CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Research Matters

Issue 33 / Spring 2022

Special issue: Comparative Judgement in educational assessment

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Proud to be part of the University of Cambridge

Cambridge University Press & Assessment unlocks the potential of millions of people worldwide. Our qualifications, assessments, academic publications and original research spread knowledge, spark enquiry and aid understanding.

Citation

Articles in this publication should be cited using the following example for article 1: Benton, T., Gill, T., Hughes, S., & Leech, T. (2022). A summary of OCR's pilots of the use of Comparative Judgement in setting grade boundaries. *Research Matters: A Cambridge University Press & Assessment publication, 33*, 10–30.

Credits

Reviewers: Matthew Carroll, Vicki Crisp, Melissa Mouthaan, Nicky Rushton, Sylvia Vitello, Joanna Williamson

Editorial and production management: Lisa Bowett

Additional proofreading: Alison French

Cambridge University Press & Assessment is committed to making our documents accessible in accordance with the WCAG 2.1 Standard. We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, contact our team: Research Division, researchprogrammes@cambridgeassessment.org.uk

If you need this document in a different format contact us, telling us your name, email address and requirements and we will respond within 15 working days.

All details are correct at the time of publication in March 2022

Contents

Research Matters / 33

A Cambridge University Press & Assessment publication

Foreword

In this issue, Tom Bramley has used his editorial discretion to write a very detailed and cogent editorial. I shall exercise my discretion by balancing that with an unusually short foreword. We have written before that in respect of Comparative Judgement we are in part dealing with a paradigm shift. And typically these take time. While I don't entirely agree with Pierre Azoulay's brutal incarnation of Max Planck's argument – "Does science advance one funeral at a time?" – these insights from the philosophy and history of science suggest that radically new approaches and theorisations take time to become both established and elaborated. In this edition of *Research Matters* we are seeing significant refinement in both application and thinking associated with Comparative Judgement. Genuinely ground-breaking, the wide-ranging studies and projects examine its limits and processes as well as its relation to existing assessment approaches. There's one aspect of this edition which I really commend – it not only explores the characteristics of Comparative Judgement through carefully designed empirical work, it increases our understanding of the processes of human judgement within it. Many studies in assessment grapple with the question of "what are the measurement characteristics of the assessment?" without engaging with what might actually be happening – the mechanisms at play. The studies in this volume range freely over both – and that is extremely valuable for establishing the extent to which Comparative Judgement can both represent a new paradigm and offer new, more effective techniques in public testing and assessment.

Tim Oates, **CBE** Group Director, Assessment Research and Development

Research Matters / 33

A Cambridge University Press & Assessment publication

Editorial – the CJ landscape

Eleven years ago in *Research Matters*, Bramley & Oates (2011) described the "state of play" regarding research into Comparative Judgement (CJ). At the time it was still being referred to as a "new" method, at least in terms of its application in educational assessment. (The technique of paired comparisons in psychology has been around since the 19th century!) It is still not a mainstream technique, but much more is now known about its strengths and weaknesses. Commercial CJ products and services are being used at scale in England to assess the writing of schoolchildren, and the underlying theory is being actively researched and evaluated by pockets of researchers around the world. Recently a collection of papers devoted to CJ has been published in the journal *Frontiers in Education*, and our researchers have contributed to that as well as this special issue of *Research Matters*. In this editorial we give an overview of what we see as the current CJ landscape and some of the key research questions and practical issues.

The foremost distinction (in our view) remains that between CJ applications where the aim is to replace conventional marking and CJ applications where the aim is to link two tests or exams that have been marked in the conventional way. These two applications can be very different, and hence may require different criteria for evaluation.

In the "CJ instead of marking" context, the key questions are around the reliability and validity of the resulting scores (sometimes referred to as "measures"); the feasibility and cost of using CJ; and the transparency of the process from the point of view of the individual who will ultimately receive the score and may have an important decision made about them on its basis.

In the "CJ for linking tests" context, the key questions are around the reliability and validity of the resulting linking, not the measures for individual pieces of work. In fact, the most significant recent development in CJ theory in this context has been the realisation that the linking can be estimated without even needing to estimate measures for individual scripts. This "simplified pairs/ranks" approach (Benton, 2021) can include many more scripts than the original CJ approach and achieve similar or better accuracy with less resource (the main cost of CJ in this context is judge time).

Recently, we have begun to explore the possibilities of a third context for CJ – that of moderating non-exam assessment such as coursework. This application combines some of the features of each of the above two contexts – because only samples of work from each school are considered in traditional moderation it is more feasible to apply CJ than if every single piece of work needed to be considered. Instead of linking standards across two tests, here the challenge is to link standards across schools that may be applying the same marking criteria, but at different levels of stringency. The final article in this issue by Carmen Vidal Rodeiro and Lucy Chambers explores the practical feasibility of using CJ for moderating portfolios.

In all contexts it is important to understand how the judges are making their comparative judgements, because we need to have some confidence that even if CJ is giving the right result, it is doing so for the right reasons! One of the presumed advantages of CJ in the "instead of marking" context is that the quick, holistic and intuitive comparative judgements allow the implicit expertise of the judges to be given free rein, without bogging them down in the details of rules for assigning numerical values to qualities of student responses that are hard to specify precisely. However, in social contexts a "first impression" is also a quick, holistic and intuitive judgement about a person – and one which can be adversely affected by various biases and stereotypes held by the judge. We need to be confident that the assessment equivalents of such biases and stereotypes do not affect CJ in a way that would be deemed unfair. In the "linking two tests" context (where high accuracy for individual scripts is not needed) there is the concern that judges will not be able to allow for differences in difficulty between the two tests or exams when comparing responses to the questions. Will they be biased in favour of good answers to easier questions and against weaker answers to harder questions? (It should be recognised that any method of linking two tests that relies on expert judgement has to deal with this issue, which is practically equivalent to asking whether expert judgement can be used at all!)

There are various ways to try and investigate the processes by which judges reach their judgements in CJ. One is to observe them as they make their judgements, perhaps asking them to "think aloud" as they do so. In this special issue this approach is taken in the article by Carmen Vidal Rodeiro and Lucy Chambers (looking at the feasibility of using CJ to help with coursework moderation); and in the article by Tony Leech and Lucy Chambers (observing judges making judgements in a "linking two tests" context). Another way is to ask judges to answer questions retrospectively about how they made their judgements, as done by Emma Walland in her article comparing traditional marking with paired comparison CJ and rank-ordering CJ. A third way is to manipulate experimentally some "construct-irrelevant" features of scripts to check that changes that shouldn't make a difference don't make a difference. This was done by Bramley (2012) and more recently by Chambers and Cunningham (2021). One finding from these experiments is that missing responses (questions not attempted and left blank) are perceived as worse than incorrect responses, which implies CJ may be more suited to assessments where the quality of the response is assessed, rather than whether it is

correct or incorrect. This implication is further supported by evidence from observations (this issue) that even when instructed not to mark responses in the traditional way, some judges still find that this is the only way that they can make a comparative judgement.

Because of the technological and logistical challenges with implementing CJ as an alternative to marking for high-stakes assessment, not to mention the hurdles to overcome in achieving public acceptability of the results (particularly given the new suspicion of relatively complex algorithms in deriving individual results (see for example Kelly, 2021)), the bulk of our research has focused on the application of CJ in the second of our contexts: as a means of maintaining standards. Our first article, by Tom Benton and colleagues, brings together results from no fewer than 20 CJ studies carried out by OCR piloting the use of CJ for maintaining standards. It allows us to draw some firm conclusions about the plausibility of results from CJ, the margins of error we can expect, and which particular methods and designs for collecting the judgement data are the most efficient in terms of value (reliability) for money (judge time).

Those 20 studies were also a rich source of qualitative data about CJ. Our second article, by Tony Leech and Lucy Chambers, draws from this data to answer the questions "what processes do judges use to reach their CJ decisions?" and "what features of scripts do they focus on when making decisions?" They find that in some cases judges are not able to make the holistic comparisons required by the method and resort to re-marking and "totting up marks", which suggests that CJ may not be a universal answer to all assessment problems but rather a tool to be deployed selectively and judiciously. Their taxonomy of decision-making features includes dimensions relating to the judge, the question paper, the candidate response and the CJ task itself. It should provide a useful framework for further research in this area.

Our third article, by Emma Walland, continues the theme of analysing qualitative data, but this time in the context of CJ as an alternative to marking, and investigating judges' own perceptions of the relative merits of CJ with judgements of pairs of essays and CJ with rank orders of packs of 10 essays, and their views of these two CJ approaches compared with traditional marking. The insights from this are important because it is necessary to ensure that the judges themselves believe in the validity of what they are doing if stakeholders more widely are to be convinced.

In theory CJ differs from traditional marking on a number of dimensions: the judgements are relative rather than absolute; quick rather than slow, holistic rather than atomistic. The resulting scores/outcomes are based on the judgements of many judges rather than a single judge, and are created by a complex statistical model rather than straightforward addition of marks. Recently researchers have begun to investigate which of these contrasting features of CJ are the source of any benefits it may have. Benton and Gallacher (2018) showed that if pseudo-CJ data was created from the marks assigned in a multiple marking study and analysed with the same statistical model (Bradley-Terry) as normally used for CJ data, the resulting measures had the same predictive value as measures obtained with genuine paired comparisons.

Here "predictive value" has the same sense as "concurrent validity" – it is the correlation of scores from one assessment with scores on a different but conceptually related assessment, such as a different exam paper in the same subject. Benton and Gallacher concluded that it is the combination of multiple judgements with a statistical model that is important, not the fact that the judgements are relative: "The physical act of placing two essays next to each other and deciding which is better does not appear to produce judgements that, in themselves, have any more predictive value than getting the same individual to simply mark a set of essays" (p. 27). Our fourth article, by Tim Gill, pursues the idea of comparing the predictive value of comparative judgements with the predictive value of pseudo-comparative judgements based on mark differences. In an admittedly small and opportunistic sample of four CJ studies he found that individual marks-based comparisons had better predictive value than individual CJ-based comparisons. Further research currently underway will help to clarify whether this is because marks are better at predicting concurrent marks and CJ judgements are better at predicting concurrent CJ judgements, or because marks are better (contain less random error) in general.

It is important to recognise that traditional marking does not yield numerical values that are completely error-free, even though they are usually presented as such. Our fifth article, by Joanna Williamson, explores through sophisticated modelling and simulation the extent to which using CJ to link two tests (with the simplified pairs method) depends on the accuracy of marking in the sample of scripts that are used for the linking. Reassuringly, she finds that the linking is robust both to isolated instances of very erratic marking and also to general degradation in marking quality, provided that the sample size of the study (in terms of number of paired judgements) is kept at a reasonable level.

We hope that this special issue is of interest both to seasoned practitioners with CJ and to others who may have heard of it and want to find out more about how it is being used in educational assessment.

Tom Bramley Director, Research Division

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

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://doi.org/10.3389/feduc.2022.802392

Kelly, A. (2021). A tale of two algorithms: The appeal and repeal of calculated grades systems in England and Ireland in 2020. *British Educational Research Journal*, *47*(3), 725–741. https://doi.org/10.1002/berj.3705

A summary of OCR's pilots of the use of Comparative Judgement in setting grade boundaries

Tom Benton, Tim Gill, Sarah Hughes and Tony Leech (Research Division)

Introduction

In the context of examinations, the phrase "maintaining standards" usually refers to any activity designed to ensure that it is no easier (or harder) to achieve a given grade or above in one year than in another. That is, that the level of performance that is required to achieve each grade is held constant over time. In this article we are particularly interested in how maintaining standards is achieved through decisions about where grade boundaries are positioned. In normal (non-pandemic) times, grade boundaries in GCSEs, A levels and various other qualifications are primarily decided upon via a method referred to as comparable outcomes. Very broadly, this technique is designed to reduce grade inflation by ensuring that, at a national level, grade distributions remain more or less static over time[1]. As such, it is sometimes criticised for not allowing the exam system to recognise genuine improvements in the performances of successive cohorts of candidates.

With the above criticism in mind, a few years ago, Ofqual began investigating whether alternative sources of evidence based on comparative judgement (CJ) might be used in setting grade boundaries (Curcin et al., 2019). Their research concluded that the methods were "very promising for capturing expert judgement for the purpose of standard maintaining" (p. 13). This article adds to this body of evidence with results from OCR's own trials of CJ in awarding[2].

The fundamental question in positioning grade boundaries using expert judgement is to decide whether a candidate awarded a certain number of marks has demonstrated the performance required to deserve a particular grade – particularly with respect to the level of performance that has been required on different assessments to achieve that grade in the past. All attempts to use CJ in

1 See https://dera.ioe.ac.uk/15397/1/2012-05-09-maintaining-standards-in-summer-2012.pdf for further discussion.

2 In our context, "awarding" means the process of choosing grade boundaries so that candidates, who have already been allocated marks on their exam scripts, can be awarded grades.

standard maintaining reduce this fundamental question to a series of comparisons between scripts. For example, rather than asking examiners in the awarding meeting "is this script that was awarded 63 marks worthy of a grade B?" we might ask "is this script [that was awarded 63 marks] deserving of a higher grade than this script from last year [that was awarded, say, 62 marks on a different assessment]?". Expert judges answer the latter question based on the content and quality of responses rather than the marks themselves (marks are typically removed from scripts and not shared with judges) and the results of many such comparisons are used to determine the location of grade boundaries. The use of a CJ method in standard maintaining forces decisions to focus on the quality of responses rather than be swayed by other sources of evidence such as previous grade boundaries or statistical data. These alternative sources of evidence would only be allowed to influence the final grade boundary decision at a separate stage later on (Bramley & Benton, 2015).

Ofqual's interest in the use of CJ in awarding was itself inspired by research conducted over the past 20 years within Cambridge Assessment. In particular, the specific method they trialled was originally suggested by Bramley (2005) and has previously been evaluated by (among others) Bramley & Gill (2010) and Gill et al. (2007). The proposed approach uses the Bradley-Terry model to analyse the results of a CJ study using scripts from two different test versions (usually from different examination sessions). The analysis produces a measure of performance (a CJ "measure") for each script based on which other scripts it was deemed superior to, and which it was deemed inferior to, over a number of pairwise comparisons. Crucially, these CJ measures are located on the same scale for each of the two different tests, thus providing a mechanism to map the marks from one test onto equivalent marks on the other.

More recent research (Benton, Cunningham et al., 2020) has suggested an improved approach to the use of CJ in awarding, which we call "simplified pairs". The approach differs in that it calibrates tests against one another without the need to produce a CJ "measure" for each script. As a result, the method includes a larger number of scripts in each CJ study but reduces the number of judgements made about each script – ideally including each script in just a single judgement. Overall, this should provide just as robust a source of evidence for awarding as the previous approach but require substantially less time from expert judges and, therefore, be less costly.

The aim of the research was to evaluate the effectiveness of the different approaches to using CJ in practice. This incorporated studies of the use of CJ in awarding across a range of different qualification types (GCSEs, A levels, Cambridge Nationals, Cambridge Technicals) and subjects. In this article we use the data from these studies to establish: whether the use of CJ in awarding leads to plausible suggested grade boundaries, the reported precision of these estimates, and the amount of judge time required to produce them.

Description of the studies

This article makes use of data from 20 CJ studies relating to awarding. Details of these studies are given in Table 1.

The main focus of this article is on the 13 studies done as part of OCR's pilots of using CJ in awarding. These studies span six different qualifications and further details are shown at the top of Table 1. The majority of these studies were conducted long after original awarding had been completed and in none of these cases was evidence from CJ the major source of evidence for the live award. All of the studies involve calibrating assessments from two different exam sessions against one another (for example, June 2018 against June 2019). In most cases different studies within the same qualification and subject address different exam papers. However, in a few cases (studies 5, 6 and 7, studies 10 and 11, and studies 12 and 13) different CJ studies trialled different techniques on the same papers.

As well as conducting 13 pilot studies, OCR also used CJ to help set grade boundaries on seven live components from three separate qualifications that were taken in the autumn 2020 exam series – possibly the first time that CJ has been a primary source of evidence in setting boundaries in a live exam series. CJ was used for these qualifications in autumn 2020 as, due to the unusual nature of the exam series (a special extra exam series as a result of the coronavirus pandemic) the usual statistical sources of evidence for setting grade boundaries were not available. CJ was only used in autumn 2020 in subjects where previous research (e.g., Curcin et al., 2019, Benton, Cunningham et al., 2020) had suggested CJ should provide an effective approach and where a sufficient number of examples of student work were available to judges. Since these seven CJ studies were used to help set grade boundaries, there is no point comparing the suggested grade boundaries from CJ to final boundaries. However, data from these seven studies will be used to provide further evidence about the amount of time required for exercises of this type.

Table 1: Details of the 20 studies providing data for this article.

Study no.	Study source	Qualification	Subject	Paper	Study type (pack size)	Max. mark
1	OCR pilot	AS level	Geography	Paper 1	Simplified Ranks (4)	82
2	OCR pilot	AS level	Geography	Paper 2	Simplified Ranks (4)	68
3	OCR pilot	AS level	Sociology	Paper 1	MC PCJ	75
4	OCR pilot	AS level	Sociology	Paper 2	Simplified Pairs	75
5	OCR pilot	GCSE	English Language	Paper 1	MC PCJ	80
6	OCR pilot	GCSE	English Language	Paper 1	MC RO (4)	80
7	OCR pilot	GCSE	English Language	Paper 1	Simplified Pairs	80
8	OCR pilot	GCSE	English Language	Paper 2	MC PCJ	80
9	OCR pilot	Cambridge Technical (L3)	Business	Paper 1	Simplified Pairs	90
10	OCR pilot	Cambridge Technical (L3)	Digital Media	Paper 2	Simplified Pairs	80
11	OCR pilot	Cambridge Technical (L3)	Digital Media	Paper 2	Simplified Ranks (8)	80
12	OCR pilot	Cambridge National (L2)	Child Development	Paper 1	Simplified Ranks (4)	80
13	OCR pilot	Cambridge National (L2)	Child Development	Paper 1	Simplified Ranks (6)	80
14	OCR live	A level	English Literature	Paper 1	Simplified Pairs	60
15	OCR live	A level	English Literature	Paper 2	Simplified Pairs	60
16	OCR live	A level	Psychology	Paper 1	Simplified Pairs	90
17	OCR live	A level	Psychology	Paper 2	Simplified Pairs	105
18	OCR live	A level	Psychology	Paper 3	Simplified Pairs	105
19	OCR live	GCSE	English Language	Paper 1	Simplified Pairs	80
20	OCR live	GCSE	English Language	Paper 2	Simplified Pairs	80

The studies in Table 1 encompass four different types of data collection designs:

- Multiple comparison pairwise comparative judgements (**MC PCJ**). As suggested by the name, these studies collected data using pairwise comparative judgements. Each script was included in many pairs so that, if desired, it was possible to generate measures of script quality using a Bradley-Terry model.
- Multiple comparison rank ordering (**MC RO**). These studies collected data by asking judges to rank scripts within packs of more than two from best to worst. Each script was included in several packs so that it was possible, if

desired, to generate measures of script quality using a Plackett-Luce model[3].

- **Simplified pairs.** Data was collected by pairwise comparisons of scripts from different versions. The majority of scripts were only included in a single paired comparison and logistic regression was used to generate estimated grade boundaries.
- **Simplified ranks.** Data was collected by asking judges to rank scripts within packs of more than two. The vast majority of scripts were only included in a single pack and logistic regression was used to generate estimated grade boundaries.

For more information on these different types of studies, including the precise calculations used to produce estimated boundaries and confidence intervals, see Benton, Cunningham et al. (2020). The four types of study listed above really only vary in two respects. Firstly, whether judges are asked to pick which out of a pair of scripts is superior (PCJ or "pairs"), or whether they are asked to rank larger groups of scripts (RO or "ranks"). Secondly, whether each script in the study is judged many times (an "MC" design) or whether each script is usually only included in a single pack or pair (a "simplified" design). Note that, although this typology may give the impression of these designs being qualitatively distinct, as described by Benton, Cunningham et al. (2020), all of them can be analysed in essentially the same way based around logistic regression of judges' decisions on the marks awarded to the scripts being compared. For the purposes of this article, we will refer to this approach to analysis as the "universal method". Although for study types with the prefix "MC" it is possible to fit a Bradley-Terry model to the data and apply the approach to awarding described by Bramley (2005), this is not the approach that was used. Having said this, it is worth noting that, for these data sets, where different analytical approaches are possible, in most cases they lead to similar recommended grade boundaries.

The development of the universal method is important as it allows us to avoid making a hard distinction between MC studies designed for use with the Bradley-Terry model and simplified approaches. Rather, all CJ studies relating to awarding can be thought as belonging to a single continuum in terms of the size of packs presented to judges and the number of packs each script is included in and can all be analysed in essentially the same way. In particular, due to the lack of available scripts in autumn 2020, for the OCR live studies, scripts from the 2020 series were used multiple times, whereas those from June 2019 were used just once. Nevertheless, the universal method could seamlessly handle this novel design.

