Research Matters / 39
A Cambridge University Press & Assessment publication

ISSN: 1755-6031

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

Is one comparative judgement exercise for one exam paper sufficient to set qualification-level grade boundaries?

Tom Benton

To cite this article: Benton, T. (2025). Is one comparative judgement exercise for one exam paper sufficient to set qualification-level grade boundaries? *Research Matters: A Cambridge University Press & Assessment publication, 39*, 26–38. https://doi.org/10.17863/CAM.116168

To link this article: https://www.cambridgeassessment.org.uk/Images/research-matters-39-is-one-comparative-judgement-exercise-for-one-exam-paper-sufficient-to-set-qualification-level-grade-boundaries.pdf

Abstract:

This research draws on evidence from three qualifications taken in autumn 2020, when comparative judgement (CJ) was used as a key source of data in setting grade boundaries. In these cases, a separate CJ exercise was completed for each individual paper in the qualification so that standards could be maintained from a previous series. In this article, we explore what would have happened had we relied on a single CJ exercise on one paper to maintain standards in the whole qualification. We first examine whether evidence from different papers provides a consistent picture of changes in cohort ability between series. We then explore the impact of relying on evidence from one paper only on the precision with which we can identify appropriate qualification-level grade boundaries using CJ.

Is one comparative judgement exercise for one exam paper sufficient to set qualification-level grade boundaries?

Tom Benton (Research Division)

Introduction

In high-stakes qualifications such as GCSEs and A Levels in England, grade boundaries refer to the minimum score a candidate requires to achieve a given grade. Historically, the way in which these have been determined has always included an element of human expert judgement (Benton & Elliott, 2015). Most commonly this has been in the form of expert examiners in the relevant subjects reviewing examination scripts with scores close to suggested grade boundaries and indicating whether or not they feel they are of sufficient quality to be awarded the focus grade. However, for some time, a number of authors have argued that comparative judgement (CJ) may provide a more effective means of incorporating inputs from subject experts into the process of determining grade boundaries (e.g., Bramley, 2007; Curcin et al., 2019). The awarding body OCR conducted a substantial programme of research investigating this technique, which is described in Benton et al. (2022).

In the context of awarding, a typical CJ exercise involves judges (usually examiners) comparing many pairs of scripts. Each pair consists of one script from the current exam series and one from a previous series. Note that the scripts from different series will consist of responses to different sets of questions. Allowing for this fact, judges must decide which script in each pair displays superior overall performance in the subject. Several hundred such comparisons are completed. The decisions from expert judges are then analysed with respect to the scores that were awarded to the different scripts with the aim of identifying pairs of scores in each series that, based on expert judgement, display equivalent levels of performance. For the CJ exercises in this article, we define two scores as equivalent if, when scripts with the two scores from the two respective series are compared, judges are equally likely to select either one as superior. For further details of the method see Benton (2021) or Benton et al. (2020). A slightly different methodology for using CJ in awarding is described by Curcin et al. (2019).

At present, if CJ is used in awarding, a separate exercise is carried out for each component of which a qualification is comprised. For example, suppose

a qualification consists of two exam papers. We typically wish to set grade boundaries on the qualification that reflect performance standards at some point in the recent past (often within the prior year). In order to do this, we would complete one CJ exercise for paper 1 to identify grade boundaries that reflect equivalent levels of performance to grade boundaries on the previous version of this same paper in a previous exam series. We would also complete a separate exercise for paper 2 to identify the appropriate boundaries on this paper. Finally, boundary scores at each grade would be added together across papers to set grade boundaries for the whole qualification. For example, if 10 marks were needed to achieve a grade E on paper 1, and 12 marks were needed for a grade E on paper 2, then we would know that, for the qualification as a whole, the grade E boundary should be 22.

