Cambridge Assessment

Does comparative judgement of scripts provide an effective means of maintaining standards in mathematics?

Research Report

Tom Benton
Tony Leech
Sarah Hughes

11th November 2020

Author contact details:

Tom Benton
Assessment Research and Development,
Research Division
Cambridge Assessment
The Triangle Building
Shaftesbury Road
Cambridge
CB2 8EA
UK

benton.t@cambridgeassessment.org.uk

http://www.cambridgeassessment.org.uk

As a department of Cambridge University, Cambridge Assessment is respected and trusted worldwide, managing three world-class examination boards, and maintaining the highest standards in educational assessment and learning. We are a not-for-profit organisation.

Cambridge 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: researchdivision@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.

How to cite this publication:

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, UK: Cambridge Assessment.

Acknowledgements

We would like to thank all of the judges that took part in this study for their help as well as the Assessment Processing Team for managing the CJ exercises and judges. Thanks also to Will Hornby for advising on and supporting this work.

Contents

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 in one year than in another. Specifically, it tends to mean activities associated with setting examination grade boundaries. Benton et al (2020) describes a method to inform grade boundaries called *simplified pairs* that allows us to map scores between test versions using comparative judgement (CJ). The method requires fewer paired comparisons than previously suggested methods of using CJ to set grade boundaries (e.g. Bramley, 2005) and simplifies the approach to analysis. This method will be the main focus of this report.

The majority of trials of using CJ in awarding have focussed upon examinations requiring extended responses such as essays. For example, the trials described in Curcin et al (2019) focussed upon English language, English literature, psychology and media studies. Many earlier studies of CJ in awarding also focussed on English language and the humanities (see, for example, Gill, Black & Bramley, 2007). However, several studies exploring the effectiveness of CJ approaches in mathematics and the sciences have also been completed in the past. For example, Bramley (2012) applied such an approach to a chemistry examination, Raikes, Scorey & Shiell (2008) and Novakovic & Suto (2010) applied the method to biology examinations, and Jones et al (2016) used CJ to explore changes in A level mathematics standards over time.

More broadly, other studies have examined the validity of using CJ as an alternative to marking for mathematics assessments (for example, Jones and Inglis, 2015). Relatedly, Jones, Swan and Pollitt (2015) investigated the use of CJ for the assessment of problem-solving, concluding that this was a valid approach, but was perhaps particularly appropriate with tasks designed specifically to be judged comparatively.

None of these studies revealed any major difficulties in applying CJ to explore standards over time for examinations of short answer questions. In particular, they have shown that when CJ is applied to awarding science and mathematics assessments:

- CJ decisions tend to be reasonably consistent with the marks awarded to candidates[1]
- Different groups of expert judges tend to provide similar results

Whilst previous studies have shown that the same CJ techniques for awarding can be applied both in examinations requiring a few extended responses from each candidate (such as English literature), and in those consisting of a large number of shorter questions (such as mathematics), it has not been determined whether such techniques yield accurate answers. In particular, a recurring concern within assessment circles is that raised by Baird (2007) that "examiners cannot adequately compensate in their judgements of candidates' work for the demands of the question papers" (page 142). In particular, the concern is that, as suggested by Good & Cresswell (1988), subject experts will be more impressed by a candidate achieving a high score on an easy paper than by a candidate achieving a (statistically equivalent) lower score achieved on a harder paper. One experiment presented by Benton et al (2020) for an English literature examination suggested that such a concern need not always apply. However, no published research has directly explored the applicability of this concern to mathematics or science exams when grade boundaries are set using CJ.

[1] Usually evidenced using a correlation between marks and CJ measures of script quality. However, this is sometime also evidenced directly (Raikes et al, 2008) or by the R^2 values of a regression of CJ quality measures on awarded grades (Jones et al, 2016).

In order to evaluate any method of setting grade boundaries, it is necessary to compare the outcomes with an empirical estimate of equivalent grade boundaries, such as would be provided by statistical equating (Kolen & Brennan, 2004). In other words, we need to compare a mapping of scores from one assessment to another as indicated by a CJ method, with a 'known' mapping based on statistical equating of actual assessment results. Crucially, however, in order to establish these 'known' differences in difficulty, it is essential that at least one of the following three conditions be satisfied in the study: that some of the students take both of the assessments being equated; that some of the same items are present in both assessments; or, that some students taking either assessment version also take a common anchor assessment.

Following the research presented by Benton et al (2020) we will use equating between assessments completed by the same students to provide a benchmark against which the results of a CJ exercise can be compared. Specifically, our study will compare the results of a simplified pairs approach to the outcome of statistical equating using the equipercentile method (Kolen & Brennan, 2004). This will give us a direct estimate of the accuracy of the simplified pairs method. In order for results to specifically address the concern raised by Baird (2007) and Good & Cresswell (1988), the experiment presented in this report focusses on creating a mapping of scores between two assessments known to have substantially different difficulties.

Choice of assessments

The first step in this study involved selecting two mathematics assessments to be used for the analysis. The assessments chosen for this study were created by splitting a single 100-mark GCSE Mathematics exam into two 50-mark examinations. The original full-length assessment for analysis was chosen as it was taken by a very large sample of students thus allowing us to undertake equipercentile equating as a comparator to the simplified pairs method. In addition, splitting the full-length paper into two half-length assessments was relatively easy to do, as separate questions were mostly printed on distinct pages of the exam booklet. This made it straightforward to create PDF images of candidate responses that were restricted to a particular set of questions relating to one of our half-length assessments.