Further details on the designs of the different studies are given in Table 2. This table brings out the features more clearly. It shows that simplified studies (both pairs and ranks) tend to use far more scripts from each series (usually hundreds) than MC approaches. However, as shown by the final three columns, they tend to use fewer resources. The final three columns represent three different ways of representing the total sizes of the tasks. Most transparently, one column

3 The Plackett-Luce model is equivalent to the Bradley-Terry model but can handle pack sizes larger than two avoiding the needs to convert rankings to pairs (as has been done for some previous research).

simply shows the total number of packs that needed to be judged in each study. However, since it obviously takes longer for a judge to rank a pack of 8 scripts than a pair of 2, further measures are needed. The second to last column calculates the total number of decisions needed. For example, a pack of 2 requires only 1 decision (who is better), whereas a pack of 8 requires 7 decisions (who is first, who is second, and so on). As will be shown later, this measure is the one most closely associated with the time required from judges to complete a study. The final column represents the size of the study in terms of the total number of pairs considered – for example, a single pack of 8 might be considered as providing information on 28 pairs of scripts. It can be seen that simplified studies tended to require fewer resources than MC studies and, as a result, they were usually completed by 5 or 6 judges whereas MC studies typically (though not always) used 10 or more.

Note that, in addition to the studies detailed in Table 1, an additional two recent experimental studies have been conducted with designs that allow a comparison between CJ methods and direct statistical equating between assessments using common pupils. Details on these studies can be found in Benton, Cunningham et al. (2020) and Benton, Leech et al. (2020). These will not be discussed further within the current article.

The remainder of the article is organised as follows:

- The next section will focus on the 13 OCR pilots of the use of CJ in awarding and assess the plausibility of the resulting recommendations regarding grade boundaries.
- The following section will consider how the level of precision associated with these grade boundary recommendations compares to previous pilots conducted by Ofqual.
- Drawing on both sets of data (pilots and live awarding), the final section will review the evidence regarding the amount of time needed from judges for studies of different types.

Table 2: Further details on the designs of different studies.

Study no.	Qual.	Subject	Study type	No. scripts		Pack size	No. judges	No. packs	No. decisions	No. pairs
				Series 1	Series 2					
1	AS level	Geography	Simp. Ranks	190	190	4	6	95	285	570
2	AS level	Geography	Simp. Ranks	194	194	4	6	97	291	582
3	AS level	Sociology	MC PCJ	70	70	2	21	1324	1324	1324
4	AS level	Sociology	Simp. Pairs	289	282	2	5	289	289	289
5	GCSE	English Language	MC PCJ	57	70	2	13	999	999	999
6	GCSE	English Language	MC RO	70	70	4	8	169	507	1014
7	GCSE	English Language	Simp. Pairs	291	291	2	5	291	291	291
8	GCSE	English Language	MC PCJ	57	72	2	15	1161	1161	1161
9	Cam. Tech. L3	Business	Simp. Pairs	256	249	2	6	284	284	284
10	Cam. Tech. L3	Digital Media	Simp. Pairs	227	235	2	6	314	314	314
11	Cam. Tech. L3	Digital Media	Simp. Ranks	164	164	8	6	41	287	1148
12	Cam. Nat. L2	Child Development	Simp. Ranks	190	190	4	9	95	285	570
13	Cam. Nat. L2	Child Development	Simp. Ranks	103	174	6	6	58	290	870
14	A level	English Literature	Simp. Pairs	466	91	2	6	466	466	466
15	A level	English Literature	Simp. Pairs	414	97	2	5	414	414	414
16	A level	Psychology	Simp. Pairs	498	66	2	6	498	498	498
17	A level	Psychology	Simp. Pairs	500	53	2	6	500	500	500
18	A level	Psychology	Simp. Pairs	500	51	2	6	500	500	500
19	GCSE	English Language	Simp. Pairs	350	291	2	6	350	350	350
20	GCSE	English Language	Simp. Pairs	350	345	2	6	350	350	350

Does CJ yield plausible grade boundaries?

In this section we explore the accuracy of the grade boundary estimates from CJ exercises. This is in terms of both how they compared with the actual boundaries as decided in the awarding meetings and how confident we were in the estimates (as measured by their standard errors).

For this analysis, we used data from the 13 CJ exercises which were part of the OCR pilots. This meant it was possible to compare the CJ grade boundary estimates with the actual grade boundaries. The majority of these trials were conducted well after grade boundaries had been set and could not have influenced the awarding decisions. However, for two of these trials (the MC PCJ trials for GCSE English Language) the studies were conducted prior to awarding and results were seen by the assessment manager. Nonetheless, at the time, statistical alternatives were available to inform grade boundaries and the results of the CJ exercises were not the primary drivers of decisions.

In each study, the aim of analysis was to recommend grade boundaries in series 2 (the more recent exam series) of the assessment that were of equivalent difficulty to existing grade boundaries in series 1 (the previous exam series). The results of the analysis are summarised in Figure 1. This shows, for each CJ exercise, across a number of key grades, the difference between the recommended grade boundary based upon the CJ study and the actual final grade boundary for the series 2 papers. Confidence intervals are shown based on the uncertainties around the CJ estimates. All differences between suggested and actual boundaries are presented as a percentage of the total available marks on each assessment.

The difference between the CJ estimated boundaries and the actual boundaries varied from -8 per cent of marks (study 8 (English Language, MC PCJ), grade 4) to 8 per cent of marks (study 7 (English Language, Simplified Pairs), grade 9). The mean difference between estimated and actual grade boundaries was -1 per cent of marks and there was no evidence that the CJ estimates were more likely to be systematically higher or lower than the actual boundaries.

The confidence intervals in Figure 1 give an indication of when the difference between the actual outcome and the CJ outcome was statistically significant (i.e., where the confidence intervals do not contain zero). There were six such instances spread across four different assessments. Further details on the differences, in raw marks rather than as a percentage of marks, and after allowing for rounding, are as follows:

- Study 1 (AS level Geography, Simplified Ranks) grade A. The confidence interval for CJ suggested a boundary on the series 2 paper of between 38 and 46 marks. The actual boundary was 48.
- Study 8 (GCSE English, MC PCJ) grades 4 and 1. CJ suggested that the grade 4 boundary should be between 23 and 33 marks and the final boundary was at 34. Similarly, CJ suggested that the grade 1 boundary should be between 1 and 7 marks and the final boundary was 8 marks.
- Study 9 (Cambridge Technical Business, Simplified Pairs) grades D and P. CJ suggested the grade D boundary should be between 53 and 60 marks and the final boundary was 62. Similarly, CJ suggested the grade P boundary should be between 24 and 31 marks and the final boundary was 32.
- Study 10 (Cambridge Technical in Digital Media, Simplified Pairs) grade D. CJ suggested the boundary should be between 56 and 63 marks and the final boundary was 54.

From the above descriptions it can be seen that, even where suggestions from CJ were significantly different from those used in practice, a change to the grade boundary of no more than 2 marks would be sufficient to bring the result within the confidence interval. These results are also encouraging for the use of CJ in that they show clear cases where the use of CJ would likely have an impact on decisions about boundaries. If no such cases were identified, then there would be little point in adopting CJ. However, it is also encouraging that the scale of change being suggested to grade boundaries (up to 2 marks) is not so large as to be implausible.

There were a few assessments (GCSE English Language paper 1, Cambridge Technical in Digital Media and Cambridge National in Child Development) which were analysed multiple times, using different CJ methods. The results of these were compared to see if there were any interesting differences between methods.

For GCSE English Language paper 1 (studies 5, 6 and 7), the boundary estimates from MC PCJ and MC RO were very similar, within 1 mark at grades 9 and 7 and within 2 marks at grades 4 and 1. In contrast, the estimates from simplified pairs were very different, up to 8 marks higher at grades 9 and 7, and up to 4 marks lower at grade 1. However, due to the wide confidence intervals at certain grades for the simplified pairs method, these differences were not statistically significant. It is acknowledged that the design of this simplified pairs study (which was the very first one ever undertaken by Cambridge Assessment) did not include a wide enough range of marks to provide accurate estimates at different grade boundaries. This is why the confidence intervals were so wide for grades 9 and 7.

For the Cambridge Technical in Digital Media (studies 10 and 11), the estimated boundaries for simplified pairs and simplified ranks were close to each other, differing by around 2 marks at both grades D and P. The confidence intervals for the two methods comfortably overlap with each other at each grade.

Finally, for the Cambridge National in Child Development (studies 12 and 13), the estimates for grades D2 and P2 were very similar for both methods (simplified ranks with packs of 4 scripts or with packs of 6 scripts). There was a slightly larger difference at grade P1, although only 1.5 marks.

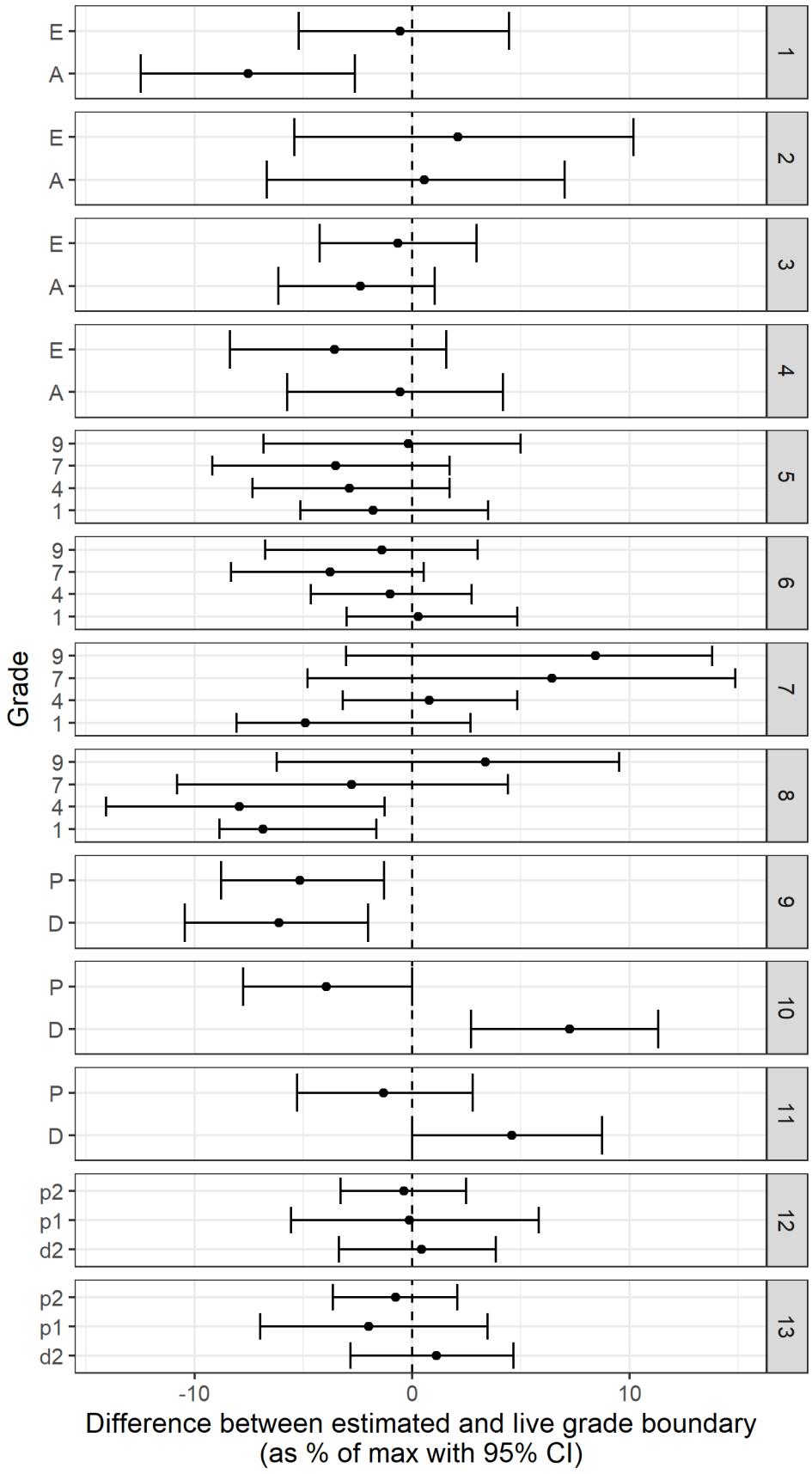

Figure 1: Plots of differences between estimated and live grade boundary for the 13 OCR pilot studies. 95 per cent confidence intervals for the differences are also shown.

Figure 2 compares the actual and estimated grade boundaries in a different way. For each of the 36 grade boundaries being investigated, Figure 2 shows how the actual change in grade boundaries between the two exam series in the study relates to the amount CJ suggested grade boundaries should shift between series 1 and series 2. For simplicity, these changes are shown in raw marks rather than as a percentage of maximum available mark. As can be seen, for the assessments considered in this article, grade boundaries only changed a small amount between series 1 and series 2. No grade boundary moved by more than 3 marks and 12 remained completely static between series[4]. Nonetheless, where boundaries shifted between series, the suggested direction of the shift from CJ was relatively consistent with what happened in practice. In particular, in only two cases did CJ suggest the boundary should rise when, in fact, it was lowered, and in only one case did CJ suggest lowering a boundary that was actually raised.

It is also clear from Figure 2 that the range of suggested boundary changes from CJ is somewhat wider than the range of changes in practice. However, given the fairly wide margins of error around the CJ estimates (see Figure 1), this is not particularly unexpected. Furthermore, the regression line in Figure 2 suggests that, on average, suggested boundary changes from CJ are close to those enacted in practice.

4 This level of consistency is not typical of all qualifications. For example, between 2015 and 2016, OCR's GCSE grade boundaries changed by an average of 4 per cent of the available maximum marks.

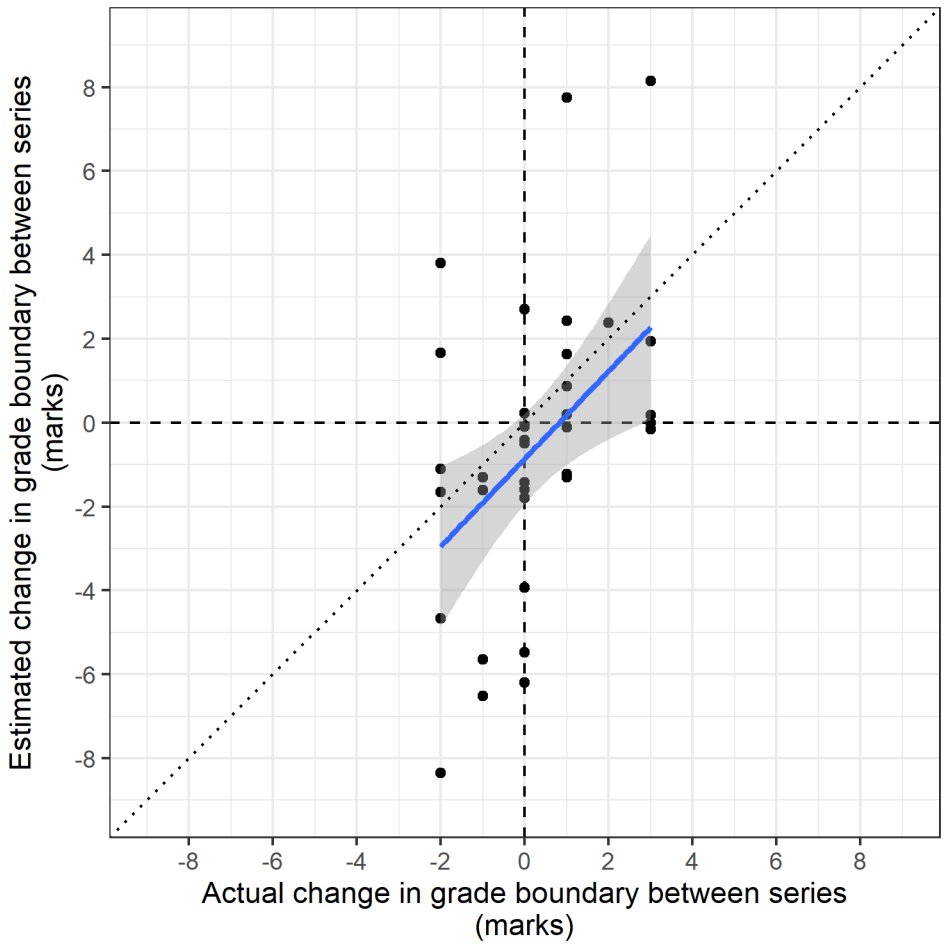

Figure 2: Relationship between actual and estimated changes in grade boundaries between series.

The solid blue indicates a regression line and the grey shaded area a 95 per cent confidence interval for the line. The dotted diagonal line represents a line of equality.

Figure 3 presents data on the precision of the estimates from the CJ exercises. Two different measures are shown for each exercise. Firstly, the average estimated standard error (SE) of each CJ grade boundary estimate[5] within each study (shown on the y-axis). To allow greater comparability across different studies, the figure presents the SE as a percentage of the maximum mark on the paper.

As well as producing estimates at each individual boundary, CJ can generate an overall estimate of the relative difficulty of two assessments. The second measure of precision (shown on the x-axis) is the SE of this estimate of the overall difference in difficulty between the series 1 and series 2 papers. Again, the figure presents the SE as a percentage of maximum mark.

Figure 3 compares the two measures of precision for each of the 13 CJ studies. The dotted line shows the line of equality, and the different markers indicate different

5 Calculated by dividing the range of the 95 per cent confidence intervals (Upper CI – Lower CI) by 3.92 (2 x 1.96).

study types. This figure shows that, for all the studies apart from one, the mean SE across grade boundaries was higher than the SE of the overall difference in difficulty. This indicates that there is a gain in precision to be made if we are willing to assume a constant change in difficulty across all grade boundaries. The results shown in Figure 1 suggest that this assumption is plausible. Specifically, the confidence intervals surrounding the recommended levels of adjustment at different boundaries within an assessment tend to overlap. Since changes at different boundaries are not independent, we need to be careful not to overinterpret this fact. However, from a pragmatic perspective, this does show that it is possible to pick a single adjustment figure that is consistent with the recommendations at the different boundaries.

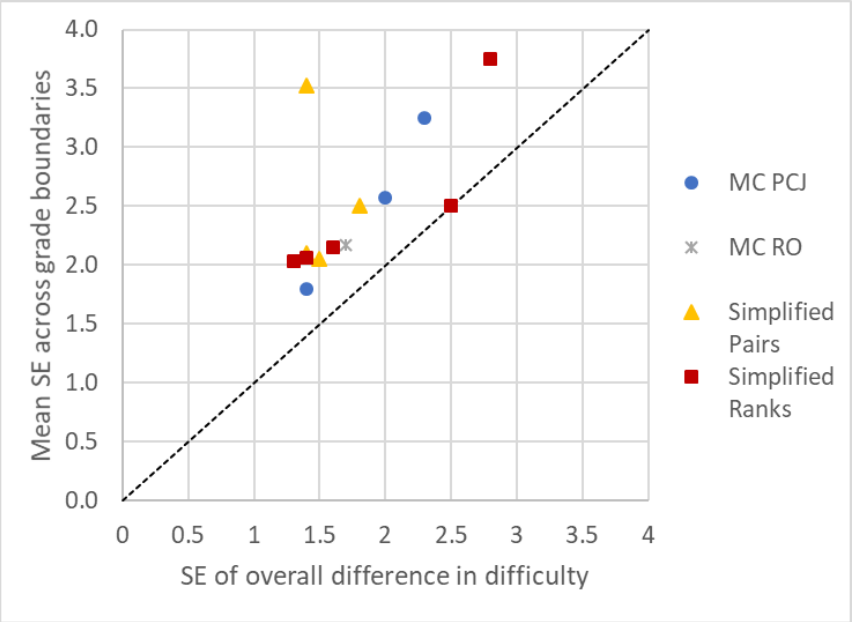

Figure 3: Comparison of the SE of the overall difference in difficulty with the mean SE of the grade boundary estimates.

Table 3 compares the precision of the different study types, showing the mean SE of the overall difference in difficulty and the mean SE of the grade boundary estimates. This shows that there were not large differences between the different study types. Looking at the SE of the overall difference, the lowest mean was for Simplified Pairs (1.53) and the highest was for Simplified Ranks (1.92). For the SE of the grade boundary estimates, the lowest mean was for MC RO (2.18) and highest mean for Simplified Pairs (2.74). However, as noted previously, the design of one of the simplified pairs studies (study 7) didn't include a wide enough range of marks to provide accurate estimates at different grade boundaries. With this study removed, the mean SE of grade boundary estimates for simplified pairs studies was 2.22.

Table 3: Mean precision of CJ exercises (as a percentage of the maximum mark).

Study type	No. of studies	Mean SE overall	No. of grade boundary estimates	Mean SE of grade boundary estimates
Simplified Ranks	5	1.92	12	2.43
Simplified Pairs	4	1.53	10	2.74
MC PCJ	3	1.90	10	2.69
MC RO	1	1.70	4	2.18

These results demonstrate that the precision of studies using simplified methods did not seem to be substantially worse than those using multiple comparison methods but had the advantage of using far fewer resources (see Table 2).

How does achieved precision compare to previous pilots of CJ in awarding?

In order to appraise the levels of precision reported in the previous section we compare against reported precisions for previous pilots of the use of CJ in awarding. In order to do this, we make use of the precision of estimates of 77 grade boundaries across 23 CJ studies conducted by Ofqual and reported in Curcin et al. (2019)[6].

All standard errors were converted to percentages of the maximum mark available and are summarised in Figure 4. As can be seen, the average level of precision achieved in the OCR pilots was similar to (or perhaps slightly lower than) that achieved in Ofqual's pilots of CJ in awarding. This indicates that OCR's recent pilots have achieved similar levels of reported precision to previous uses of CJ in awarding. Furthermore, according to Table 5 of Curcin et al. (2019), on average, Ofqual's studies required over 1000 paired comparisons – substantially more than the number used within the simplified methods (see Table 2). In other words, simplified methods have allowed us to achieve similar levels of precision to previous studies while using substantially fewer comparisons.

