Research Matters / 33
A Cambridge University Press & Assessment publication

ISSN: 1755-6031

Journal homepage: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/

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

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

To cite this article: 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.

To link this article: https://www.cambridgeassessment.org.uk/Images/research-matters-33-a-summary-of-ocrs-pilots-of-the-use-of-comparative-judgement-in-setting-grade-boundaries.pdf

Abstract:

The rationale for the use of comparative judgement (CJ) to help set grade boundaries is to provide a way of using expert judgement to identify and uphold certain minimum standards of performance rather than relying purely on statistical approaches such as comparable outcomes. This article summarises the results of recent trials of using CJ for this purpose in terms of how much difference it might have made to the positions of grade boundaries, the reported precision of estimates and the amount of time that was required from expert judges.

The results show that estimated grade boundaries from a CJ approach tend to be fairly close to those that were set (using other forms of evidence) in practice. However, occasionally, CJ results displayed small but significant differences with existing boundary locations. This implies that adopting a CJ approach to awarding would have a noticeable impact on awarding decisions but not such a large one as to be implausible. This article also demonstrates that implementing CJ using simplified methods (described by Benton, Cunningham et al, 2020) achieves the same precision as alternative CJ approaches, but in less time. On average, each CJ exercise required roughly 30 judge-hours across all judges.

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.