An initial split of the original full-length assessment into two half-length assessments was created by taking alternate pages in the test booklet. After this, a few slight manual amendments were made to this initial split to ensure that both halves contained the same number of marks. Further details on the two half-length assessments are displayed in Table 1. Each half-length assessment contained 10 questions worth a total of 50 marks. The mean scores of each question were calculated based on the responses of 16,345 candidates and are also displayed. As can be seen, the total of these mean question scores indicates that Half 2 was roughly 5 marks harder than Half 1.

Table 1: Details of questions included in each half-length assessment in the study

	Mean question scores		Max question scores	
Question	**Half 1**	**Half 2**	**Half 1**	**Half 2**
Q1	3.34		4	
Q2	0.85		1	
Q3		4.32		7
Q4	7.86		9	
Q5		4.44		6
Q6	3.40		6	
Q7		3.69		6
Q8	2.02		5	
Q9	2.40		6	
Q10		1.89		5
Q11		1.13		4
Q12		3.15		4
Q13	4.30		5	
Q14		1.92		3
Q15	2.74		7	
Q16		1.32		3
Q17	1.22		3	
Q18		1.74		6
Q19	2.05		4	
Q20		1.64		6
Total	**30.19**	**25.22**	**50**	**50**

Collecting paired comparison data

Scripts were randomly sampled for use within a simplified pairs exercise. Different samples of students were used to provide script images for the Half 1 assessments and for the Half 2 assessments. For each half-length assessment, scripts with scores between 10 and 45 (out of 50) on the relevant half were selected with an approximately uniform distribution of marks within this range. Scripts from each half were randomly assigned to pairs subject to the restriction that the raw scores of each half-script within a pair had to be within 15 marks of one another.

Once the scripts had been chosen and paired, 6 experienced markers were recruited from OCR to act as judges of the exam scripts for the simplified pairs method. These judges were each issued with 50 pairs of half-scripts in the form of PDFs. Each pair contained one PDF constructed from the pages relating to Half 1 and one PDF containing pages relating to Half 2. The judges were then asked to indicate which of the two half-scripts "is better based on overall quality?". To help them make judgements, each judge was provided with the original mark schemes from the full-length examination. Judgements were made on-screen using the Cambridge Assessment Comparative Judgement Tool (https://cjscaling.cambridgeassessment.org.uk/). No marks or other annotations were visible to the judges on any of the scripts.

Full details of the design of the study (and the results of judgements too) are shown in Figure 1. The horizonal axis denotes the mark awarded to the first half-script within a pair, and the vertical axis the mark awarded to the second half-script. As can be seen, the scripts that were paired within the study covered a range of mark differences. Some pairs were included where the score on Half 1 was higher and some where the score on Half 2 was higher. Further descriptive statistics of the scripts that were included in the study are displayed in Table 2. As can be seen, although it was known by the researchers beforehand that Half 2 was substantially harder the Half 1, scripts with a similar distribution of scores were included from each half.

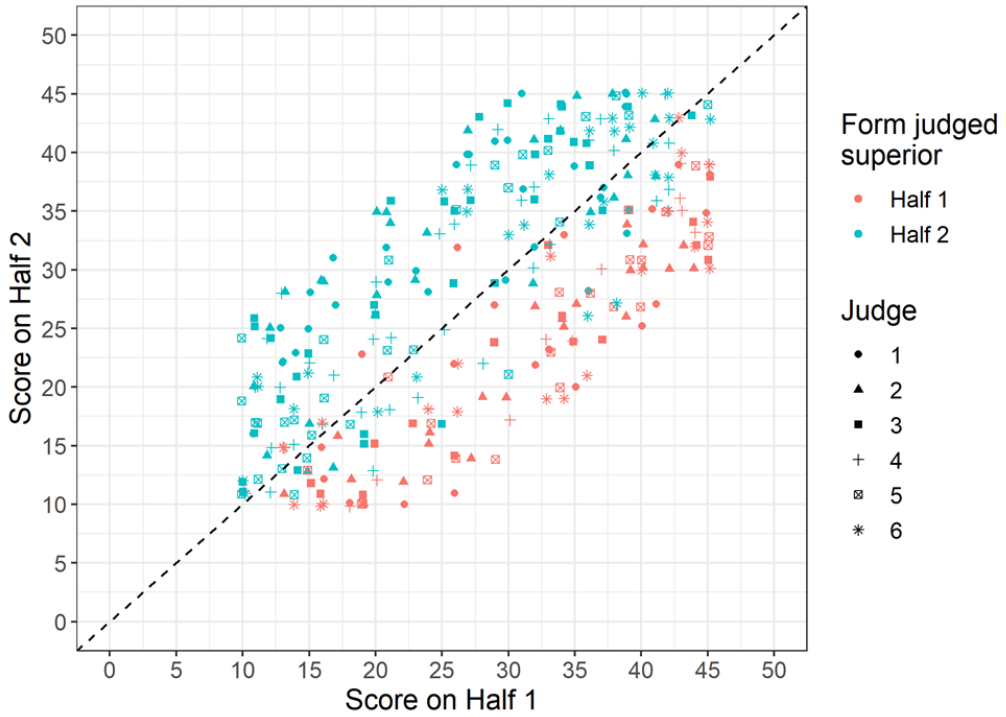

Figure 1: The design of the simplified pairs study. The locations of the points show which scores on Half 1 were paired with which scores on Half 2.