6 This represents all of Ofqual's studies undertaken using similar methods to the ones described in this report. A small number of studies using teachers (rather than examiners) for PCJ and also the (somewhat unsuccessful) trials of the "pinpointing" method are excluded. Standard errors are calculated by dividing reported values for "CI_2SD" in the Ofqual report by 2. Note that the standard errors in Curcin et al.'s report are based upon a bootstrapping procedure. While this approach differs from that used in OCR's pilots, the reported results still provide a benchmark for the perceived level of precision from previous studies.

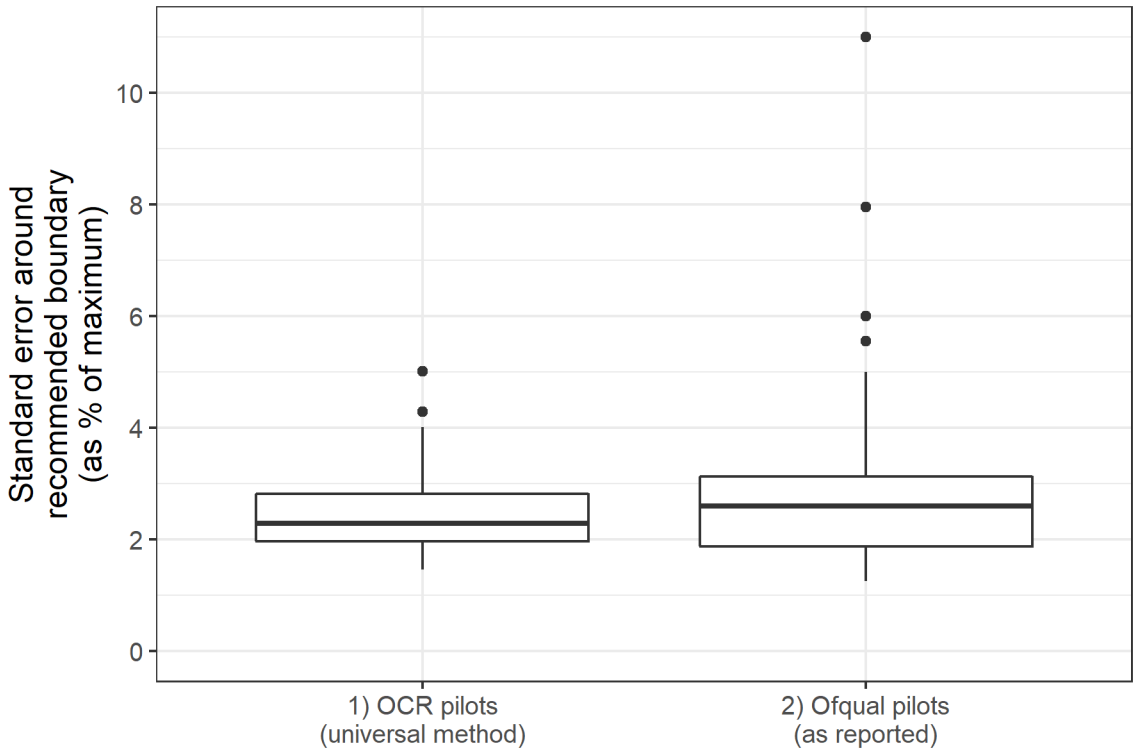

Figure 4: Boxplot of standard errors (as % of maximum available mark) around estimated grade boundaries in OCR's recent pilots and in pilots reported by Ofqual in Curcin et al. (2019).

How long do studies take?

This section looks in detail at the amount of time spent on CJ studies. Ideally, we would want these studies to take as little time as possible, but not at the expense of the accuracy of grade boundary estimates resulting from the exercise. As well as investigating the overall time spent, we also look at the average time spent in making individual CJ decisions and the average time spent in ranking packs of different sizes.

This investigation focused on the 13 exercises that were part of the OCR awarding trials and also the seven exercises that were used by OCR in live awarding in the autumn 2020 examination session. This was made possible because the online CJ tool used for data collection provided an accurate measure of how long judges spent on each exercise. The 20 exercises explored in this section included at least one from each of the four different methods of data collection, as described earlier.

The amount of time spent (recorded in seconds) on each pack (or pair) was measured by the CJ tool and included in the study results. For easier interpretation, we converted this into minutes, and then calculated the "robust"

mean[7] time per pack (or pair) for each study. To calculate the overall time spent on each study, we multiplied the robust mean by the number of packs in the study. This total was then converted into hours.

Study time by study type

We start with a simple breakdown of overall study time by study type. Figure 5 shows the total study time (in hours) for each of the CJ studies, grouped by study type. Table 4 shows the mean, minimum and maximum time by study type.

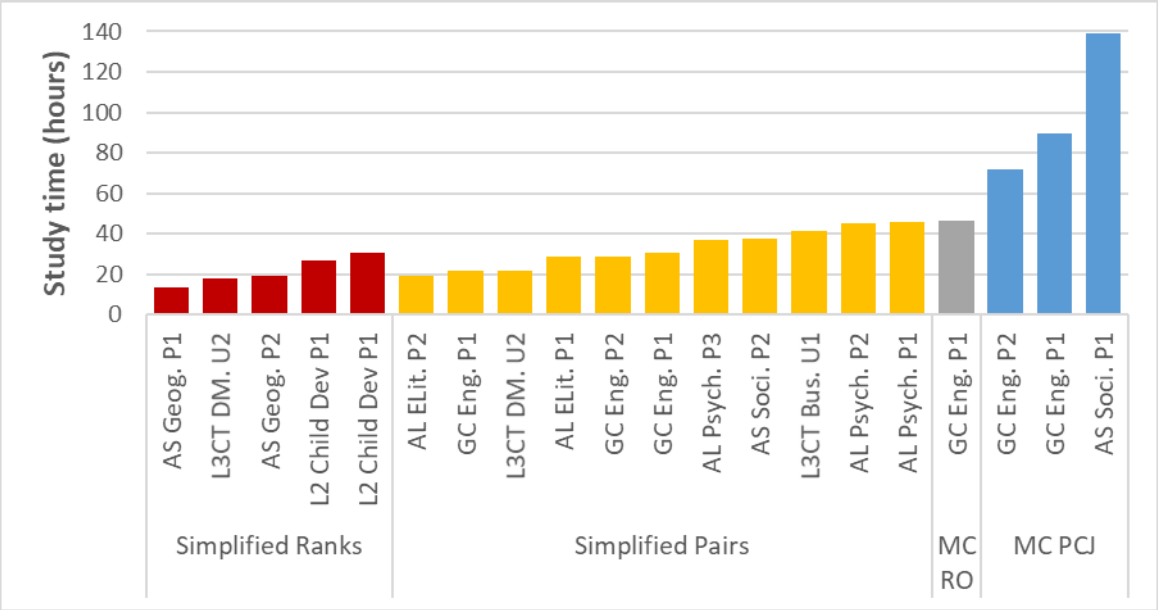

Figure 5: Time spent on each study, grouped by study type.

Table 4: Summary statistics for time spent on CJ studies (in hours), by study type.

Study type	No. of studies	Mean	Min.	Max.
MC PCJ	3	100.0	71.6	139.0
MC RO	1	46.5	46.5	46.5
Simplified Pairs	11	32.5	19.4	45.8
Simplified Ranks	5	21.4	13.3	30.6

Figure 5 shows that the study taking the longest time (139 hours) took more than 10 times as long as the shortest (13 hours). Two clear patterns can be seen in this data. Firstly, all of the simplified studies took less time than any of the multiple comparison studies. This was not surprising as, in the simplified studies, each script

7 This type of mean gives less weight to outliers, which otherwise might distort results. We used this measure because each study had a few packs with very unlikely looking apparent times. These were likely to be occasions when the judge stopped for a break during the task, but left the task window open so that the tool continued to record the time.

was only involved in one comparison (pack), whereas in the MC studies, each script was included in many comparisons. Given the numbers of scripts included, this results in MC studies having a greater number of comparisons in total. The shorter overall study times for simplified studies are of interest because, as shown earlier, we know that simplified studies are not associated with reduced precision. Secondly, studies that involved ranking of (more than two) scripts tended to take less time than those involving paired comparisons. This suggests that it was quicker for the judges to generate estimates with reasonable precision through ranking multiple scripts in one pack than through paired comparisons. However, it is worth noting there were some simplified pairs studies that took less time than some simplified ranks studies.

Study time by number of decisions made

Another way to categorise the different studies is by the overall number of decisions that the judges were required to make. We calculated this by multiplying the number of decisions per pack by the number of packs, where there were n-1 decisions for a pack of size *n* (e.g., for a pack of 8 there were 7 decisions to be made about the order of the scripts). We expected that the more decisions overall, the longer the total time taken on average. Figure 6 plots the total number of decisions against the total time taken. Each symbol and colour represents a different study type, and there is an overall line of best fit.

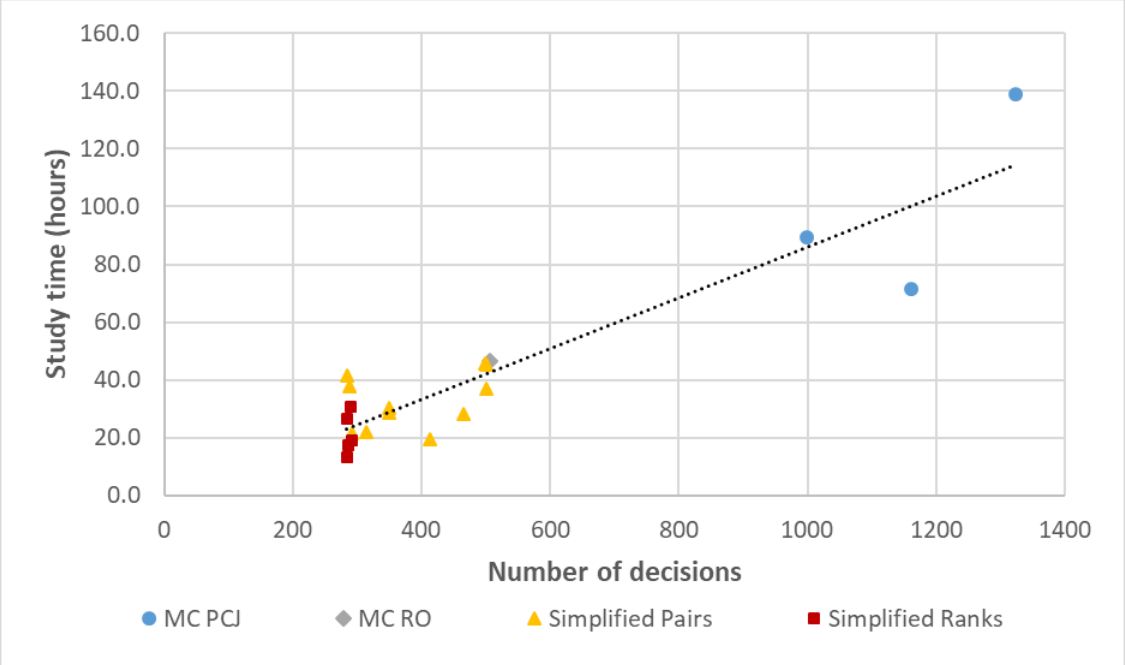

Figure 6: Study time, by total number of decisions made.

This shows that, overall, there was a clear positive relationship, with more decisions associated with a longer study time. Furthermore, the chart shows that the differences in the numbers of decisions required largely explain the differences in time required between techniques shown in Table 4. The line of best fit indicates that every 11 decisions within a study (e.g., every 11 pairs) will add approximately an hour to the required total time – that is, every decision in a study requires between 5 and 6 minutes.

However, within each study type the relationship was less clear. For example, all the simplified ranks studies involved very similar numbers of decisions (between 285 and 291), but still had a substantial range of overall duration (between 13.3 and 30.6 hours). This suggests that there were other reasons for the differences in the study time, possibly relating to the nature of the scripts involved.

Average time per pack, by pack size

As well as looking at the overall time, we also investigated the average time spent per pack, by the size of the pack. Figure 7 presents the (robust) mean time spent per pack for each of the studies, ordered by the pack size.

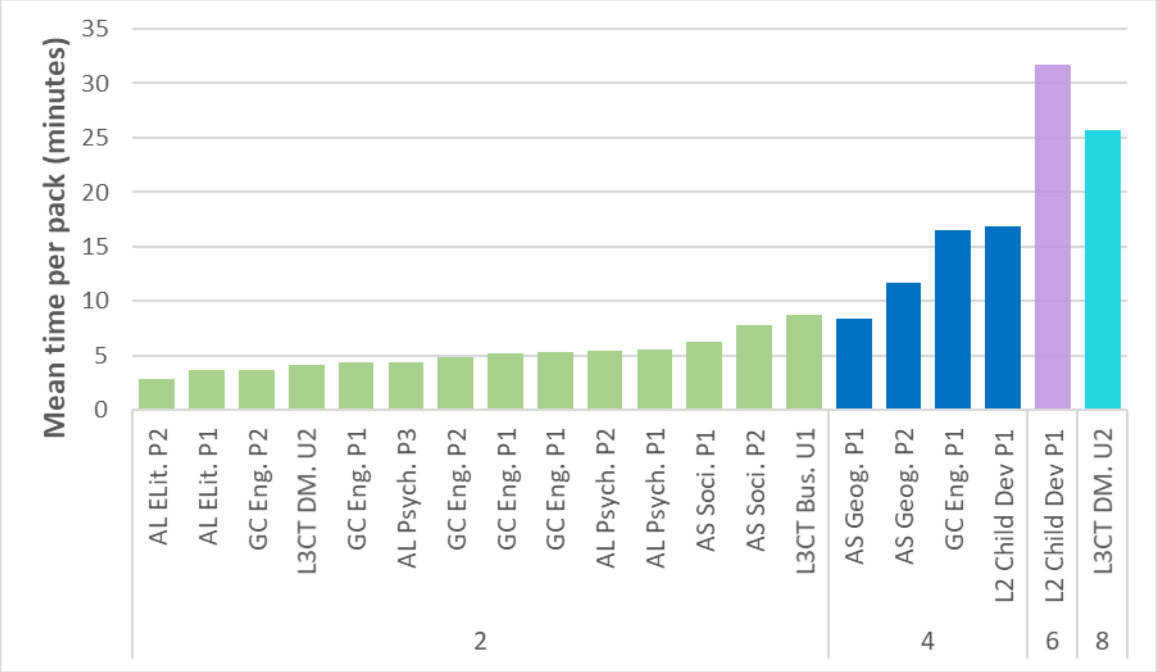

Figure 7: Mean time spent per pack, by pack size.

As expected, the larger the pack, the longer the time spent on average. With a pack size of 2, the robust mean time per pack varied between 2.8 and 8.8 minutes. These times are in line with what would be expected from previous research on the time required for paired comparisons (e.g., Curcin et al., 2019, p. 80). For packs of 4 scripts the mean varied between 8.4 and 16.9 minutes. Packs of 6 or 8 scripts took considerably longer.

In theory we might expect the time per pack to increase linearly with the number of decisions required within each pack. That is, a pack of 4 to require 3 times as long as a pack of 2, a pack of 6 to require 5 times as long and a pack of 8 to require 7 times as long. Very broadly, the data reflects this expectation.

Conclusion

A vast amount of trialling of the use of CJ in setting grade boundaries has been conducted over the past 20 years. This includes numerous previous studies by Cambridge Assessment, a large number of trials by Ofqual, the 13 pilot studies

by OCR described in this article, and 7 applications of the method by OCR during live awarding. With such a large number of research studies completed it might be argued that, as an assessment research community, we really should be in a position to make a call as to whether the method should be applied in practice or not.

With this in mind, the current synthesis of CJ studies suggests the following encouraging results:

- The grade boundaries suggested by CJ are plausible. For the 13 pilots OCR has recently completed looking at CJ in awarding, there were no instances of the actual grade boundaries that had been set in practice being more than 2 marks outside the confidence interval suggested by CJ. In most cases the actual grade boundaries were within the range suggested by CJ. To put this another way, the use of CJ would likely have some impact on grade boundaries but not so large an impact as to lead to implausible results.
- The precision of boundaries from CJ indicated that this could be an informative source of evidence. Specifically, the confidence intervals suggested we could estimate the relative overall difference in difficulty between two assessments to a precision of +/- 4 per cent of the paper total. The precision of recommendations at individual grade boundaries was marginally worse (confidence intervals of around +/- 5 per cent on average). The level of precision in OCR's pilots was similar to (or perhaps slightly better than) what had been achieved in previous studies of CJ in awarding.
- The development of simplified methods (simplified pairs and ranks) has improved the efficiency of CJ for awarding. In particular, the analysis in this article shows that we have been able to achieve similar precision to previous uses of CJ while requiring far less time for judges. A typical simplified study tends to require about 30 hours of judge time usually spread across 6 judges. In contrast, the MC studies in our pilots used between 46 and 140 hours.

Despite the encouraging results in this article and in previous studies on the use of CJ in awarding, there are some barriers to the widespread uptake of CJ for awarding.

Firstly, while studies comparing estimated and actual grade boundaries can be used to indicate plausibility, they do not allow an assessment about whether the results from CJ are actually correct. In particular, where differences are seen it could either be because of a problem with the CJ method or with the way in which boundaries were set in practice (in our cases, largely reliant on comparable outcomes). While some experimental studies (Benton, Cunningham et al., 2020, Benton, Leech et al., 2020) have endeavoured to identify the accuracy of CJ in an absolute sense, these are relatively rare. This gap in the research leaves ongoing concerns about the extent to which grade boundaries suggested by CJ can be trusted – particularly in more objectively marked subjects such as mathematics and science.

Secondly, while the development of simplified methods has significantly reduced

the cost of CJ studies of this type, each CJ exercise requires about 30 person-hours of time typically realised as needing 5 hours of time from each of 6 expert judges. To award a whole qualification this time requirement is multiplied by the number of assessment components that the qualification is comprised of. Thus, while achievable, the amount of time needed from examiners, and hence the cost, is still higher than the current, more confirmatory, procedure for the inclusion of expert judgement in awarding.

References

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

Benton, T., 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 Judgement. *Journal of Applied Measurement*, *6*(2), 202–223.

Bramley, T., & Benton, T. (2015). The use of evidence in setting and maintaining standards in GCSEs and A levels.

Bramley, T., & Gill, T. (2010). Evaluating the rank ordering method for standard maintaining. *Research Papers in Education*, *25*(3), 293–317. https://doi.org/10.1080/02671522.2010.498147

Curcin, M., Howard, E., Sully, K., & Black, B. (2019). *Improving awarding: 2018/2019 pilots*. Ofqual report Ofqual/19/6575. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/851778/Improving_awarding_-_FINAL196575.pdf

Gill, T., Bramley, T., & Black, B. (2007). An investigation of standard maintaining in GCSE English using a rank-ordering method. Paper presented at the British Educational Research Association Conference, 5–8 September in London, UK.

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.

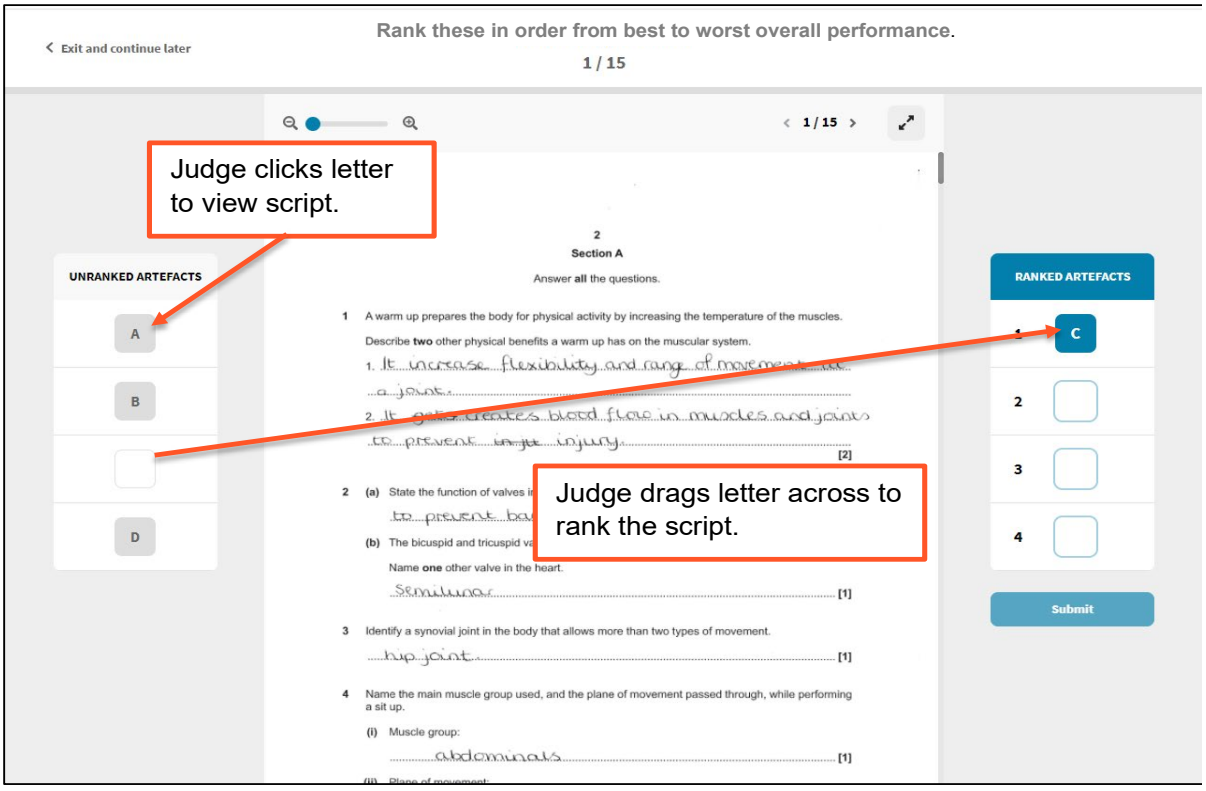

Figure 1: Annotated screenshot of judge's view of CJ software.