As noted by Benton et al. (2022), the use of CJ in awarding requires substantial time from expert judges. With this in mind, it is of interest to explore ways of making the application of CJ in awarding more efficient. One possibility for increased efficiency would be to only conduct a single CJ exercise for a single component. In theory, the outputs of the CJ exercise provide a complete mapping of scores on the previous version of the component to equivalent scores on the current version. Therefore, once this mapping is applied, we have access to a measure of the abilities of the students taking the alternative versions of the qualification (from different exam series) on the same scale — that is, essentially an anchor test. We might then use the scores on this anchor component to perform complete test equating of scores on the different qualification versions using any equating method we choose from those applicable to a non-equivalent anchor test (NEAT) design (Kolen & Brennan, 2004).

Although it is clear that the above approach would require less time from expert judges than completing a CJ exercise for every component, it is not known whether it would provide accurate results. In this article, we explore this issue in more detail. Firstly, we examine the consistency of evidence from different individual components regarding changes in cohort ability between series. This is of interest as, if we are to rely on evidence from a single component, it is important that changes in performance in one component are indicative of changes in performance on the qualification as a whole.

Having done this, we then explore in detail whether the results from separate CJ exercises on different components necessarily lead to the same grade boundaries at qualification level. Furthermore, we evaluate the impact of using CJ from a single component upon the standard errors of estimated grade boundaries compared to the current approach of combining several separate CJ exercises. That is, even if we assume that a single component is sufficient to indicate how the performance of a cohort has changed, what is the impact of using just one component on the precision of the technique?

Data

This research re-analysed data from three qualifications from Benton et al. (2022) where all of the components comprising a qualification were included in separate

CJ exercises. Further details of the CJ exercises conducted for each of the three qualifications, referred to as "projects" for the purposes of this article, are given in Table 1. The projects all relate to qualifications that were awarded using CJ in autumn 2020. In each of these cases, a "simplified pairs" CJ approach (Benton, 2021; Benton et al., 2020) was taken. For all three of the projects, the aim was to carry forward performance standards from June 2019 to set grade boundaries in autumn 2020.

Table 1 shows the amount of available data in each of the CJ exercises. Note that, due to the unusual nature of the autumn 2020 exam series,[1] very few scripts were available. As such, fewer scripts from this series than from June 2019 were included in the study. Because of this, each script from June 2019 was included in a single paired comparison whereas scripts from the autumn series were often included in multiple pairs.

Table 1: Details of the CJ exercises

Project ID	Qualification	Subject	Paper	Max. score	N scripts		N judges	N pairs
					June 2019	Autumn 2020		
1	A Level	English Literature	1	60	466	91	6	466
			2	60	414	97	5	414
2	A Level	Psychology	1	90	498	66	6	498
			2	105	500	53	6	500
			3	105	500	51	6	500
3	GCSE	English Language	1	80	350	291	6	350
			2	80	350	345	6	350

As well as the data from the CJ exercises, data on the scores achieved in each exam paper by all those who took the qualifications was used for some analyses.[2]

Relationship between scores in different components

Intrinsically, any justification for using a single component as an anchor for an entire qualification is dependent upon the different components measuring broadly the same abilities. Of course, different components tend to cover different topic areas within a subject. However, we would hope that they all rely on broadly the same underlying set of knowledge and skills. Some evidence of this can be provided by looking at the correlations between scores on different components for the full cohort of candidates that took each qualification. These are shown in Table 2. As can be seen, within each series, scores for all of the components within a qualification display fairly strong correlations with one another. These

1 The autumn 2020 exam series was specially arranged to allow students who were unhappy with the grades they were awarded by their school during the pandemic to sit formal exams. Autumn exams are not normally made available for GCSEs or A Levels, except for GCSEs in English Language and Mathematics.

2 The data used in this research was collected as part of the operational marking and processing of candidates' examination scripts. Data has been stored and used in line with Cambridge University Press & Assessment's Data Privacy notice (https://www.cambridge.org/legal/candidate-privacy-notice).

correlations provided initial evidence that it was reasonable to conduct the analyses reported in this article in order to explore the potential of using a CJ exercise on just one component to inform awarding decisions.

Table 2 also shows the total number of candidates in each series. From this we can see that the amount of available data in the autumn series was quite low for two of the qualifications.