The colours of the points in Figure 1 denote whether judges selected the half-script from Half 1 or Half 2 as superior. Very broadly speaking, Figure 1 shows that the half-script with the higher score was more likely to be selected than that with the lower score. More detailed analysis of the relationship between assessment scores and judges' decisions will be shown later in this report.

Table 2: Descriptive information on scripts from each half included in the simplified pairs study

	Assessment – Half 1	**Assessment – Half 2**
Number of scripts	300	300
Mean score (out of 50)	27.8	27.8
Standard Deviation	10.8	10.4
Minimum score included	10	10
Maximum score included	45	45

Once each judge had made their decisions on all of their 50 allocated pairs, these decisions were analysed using the method described Benton et al (2020). The results of this analysis were then compared to the 'known' difference generated by the statistical equating of the actual scores given to each assessment when it was previously sat by a large number of students. This comparison subsequently provided us with a means of observing the accuracy of the results of the simplified pairs method.

Results

Overall difference in difficulty

To begin with, we use mean equating to show the average empirical difference in the difficulty of the two half-length assessments based on all students that completed the Maths assessment (i.e. not just those with scripts included in the simplified pairs study). Table 3 reiterates the point we have already made that Half 2 was substantially harder than Half 1. For the same set of pupils, the difference in mean scores was 5 marks (i.e. 10 per cent of the maximum number of marks available). The confidence interval for the differences in means indicates a high degree of precision in his estimate for the true differences in difficulty of the two half-length assessments.

Table 3: Results from mean equating of the actual scores of pupils taking the two half-length assessments.

	Half 1	**Half 2**
Number of pupils	16,345	16,345
Mean score	30.19	25.22
SD	9.78	9.71
Difference in means (Half 2 – Half 1)	-4.96	
SE of difference in means	0.04	
Confidence interval for difference in means	[-5.04, -4.88]	

Simplified pairs results

Having determined, using mean equating, that Half 2 is harder than Half 1 by about 5 marks on average, we next generate an estimate of the difference in difficulty using the data from expert judges as captured by the simplified pairs study. This analysis is shown in Figure 2.

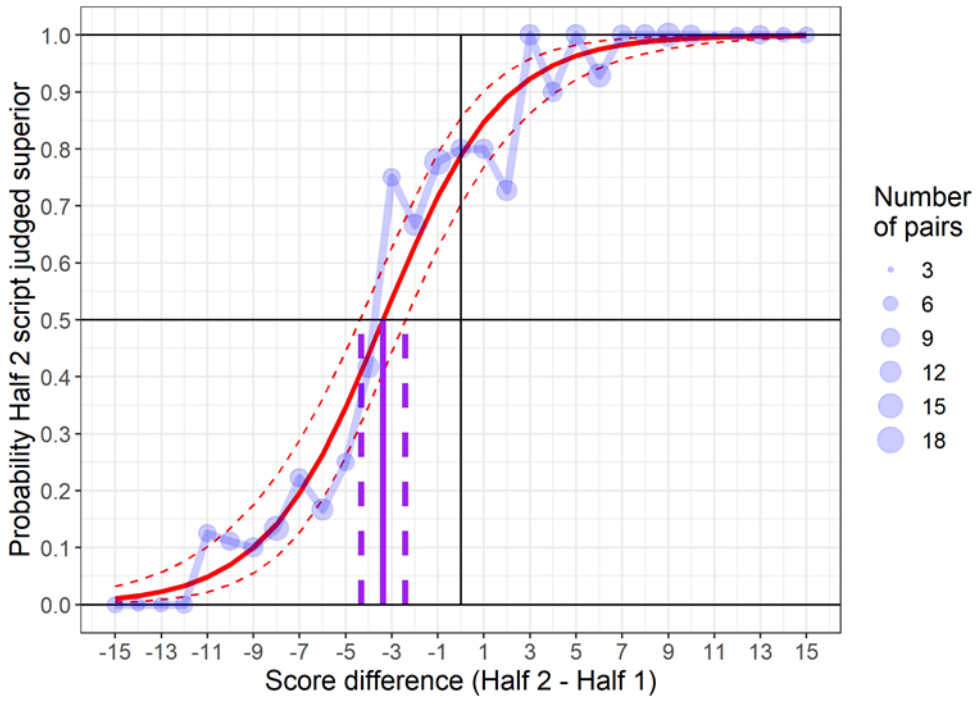

Figure 2: Graphical depiction of the results of using simplified pairs to gauge the relative difficulty of two assessment versions.

Figure 2 displays the results of analysis using the simplified pairs method to evaluate the relative overall difficulty of the two half-length assessments. For each difference in marks where comparisons were made (the horizontal axis), the blue line and points in Figure 2 show the proportion of pairs where the script from Half 2 was judged to be superior. Larger points depict mark differences where a larger number of pairs were included in the study. As can be seen, the proportion of pairs where Half 2 is deemed superior tends to increase with the extent to which the mark on the Half 2 script exceeds the mark on the Half 1 script.