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.

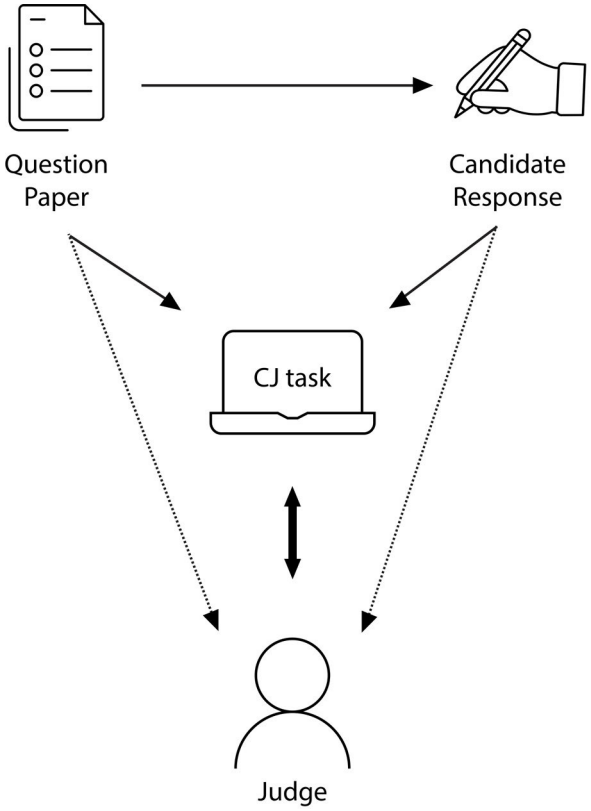

Figure 2: Dimensions of judge decision-making.

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

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.

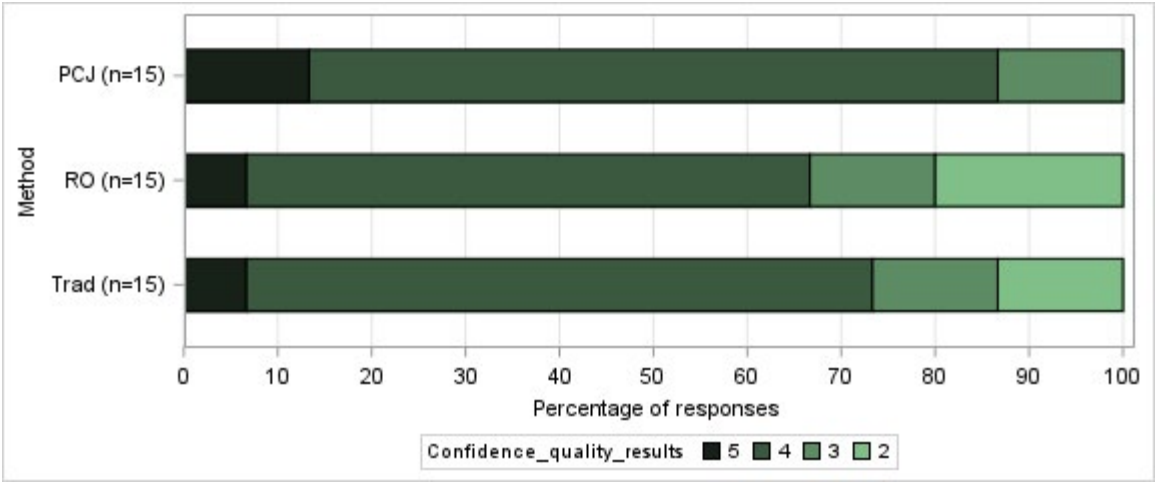

Figure 1: Participants' responses about their confidence in the quality of the results produced by the methods on a scale from 1 to 5. Darker shading represents more positive responses (increased confidence). 5 was "very confident", 4 was "fairly confident", 3 was "not sure", 2 was "fairly unconfident"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

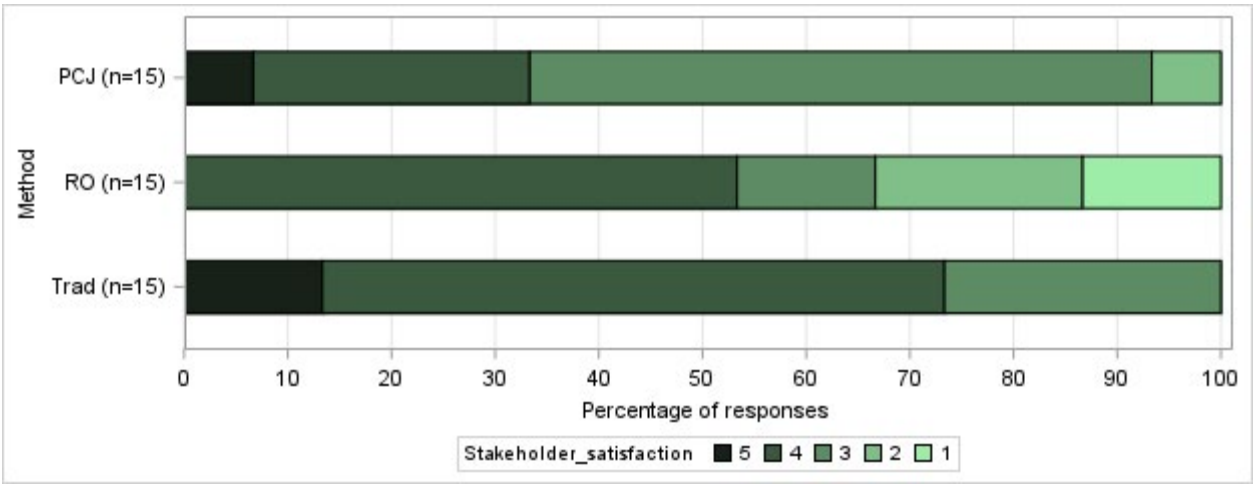

Figure 2: Participants' responses about their opinions on stakeholder satisfaction with the results produced by each of the methods on a scale from 1 to 5. Darker shading represents more positive responses (greater satisfaction). 5 was "very satisfied", 4 was "fairly satisfied", 3 was "not sure", 2 was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

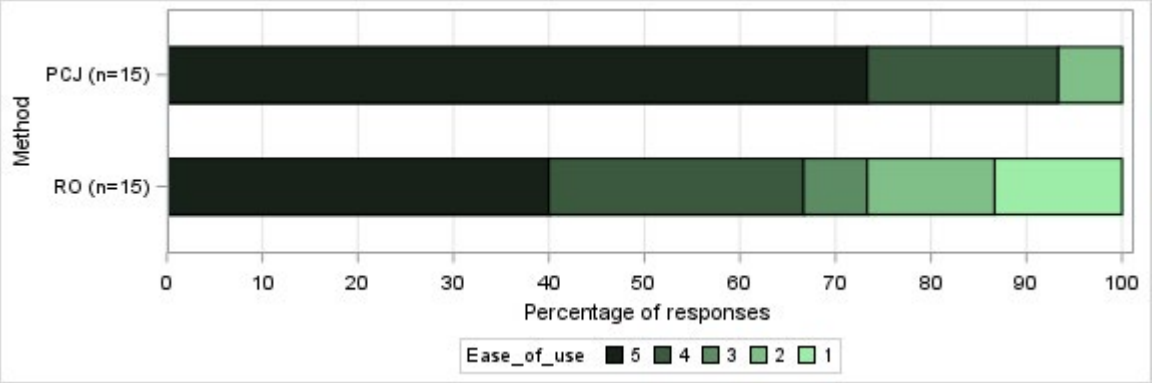

Figure 3: Participants' responses about the ease of use of each of the methods in comparison with analytical marking on a scale from 1 to 5. Darker shading represents more positive responses (easier to use). 5 was "much easier to use", 4 was "a little easier to use", 3 was "much the same", 2 was "a little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

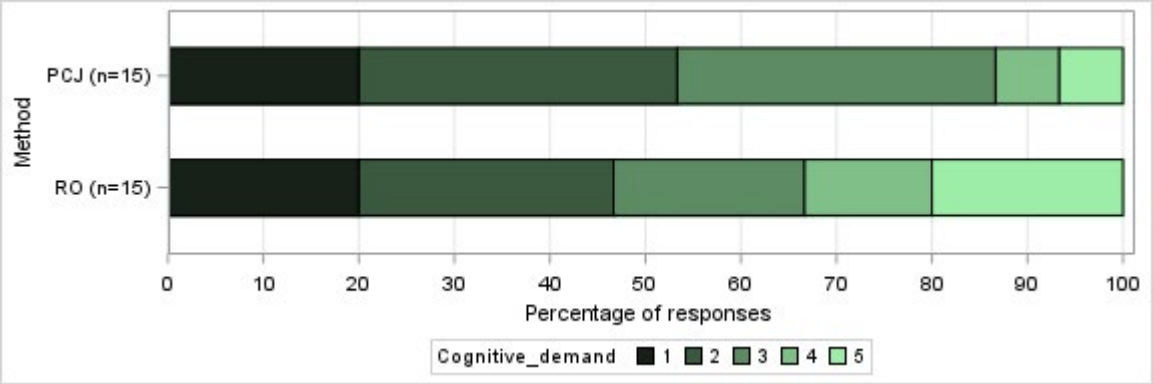

Figure 4: Participants' responses about the cognitive demand of each of the methods on a scale from 1 to 5. Darker shading represents more positive responses (less cognitively demanding). 5 was "much more cognitively demanding", 4 was "a little more cognitively demanding", 3 was "much the same", 2 was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

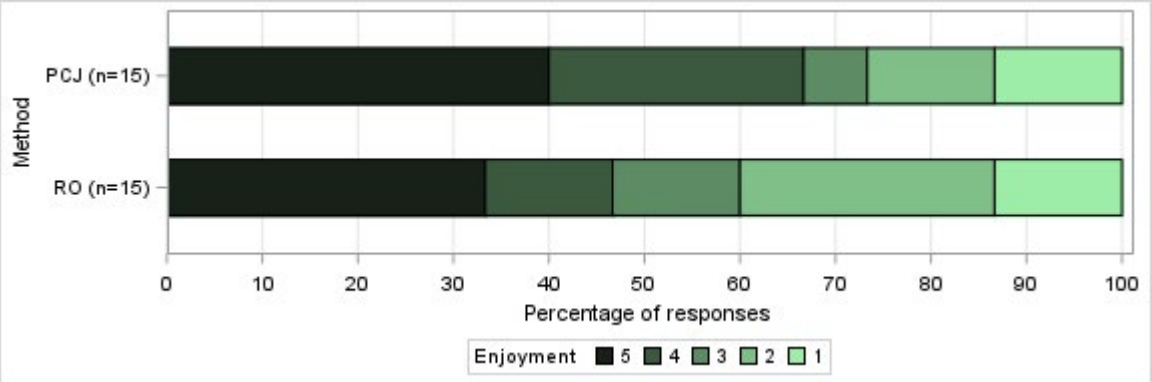

Figure 5: Participants' responses about their enjoyment of each of the methods in comparison with analytical marking on a scale from 1 to 5. Darker shading represents more positive responses (more enjoyment). 5 was "much more enjoyable", 4 was "a little more enjoyable", 3 was "much the same", 2 was "a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

The concurrent validity of Comparative Judgement outcomes compared with marks

Tim Gill (Research Division)

Introduction

In Comparative Judgement (CJ) exercises, examiners are asked to look at a selection of candidate scripts (with marks removed) and order them in terms of which they believe display the best quality. The comparisons can either take the form of ranking of pairs of scripts ("paired CJ" or "PCJ") or of ranking of more than two scripts ("rank ordering" or "RO"). By including scripts from different examination sessions, the results of these exercises can be used to help with maintaining standards.

Results from previous CJ studies have demonstrated that the method appears to be valid and highly reliable in many contexts, including for marking of essays (Steedle & Ferrara, 2016) and standard maintaining (Benton, Leech & Hughes, 2020; Curcin et al., 2019). However, it is not entirely clear why CJ works as well as it does. Proponents of the method argue that it is because of the physical and judgemental processes involved in making comparative judgements. That is, the physical act of placing two scripts next to each other and deciding which is better based on an intuitive, holistic and relative judgement of quality. In particular, they argue that it is the relative aspect of the judgement that is important, because humans are better at making relative than absolute judgements (Laming, 1984). An alternative explanation, proposed by Benton & Gallacher (2018), is that the CJ method works well because CJ exercises capture a lot of individual paired comparison decisions quickly. In their study, they found that the predictive validity of scores derived from a CJ exercise was no better than the predictive validity of pseudo-CJ scores derived from comparing marks. This would suggest that CJ works well because of the number of judgements involved, not because the judgements come from the physical act of putting scripts next to each other and making a holistic relative comparison.

The analysis presented in this article adds to the research on this question by comparing the concurrent validity of the outcomes of CJ paired comparisons with the concurrent validity of outcomes based on the original marks given to scripts.

The focus here is on the validity of the outcomes of individual paired comparisons (the smallest building block within the CJ process), rather than the validity of

scores allocated to scripts by a statistical model (such as the Bradley-Terry model) following multiple comparisons. The aim is to discover whether the decisions of a human judge directly comparing two pieces of work have more validity than those based on comparing the marks of two scripts, when these are derived independently and (usually) by different markers. As such, this research provides direct evidence on whether the idea that humans are better at making relative rather than absolute judgements (Laming, 1984) applies in the context of educational assessment when absolute judgements are supported by a mark scheme. Previous research in the context of awarding (Gill & Bramley, 2013) found that examiners were better at making relative judgements of quality than absolute judgements.

Data and methods

For this research, we re-used data from several previous CJ studies undertaken by Cambridge Assessment. All of these were experimental trials of the CJ method, with the aim of determining whether CJ had the potential to be used in standard-maintaining exercises in GCSEs and AS or A level qualifications in England. Each of these CJ studies used exam scripts taken from qualifications offered by the OCR awarding body (either GCSEs or AS levels). In all cases, the method was similar: either five or six examiners were asked to make comparisons of exam scripts (either in pairs or in packs of four) and to order the scripts from best to worst, in terms of the overall quality of the work. In most of the studies, at least some of the paired comparisons involved scripts from the same exam paper, but a version taken in a different exam session and the results of the comparisons were then analysed statistically to give an indication of the relative difficulty of the two papers. In total, there were 20 datasets which were all analysed separately. Details of these are presented in Table 1.

Most of these CJ studies asked examiners to make comparisons between pairs of scripts, but there were three which asked examiners to rank order packs of four scripts instead. For these studies, the rank ordering outcomes were converted into paired comparisons data (i.e., 1st beats 2nd, 1st beats 3rd, 1st beats 4th, 2nd beats 3rd etc.).

To compare the concurrent validity of CJ decisions with decisions based on the marks we needed the original marks given to the scripts and a measure of concurrent validity. Each CJ dataset contained the centre and candidate numbers of each candidate included in the paired comparisons, the original mark given to each script by the original examiner in the live exam session and the outcome of the paired comparison (i.e., which script was judged to be better). Candidates were matched (using centre and candidate numbers) to their marks achieved on other component(s) in the same qualification. These marks were used as the measure of concurrent attainment. Where all candidates within a study took more than one other component in the same qualification, marks were summed and the total used.

Some of the previous CJ studies only included paired comparisons between scripts from the same exam paper taken in different sessions, while others also

included some comparisons between scripts from the same paper taken in the same session. For these latter studies, the datasets were split, so that the comparisons of scripts from the same exam session were analysed separately from the comparisons of scripts from different exam sessions. For example, we created three different sets of data for component AS level Geography Paper 1: comparisons between scripts from June 2018 and June 2019; June 2018 only comparisons; and June 2019 only comparisons.

In each dataset, the scripts were labelled as being either from the version 1 ("v1") paper or from the version 2 ("v2") paper. Every paired comparison included one v1 script and one v2 script. For the analysis of paired comparisons of scripts from different exam sessions, the scripts from the earlier session were designated as v1 and scripts from the later session as v2. For the analysis of paired comparisons of scripts from the same session, we needed to decide arbitrarily which of each pair of scripts would be the v1 script and which would be the v2 script. This was done by sorting each pair by the centre and candidate number and choosing the first script as the v1 script.

Table 1: Details of CJ study datasets used in the analysis.

Qualification and subject	Paper(s)	v1 exam session	v2 exam session	Pairs (PCJ) or Rank Order (RO)?	No. of judges	No. of scripts	No. of comparisons
AS Geography	Paper 1	June 18	June 19	RO	6	400	400
AS Geography	Paper 1 June 18	June 18	June 18	RO	6	200	100
AS Geography	Paper 1 June 19	June 19	June 19	RO	6	200	100
AS Geography	Paper 2	June 18	June 19	RO	6	400	400
AS Geography	Paper 2 June 18	June 18	June 18	RO	6	200	100
AS Geography	Paper 2 June 19	June 19	June 19	RO	6	200	100
AS Sociology	Paper 1	June 18	June 19	Pairs	22	140	1337
AS Sociology	Paper 2	June 18	June 19	Pairs	5	569	289
GCSE Eng Lang	Paper 1 PCJ	June 19	Nov 19	Pairs	14	124	517
GCSE Eng Lang	Paper 1 June 19 PCJ	June 19	June 19	Pairs	14	57	210
GCSE Eng Lang	Paper 1 Nov 19 PCJ	Nov 19	Nov 19	Pairs	14	70	303
GCSE Eng Lang	Paper 1 RO	June 19	Nov 19	RO	9	141	772
GCSE Eng Lang	Paper 1 RO June 19	June 19	June 19	RO	9	70	193
GCSE Eng Lang	Paper 1 RO Nov 19	Nov 19	Nov 19	RO	9	70	176
GCSE Eng Lang	Paper 1 SP	June 19	Nov 19	Pairs	5	570	285
GCSE Eng Lang	Paper 2 PCJ	June 19	Nov 19	Pairs	15	129	555
GCSE Eng Lang	Paper 2 PCJ June 19	June 19	June 19	Pairs	15	57	235
GCSE Eng Lang	Paper 2 PCJ Nov 19	Nov 19	Nov 19	Pairs	15	72	371
GCSE Maths	Paper 1	June 19	June 19	Pairs	6	600	300
GCSE Eng Lit	Paper 1 / Paper 2	June 16	June 16	Pairs	6	572	286

Table 1 includes three different datasets for GCSE English Language Paper 1. This is because they were taken from a Cambridge Assessment research project investigating which method of paired comparative judgement (PCJ), rank ordering

(RO) or simplified pairs (SP)[1] was most helpful for identifying grade boundaries (see Benton et al., 2022, this issue). Therefore, three different CJ exercises were undertaken. For GCSE Maths, the v1 and v2 sessions were the same because this study involved splitting the June 2019 paper into two halves and making comparisons between scripts from each half (see Benton, Leech & Hughes, 2020). Similarly, for the GCSE English Literature exercise, the v1 and v2 sessions were the same since comparisons were made between different papers in the same session (see Benton, Cunningham, Hughes & Leech, 2020).

To generate the measures of concurrent validity, the following process was undertaken for each dataset:

- For every paired comparison, a variable (called "v2CJsuperior") was created and was given a value of 1 if the v2 script was judged superior, and 0 otherwise.
- A variable (called "v2marksuperior") was created and was given a value of 1 if the V2 script was given a higher mark by the original marking, and 0 otherwise. For studies where the v1 and v2 were from different exam sessions, marks were converted to Uniform Mark Scale (UMS) marks so that they were directly comparable[2]. For the two studies (GCSE English Literature and GCSE Maths) where the papers being compared were from the same exam session, all candidates took both papers (or half papers in the case of GCSE Maths) being compared. This meant it was possible to use statistical equating (using the equipercentile method) to find the equivalent marks on v2 for each mark on v1.
- For the candidates in each CJ exercise, the total marks achieved in the other component(s) in the same specification in the same session were found ("concurrent marks"). For studies where the v1 and v2 scripts were from different exam sessions (and therefore the concurrent marks were also from different exam sessions), the marks were converted to UMS so that they were directly comparable. These variables were called "v1concurrentmark" and "v2concurrentmark".
- Pearson correlation coefficients[3] were calculated between both "v2CJsuperior" and "v2marksuperior" and the differences in candidate mark on the concurrent assessment(s) (v2concurrentmark-v1concurrentmark).

1 The Simplified Pairs method of CJ enables the mapping of marks between different tests without the need to estimate values on a common scale by fitting a statistical model (such as the Bradley-Terry model) to the experts' judgements. See Benton, Cunningham et al. for a more detailed description of this method (2020).

2 UMS marks are on a common scale, so that they can be directly compared between exam series (see https://ocr.org.uk/students/getting-your-results/calculating-your-grade/). If we had not done this it would mean that, if the two exams differed in difficulty, it would not be possible to say which script was judged to be superior according to the raw marks. As it happens, the differences in difficulty were all very small, meaning that there were very few instances of the order of pairs of marks changing after converting to UMS.