Table 2: Correlations between components within each qualification in each exam series

Project ID	Qualification	Subject	Papers	Correlation in...		N candidates	
				June 2019	Autumn 2020	June 2019	Autumn 2020
1	A Level	English Literature	1 and 2[3]	0.58	0.62	9677	119
2	A Level	Psychology	1 and 2	0.69	0.73	5567	70
			1 and 3	0.66	0.69		
			2 and 3	0.72	0.78		
3	GCSE	English Language	1 and 2	0.81	0.84	13 199	496

Summary of CJ results for individual components

All of the results in subsequent sections are derived from the mappings from June 2019 to autumn 2020 scores that were identified for each individual component using CJ. These mappings are displayed in Figure 1. For each score on each June 2019 paper, the solid black line in each chart shows the score on the autumn 2020 paper that was estimated to be equivalent. An "equivalent" score means one where expert judges would be equally likely to consider a script with this score better than, or worse than, a June 2019 script with the associated score on the x-axis. Figure 1 also shows 95 per cent confidence intervals (CI) as dotted lines, calculated using the method described in Benton et al. (2020). A faint grey line of equality is included in each chart to aid interpretation.

For the English Literature papers, Figure 1 shows the papers in autumn 2020 were perhaps slightly easier than those in June 2019 at the top end of the score distribution — that is, the equivalent scores are significantly above the line of equality. However, at the lower end, the autumn 2020 exams were perhaps easier. Having said this, due to a lack of scripts with low scores in autumn 2020, the confidence intervals are very wide at the lower end.

Figure 1 suggests that, for Psychology, paper 1 was of very similar difficulty in autumn 2020 and June 2019, paper 2 was slightly harder in autumn 2020, and paper 3 was slightly easier. Finally, for English Language, Figure 1 suggests that paper 1 was of similar difficulty in autumn 2020 and June 2019 but that paper 2 was slightly harder.

3 In an ordinary series, A Level English Literature also includes an additional component of non-examined assessment. However, due to the unusual nature of the autumn 2020 series, the qualification was awarded without this element on this occasion.

