B vs low B". Across the surveys, some different sets of issues were mentioned, including performance on different sections of papers, different skills (both generic and subject-specific) and different types of questions. While differences in the ability of individual candidates in these areas would of course be evident in normal marking, what is new in the CJ context is the fact that 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.

There is a potential issue here, inasmuch as the idea of a holistic judgement implicitly relies on it being possible to understand the whole paper and have a singular conception of "better performance" which determines which of the scripts is superior. This is not the situation in current exam papers generally, as they are built to be marked, so the superior candidate is the one who receives the highest number of marks – marks which might have been earned on any combination of different items, some answered more and some less well, but where the relative contribution of each performance on each item is identified clearly by the number of marks it is awarded. Without this identification, it is more difficult for judges to determine which skill should be more highly regarded, and certainly for judges to be consistent with each other on this matter. This issue is also present in current judgemental approaches to standard maintaining where (non-comparative) judgements are made of papers around grade boundaries to see if they meet a putative standard of, say, "A-grade-ness". But it is not clear in this context which specific skills or knowledge meet these criteria and which do not, as these standards are mostly general and implicit, and may differ between judges.

Conclusion

We have seen that in some important respects not all the work of CJ judges in the studies described involves a true holistic judgement, which has important validity

implications. On the one hand, it seems that judges are able to compare scripts against each other directly, and that they find this straightforward, which is encouraging. Moreover, while judges use different methods to judge, this does not seem to present a major problem for the CJ outcomes (see Benton et al., 2022, this issue, for details).

However, on the other hand, it seems clear that for many judges, at least in tasks where there were many short-answer items, it was difficult to make a holistic judgement and judges instead essentially re-marked the papers and totted up the marks. This has implications for judges being able to properly take account of differences in difficulty between different papers, an essential element of the rationale for comparative judgement in standard maintaining. Indeed, it could be questioned whether all judges are even trying to take account of differences in assessment difficulty – the very purpose of the whole exercise. Since this is a new technique for most judges, who are experienced instead in marking, more support and training for CJ judges would be necessary to try to ensure that they are able to make holistic judgements and comfortable that they are doing the right thing.

The question raised here of whether CJ judges can make holistic judgements and thereby make decisions comparing two different papers raises the broader issue of how well this is actually possible in current standard-maintaining procedures. Script judgements in current procedures are nominally holistic and based on a whole-paper view. The same challenges of different judges' styles and responses to question paper and candidate level differences are therefore also present as in CJ. In the current procedures, a small number of scripts that are very similar in quality (as they are chosen to be just a couple of marks away from statistically recommended grade boundaries) are judged. CJ gives the judgemental element of the process of standard maintaining more safeguards, including a greater number of scripts to look at, more judges, scripts chosen from across the mark range and a statistical method that leads to it being possible to determine a quantifiable difference in difficulty between the two assessments.

References

Benton, T. (2021). Comparative Judgement for Linking Two Existing Scales. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.775203

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, UK: Cambridge Assessment.

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

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

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

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

Bramley, T. (2010). 'Key discriminators' and the use of item level data in awarding. *Research Matters: A Cambridge Assessment publication, 9,* 32-38

Bramley, T. (2012). The effect of manipulating features of examinees' scripts on their perceived quality. *Research Matters: A Cambridge Assessment publication*, *13*, 18–26.

Bramley, T., & Oates, T. (2011). Rank ordering and paired comparisons – the way Cambridge Assessment is using them in operational and experimental work. *Research Matters: A Cambridge Assessment publication*, *11*, 32–35.

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

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

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