3 With one binary variable and one continuous variable this is equivalent to a point biserial correlation.

- A multiple logistic regression was undertaken of "v2CJsuperior" on the two concurrent marks. The pseudo R-squared value was recorded[4], as a measure of the model fit.
- A multiple logistic regression was undertaken of "v2marksuperior" on the two concurrent marks, and the pseudo R-squared value recorded.

By comparing the correlation coefficients and the pseudo R-squared values, it was possible to determine whether the individual decisions based on marks had higher concurrent validity than those derived using CJ. The correlation coefficients indicate the strength of the relationship between wider candidate ability (as measured by the marks on assessments taken concurrently) and which candidate was judged to be better by either the paired comparison or the marks. As the value of v2concurrentmark-v1concurrentmark increases we would also expect the likelihood of the v2 script winning to increase.

The purpose of undertaking the logistic regressions was to allow for the possibility that the UMS had not completely controlled for difficulty. The pseudo-R square measure can be thought of as an indication of how well the outcome (which script was better according to either CJ or marks) was predicted by the independent variables (marks on concurrent components). A higher pseudo-R square value for the prediction of the CJ outcome would be an indication of better concurrent validity for the CJ outcome than for the marks outcome.

As shown in Table 1, most of the data came from CJ exercises which were comparing scripts from different exam sessions (hence the need for two separate concurrent marks in the above description). However, there were several datasets where all the data came from a single session, so that the concurrent marks were directly comparable. For these, it was only necessary to calculate and compare the correlation coefficients.

Although the main focus of this research was on the validity of the outcomes of individual paired comparisons, a further analysis was undertaken to compare the concurrent validity of the CJ "measure" (see below for an explanation of the term "measure") with the concurrent validity of UMS marks. If the concurrent validity of CJ is substantially improved by using the measure instead of the outcomes of the individual paired comparisons, then this will be a further indication that it is the way in which CJ incorporates the many judgements that makes the method successful. For this analysis we just used data from the studies where each script was involved in multiple comparisons (AS level Sociology Paper 1, GCSE English Language Paper 1 PCJ and RO, and GCSE English Language Paper 2). For these studies, the paired comparison data was analysed using the Bradley-Terry model (Bradley & Terry, 1952). This generated a measure of quality for each script, based on the number of times each script was judged superior across the multiple comparisons it was included in. Pearson correlation coefficients were calculated between the measure and the UMS marks on the concurrent component, and these were compared with correlations between UMS marks on the component of interest and the UMS marks on the concurrent component.

4 Proc Logistic in SAS software reports the Cox & Snell (1989) calculation of R-squared.

Results

Table 2 presents the results of the correlations and the pseudo R-squared values for each dataset. For further details about the logistic regression (including the regression equation and some example output from one dataset), see the Appendix.

Table 2: Correlation coefficients and pseudo R-squared values for CJ study datasets.

Paper	Corr between concurrent marks and CJ outcome	Corr between concurrent marks and marks outcome	Pseudo R-square for CJ outcome	Pseudo R-square for marks outcome	Decision with higher concurrent validity
AS Geography Paper 1	0.37	0.38	0.14	0.15	Marks-based
AS Geography Paper 1 June 18	0.41	0.44	n/a	n/a	Marks-based
AS Geography Paper 1 June 19	0.36	0.48	n/a	n/a	Marks-based
AS Geography Paper 2	0.34	0.27	0.12	0.08	CJ-based
AS Geography Paper 2 June 18	0.37	0.33	n/a	n/a	CJ-based
AS Geography Paper 2 June 19	0.47	0.20	n/a	n/a	CJ-based
AS Sociology Paper 1	0.52	0.58	0.28	0.35	Marks-based
AS Sociology Paper 2	0.22	0.39	0.07	0.16	Marks-based
GCSE Eng Lang Paper 1 PCJ	0.57	0.66	0.33	0.44	Marks-based
GCSE Eng Lang Paper 1 PCJ June 19	0.63	0.74	n/a	n/a	Marks-based
GCSE Eng Lang Paper 1 PCJ Nov 19	0.48	0.60	n/a	n/a	Marks-based
GCSE Eng Lang Paper 1 RO	0.37	0.47	0.14	0.23	Marks-based
GCSE Eng Lang Paper 1 RO June 19	0.41	0.47	n/a	n/a	Marks-based
GCSE Eng Lang Paper 1 RO Nov 19	0.33	0.50	n/a	n/a	Marks-based
GCSE Eng Lang Paper 1 SP	0.37	0.40	0.16	0.18	Marks-based
GCSE Eng Lang Paper 2 PCJ	0.51	0.61	0.25	0.38	Marks-based
GCSE Eng Lang Paper 2 PCJ June 19	0.50	0.54	n/a	n/a	Marks-based
GCSE Eng Lang Paper 2 PCJ Nov 19	0.58	0.63	n/a	n/a	Marks-based
GCSE Maths Paper 1	0.56	0.59	n/a	n/a	Marks-based
GCSE Eng Lit Paper 1 / Paper 2	0.45	0.38	n/a	n/a	CJ-based

The "n/a" in the table indicates CJ exercises where all the data came from the same session and so it was not necessary to run a logistic regression model. The final column in the table indicates which decision (CJ-based or mark-based) had higher concurrent validity, according to the results of the correlations and the pseudo-R squares.

Figures 1 and 2 illustrate the relationships visually for two of the datasets (GCSE English Language Paper 1 PCJ, with a relatively high correlation and pseudo-R squared, and AS level Geography Paper 2, with a relatively low correlation and pseudo-R squared). The figures compare the range of mark differences in the concurrent attainments (v2concurrentmark-v1concurrentmark) by whether the V2

script was judged superior and by the judgement type (CJ or marks).

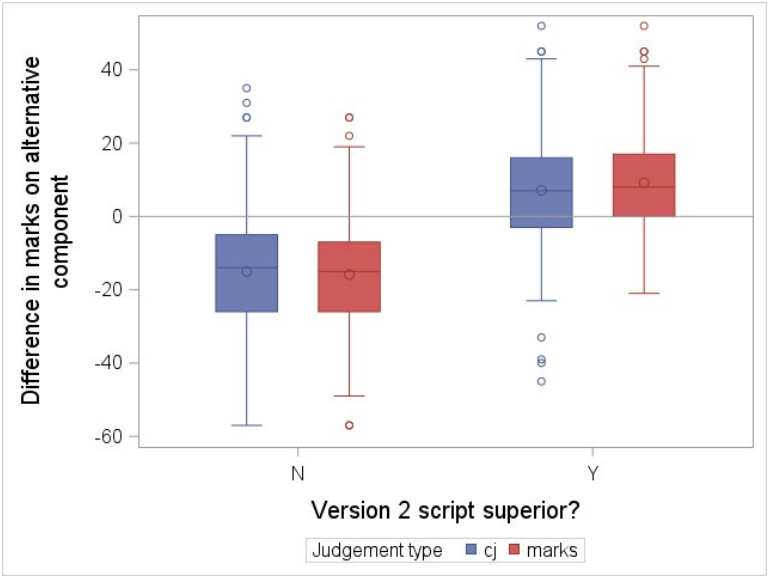

Figure 1: Distribution of v2concurrentmark- v1concurrentmark by superiority of v2 script and by judgement type (GCSE English Language, Paper 1, PCJ).

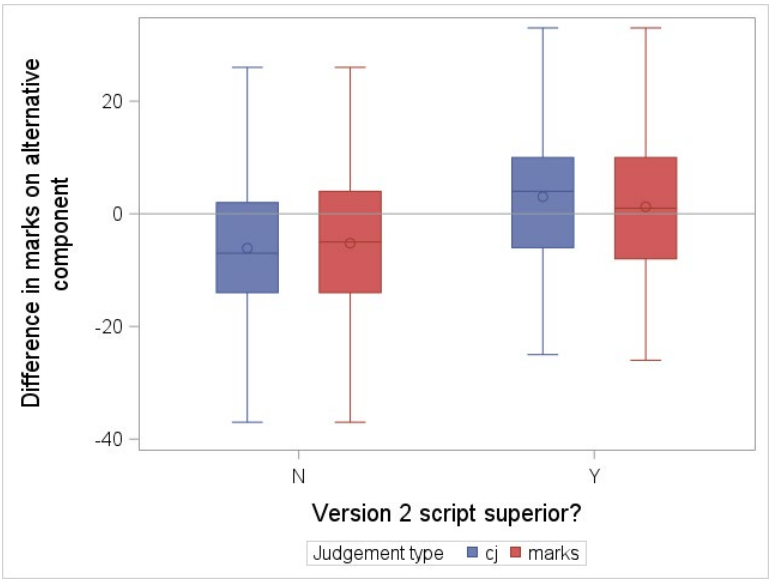

Figure 2: Distribution of v2concurrentmark–v1concurrentmark by superiority of V2 script and by judgement type (AS level Geography, Paper 2).

For example, Figure 1 shows that for V2 scripts judged to be superior according to CJ, the average difference in marks on concurrent components was around 10 marks. In contrast, when the V2 script was judged to be inferior, the average difference was around -15 marks. Figure 2 shows a much smaller difference in the average mark differences, being around 2 marks for V2 judged superior and around -5 when V2 was judged inferior.

In Figure 1, the red boxes are slightly further apart than the blue boxes, indicating a stronger relationship between the marks-based decision and the mark difference than between the CJ-based decision and the mark difference. This implies that the marks-based decision had higher concurrent validity. In contrast,

the blue boxes were further apart than the red boxes in Figure 2, implying that the CJ-based decision had higher concurrent validity.

Table 2 shows that for 16 out of the 20 data sets analysed, marks-based decisions had higher concurrent validity than CJ-based decisions. All but one of the pseudo R-squared values was higher for marks than for CJ. The only exception was AS level Geography, Paper 2, which had an R-squared of 0.12 for the CJ outcome model, compared with 0.08 for the marks model. For the 12 datasets which only included comparisons within the same session (and therefore with no logistic regression undertaken), there were only three occasions where the correlation coefficient was higher for the CJ outcome than for the marks outcome. These were for component AS level Geography, Paper 2 (both the 2018 only and the 2019 only datasets) and for the comparison between GCSE English Literature, Papers 1 and 2.

The AS level Geography, Paper 2 study used rank ordering, but otherwise the results showed no evidence of any different pattern for rank ordering studies compared with paired comparison studies.

Table 3 shows the correlation coefficients between the script measures (generated using the Bradley-Terry model) and the UMS marks on the concurrent component. It also shows the correlations between UMS marks on the component of interest and UMS marks on the concurrent component.

Table 3: Comparison of correlation coefficients of script measures and UMS with concurrent component UMS.

Component	No. of scripts	Corr between script measure and concurrent UMS	Corr between UMS and concurrent UMS
AS Sociology Paper 1	139	0.67	0.66
GCSE Eng Lang Paper 1 PCJ	124	0.77	0.84
GCSE Eng Lang Paper 1 RO	137	0.68	0.81
GCSE Eng Lang Paper 2 PCJ	129	0.71	0.77

These results mainly follow the pattern seen in Table 2, with higher correlations for marks-based outcomes (UMS) than for CJ-based outcomes (script measure). The only exception to this was for AS level Sociology, where the correlation between the script measure and concurrent component UMS was very slightly higher. This contrasts with the results from Table 2, where the correlation between the CJ outcome and concurrent component UMS (0.52) was lower than between the marks-based outcome and concurrent component UMS (0.58).

Having seen that individual decisions based on marks had higher concurrent validity than those based on CJ (Table 2), we had hoped that the additional analysis in Table 3 would illustrate how this is overcome by the way CJ incorporates many judgements. This effect was visible in only one of the four studies. Specifically, we found that for AS Sociology Paper 1, although the concurrent validity of individual CJ decisions was lower than that of marks-based decisions (Table 2), the concurrent validity of CJ estimated measures was higher than that of the original marks. However, the expected effect was not visible in the

other papers. Our expectations may have been confounded elsewhere because, although the CJ validity benefits from combining many judgements, the concurrent validity from marks also increased, for a different reason – namely that, analysing it in this way used the marks awarded to scripts, not just which of a pair is higher.

To think of this another way, it is clear that our earlier analysis provided a straightforward like with like comparison. Individual choices between two scripts based on judges' opinions were compared to individual choices based on marks. However, in this additional analysis we are comparing scores on one scale based upon multiple pairwise comparisons of each script (and different numbers of these for different components) to scores on an entirely different scale based on detailed marking. As such, meaningful interpretation is much harder.

It should be remembered that, in this section, we only have results from a relatively small number of studies, each of which only incorporates a fairly small number of scripts. As such it is important that we do not overinterpret these particular findings.

Conclusion

The main conclusion from this analysis is that the concurrent validity of the decision based on marks was generally higher than the concurrent validity of the CJ decision. Two possible reasons for this finding suggest themselves: firstly, CJ decisions reward different skills to marks (and ones that are less related to marks on other components). This may be because of the different processes involved. In CJ, the judges make holistic and relative judgements of quality, without direct reference to a mark scheme. In contrast, in live marking, the total mark is an absolute judgement of quality based on the summation of marks given for responses to individual items, with direct reference to the mark scheme. An alternative explanation is that individual CJ decisions are of lower quality than decisions based on marks. In other words, judges are less able to make reliable judgements of the relative qualities of scripts when using the quick holistic approach required of comparative judgements.

This finding adds further evidence in favour of the contention in Benton & Gallacher (2018) that it is not the physical process of making intuitive, holistic and relative judgements of quality that makes CJ successful, but rather that it is able to capture many individual paired comparison decisions quickly.

The results here contrast with a previous study evaluating examiners' holistic judgements of script quality (Gill & Bramley, 2013), which found that examiners were better at making relative judgements of quality than absolute judgements. The results of the current research suggest that the absolute judgements (i.e., marks) were better than the relative judgements (CJ). This difference may be because in practice marking also involves some form of relative judgement, versus a fixed mark scheme. This differs from the context of the previous study (Gill & Bramley, 2013) where the absolute judgements were made without access to the mark scheme and therefore dependent only on the judges' own idea of what grades should look like.

This research was opportunistic, in that it used already available datasets. Further research which is designed to answer a specific research question would be worthwhile. For example, it would be interesting to investigate which of CJ decisions or marks-based decisions in one component is a better predictor of CJ decisions in a related component. If CJ decisions are better then this would suggest that they are indeed rewarding different skills to marks.

References

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

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

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

Cox, D. R., & Snell, E. J. (1989). *The Analysis of Binary Data*, (2nd ed.). Chapman and Hall.

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

Gill, T., & Bramley, T. (2013). How accurate are examiners' holistic judgements of script quality? *Assessment in Education: Principles, Policy & Practice, 20*(3), 308–324. https://doi.org/10.1080/0969594X.2013.779229

Laming, D. (1984). The relativity of 'absolute' judgements. *British Journal of Mathematical and Statistical Psychology*, *37*(2), 152–183. https://doi.org/10.1111/j.2044-8317.1984.tb00798.x

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

Appendix – details of logistic regression

Logistic regression equation:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 v1concurrentmark_i + \beta_2 v2concurrentmark_i$$

Where p_i is the probability that in comparison "i" the version 2 script was judged superior, v1concurrentmark and v2concurrentmark are the independent variables and β_1 and β_2 are the regression coefficients.

Table A1: Example output from logistic regression (AS level Geography Paper 1, dependent variable = CJ-based decision)

	Parameter estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	0.0771	0.5441	0.0201	0.8874
v1concurrentmark	-0.0652	0.0114	32.6160	<0.0001
v2concurrentmark	0.0679	0.0122	31.1551	<0.0001

Table A2: Example output from logistic regression (AS level Geography Paper 1, dependent variable = marks-based decision)

	Parameter estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	0.7557	0.5456	1.9182	0.1661
v1concurrentmark	-0.0800	0.0118	46.1882	<0.0001
v2concurrentmark	0.0549	0.0119	21.2937	<0.0001

How are standard-maintaining activities based on Comparative Judgement affected by mismarking in the script evidence?

Joanna Williamson (Research Division)

Introduction

Providing evidence that can inform awarding is an important application of Comparative Judgement (CJ) methods in high-stakes qualifications. The process of marking scripts is not changed, but CJ methods can assist in the maintenance of standards from one series to another by informing decisions about where to place grade boundaries or cut scores. The research described in this article set out to increase understanding of the risks associated with this use of CJ. Specifically, the research explored how robust the outcomes of CJ-based awarding activities would be to mismarking in the script evidence.

In recent years, Ofqual has investigated various CJ methods for identifying cut scores in standard maintaining, and Curcin et al. (2019) reported the results of a large-scale pilot of several variants. This article focuses on the "simplified pairs" method (Benton et al., 2020), an example of the "universal method" discussed by Benton et al. (2022, this issue). Like other CJ methods, simplified pairs (SP) harnesses the information from paired comparisons in order to put the scores from two different assessments onto a common scale, but it does so without the need to fit a Bradley-Terry model and without the need to include individual scripts in multiple comparisons. Previous research has shown SP to be an efficient method, and comparisons with statistical equating have provided further evidence of the ability of SP to correctly determine the relative difficulty of two assessments, as well as for the ability of judges to account for the difficulty of different assessments in their comparisons (Benton et al., 2020).

In this article we explore the extent to which the SP method would be robust to mismarking in the sample of scripts used for the comparison exercise. In a particularly extreme case (e.g., if every script sampled from one assessment happened to be marked by a particularly harsh examiner, who undermarked by 10 marks), it is clear that the relationship estimated between scores on assessment A and assessment B would reflect this. More realistically, we know that mismarking can occur in live assessments, and quality of marking can vary, and it is therefore desirable to know how CJ-based awarding activities may be affected.

The simplified pairs method

In a typical application of SP for standard maintaining, there are two assessments (form A and form B), and existing grade boundaries or cut scores for form A. The SP method is applied in order to find the scores on form B that represent an equivalent level of performance to the grade boundary scores on form A. In the most straightforward case, we assume a fixed overall difference in difficulty between the two assessments, and the purpose of SP in this context is to find the difference d such that for scores x_A and x_B representing equivalent levels of performance on forms A and B respectively, $x_B = x_A + d$.

In an SP study, judges are asked to compare pairs of scripts, always comparing one form A script with one form B script, and decide which one is superior. Scripts from the extremes of the score distribution are excluded from the judging process, since where candidates have answered everything (or nothing) correctly, there is no basis for judging either to be superior. Scripts are sampled from a sub-range (e.g., those with scores between 20 and 90 per cent of the total available score), and paired for comparison in such a way that pairs include a wide range of score differences – Benton et al. (2020) recommend differences should span at least -20 to +20 per cent of the maximum available score. A typical SP study uses each script only once, to maximise the new information gained from each judgement, and can include several hundred pairs of scripts (Benton et al., 2020, pp. 5–6). This contrasts with typical CJ study designs, which would involve a smaller set of scripts from each assessment, that are then judged multiple times.

The overall difference in difficulty between form A and form B is found via logistic regression analysis of the judges' decisions. For the *i*th pair of scripts judged by judge *j*, the decision is represented by the outcome variable y_{ij}, where $y_{ij} = 0$ if the form A script is judged superior, and $y_{ij} = 1$ if the form B script is judged superior. The difference between the form A script score and form B script score is the independent variable and is notated d_{ij}, so that the modelled relationship is the following:

$$\text{log odds}\,(y_{ij} = 1) = \beta_0 + \beta_1 d_{ij}$$

where β_0 and β_1 are the intercept and slope in the linear relationship between score difference d_{ij} and the log odds[1] of the event $y_{ij} = 1$ (the event that the form B script is judged superior) in the logistic regression model. Since scores on form A and form B are considered equivalent when scripts with those scores have an equal probability of being judged superior, the overall difference d is d_{ij} where $P(y_{ij}=1) = 0.5$. Figure 1 gives a graphical example of this analysis: the blue markers and blue line show the percentage of script pairs at each mark difference where the form B script was judged to be superior to the form A script. The solid red line shows the fitted logistic regression line, and the dotted red lines show its 95 per cent confidence interval. The purple lines show d, the estimated overall difference in difficulty between form B and form A (in this example, 8 marks) and its estimated confidence interval.

1 The log odds or logit of the event $y_{ij}=1$ is $ln\left(\frac{p}{1-p}\right)$, where p is the probability that $y_{ij}=1$.

If the estimated relationship between script mark differences and judgement of superiority is very weak, the slope of the fitted logistic regression will be shallow and – in extreme cases – the SP analysis may result in 'flatlining'. This term describes a result such as that shown in Figure 2, where the dotted red lines representing the upper and/or lower 95 per cent confidence intervals for the logistic regression line fail to intersect the line y=0.5 at all. This indicates "a complete failure of the CJ method" (Benton et al., 2020, p. 8) – the relationship between script marks and judges' CJ decisions is so weak that it is impossible to produce a reliable confidence interval for the estimated difference in difficulty, meaning that the CJ method is unable to produce the evidence sought for awarding. The occurrence of mismarked scripts is a factor that can weaken the estimated relationship between mark differences and judgements of script superiority. It is, therefore, important to investigate quantitatively how robust SP analyses are to changes in the quality of marking in the selected script evidence.