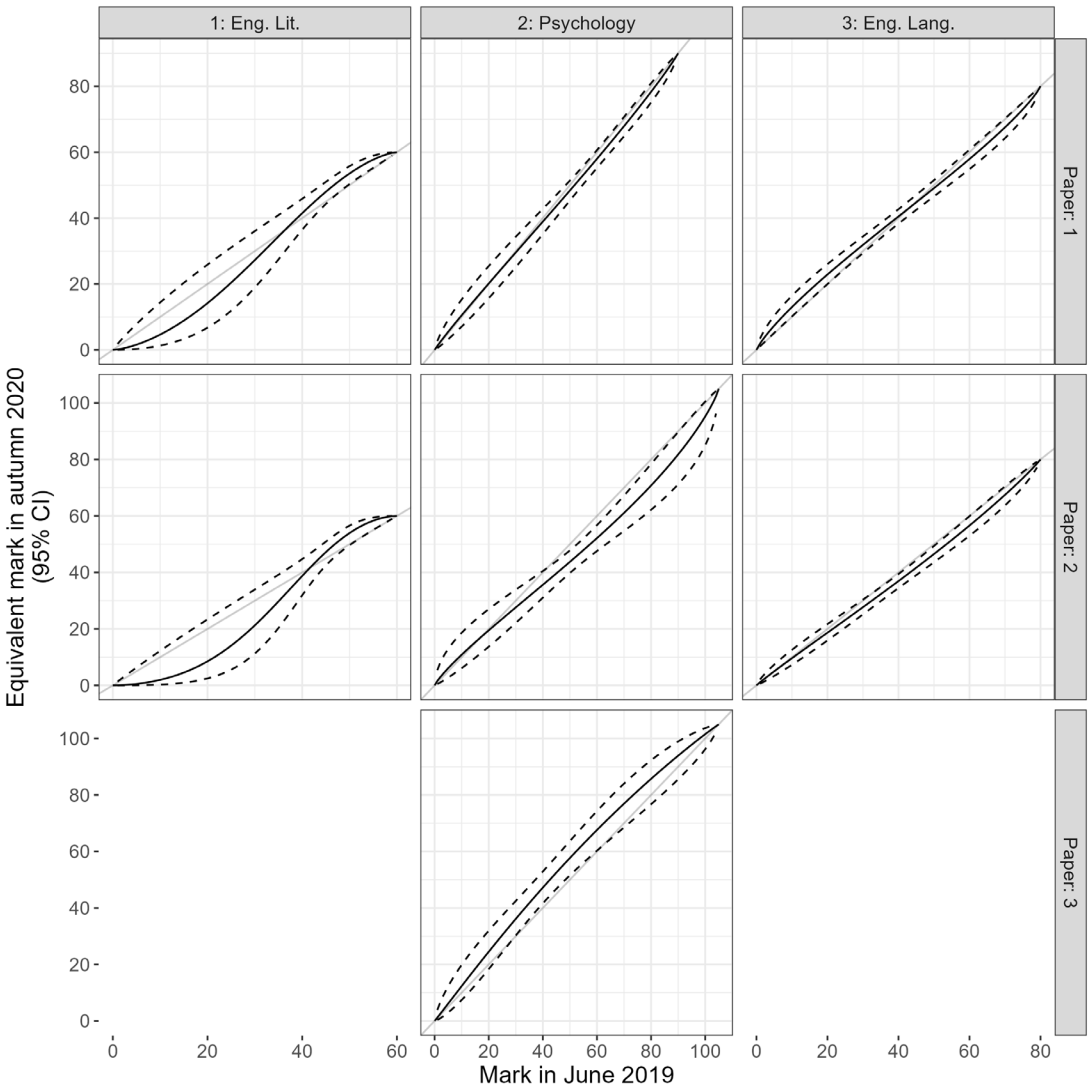

Figure 1: Mapping for each component of each subject based on analysis of CJ data.

Table 3 shows the same results in tabular form for the key grade boundaries[4] on each June 2019 paper. For ease of presentation, equivalent scores on the autumn 2020 papers have been rounded to whole numbers and confidence intervals are presented in terms of each score being correct plus or minus a particular (rounded) value. Across the grades and papers in Table 3, the majority of equivalent scores were identified with a precision of no worse than plus or minus 3 marks. However, some were larger; for example, the widest confidence interval (English Literature, paper 2, grade E) had a precision of plus or minus 12 marks.

4 Key grades are those where regulations require awarding organisations to make explicit decisions about boundaries. The key grades are grades E, A and A* for A Level, and grades 1, 4, 7 and 9 for GCSEs. All other grade boundaries are usually set by linear interpolation between these.

Table 3: Summary of score mappings and confidence intervals based upon CJ for individual components at key grades

Project ID	Subject	Grade	Paper 1		Paper 2		Paper 3	
			June 2019 boundary	Autumn 2020 equivalent	June 2019 boundary	Autumn 2020 equivalent	June 2019 boundary	Autumn 2020 equivalent
1	English Literature	E	26	22 +/- 9	27	17 +/- 12	-	-
		A	53	56 +/- 2	54	58 +/- 3	-	-
		A*	56	58 +/- 2	57	59 +/- 1	-	-
2	Psychology	E	29	29 +/- 5	22	21 +/- 7	32	38 +/- 6
		A	69	67 +/- 3	62	54 +/- 5	69	76 +/- 8
		A*	75	73 +/- 3	71	62 +/- 6	77	83 +/- 8
3	English Language	1	8	11 +/- 3	8	8 +/- 2	-	-
		4	34	35 +/- 2	36	33 +/- 2	-	-
		7	53	52 +/- 3	54	50 +/- 3	-	-
		9	65	63 +/- 3	66	63 +/- 4	-	-

Evidence about cohort ability from different components

Using the mappings shown in Figure 1, it was possible to transform the scores for all candidates who took each paper in June 2019 to the equivalent scores on the autumn 2020 papers. Having done this, we can compare the performance of candidates on each paper between series.