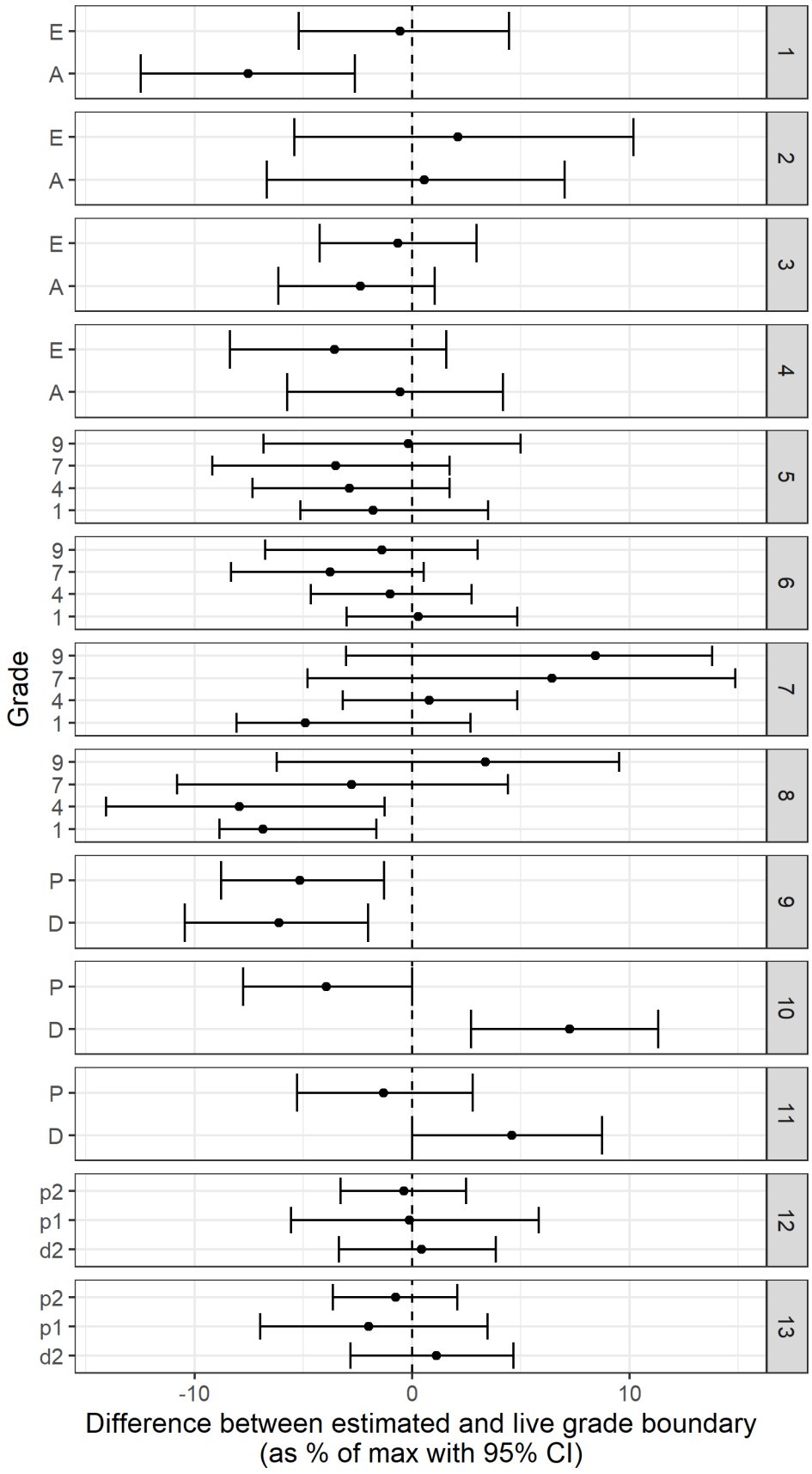

Figure 1: Plots of differences between estimated and live grade boundary for the 13 OCR pilot studies. 95 per cent confidence intervals for the differences are also shown.

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.

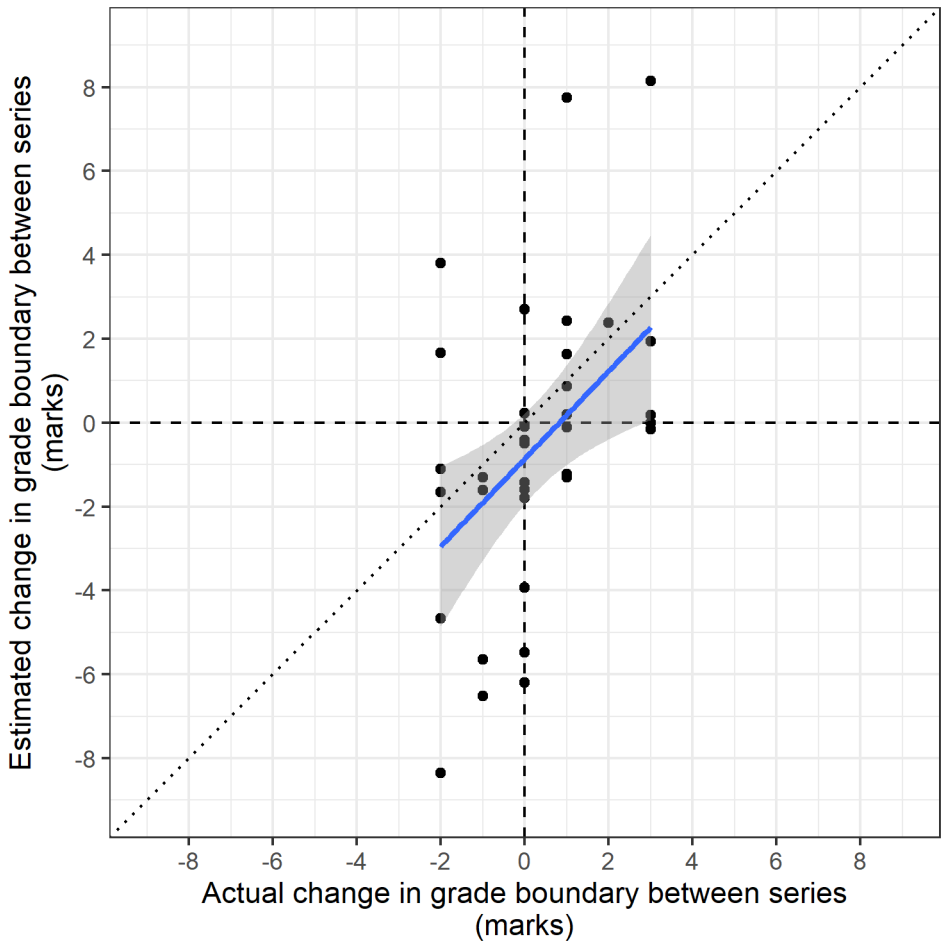

Figure 2: Relationship between actual and estimated changes in grade boundaries between series.