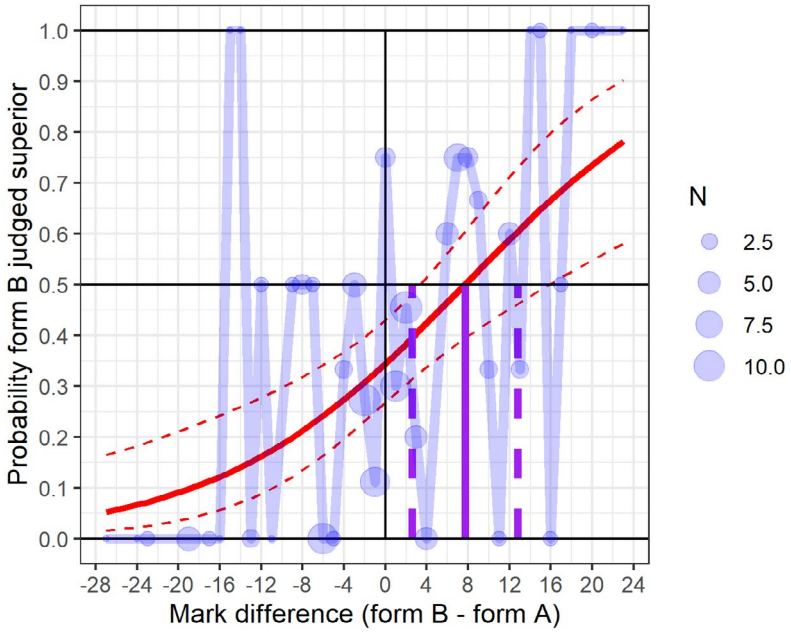

Figure 1: Example of a successful simplified pairs analysis.

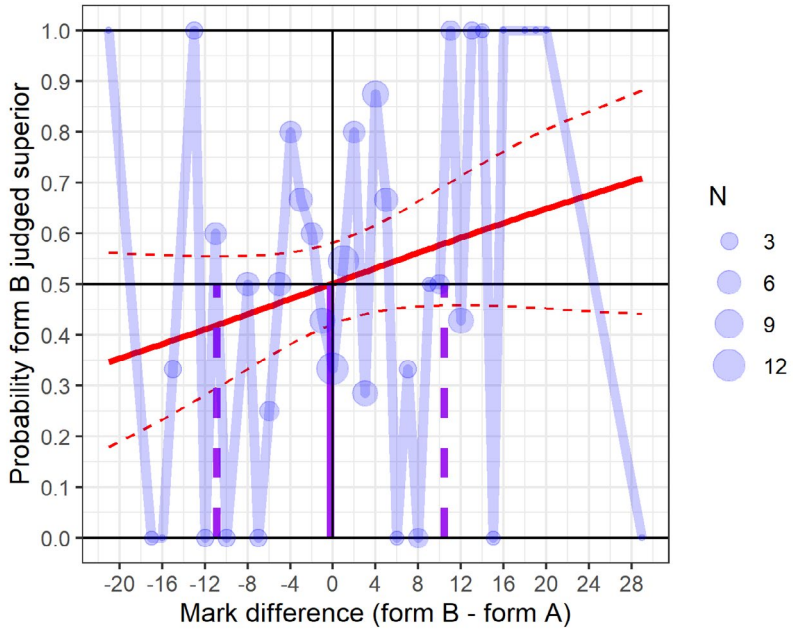

Figure 2: Example of a flatlining simplified pairs analysis.

Research overview

The overarching research question was addressed via three specific sub-questions, to explore robustness against mismarking in slightly different scenarios:

1. What is the impact on SP outcomes of large, one-off marking errors in the script evidence?
2. How many moderately sized marking errors can occur in the script evidence before SP analyses fail?
3. What is the impact on SP outcomes of a degradation in marking quality?

The first two questions were addressed using simulations based on data from previous SP studies, while the final question was addressed by simulating a large number of SP studies from scratch. All data simulation and analysis was carried out in R (R Core Team, 2021).

Impact of single large marking errors

The first set of simulations explored the impact on SP analyses of single large marking errors in the script evidence – such as could be introduced by a transcription error on a script (e.g., recording 13 as 31). These simulations were based on data from three real-life SP studies comparing different versions (forms) of various GCSE and AS level components.

To simulate a large one-off marking error in one of these SP studies, the mark difference for a single pair of scripts was manually altered (without changing the judge's decision) before re-running the SP analysis. To investigate the range of outcomes that such an error could cause, this was repeated, in turn, for every paired judgement in the dataset. For each SP study, we investigated four variants

of large errors, so each of the original SP studies therefore resulted in 4*n* simulated SP studies, where *n* was the number of pairs in the original study. The four types of large error were generated by altering the mark difference of the "marking error" script pair to one of the following values:

1. The largest positive mark difference between paired scripts in the study.
2. The largest negative mark difference between paired scripts in the study.
3. 70 per cent of the component maximum mark.
4. -70 per cent of the component maximum mark.

Figure 3 shows the distributions of estimated mark differences *d* for one of the original SP studies, under the four simulation conditions. The estimated difference between form B and form A in the original study (i.e., before deliberately introducing error) was -3.38 marks, and this value is shown by the vertical dotted line in each panel. The largest positive mark difference (form B script–form A script) between paired scripts in the original study was 15 marks, the largest negative mark difference was -15 marks, and the component maximum mark was 50 marks. A script pair selected as the "marking error" pair therefore had its mark difference altered to 15 marks, -15 marks, 35 marks and -35 marks in the four simulation conditions respectively. It is worth noting that the "error" introduced could therefore change the direction as well as the magnitude of the actual mark difference for the pair. It is clear from Figure 3 that the estimated mark differences from the simulated studies were all close to the originally estimated mark difference. While the shape of the distribution differed according to which particular large error was simulated, in all cases the estimated differences were very close to the originally estimated difference d in absolute terms. Although the values appear spread out along the x-axis, the scale is very fine-grained, and all estimates from the simulated studies were within a fifth of a mark of the originally estimated value for d.

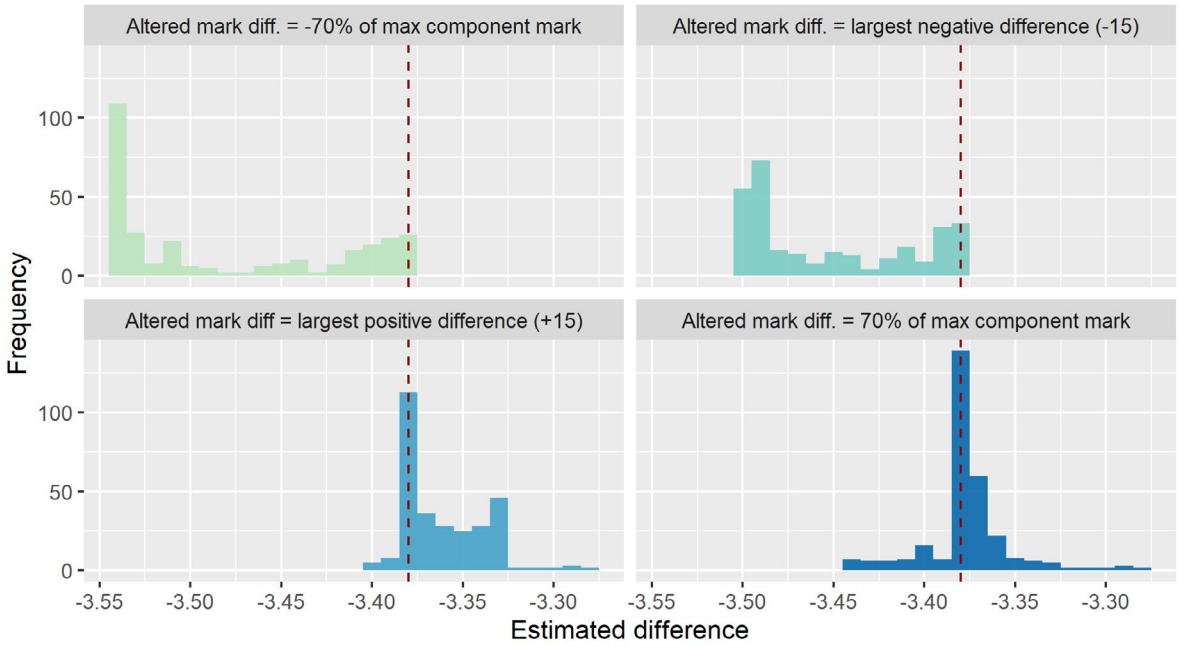

Figure 3: Estimated difficulty differences from simulating one-off large marking errors, Assessment 2 (reference line shows the original estimated *d*, before simulation of marking error).

For all four of the original SP studies, the estimated difficulty differences and associated standard errors changed little when a single large marking error was simulated. Table 1 summarises the range of outcomes from the simulated SP studies, in comparison with the original SP study results. In all cases, the estimated difference d was very close to the estimated difference from the original study (i.e., before simulating a large marking error), and the standard errors of estimates increased only moderately.

Table 1: Summary of single large marking error simulations in comparison with original studies.

Component	Max. mark	Pairs (*n*)	Original study *d* (SE)	Min *d* (SE) from simulated studies	Median *d* (SE) from simulated studies	Max *d* (SE) from simulated studies
Assessment 1 (English Language)	80	292	1.38 (1.14)	1.25 (1.10)	1.43 (1.16)	1.72 (1.36)
Assessment 2 (Maths)	50	300	-3.38 (0.49)	-3.54 (0.48)	-3.39 (0.49)	-3.28 (0.54)
Assessment 3 (Sociology)	75	289	-2.61 (1.33)	-3.04 (1.29)	-2.65 (1.35)	-2.43 (1.53)

How many marking errors can occur before SP fails?

The second set of simulations made use of data from the same three real-life SP studies (Table 1), but this time simulated the occurrence of multiple moderately large marking errors. The purpose of these simulations was to explore how many such errors could occur before the SP method broke down.

For each original SP study, the simulations were carried out as follows:

1. Randomly select *n* pairs from the original study.
2. Add a fixed "marking error" *e* to the observed mark difference for each of these pairs[2].
3. Re-run the SP analysis.
4. Retain/calculate:
    a. whether the analysis flatlined or not
    b. whether the 95 per cent confidence interval for the estimated overall difference d includes the value estimated in the original study (pre-error d)
    c. the difference between the estimated d and the value estimated in the original study (pre-error d).

These steps were carried out for two values of "marking error" *e*, equal to 10 per cent of component maximum mark, and -10 per cent of component maximum, and 1000 studies were simulated for each combination of conditions. For each *n* investigated, 6000 simulations were therefore carried out (3 original studies x 2 values of "marking error" x 1000 repetitions). The simulations were carried out at values of *n* from 10 up to 150. To give some context to the "marking errors" in this set of simulations, the value of 10 per cent of component maximum mark was chosen as a marking error that would be moderately large but of the magnitude that could occur in real life assessment scenarios. In the case studies presented by Ofqual (2014, pp. 31–32), for example, which analyse mark changes following enquiries about results for Geography A level and French A level, 1 per cent of mark changes made were of a magnitude of 10 per cent of the total raw marks, or larger.

As in the simulation of single large marking errors, the results showed that the SP studies were robust. Figure 4 shows the proportion of simulated studies for which the 95 per cent confidence interval for d contained the original (pre-error) estimate, according to number of marking errors introduced. The proportion only fell below 1 once the number of pairs of scripts containing marking error was large: around 50 pairs (out of 300) for Assessment 2, and only after 75 pairs for the other two studies.

Figure 5 shows how the estimated overall differences d deviated from the original (pre-error) estimates as more marking errors were introduced. The mean size of these deviations (expressed as percentages of component maximum) increased linearly, and at a moderate rate: for simulations adding marking errors to 50 pairs of scripts, the average deviation from original d was up to 2 per cent of the component maximum mark. The size of the deviations in d increased at a

2 This method (adding "error" to pairs of scripts selected on the basis of their original marks) results in a set of script pairings with a different distribution of mark differences than if scripts were selected on the basis of observed marks that already included large marking errors. Most obviously, the added "error" may cause mark differences to fall outside the original range of mark differences. The method used here should produce similar or worse outcomes (i.e., overestimate rather than underestimate risk).

higher rate when the sign of the marking errors introduced matched the sign of the original difference d. For Assessment 1, for example, the originally estimated overall difference was positive (1.38 marks), and the mean size of deviations in d increased faster for marking errors of +10 per cent than for marking errors of -10 per cent. The results show that, across all cases studied, at least 25 script pairs would need to contain such a marking error in order to alter the estimate by at least 1 per cent of the maximum.

None of the simulated SP studies resulted in flatlining.

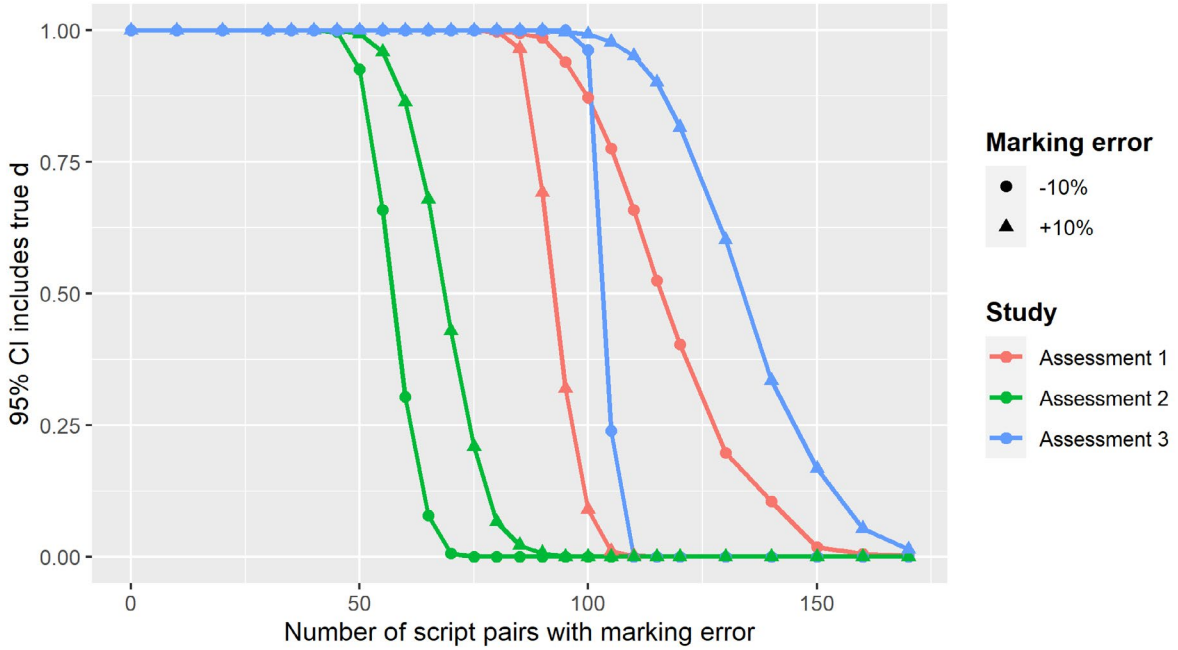

Figure 4: Proportion of simulated SP studies on target, by number of script pairs containing marking error.

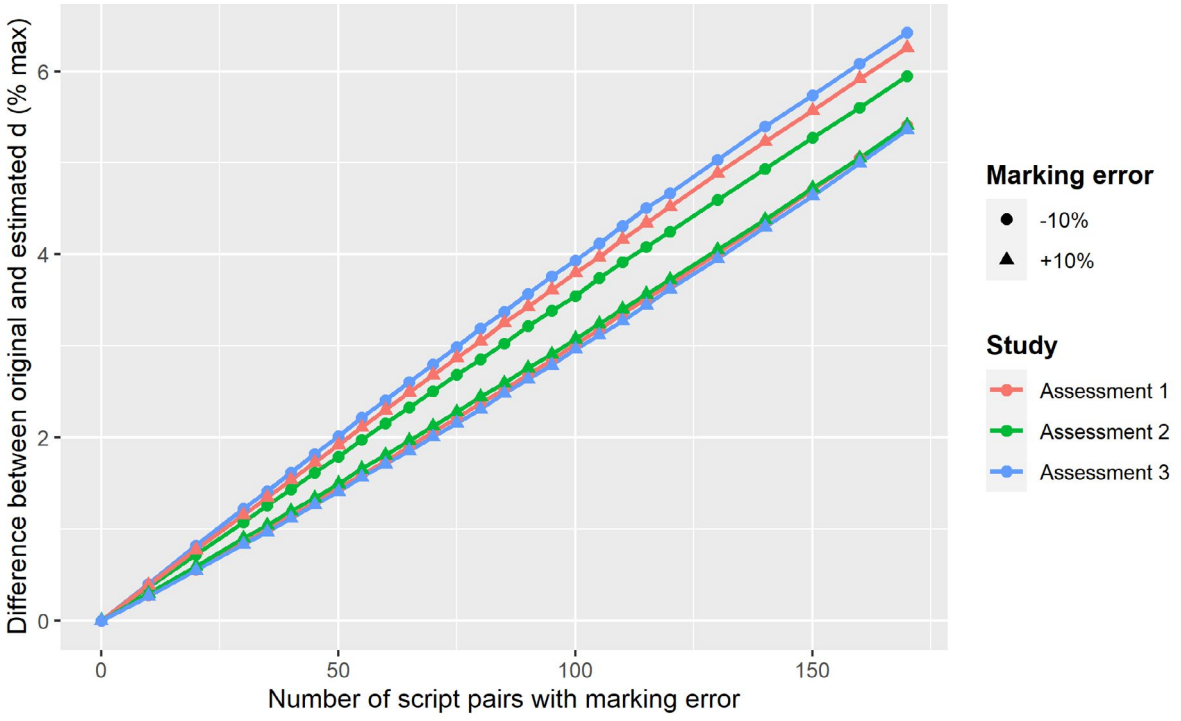

Figure 5: Mean absolute difference between original and estimated *d*, as a percentage of maximum mark.

The impact of progressively degrading marking quality

The third and final research question was addressed by simulating a large number of SP studies from scratch. The purpose of these simulations was to investigate the impact on SP results of progressively degrading quality of marking. These simulations differed from the earlier simulations by focusing on the overall relationship between awarded marks and script quality, rather than on single large marking errors or a fixed number of over- or under-marked scripts. The simulated SP data therefore needed to contain plausible data on mark differences, and simulated comparative judgements for these mark differences, and we needed to simulate how the relationship between mark differences and judgements would vary if marking quality decreased.

In the section below, we first explain the model relating marks and CJ measures, and how this relationship varies with marking quality. We then describe how the relationship between mark differences and CJ judgements can be expected to vary as marking quality varies, which is the foundation for the simulations. Finally, we explain how specific values for the key parameters were chosen.

Simulating SP study data

Throughout this section, we label all marks as x_i and all true CJ measures as θ_i. The CJ measures θ_i are the holistic measures of script quality that would result from analysing the outcomes of paired script comparisons using a Bradley-Terry model (Bradley & Terry, 1952). By "true" CJ measures, we mean the CJ measures if they were measured without error (i.e., with an extremely large number of comparisons for each script). The CJ measures are on a logit scale, which means that the difference between two scripts' measures $(\theta_j - \theta_i)$ is equal to the log of the odds of script *j* being judged higher quality than script *i* in any single paired comparison. For the time being we ignore differences in difficulty between different versions of assessments that may be included in a CJ exercise.

Following the approach in Benton and Elliott (2016) and Bramley and Gill (2010) we assume that over the range of interest[3], the relationship between marks and CJ measures can be summarised in the form:

$$\theta_i = \beta x_i + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$. There are two parts to the relationship between marks and measures:

1. First is "σ" (the standard deviation of the normally distributed residuals), which expresses the extent to which scripts with the same mark may have different "true" CJ measures. This might be because marking and CJ in fact measure slightly different constructs – so that even if scripts were marked perfectly and even if we included each script in a huge number of pairwise comparisons, we still wouldn't achieve a perfect correlation between marks

3 As previously noted, SP studies – like other CJ studies – exclude scripts from the extremes of the mark distribution, where the linear regression relationship would be affected by the floor and ceiling effects of the fixed total mark range.

and measures. It might also be a result of marking error. Higher levels of marking error will result in a larger value of σ.

2. Second is the coefficient "β", which expresses the strength of the association between marks and the decisions made by judges. Even if σ =0 (meaning that CJ and marking measure the same construct, and there is no marking error) it is likely that individual judges' decisions will not correspond perfectly to the marks that were awarded. However, the higher β is, the stronger the association. The CJ measures (θ_i) are constrained to have a mean of zero and the unit size (the logit) is directly related to judges' discrimination between scripts: a difference between two scripts of zero logits means that the scripts are equally likely to be judged superior (i.e., the probability of script *j* being judged superior is 0.5), and a difference of 1 logit between scripts means that the higher-rated script is judged superior with a probability of just over 0.7. When the coefficient β is higher, the same level of discrimination (e.g., a 1 logit difference) is associated with a smaller mark difference than when the coefficient β is lower. Alternatively, seen from the perspective of marks, a higher value of β means that the same mark difference between scripts corresponds to a higher probability of the higher-rated script being judged superior than when β is lower. Assuming a fixed level of reliability for CJ itself, then lower marking reliability would result in a lower value for β.

The logistic model describing CJ judgements tells us that for true CJ measures θ_i, the probability of script *j* being judged superior to script *i* is:

$$P(j \text{ beats } i) = \frac{\exp(\theta_j - \theta_i)}{1 + \exp(\theta_j - \theta_i)}$$

Via transformation and substitution (shown step by step in the Technical Appendix), we can re-express the likelihood of a script *j* "win" in terms of the mark difference between the scripts compared, and the two parameters β and σ reflecting marking quality. This means that the slope of the logistic regression linking mark differences and the probability of judges deciding script *j* is superior to script *i* (for brevity, written "GLM slope" from here on, for Generalised Linear Model slope) is given by:

$$GLM\ slope = \frac{1.7\beta}{\sqrt{1.7^2 + 2\sigma^2}}$$

Once a plausible value for the GLM slope is chosen, this value, together with a suitable set of mark differences (consistent with the methods used to sample pairs of scripts for an SP study) is sufficient to simulate a dataset of SP judgements.