It can be seen immediately from Figure 2 that judges tended to infer for themselves that Half 2 was harder than Half 1. For example, where both half-scripts had been awarded the same mark (i.e. the mark difference was zero), Half 2 was selected as superior for around 80 per cent of pairs. As such, the results from experts match the empirical analysis with regard to which of the two assessments was more difficult.

The formal analysis within a simplified pairs study is done using logistic regression. This is represented by the solid red line in Figure 2 which smoothly captures the relationship between mark differences and the probability of a Half 2 script being judged superior. The main purpose of analysis is to identify the mark difference where this fitted curve crosses the 0.5 probability. This happens at a mark difference of -3.4. This implies that a Half 2 script will tend to be judged superior to a Half 1 script wherever the mark difference exceeds -3.4. In other words, based on expert judgement we infer that Half 2 is 3.4 marks harder than Half 1. A 95 per cent confidence interval for this value (the dashed vertical lines) indicate that the judged difference in difficulty is between -2.4 and -4.3 marks. It should be noted that the size of this confidence interval, of essentially plus or minus a single mark, is very narrow compared to previous published examples of both simplified pairs (Benton et al, 2020) or other kinds of CJ in awarding (Curcin et al, 2019). This is due to the fact that the relationship between mark differences and judges' decisions depicted in Figure 2 is much stronger than in many previous applications – leading to increased precision.

We saw earlier (Table 3) that, based on a comparison of the mean scores of students that did both tests, the true overall difference in difficulty between the two assessments is 5 marks. As such, the estimated difference based on expert judgement (via simplified pairs) falls a little short of the true difference at only 3.4 marks. Furthermore, the confidence interval for the simplified pairs estimate does not overlap with the empirical difference. This indicates that we cannot dismiss the differences in results from mean equating and simplified pairs as being purely due to sampling error. Nonetheless, it is clear that the exercise has correctly identified the direction of difference in difficulty and that the estimate is close to the correct answer. Thus, the initial evidence here suggests that, whilst expert judgement cannot be expected to exactly replicate the results of statistical equating, it can still provide a useful source of evidence.

Judge fit

Before exploring more detailed results on the relative difficulty of the two tests across the score range, we first review the data on the differences between results for the 6 judges involved in the study. Statistics on the decisions made by each of the judges are shown in Table 4 and a visual depiction of how the fitted logistic curves differed between judges is shown in Figure 3.

All six judges selected Half 2 as being superior more than 50 per cent of the time and, similarly, each of the logistic curves for separate judges intersects the 0.5 probability line at mark differences below zero. This indicates a unanimous suggestion across judges that Half 2 was a harder assessment than Half 1.

Each judge displayed strong point biserial correlations between the differences in marks for the half-scripts being compared and the final decision they made about which was superior. In fact, the range of point biserials here (from 0.63 to 0.82) is substantially higher than the range shown for either example in Benton et al (2020). Specifically, the earlier report showed a range of point biserials between 0.24 and 0.53 for AS Sociology assessments, and between 0.33 and 0.62 for assessments from GCSE English Literature. This reiterates the strong relationship between mark differences and judges' decisions in the current study.

For consistency with other CJ studies, judge fit is also calculated using INFIT and OUTFIT (see Wright and Masters, 1990). None of the values are high enough to warrant serious concern over any of the judges. The highest values occur for the two judges (judges 1 and 6) with logistic curves (Figure 3) that suggest the smallest estimated difference in the difficulty of the tests. However, since the decisions within the exercise are to some extent a matter of opinion (see Benton et al, 2020) we tend to prioritise information from point biserials over judge "fit".

For interest we might note that across different judges the median time required per judgement was between 2 and 6 minutes. There was no obvious relationship between the speed at which judges made decisions and their fit.

Table 4: Judge fit and speed for each of the 6 judges.

Judge	Number of pairs	Proportion of pairs with Half 2 selected	INFIT	OUTFIT	Point biserial correlation between difference in marks and selecting half 2	Median time per judgement (minutes)
1	50	0.62	1.53	1.58	0.73	3.5
2	50	0.56	0.58	0.26	0.82	5.1
3	50	0.70	0.73	0.34	0.77	2.2
4	50	0.72	1.10	0.74	0.63	4.2
5	50	0.58	0.68	0.43	0.82	5.6
6	50	0.58	1.34	1.43	0.63	4.2