This comparison is shown in the form of boxplots in Figure 2. The top and bottom of each box in the figure represents the 25th and 75th percentiles of the total scores on each paper. The central lines within each box represent the medians and the whiskers represent the range of scores that were seen excluding outliers. Note that the scores from June 2019 have been transformed using the results from the CJ studies (Figure 1) so that, theoretically, if the CJ exercise has worked as intended, scores in the different series are directly comparable as if candidates had taken the exact same versions of each assessment. Therefore, differences in performance in Figure 2 potentially indicate differences in the subject ability of the candidates entering in the different series.

From Figure 2 we can see that, in every paper, the performance of candidates was stronger in June 2019 than autumn 2020. This is unsurprising since the autumn series was mainly intended for candidates who had not achieved the grades they wanted during summer 2020. As such, it is expected that autumn 2020 would attract entries from weaker candidates.

For English Language the difference in the performance of candidates in the two series is very consistent across each paper. Specifically, in each paper, the median performance in June 2019 is just slightly above the 75th percentile of performance in autumn 2020. For English Literature differences are also fairly consistent in that, for each paper, the median performance in June 2019 is between the median and 75th percentile of performance in autumn 2020.

However, for Psychology the three papers show very different patterns. Based on paper 1, candidates in June 2019 were only slightly stronger than those in autumn 2020. Differences on paper 2 appear a little larger and on paper 3 the difference appears huge with the median performance in June 2019 well above the 75th percentile of performance in autumn 2020.

Overall, the results here show that different components can potentially lead to different conclusions about the relative strengths of groups of candidates. For this reason, it is good that we have direct evidence on relative performance levels in all of them rather than relying on a single component. We explore the impact of these differences on grading the qualification overall in the next section.