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.

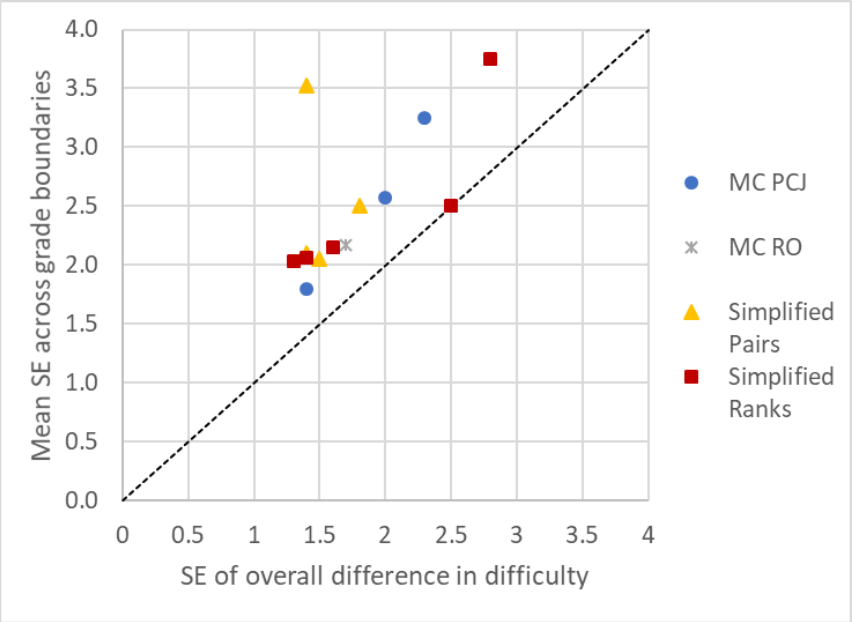

Figure 3: Comparison of the SE of the overall difference in difficulty with the mean SE of the grade boundary estimates.

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.