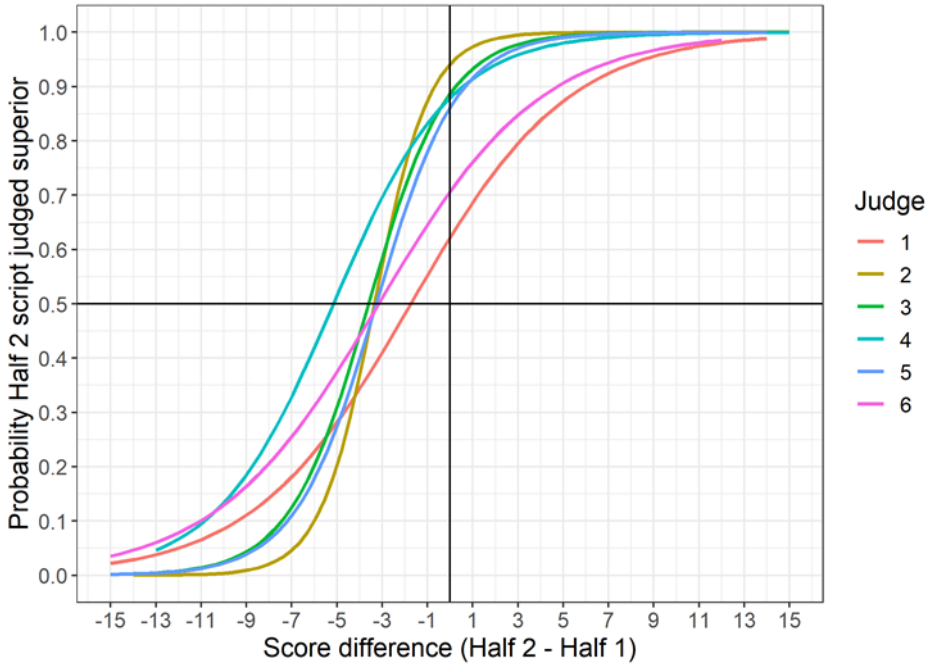

Figure 3: The relationship between differences in marks on half-length assessments within a pair and the likelihood of different judges selecting Half 2 as being superior.

Equating across the score range

Both equating and simplified pairs methods were used to estimate the mapping of scores from Half 1 to Half 2 across the available range. For empirical equating, this was done using the equipercentile approach using the R package *equate* (Albano, 2016). For simplified pairs, the mapping was created using logistic regression of judges' decisions on the (transformed) marks of each script in a pair. Further details of this latter method are provided in Benton et al (2020). A comparison of both mappings is provided in Figure 4.

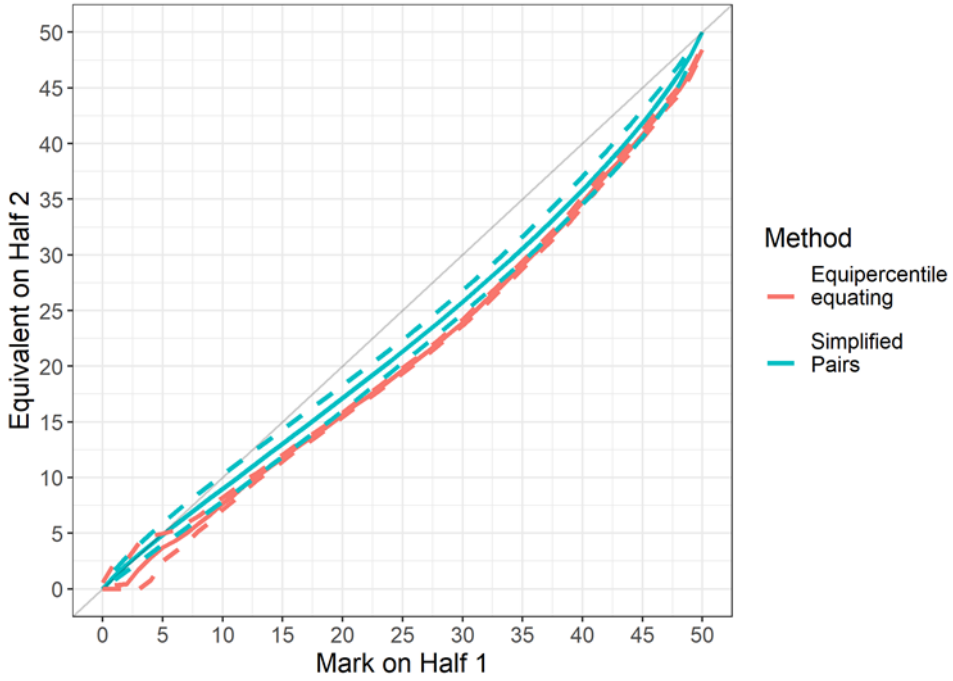

Figure 4: Results of both equating and simplified pairs across the score range.

Figure 4 shows us the equivalent mark on Half 2 for any given mark on Half 1 for both the equipercentile equating and the simplified pairs method. As is displayed in the graph's key, the

solid blue line represents the simplified pairs method and the solid red line represents equipercentile equating. The dashed lines represent the corresponding confidence intervals for each of the methods. For reference, a straight diagonal line of equality is also provided.

The results from empirical equating (the red line) confirm that Half 2 is harder than Half 1. This difference in difficulty is particularly visible for marks between 25 and 45 marks on Half 1. A similar pattern is also visible from the results of simplified pairs (the blue line) indicating a reasonable level of agreement between the two techniques.

Figure 5 presents the same results in a different way in order to allow a closer inspection of the differences between the two sets of score mappings. For this figure, the difference between scores on Half 1 and equivalent scores on Half 2 is displayed on the vertical axis. The extra detail in this chart allows us to see that estimates from the two methods are within two marks of one another across the entire score range. We can also see that, although the confidence interval for simplified pairs does not always encompass the estimate from equating, the lower confidence interval for simplified pairs is always within a mark of the estimate from statistical equating.