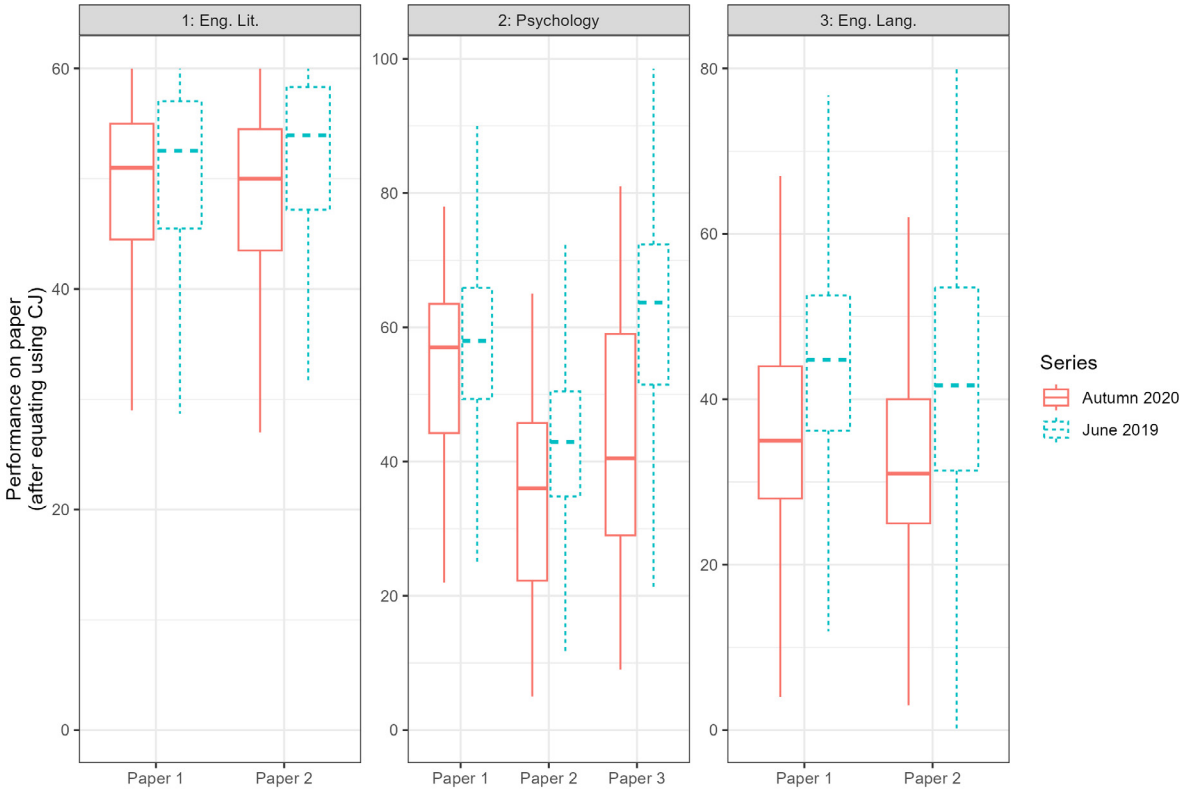

Figure 2: Boxplots showing the distribution of the scores on each paper in June 2019 and autumn 2020. Scores from June 2019 have been transformed using the results from the CJ analysis (Figure 1) so that they are, theoretically, directly comparable with those from autumn 2020.

Impact of using a single component on overall grade boundaries

For each subject, the results from the CJ exercises were used to identify qualification grade boundaries for autumn 2020 equivalent to grade boundaries from June 2019. Note that, due to the impact of the pandemic, and discussions over how the standards required in autumn 2020 exams should relate to those from grades awarded purely via teacher assessment in summer 2020, these do not match the June 2019 boundaries that were used in the actual awarding exercises for these qualifications. However, for the purposes of this research, we will imagine that standards from June 2019 were carried forward to autumn 2020 in a straightforward manner.

Overall grade boundaries for autumn 2020 for each subject were calculated by adding the relevant grade boundaries from different components together — that is, the estimated boundaries in Table 3. Standard errors of estimated grade boundaries at whole qualification level were calculated by taking the square root of the sum of the squared standard errors across the constituent components.

Next, we generated grade boundaries for each whole qualification but based upon each separate available component on its own. The following process was used to achieve this:

1. Collate a dataset for all candidates in either June 2019 or autumn 2020 (i.e., not just those candidates in the CJ exercise) with scores on all constituent components with the following pieces of data:
    a. Each student's score on the component of interest.
    b. Each student's total score on the entire qualification.
2. For each student in the June 2019 data from step 1, replace their component score on the component of interest with the equivalent score from the autumn series as defined in Figure 1. Note that their total score on the whole qualification should not be adjusted.
3. Adjusted component scores from June 2019 and unadjusted scores on the same component in autumn 2020 are now treated as if they are interchangeable. As such, these two sets of scores are treated as an anchor test to allow whole qualification scores from June 2019 to be equated to equivalent scores at whole qualification level in autumn 2020.
4. Use the results from step 4 to identify equivalent values in autumn 2020 to the June 2019 grade boundaries at whole qualification level.

For step 3, chained equipercentile equating was used (Kolen & Brennan, 2004, p. 145). Briefly, equipercentile equating means that, where two assessments have been taken by the same group of candidates, equivalent scores are identified as those that are at the same percentile in the distribution. The "chained" element means this was applied in two steps – first to map grade boundaries on the whole qualification in June 2019 to equivalent points on the anchor, and then to map these anchor points to appropriate positions in the autumn series.

Note that loglinear models were used to smooth the score distributions before the chained equipercentile procedure was applied. Specifically, the empirical score distributions were replaced with smooth versions that retained the mean, standard deviation, skewness and kurtosis of scores on each paper in each series. This method was necessary to address the small sample sizes in the autumn 2020 series. Without smoothing, the large gaps between the scores that actually occur in the data could manifest themselves in some unusual results. In a normal summer exam series, most subjects have entries from rather larger numbers of candidates and such issues do not occur. Thus, the use of smoothing helps ensure the results here are more indicative of what might happen in practice more widely.

In addition to producing estimates of grade boundaries at whole qualification level based upon each separate component, it was of interest to produce standard errors from using a single CJ exercise. These standard errors are intended to show how the precision of a CJ exercise for a single component (Table 3, Figure 1) manifests itself when applied to setting grade boundaries for an entire qualification. The standard errors do not incorporate the uncertainty in the equating process itself (step 3). In most practical situations, with larger sample sizes, this source of uncertainty would be trivial compared to the uncertainty stemming from the CJ exercise in any case.