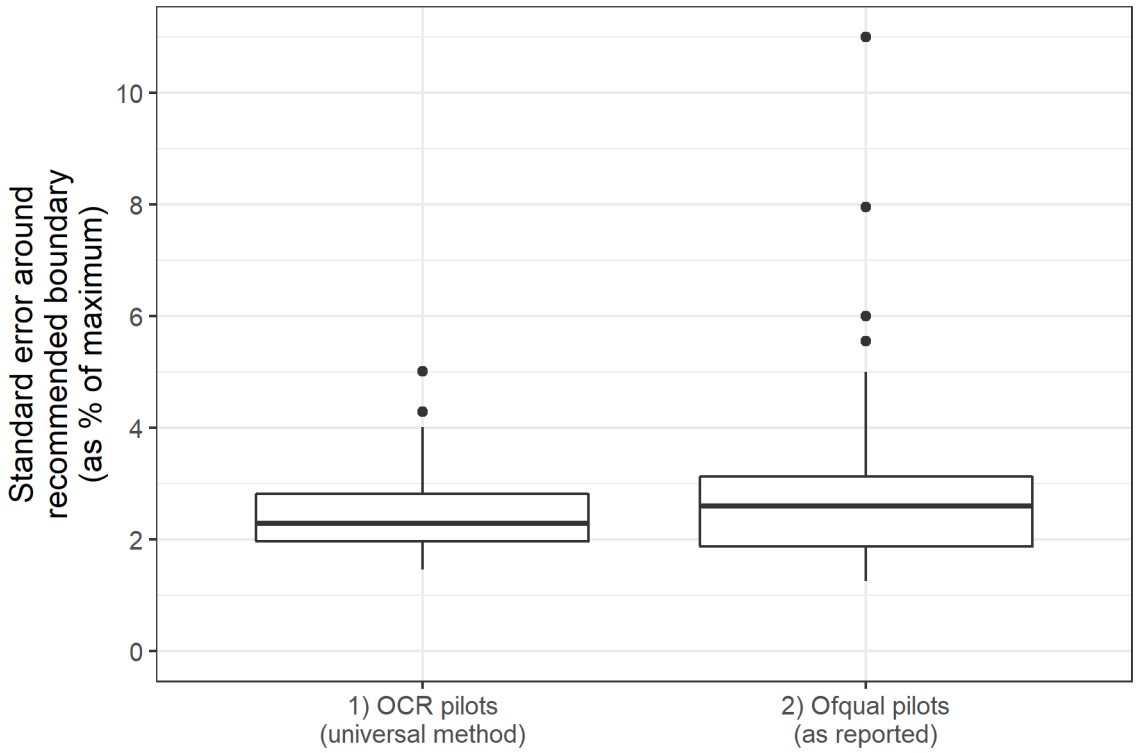

Figure 4: Boxplot of standard errors (as % of maximum available mark) around estimated grade boundaries in OCR's recent pilots and in pilots reported by Ofqual in Curcin et al. (2019).

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.

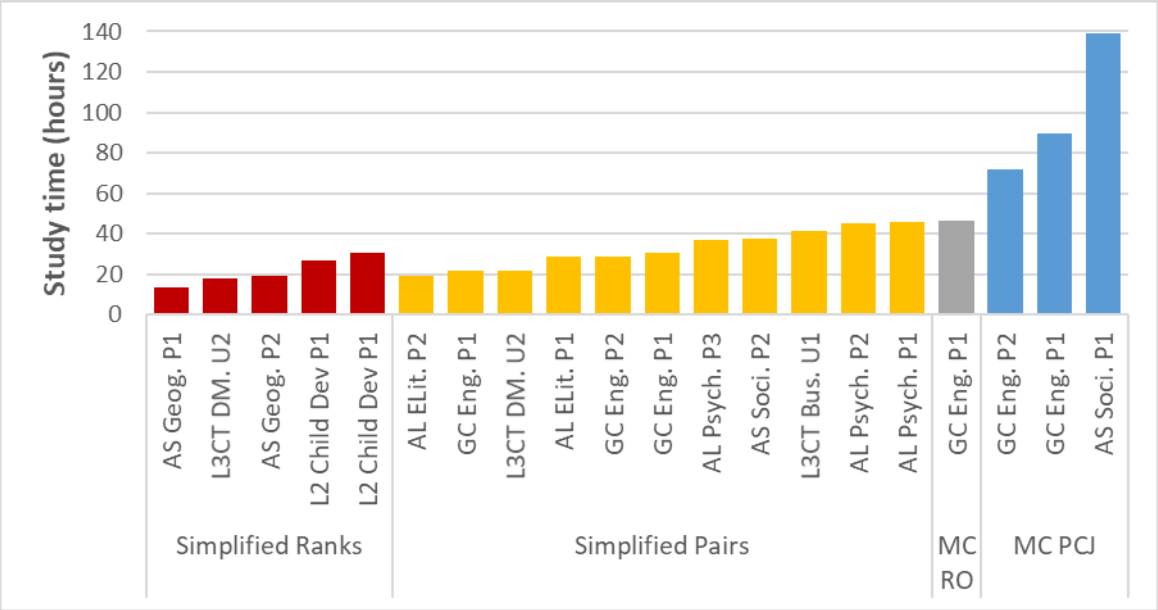

Figure 5: Time spent on each study, grouped by study type.

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.