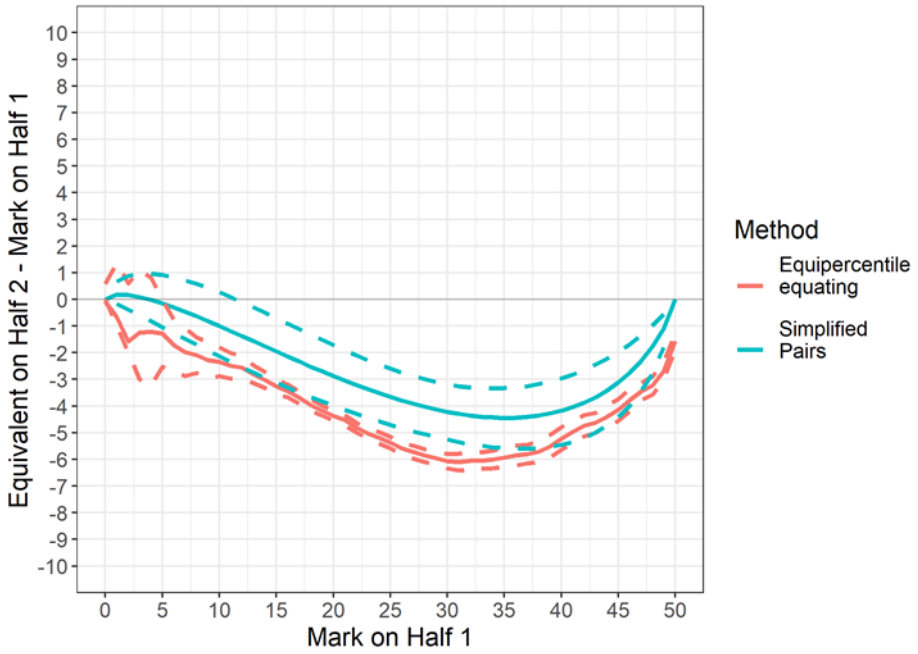

Figure 5: Results of the equating across the score range presented as differences between Half 1 scores and equivalent scores on Half 2.

Judge survey results

As well as collecting paired comparisons data from the six judges, we also invited them to take part in a survey to tell us how they found the task, and to gain a sense of how they thought they made their judgements. The survey was administered to judges via SurveyMonkey™ a short time after they had finished their judgements and took approximately ten minutes to complete.

When asked how straightforward they believed it was to make a holistic judgement, five out of the six judges said it was at least somewhat straightforward, with two of them believing it to be entirely straightforward. The sixth considered the process to be “not very straightforward”, noting that given that maths papers contain lots of questions of differing demand, making a holistic

judgement of maths papers was in their view very difficult. They highlighted that it would be easier to compare two responses to the same question, or two sets of questions of the same standard. Another judge, who had difficulties making holistic judgements initially, nonetheless said that this grew easier over time.

The judges appeared to have different specific processes for making their judgements. Some focused on the number of answers correct, while others attempted to match questions on each half of the paper by either a) their perceived difficulty or b) the skills required to answer them, and then tried to determine which script was superior. Candidates' working was considered by two judges to be a significant discriminator, while another highlighted communication. Many of the judges appeared to utilise many of these different processes at the same time.

The judges were asked directly which of the two half-length assessments they believed to have the more difficult questions. Encouragingly, four out of the six correctly identified Half 2 as the more difficult, while another saw the two halves as very similar in difficulty. The sixth noted only that one half was more difficult than the other but did not specify which. One judge noted that Half 2 contained nearly all the content that would be expected of a grade 8 candidate, but very little targeted at grades 5 or 6. This indication, especially when seen in the context of the fact that the outcome of the simplified pairs exercise did correctly approximate the empirical difference in difficulty between the two half-scripts, may somewhat counter the view that examiners struggle to make judgements about paper difficulty (e.g. Good and Cresswell, 1988). It may be the case that it is easier to make such judgements of difficulty when the judgemental task is simpler – i.e. that comparative methods are used. More evidence, however, is clearly required on this point.

Many judges described difficulties in making judgements where a candidate's response to one half-script was better in one sense, but worse in another sense, than the other candidate's response to the other half-script in the pair. For example, one judge noted an example where one candidate performed more strongly on trigonometry, but less well on algebra, while another indicated an example where one candidate answered every question, though not entirely correctly, while the other produced correct solutions to about half the questions. Most judges suggested that the tiebreaker for them in such cases was performance on the higher tariff, "harder" questions towards the end of the paper. Its is worth noting that this same issue arises even when making comparisons within the same test (Bramley, 2012). As such we cannot expect holistic judgements of quality to match the mark scale exactly.

Judges differed in their responses to a question about whether questions worth more marks are invariably better discriminators of candidate quality. Those agreeing highlighted that high-tariff questions allow for problem-solving skills to be evidenced and are often of greater complexity, while those opposed noted that some high-tariff questions can be quite routine, and can be prepared for, while communication issues can be more revealing in low-tariff questions.

There was more agreement than disagreement, however, with the contention that certain types of questions were better discriminators than others. That said, when asked what these better-discriminating questions might be, judges offered varied opinions, including multi-part questions, knowledge and understanding questions, and data analysis. Perhaps unexpectedly, there was also no strong agreement between judges as to whether they believed they did consistently focus on particular types of questions in their judgements, some suggesting that unstructured questions might be a useful tiebreaker but others attempting to make holistic judgements based on all types of questions across the paper.