Choosing values for β and σ

To simulate the SP studies, we estimated values of β and σ using data published in the appendices of Curcin et al. (2019). Using data from 20 pairwise comparison

studies[4] we used linear regression to estimate the relationship between marks (as a percentage of the total) and CJ measures of the holistic quality of papers. Across all 20 linear regressions, the median coefficient for β was 0.13 and the median value of σ (the standard deviation of estimated residual variance in the regression) was 1.3. Using these values with the GLM slope formula above, the expected slope of the logistic regression between mark difference (as a percentage of maximum mark) and judges' decisions would be the following:

$$GLM\ slope = \frac{1.7 * 0.13}{\sqrt{1.7^2 + 2(1.3)^2}} = 0.09$$

For the purposes of simulating a realistic SP study, 0.09 is therefore a reasonable value for the GLM slope of simulated data. The focus of this research, however, was on the extent to which the SP method would be robust to decreases in marking quality. Higher levels of marking error will result in higher values for σ and lower values for β, and hence smaller slope values.

In general, then, the simulations explored slope values lower than 0.09. In order to link slope values to a (quantified) degradation in marking quality, we calculated the values of σ and β (and hence, slope) that would correspond to specific decreases in marking reliability for a given SP study. This was done via substituting in marks x_i^* with added marking error, in the following way:

$$x_i^* = \rho x_i + \epsilon_i \sqrt{\left(1 - \rho^2\right)}$$

where var(ϵ_i)=400, and ρ represents the level of marking degradation – so that if the original marks x_i are perfectly reliable, then x_i^* would have marking reliability of ρ^2. The variance of ϵ_i in these simulated error-affected marks is set at 400 because a typical CJ study includes scripts with marks between 20 and 90 per cent of the available total, and roughly evenly spread (as reflected by the simulation steps in the next section). If the script marks are evenly spread between 20 and 90 per cent, their variance will be approximately 400[5].

Now, $\theta_i = \beta x_i + \varepsilon_i = \beta\rho x_i^* + \left(\beta\epsilon_i\sqrt{\left(1-\rho^2\right)} + \varepsilon_i\right)$, so we can use new values of beta and sigma to calculate the likely slope of the GLM, using $\beta^* = \rho\beta$ and $\sigma^{*2} = \sigma^2 + 400\beta^2\ (1 - \rho^2)$.

Simulation steps

We simulated a large number of SP studies from scratch. Varying levels of marking quality degradation were simulated via varying the GLM slope linking

4 Data from the rank ordering, "pinpointing" paired comparisons, and teacher paired comparison studies were not included. The rank ordering studies were analysed as pairs (and this may not be accurate), while the "pinpointing" and PCJ with teachers do not reflect Cambridge University Press & Assessment's normal practice.

5 The variance of a single-variable uniform distribution between values min and max is $\frac{1}{12}\ (max - min)^2$, see https://reference.wolfram.com/language/ref/UniformDistribution.html

mark differences and probability of script 2 "win". As shown above, this slope is dependent on both marking reliability and the strength of the relationship between marks and CJ measures.

The steps carried out were the following:

1. Simulate data from an SP "study" comparing two assessments (form A and form B) with 300 pairs of scripts, on a 0–100 mark scale:
    a. Simulate 300 script 1 marks from form A, sampled uniformly between 20 and 90 marks.
    b. Simulate 300 script 2 marks from form B, in the same way as for the script 1 marks.
    c. Pair "scripts" from form A and form B and calculate the mark difference (script 2–script 1). Scripts were paired so that the mark differences were approximately normally distributed around zero and 90 per cent of mark differences lay between -30 and 30 marks. The maximum mark differences ranged from ~-60 to ~60.
    d. Simulate a paired comparison decision for each pair of scripts by random draw from a binomial distribution, with the probability of success (script 2 "win") for each judgement being given by the logistic function of g*(mark difference - d), where g is the GLM slope and d is the overall difficulty difference (in marks) between form A and form B.
2. Analyse the simulated SP data using logistic regression.
3. Retain/calculate:
    a. estimated difficulty difference in marks (d)
    b. 95 per cent confidence intervals for d
    c. whether the estimated slope flatlined or not.

A simulated study was recorded as flatlining whenever either boundary of the 95 per cent confidence interval for the predicted probability of a script 2 "win" failed to intersect the line y=0.5 within the study's range of mark differences. This would occur, for example, if all lower bounds of the 95 per cent confidence intervals were lower than 0.5, or all upper bounds of the intervals were above 0.5, for the study's range of mark differences.

The simulation steps were carried out for two levels of true mark difference between form A and form B (d=0 and d=10), and for slope values ranging from 0.01 to 0.09, with 5000 "studies" simulated per condition. The entire set of simulations was then repeated for a simulated study size of 150 pairs of scripts, to give a sense of the impact on smaller SP studies. A true mark difference of 10 marks (i.e., 10 per cent of the mark range) between the two assessments compared is a fairly large difference, and the purpose of simulating at d=10 was to explore outcomes for a difference at the upper end of normal variation.

Results

As GLM slope value decreased, that is, the simulated relationship between mark difference and judges' decisions weakened, the proportion of simulated SP studies that flatlined increased (Figure 6). The size of confidence intervals for the estimated d increased (Figure 7) along with the variability of estimates, although estimates for d remained on target until the very lowest slope values (Figure 8). In comparison with the full SP studies using 300 pairs, outcomes deteriorated sooner when the number of pairs per simulated SP study was reduced to 150. Outcomes were better for the SP studies with no overall difficulty difference (d=0) than for those with an overall difference of 10 marks.

At a slope value of 0.09 (the GLM slope estimated from the median values for β and σ in the Ofqual studies), the simulated SP studies were successful: none flatlined, and the difficulty difference was estimated with confidence intervals comfortably smaller than 10 marks for 300-pair studies, and smaller than 15 marks for 150-pair studies. The "worst" values[6] in the Ofqual studies reported by Curcin et al. (2019) were σ =2 and β =0.09, which produced an estimated GLM slope of 0.046. The simulated SP study outcomes for a slope of this magnitude were slightly worse: Figure 6 shows that flatlining occurred for such studies with a non-zero difficulty difference, and for the 150-pair studies; and Figure 7 shows that 95 per cent confidence intervals for the estimated difficulty difference had a median size of around 10 marks, for the "best case" condition of no overall difficulty difference and *n*=300 pairs.

6 The study producing these values was AS Psychology specification 2, paper 1, year 1.

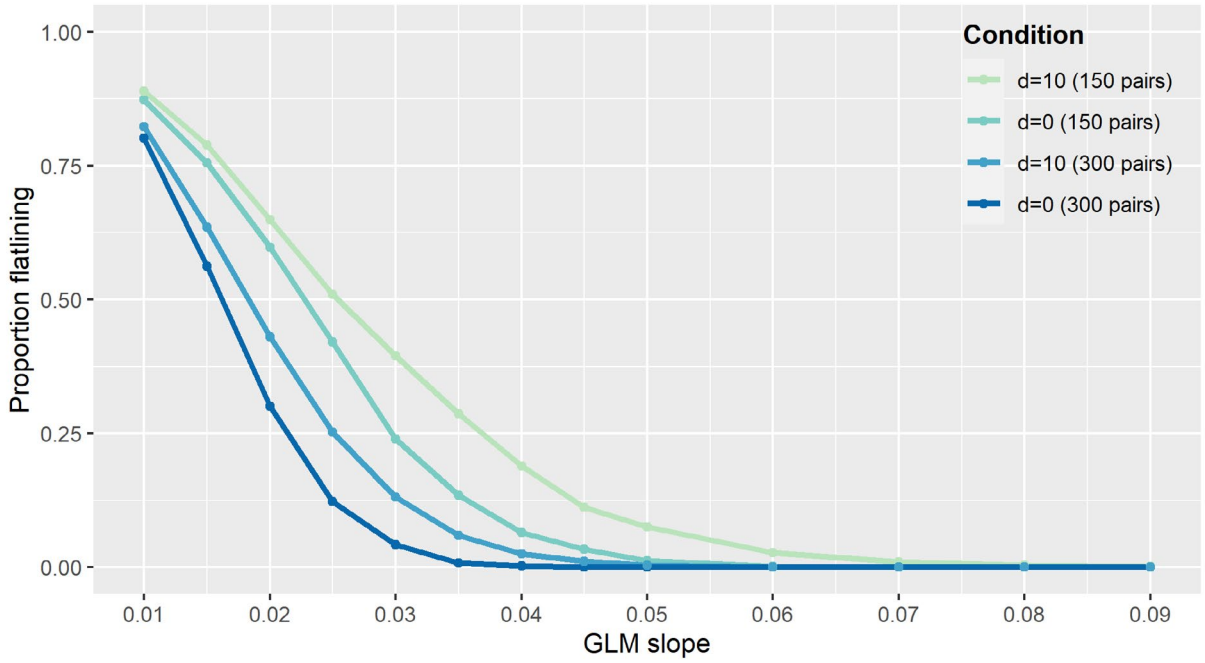

Figure 6: Proportion of SP studies flatlining.

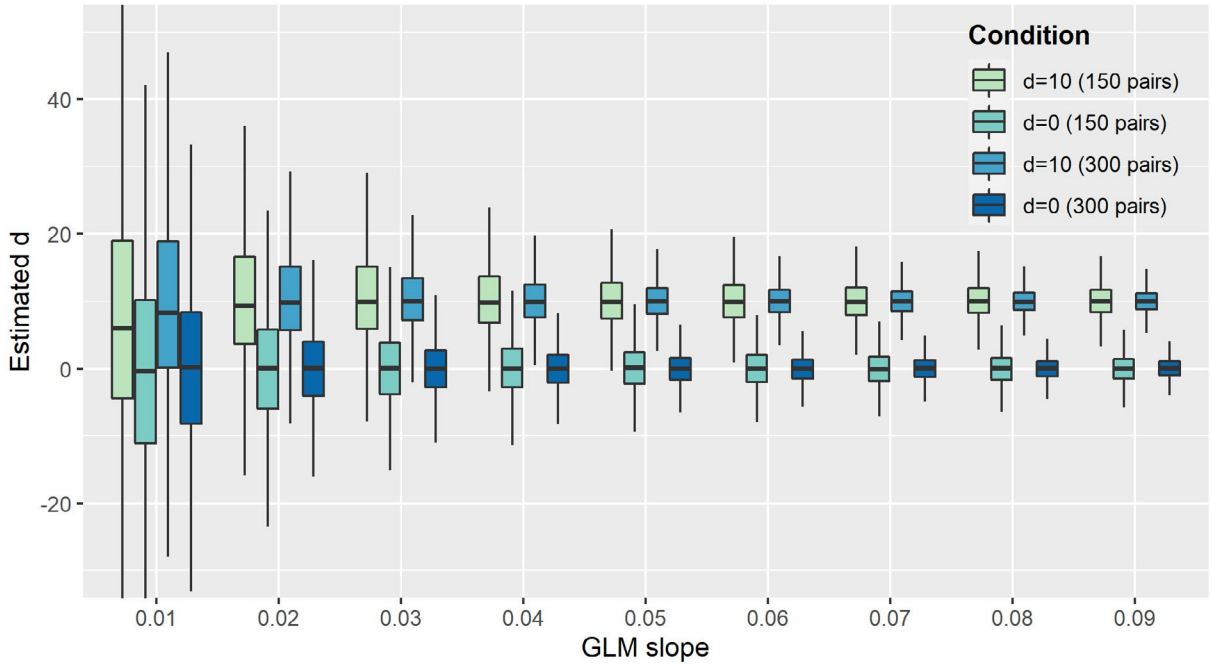

Figure 7: Distributions of confidence interval sizes (outliers not plotted; y-axis cropped).

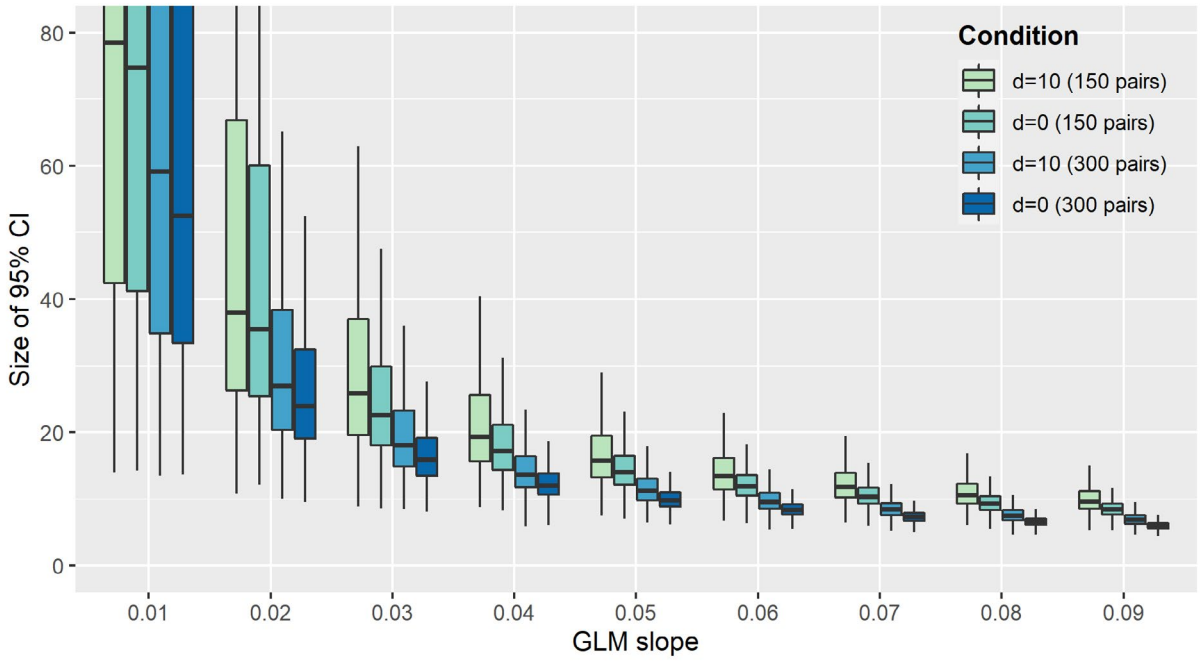

Figure 8: Distributions of estimated overall differences (outliers not plotted).

To consider the impact of specific levels of marking degradation, we simulated reductions in marking quality from the starting point of these "worst" values from the Ofqual studies (σ =2 and β =0.09, producing estimated GLM slope 0.046). Since the assessments in the Ofqual studies represent a selection of typical actual GCSE and AS level assessments (not chosen to be in any way extreme), this is a reasonable starting point to consider. Table 2 shows the estimated GLM slopes corresponding to increasing levels of marking degradation from this starting point, and the corresponding percentages of studies that flatlined at each level. Table 3 shows the median confidence interval sizes at each level of marking degradation.

For ρ =0.9, a modest degradation in marking that would result in a slope value of 0.04 (and corresponds to reliability of 0.81, if the original marking reliability is assumed to have been perfect), less than 1 per cent of 300-pair studies flatlined when the difficulty difference was zero, and 2.4 per cent flatlined when the difficulty difference was 10 marks (Table 2). The widths of the 95 per cent confidence intervals for d were about 12 marks and 14 marks respectively (Table 3). When the number of pairs per simulated SP study was reduced to 150, however, the same levels of marking degradation resulted in much more problematic outcomes: 7.3 per cent of studies flatlined when the difficulty difference was zero, and almost 20 per cent when the difference was 10 marks (Table 2). The median confidence interval sizes, meanwhile, were around 17 and 19 marks respectively (Table 3).

For higher levels of marking degradation, the results of the simulated SP studies deteriorated further. At marking degradation of ρ =0.775 (corresponding to reliability of 0.60, if original marking assumed perfect) the estimated GLM slope was 0.032. Of the 300-pair SP studies simulated with this slope, 1.8 per cent and 9.5 per cent flatlined (for d=0 and d=10 respectively), and median confidence interval sizes were around 15 and 17 marks. In the simulated 150-pair studies, the proportions flatlining were 18.4 per cent (d=0) and 32.9 per cent (d=10), and the median confidence interval sizes were around 21 and 24 marks.

Table 2: Flatlining in simulated SP studies, by condition (*n*=5000 studies per condition).

Marking degradation (ρ)	Revised slope	Percentage of studies that flatlined			
		300-pair studies		150-pair studies	
		d = 0	d = 10	d = 0	d = 10
1	0.046	0.04	0.50	2.42	9.76
0.975	0.045	0.00	1.10	3.02	11.68
0.95	0.043	0.02	1.40	4.76	14.64
0.925	0.041	0.06	1.78	5.24	16.90
0.9	0.040	0.06	2.38	7.26	19.94
0.875	0.038	0.34	3.20	9.06	22.90
0.85	0.037	0.48	4.26	11.32	23.62
0.825	0.035	1.14	5.78	13.66	27.06
0.8	0.034	1.34	7.90	16.54	30.26
0.775	0.032	1.78	9.48	18.38	32.90
0.75	0.031	2.64	11.36	23.02	36.90
0.725	0.030	4.32	13.40	25.16	40.96
0.7	0.028	6.02	16.00	30.88	43.24
0.65	0.026	10.22	22.34	38.42	50.10
0.6	0.024	15.78	30.32	45.64	54.56
0.55	0.021	24.24	38.02	54.42	61.02
0.5	0.019	34.02	46.50	62.04	68.32

Table 3: Confidence interval sizes for d in simulated SP studies, by condition (*n*=5000 studies per condition).

Marking degradation (ρ)	Revised slope	Median size of 95% CI for d			
		300-pair studies		150-pair studies	
		d = 0	d = 10	d = 0	d = 10
1	0.046	10.46	11.98	14.96	16.81
0.975	0.045	10.81	12.49	15.46	17.56
0.95	0.043	11.23	12.97	16.08	18.19
0.925	0.041	11.79	13.33	16.63	18.84
0.9	0.040	12.13	13.92	17.33	19.91
0.875	0.038	12.59	14.34	18.10	20.51
0.85	0.037	13.08	14.85	18.91	20.93
0.825	0.035	13.65	15.70	19.43	21.98
0.8	0.034	14.15	16.11	20.45	22.92
0.775	0.032	14.83	16.69	21.10	23.72
0.75	0.031	15.37	17.53	22.12	24.84
0.725	0.030	16.06	18.12	23.01	26.20
0.7	0.028	16.68	18.92	24.10	27.06
0.65	0.026	18.29	20.76	26.53	29.78
0.6	0.024	19.86	22.67	29.12	32.07
0.55	0.021	22.33	25.13	32.80	35.41
0.5	0.019	25.03	28.19	36.62	40.41

Conclusions

The research has two main conclusions. The first is that the SP method appears robust to single large marking errors, and to fairly large marking errors in quite high proportions of sampled scripts. The simulations of one-off large marking errors indicated that the estimated overall difficulty difference was affected only slightly, with numerical values very close to the originally estimated value, and only slightly increased standard errors. The simulations of multiple marking errors with a magnitude of 10 per cent of the component maximum mark, meanwhile, showed that the SP method failed only when large numbers of sampled scripts were affected – starting at around 50 out of 300 pairs. Similarly, it would take the occurrence of such marking errors in at least 25 out of 300 pairs to alter the estimated difference in difficulty between two tests by even 1 per cent of the maximum. These results are both reassuring and encouraging – the SP analyses proved robust even in the face of unusually large and unusually numerous errors in the script evidence, increasing confidence that the outcomes of SP analyses can be used to support maintenance of standards.

The second conclusion is that the SP method is more vulnerable to a general degradation of marking quality. The final set of simulations showed how SP analyses became problematic when the relationship between marks and CJ measures weakened – from whatever cause. The simulations showed that a non-extreme degradation in marking quality, from the starting point of values seen in published CJ studies, could result in failure of analysis (flatlining) and/or very wide confidence intervals around estimated differences. Importantly, the simulations showed that the deterioration in outcomes occurred much sooner for smaller studies (n=150 pairs), and when the actual overall difference between assessments was non-zero. Reducing the sample size in operational SP studies would, therefore, represent a substantial increase in risk to the success of the SP analysis and its ability to provide useful information for standard maintaining. In practical terms, SP analyses for a reduced sample size such as n=150 pairs have a much higher likelihood of failure than SP analyses for a full study of n=300 pairs, which would more than offset the advantages associated with choosing to run a smaller study. The actionable recommendation from this finding, therefore, is to avoid reducing sample sizes in operational SP studies for standard maintaining.

References

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

Benton, T., & Elliott, G. (2016). The reliability of setting grade boundaries using comparative judgement. *Research Papers in Education*, *31*(3), 352–376. https://doi.org/10.1080/02671522.2015.1027723

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

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

Bramley, T., & Gill, T. (2010). Evaluating the rank ordering method for standard maintaining. *Research Papers in Education, 25*(3), 293–317. https://doi.org/10.1080/02671522.2010.498147

Camilli, G. (1994). Origin of the Scaling Constant *d* = 1.7 in Item Response Theory. *Journal of Educational Statistics*, *19*(3), 293–295. https://doi.org/10.3102/10769986019003293

Curcin, M., Howard, E., Sully, K., & Black, B. (2019). *Improving awarding: 2018/2019 pilots* (Ofqual/19/6575). Ofqual. https://www.gov.uk/government/publications/improving-awarding-20182019-pilots

Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error*, Technical Report No. 15 (Office of Naval Research Contract No. 25140, NR-342-022). Stanford University: Applied Mathematics and Statistics Laboratory.