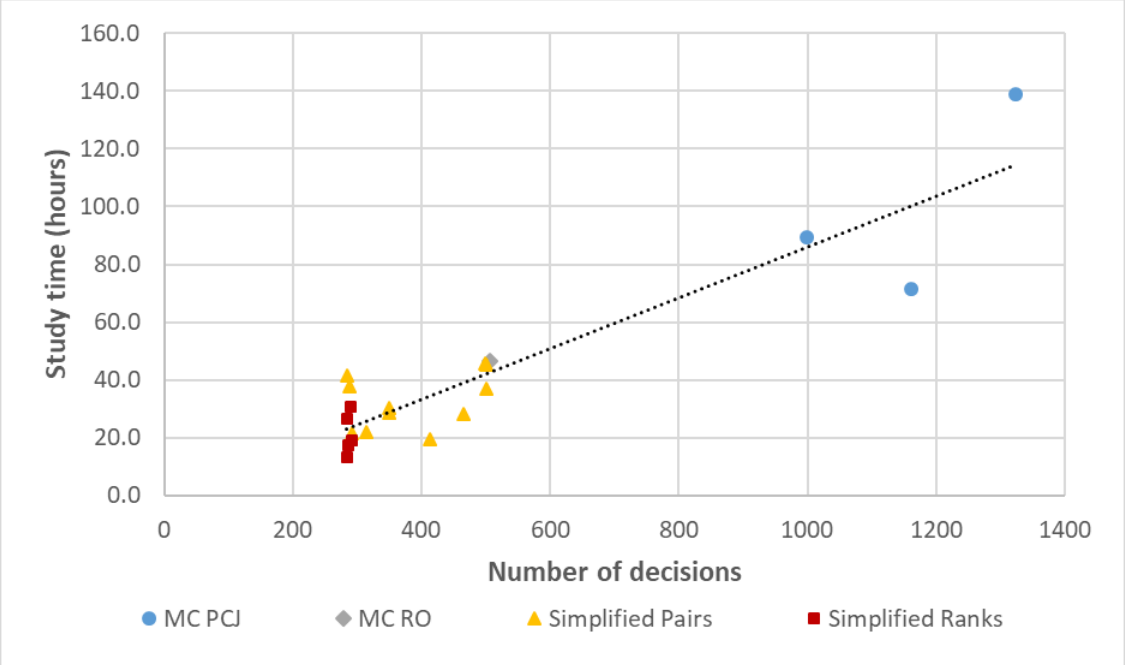

Figure 6: Study time, by total number of decisions made.

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.

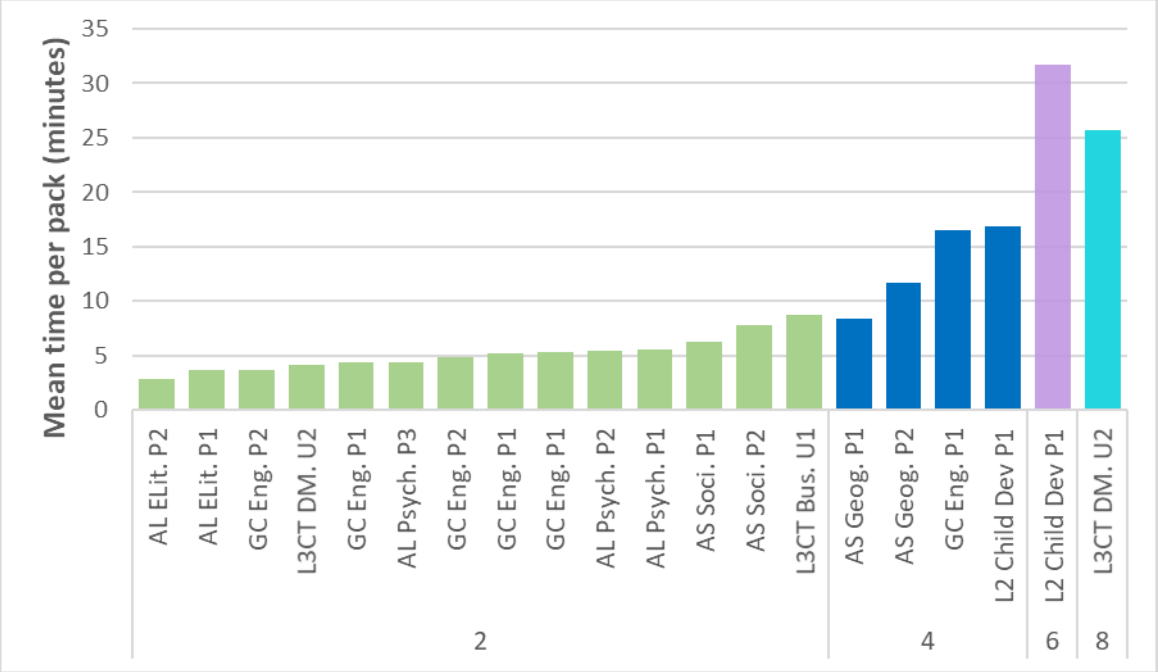

Figure 7: Mean time spent per pack, by pack size.

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.