Consequently, some validity questions arise about the nature of the judgements made. While it is generally acknowledged as a strength of marking that it involves processes that are relatively consistent across markers, does the variation in judgemental and discriminatory processes used by CJ judges and highlighted above have the potential to cause disquiet? It should certainly be remembered that, while the processes might have been different, to some degree at least, the outcomes of different judges' judgements were very similar (Table 4). Nonetheless, the issue of the different approaches used by different judges might concern stakeholders.

Likewise, if it is the case that judges attend more to certain questions (such as those worth more marks, or those more related to problem-solving than recall, for example) than others, what does this mean for validity? 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 awarding processes[2]. Again, however, this situation is likely to be mitigated by the simplified pairs approach, which collates many judgements and regresses them against the scripts' mark *difference*. Indeed, some might argue conversely 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. There is, however, undoubtedly a tension here.

Finally, two judges indicated a belief that comparative judgement methods might work less well for maths than subjects involving longer, more discursive answers such as English or history. While this view may run counter to the outcomes of the present study, and indeed also to the various studies cited in the Introduction, it is worthy of note, not least in that, for comparative judgement to be operationalised, the support of those intended to be used as judges would be vital. Principally, the concern here lies in the fact that, in many maths assessments, achieving the right answer the most times is the main objective (it "boils down to right or wrong", according to one judge). Given this, judges highlighted that it can be difficult to avoid simply re-marking the scripts. It was also suggested that the need to bear in mind many small judgements of superiority (of candidates' performance on questions testing different skills, for example) and then combine them into one overall judgement, for example, leads to more cognitive load and a more tiring task than marking.

On the other hand, most of the judges had never taken part in a comparative judgement exercise before and their experiences varied. Given this, views about the difficulty of the task given in the judge survey should not be taken to counter the empirical evidence presented above with regard to the qualified success of the method. More consideration of the validity implications of these findings, however, would be necessary.

Conclusion

This aim of this study was to evaluate the accuracy of the simplified pairs method when applied to a mathematics assessment. As part of this, it aimed to explore the level of accuracy that could be achieved in a situation when the two assessments being compared had very different levels of difficulty. The focus of the current research upon mathematics builds on the report by Benton et al (2020), which provided an initial example where using expert judgement, captured via simplified pairs, accurately identified the difference in difficulty between two English literature assessments.

[2] In fact, similar questions could be raised for all judgemental awarding processes and not just those using comparative judgement (see Bramley, 2010).

The results reported here could be summarised as a qualified success for the simplified pairs method. All of the judges in the study correctly identified which of the two assessments was harder. Furthermore, the overall estimated difference in difficulty based on simplified pairs (3.4 marks) was reasonably close to the actual overall level of difficulty based on statistical equating (5.0 marks). Finally, the level of precision of simplified pairs, as indicated by the width of confidence intervals, was improved compared to previous trials of the technique. This suggests that, for mathematics assessments, it may be possible to produce reliable results from simplified pairs using even smaller numbers of paired comparisons than for other subjects.

Having said the above, it was also clear that the small differences that were found between statistical equating and simplified pairs were statistically significant and, as such, cannot be dismissed as being merely a result of sampling error. That is, we have collected enough data to confirm that the two analysis methods are actually estimating something distinct from one another. We are confident that, were we to collect further data (i.e. include more judges or scripts), simplified pairs would still yield a slightly different answer about the relative difficulty of the two tests. This implies that, anyone hoping that simplified pairs (or any other implementation of comparative judgement) might provide a direct replacement for formal statistical equating will be disappointed as we can see that they point towards slightly different answers. This qualitative difference is, perhaps, similar to the difference between perceived difficulty and empirical difficulty of items (Benton, 2020) or the difference between equating and standard setting (Bramley, 2020). Discovering the extent to which this fundamental difference between simplified pairs and statistical equating is likely to affect the results remains an important topic for ongoing research. Other important avenues for further research include the consistency or otherwise of the judges' judgemental processes, and potential issues for the validity of the judgemental task around whether judges focus more on particular questions than others, where marking concerns the whole script.

Despite the differences between statistical equating and simplified pairs noted in the previous paragraph, the results here are still encouraging. It would be astonishing if any form of capturing expert judgement could inerrantly replicate formal statistical equating in every case. However, the fact that in experiments to date we've managed to produce results reasonably close to those from statistical equating suggests that simplified pairs could provide a useful source of evidence. Furthermore, despite the reservations in the paragraph above, it is worth noting that the level of overall accuracy achieved in the experiment is about as good as can be expected in CJ exercises of this type. For example, looking across all of the exercises described by Curcin et al (2019)[3] the median standard error around grade boundaries was 2 marks[4]. Given this expected level of precision, the fact that the present study correctly identified the difference in difficulty between the two assessments to within 2 marks can hardly be seen as a disappointment.