Ofqual. (2014). *Review of quality of marking in exams in A Levels, GCSEs and other academic qualifications: Final report*. Ofqual. https://www.gov.uk/government/publications/quality-of-marking-in-gcses-and-a-levels

R Core Team. (2021). *R: A language and environment for statistical computing*. https://www.R-project.org/

Technical Appendix: simulating SP studies

This appendix shows the derivation of the equation that expresses the slope of the logistic regression linking mark differences and judges' comparative judgements in terms of the two parameters β and σ reflecting marking quality.

As stated in the main article, we assume that over the range of interest, the relationship between marks and CJ measures can be summarised in the form:

$$\theta_i = \beta x_i + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$.

Representing the true CJ measures as θ_i, we know that the probability of script *j* being judged superior to script *i* is:

$$P(j \text{ beats } i) = \frac{\exp(\theta_j - \theta_i)}{1 + \exp(\theta_j - \theta_i)}$$

At this point, we can usefully approximate this logistic model using the probit link function. This relies on a transformation constant of 1.7 as recommended by Haley (1952) and described in Camilli (1994). Having made this approximation, we can use:

$$P(j \text{ beats } i) = \Phi\left(\frac{\theta_j - \theta_i}{1.7}\right)$$

where Φ is the cumulative distribution function for the standard normal distribution.

Combining this equation above with the equation describing the relationship between marks and CJ measures, we get the following:

$$P(j \text{ beats } i) = \Phi\left(\frac{\beta\left(x_j - x_i\right) + \varepsilon_j - \varepsilon_i}{1.7}\right)$$

We can define $\epsilon_{ji} = \varepsilon_j - \varepsilon_i$, and since the difference of two independent normally distributed variables also follows a normal distribution, we know that $\epsilon_{ji} \sim N(0, 2\sigma^2)$.

Next, we think about the nature of the probit function. What it explicitly does is calculate the following: $\Phi(y) = P(z \leq y)$, where $z \sim N(0, 1)$.

Thus,

$$\Phi\left(\frac{\beta\left(x_j - x_i\right) + \epsilon_{ji}}{1.7}\right) = P\left(z \leq \frac{\beta\left(x_j - x_i\right) + \epsilon_{ji}}{1.7}\right) = P\left(\left(1.7z - \epsilon_{ji}\right) \leq \beta\left(x_j - x_i\right)\right)$$

By the properties of normal distributions, we know that $(1.7z - \epsilon_{ji}) \sim N(0, 1.7^2 + 2\sigma^2)$. By realising that by dividing $(1.7z - \epsilon_{ji})$ by $\sqrt{1.7^2 + 2\sigma^2}$ gets us back to a variable with a standard normal distribution we can see that:

$$P(j \text{ beats } i) = \Phi\left(\frac{\beta\left(x_j - x_i\right) + \epsilon_{ji}}{1.7}\right) = \Phi\left(\frac{\beta\left(x_j - x_i\right)}{\sqrt{1.7^2 + 2\sigma^2}}\right)$$

Finally, by reversing the approximation between the logistic and normal distributions we saw to begin with (i.e., multiplying the numerator of the subject of the function by 1.7), we can say:

$$P(j \text{ beats } i) = \frac{\exp\left(\frac{1.7\beta\left(x_j - x_i\right)}{\sqrt{1.7^2 + 2\sigma^2}}\right)}{1 + \exp\left(\frac{1.7\beta\left(x_j - x_i\right)}{\sqrt{1.7^2 + 2\sigma^2}}\right)}$$

This means that the slope of the logistic regression linking mark differences and the probability of judges deciding script j is superior to script i is given by:

$$GLM \text{ slope} = \frac{1.7\beta}{\sqrt{1.7^2 + 2\sigma^2}}$$

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

Carmen Vidal Rodeiro and Lucy Chambers (Research Division)

Introduction

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

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

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

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

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

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

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

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

The overarching research question in this study was:

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

The following sub-research questions were also investigated:

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

Method

Portfolios

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

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

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

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

Judges

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

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

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

Research task

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

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

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

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

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

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

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

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

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

Findings

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

CJ data

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

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

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

Judge statistics

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

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

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

Table 1: Judge fit statistics.

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

CJ measures

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

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

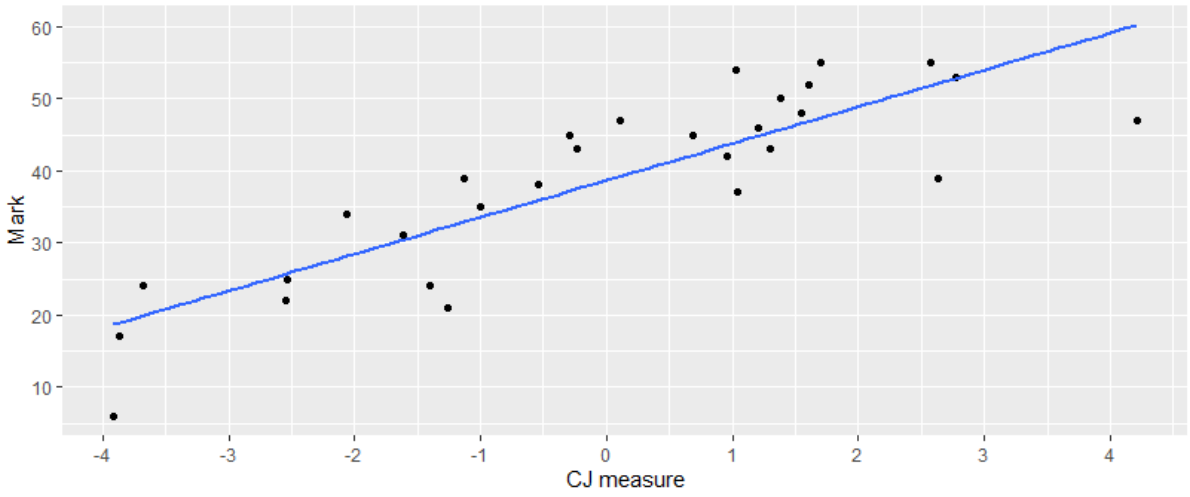

Figure 1: Portfolio marks vs. CJ measures.

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

Time required to complete the task

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

Table 2: Time required to complete the task.

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

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

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

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

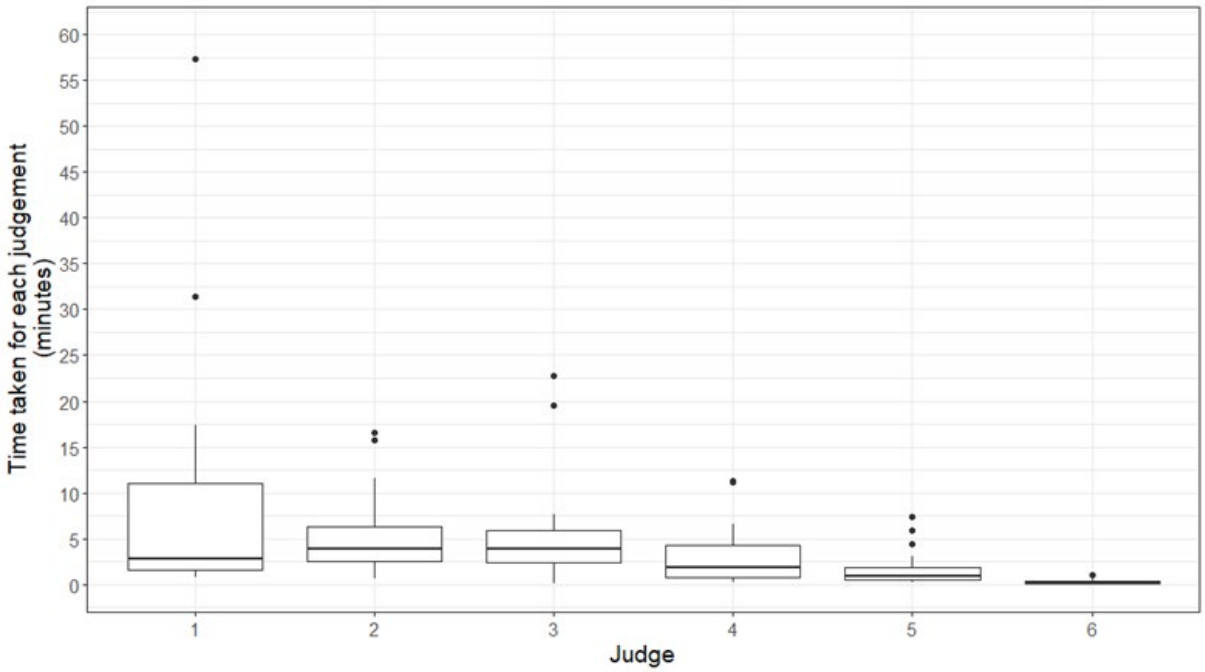

Figure 2: Time per judgement (minutes), by judge.

Observations

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

> Judge 1:

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

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

> Judge 3:

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

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

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

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

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

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

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

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

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

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

Survey and interviews

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

Use of the CJ tool and navigation

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

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

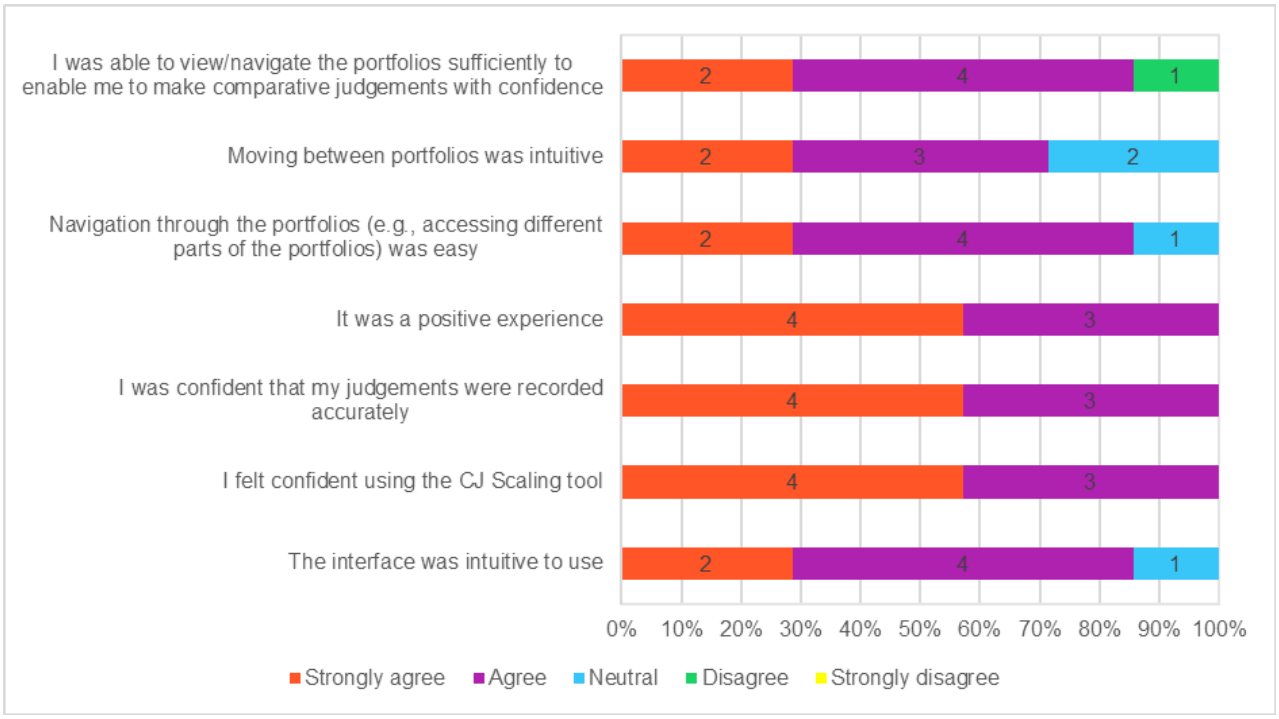

Figure 3: Judges' experiences with the CJ Scaling tool.

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

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

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

The CJ tool and judgements

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

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

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

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

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

Making holistic judgements

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

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

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

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

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

Features on which judgements were based

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

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

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

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

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

Time taken

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

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

> This method was quicker than traditional moderation.

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

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

Comparison of moderation methods

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

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

Positive sentiment

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

Neutral sentiment

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

Negative sentiment

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

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

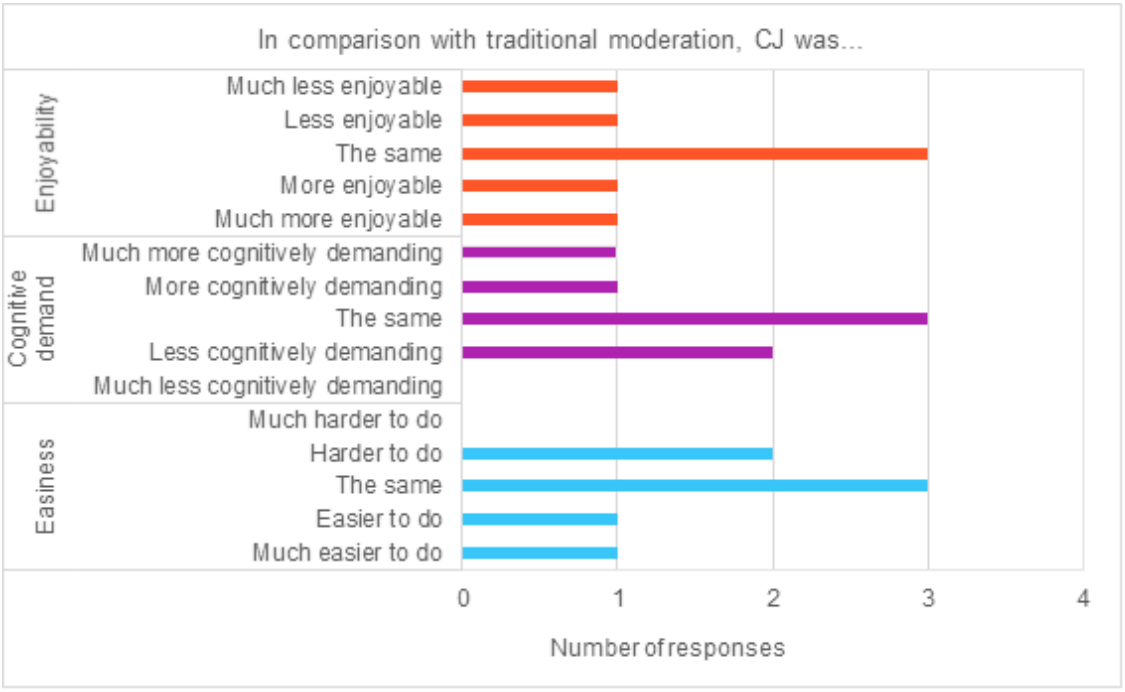

Figure 4: Responses to the prompt "In comparison with traditional moderation, Comparative Judgement was...".

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

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

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

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

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

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

Conclusions

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

Is CJ a practically feasible method for moderating NEAs?

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

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

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

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

On what basis do moderators make their judgements?

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

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

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

Are moderators confident making Comparative Judgements on portfolios?

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

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

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

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

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

Recommendations and further research

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Navigating the debate around the future of education

Over the course of the past year, Cambridge University Press & Assessment has published more than 50 blogs to help chart the future of teaching, learning and assessment around the world. These evidence-based outputs have been curated around a set of outline principles for the future of education, reflecting that it is often not possible to discuss one aspect of an education system without considering the potential implications for other parts. Explore all our outputs and contribute to the debate at:

cambridge.org/future-of-education

Why exam boards are a 'Public Good'

What have GCSEs ever done for us?

Prior to pandemic, was England getting worse?

Implementing education reform

Diary insights into teaching during lockdown

Why don't we just put our high stakes exams on screen?

Cambridge Assessment Network

Become a member of a global assessment community

Enhance your status as a recognised assessment expert by becoming a member of the Cambridge Assessment Network. Develop professionally as you take your learning to the next level. Advance through our evidence based professional standards. Work towards an award in recognition of your learning.

You'll belong to a worldwide network of professionals committed to getting assessment right.

Member benefits

- 10% discount on all training courses
- Access to professional awards
- 'Ask the Expert' sessions
- Member only online community
- Member newsletter

Launching Spring 2022

Register your interest:
www.cambridgeassessment.org.uk/membership

Cambridge Assessment Network

Postgraduate Advanced Certificate in Educational Assessment

Stand out as an assessment expert

Our University of Cambridge Postgraduate Advanced Certificate in Educational Studies: Educational Assessment is a practice-based qualification. The course is taught part-time over 15 months, giving you 90 credits at Master's level (Level 7).

You'll learn through a mix of online learning and four day schools led by experts from Cambridge Assessment and the University of Cambridge Faculty of Education.

On successful completion of the course you'll be awarded a Postgraduate Advanced Certificate in Educational Studies (PACES).

With this qualification you'll:

- develop an in-depth understanding of assessment in a supportive academic environment
- work on small-scale enquiries relevant to your own professional practice
- apply research methodologies to your professional context
- learn from other students and build a network beyond your specialist area.

Starting September 2022
Application deadline: 24 June
cambridgeassessment.org.uk/pgca

Research News

Lisa Bowett (Research Division)

Publications

The following reports and articles have been published since *Research Matters*, Issue 32:

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

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

Johnson, M., Fitzsimons, S., & Coleman, V. (2021). Development challenges in challenging contexts: A 3-stage curriculum framework design approach for Education in Emergencies. *Prospects*. https://doi.org/10.1007/s11125-022-09601-0

Vidal Rodeiro, C. L. (2021). The role of Cambridge Technicals in the post-16 qualifications landscape. *Cambridge University Press & Assessment Research Report*. Cambridge University Press & Assessment.

Conference presentations

The AEA-Europe Conference 2021 took place online from 3 to 5 November 2021, with the theme 'Assessment for Changing Times: Opportunities and Challenges'. Our researchers presented a total of six papers:

Use of eportfolios to assess hard-to-measure constructs and what makes a good examiner team leader. Emma Walland and Stuart Shaw.

The use of test accommodations in high-stakes assessments. Tori Coleman and Martin Johnson.

Equal opportunity or unfair advantage? The use of test accommodations in high-stakes assessments. Carmen Vidal Rodeiro and Sylwia Macinska.

Metaphors and the psychometric paradigm. Tom Bramley.

Does removing tiering from high-stakes examinations reduce the size of attainment gaps? Matthew Carroll.

Evaluating the simplified pairs method of standard maintaining using comparative judgement. Tom Benton.

Jackie Greatorex and Tori Coleman also presented online at the International Conference of Education, Research and Innovation, which was held remotely on the 8–9 November.

Greatorex, J., & Coleman, T. (2021, November 8–9). *Defining and understanding decolonisation in the context of the 14 to 18 curriculum in England* [Conference session]. International Conference of Education, Research and Innovation, online.

‘Changing texts’: An international review of research on textbooks

We submitted to the Swedish Textbook Authors Association our wide-ranging review of research literature on the form and function of textbooks and digital learning materials, authored by Melissa Mouthaan, Sinead Fitzsimons, Fiona Beedle and Tim Oates. Increasingly, textbooks are the focus of comment in discussions of means of reducing teacher workload. Sweden is examining improvement strategy, and the Association is concerned that the role of textbooks has been overlooked, and seeks also to understand the nature of the growing market in digital resources. While many articles cite decline in the volume of research on textbooks, we found a wealth of literature – and included consideration of popular discussion about textbooks as well as academic literature. Download our review here.

Blogs

The following blogs have been published since *Research Matters*, Issue 32:

Bramley, T. (2022, February 7). Does giving advance notice disadvantage lower-attaining students?

Greatorex, J., & Vitello, S. (2022, January 26). What is competence? A shared interpretation of competence to support teaching, learning and assessment.

Hughes, S. (2021, November 18). What do we mean by ‘digital assessment’?

Hughes, S. (2022, January 6). Why don’t we just put our high stakes exams on screen?

Johnson, M. (2021, December 24). Diary insights into teaching during lockdown.

Oates, T. (2021, November 4). What is the cost of massive change?

Oates, T. (2021, November 11). Prior to pandemic, was England getting worse?

Sharing our research

We aim to make our research as widely available as possible. Listed below are links to the places where you can find our research online:

Journal papers and book chapters: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/journal-papers-and-book-chapters/

Research Matters (in full and as PDFs of individual articles): https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/

Conference papers: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/conference-papers/

Research reports: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-reports/

Data Bytes: www.cambridgeassessment.org.uk/our-research/data-bytes

Statistics reports: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/statistical-reports/

Blogs: www.cambridgeassessment.org.uk/blogs/

Insights (a platform for sharing our views and research on the big education topics that impact assessment around the globe): https://www.cambridgeassessment.org.uk/insights/

Our YouTube channel: https://www.youtube.com/channel/UCNnk0pi7n4Amd_2afMUoKGw contains Research Bytes (short presentations and commentary based on recent conference presentations), our online live debates #CamEdLive, and podcasts.

You can also learn more about our recent activities from Facebook, Instagram, LinkedIn and Twitter.

Contents / Issue 33 / Spring 2022

Cambridge University Press & Assessment
Shaftesbury Road
Cambridge
CB2 8EA
United Kingdom

researchprogrammes@cambridgeassessment.org.uk
www.cambridge.org