The major question that remains is where evidence from simplified pairs should sit alongside statistical evidence within our normal awarding processes. Where formal statistical equating is possible, through assessments sharing common items or being taken by common pupils, this is almost certainly a more accurate approach to ensuring comparability. However, given that these requirements for equating are rarely met for high stakes examinations in England, we need to make use of other sources evidence. One such source of evidence comes from statistical information about candidates in the form of either prior or concurrent attainment. Another source of evidence comes from expert judgement which could potentially be collated using comparative judgement. Reflecting on the use of statistics alongside expert judgement more broadly, rather

[3] Excluding "pinpointing studies" and those using teachers rather than examiners to make judgements as these are less relevant to the current study.
[4] Specifically, the median of the halves of the "2SD" precisions in the Curcin et al report. Calculated across a total of 23 comparative judgement studies and 77 grade boundaries.

than specifically comparative judgement, Stringer (2012) wrote that "the fact that the two sources of evidence testify to different definitions of difficulty means that there can be no logically coherent system for reconciling the different answers they produce" (page 552). To some extent, the results from the present study support the first part of this statement even if expert judgement is collated via comparative judgement. However, we'd contend that *both* expert judgement *and* typical statistical sources of evidence such as prior attainment can lead to something slightly different from what we're truly interested in – the results we'd get from formal statistical equating. As such, there remains a place for expert judgement within the process of standard maintaining. Knowing that the estimates of grade boundaries from a method such as simplified pairs should be very close to the estimates we'd get from formal equating (if that were possible) allows us to incorporate this information in a logically coherent way. Specifically, with the likely accuracy of the process firmly in mind, simplified pairs could be used to verify whether grade boundaries set using statistical methods are plausible (Benton et al, 2020). As such, we may continue to hope that the technique could provide a mechanism whereby grade boundaries for future assessments are influenced by the actual performances of students and are not entirely driven by limited statistical information.

References

Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software*, 74(8), 1-36. https://www.jstatsoft.org/article/view/v074i08/v74i08.pdf.

Baird, J.-A. (2007). Alternative conceptions of comparability. In P. E. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/487054/2007-comparability-exam-standards-f-chapter4.pdf.

Benton, T. (2020). How useful is comparative judgement of item difficulty for standard maintaining? *Research Matters: A Cambridge Assessment publication*, 29, 27-35. https://www.cambridgeassessment.org.uk/Images/579401-how-useful-is-comparative-judgement-of-item-difficulty-for-standard-maintaining-.pdf.

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. https://www.cambridgeassessment.org.uk/Images/599225-comparing-the-simplified-pairs-method-of-standard-maintaining-to-statistical-equating.pdf.

Bramley, T. (2005). A Rank-ordering Method for Equating Tests by Expert Judgement. *Journal of Applied Measurement*, 6(2), 202–223.

Bramley, T. (2010). 'Key discriminators' and the use of item level data in awarding. *Research Matters: A Cambridge Assessment publication*, 9, 32-38. https://www.cambridgeassessment.org.uk/Images/473654--key-discriminators-and-the-use-of-item-level-data-in-reporting.pdf.

Bramley, T. (2012). The effect of manipulating features of examinees' scripts on their perceived quality. *Research Matters: A Cambridge Assessment publication*, 13, 18-26. https://www.cambridgeassessment.org.uk/Images/469827-the-effect-of-manipulating-features-of-examinees-scripts-on-their-perceived-quality.pdf.

Bramley, T. (2020). Comparing small-sample equating with Angoff judgment for linking cut-scores on two tests. *Research Matters: A Cambridge Assessment publication*, 29, 23–27. https://www.cambridgeassessment.org.uk/Images/579400-comparing-small-sample-equating-with-angoff-judgement-for-linking-cut-scores-on-two-tests.pdf.

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. https://www.cambridgeassessment.org.uk/Images/109760-an-investigation-of-standard-maintaining-in-gcse-english-using-a-rank-ordering-method.pdf.

Good, F. J., & Cresswell, M. J. (1988). Grade awarding judgements in differentiated examinations. *British Educational Research Journal*, *14*(3), 263-281.

Kolen, M.J., & Brennan R.L. (2004). *Test equating, scaling, and linking: Methods and practices*. 2nd. New York: Springer.

Jones, I., Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help?. *Educational Studies in Mathematics,* 89, 337–355.

Jones, I., Swan, M. & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education,* 13, 151-177.

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

Novakovic, N., and Suto, I. (2010) The reliabilities of three potential methods of capturing expert judgement in determining grade boundaries. *Research Matters: A Cambridge Assessment Publication*, 9, 19-24. https://www.cambridgeassessment.org.uk/Images/473651-the-reliabilities-of-three-potential-methods-of-capturing-expert-judgement-in-determining-grade-boundaries.pdf.

Raikes, N., Scorey, S. and Shiell, H. (2008) *Grading examinations using expert judgements from a diverse pool of judges.* A paper presented to the 34th annual conference of the International Association for Educational Assessment, Cambridge, UK, September 2008. https://www.cambridgeassessment.org.uk/Images/109766-grading-examinations-using-expert-judgements-from-a-diverse-pool-of-judges.pdf.

Stringer, N. S. (2012). Setting and maintaining GCSE and GCE grading standards: the case for contextualised cohort-referencing. *Research Papers in Education*, 27(5), 535-554.

Wright, B.D., & Masters, G.N. (1990). Computation of OUTFIT and INFIT Statistics. *Rasch Measurement Transactions*, 3(4), 84-85. https://www.rasch.org/rmt/rmt34e.htm.