Standard errors were estimated as follows:

A. The mappings in Figure 1 are based on the coefficients from a logistic regression (see Benton et al., 2020, for further details). Rather than using point estimates of these coefficients, these were sampled from a multivariate normal distribution with a mean at the estimated coefficients and using the variance-covariance matrix of the model parameters.
B. Apply steps 1 to 4 (above) based on a mapping derived using the logistic regression coefficients sampled in A to derive a fresh estimate of the qualification grade boundary.
C. Repeat steps A and B 500 times and use the standard deviations of the estimated boundaries across these repetitions as the standard errors.

For each boundary, 95 per cent confidence intervals were calculated by the usual approximation of multiplying the standard errors by 1.96.

The results of this analysis are shown in Table 4. Table 4 shows the estimated grade boundaries and confidence intervals at whole qualification level based on CJ evidence from each individual component only and also (the final column) based on all of the CJ evidence across all components combined. For ease of reading, all estimated boundaries and confidence interval widths have been rounded to whole numbers.

Note that the estimates from using all components need not always fall between the estimates from individual components.[5] Also note that, due to the impact of rounding, the estimated overall qualification boundaries in the final column may not perfectly match the sum of the estimated values for each paper shown in Table 3.

5 This reflects the fact that, in ordinary chained equating, if we had two possible anchor tests A1 and A2, chained equating using the sum of both anchor tests as the anchor would not give the same result as taking the average of analyses using each anchor test separately. This is because the summed anchor test will have different reliability as well as a differently shaped distribution to either of the individual anchors.

Table 4: Estimated qualification-level grade boundaries in autumn 2020 and standard errors based on different individual components

Project ID	Subject	Grade	June 2019 boundary	Estimated qualification-level grade boundary from source component (paper) and 95% confidence interval											
				Paper 1			Paper 2			Paper 3			All (Original)		
1	English Literature (Max. score =120)	E	53	43	+/-	20	40	+/-	11		-		39	+/-	15
		A	107	113	+/-	4	114	+/-	4		-		114	+/-	3
		A*	113	117	+/-	2	118	+/-	2		-		118	+/-	2
2	Psychology (Max. score =300)	E	83	73	+/-	9	87	+/-	21	88	+/-	17	88	+/-	10
		A	200	190	+/-	11	186	+/-	13	206	+/-	17	197	+/-	10
		A*	223	215	+/-	10	207	+/-	17	223	+/-	16	218	+/-	11
3	English Language (Max. score =160)	1	16	15	+/-	5	21	+/-	4		-		19	+/-	4
		4	70	68	+/-	5	67	+/-	5		-		69	+/-	3
		7	107	100	+/-	6	104	+/-	6		-		102	+/-	4
		9	131	123	+/-	7	131	+/-	6		-		126	+/-	5

The same information in Table 4 is displayed in a different way in Figure 3. Figure 3 displays estimated grade boundaries for autumn 2020 in terms of how far they moved from June 2019. It also shows 84 per cent confidence intervals for each estimated change as, according to Cumming (2009), where such confidence intervals do not overlap, we can infer that the two estimates are significantly different at the 5 per cent level.[6] As can be seen, for English Literature and English Language there clearly are no significant differences between the estimated grade boundaries from different components. That is, while different components would indeed lead to different grade boundaries at whole qualification level, the size of these differences is no larger than we would expect given the quantifiable uncertainty in the CJ methods. That said, particularly for English Literature, some of the confidence intervals are very wide, which would restrict their practical usefulness operationally.

For Psychology, larger differences in the estimated boundaries are evident. For example, based on CJ evidence from paper 2, we would set the qualification A grade boundary at 186 marks. In contrast, based on paper 3 it would be set at 206 marks. These differences reflect the discrepancies already discussed earlier (Figure 2) in the evidence from different papers regarding the size of the difference in ability between the candidates that took the qualifications in different series.

The differences in grade boundaries (Figure 3, Table 4) are close to statistically significant and had we shown results at all grades rather than the key ones only, slightly larger, and statistically significant differences would have been visible. As such, we are confident in stating that it is possible for CJ evidence from different components to lead to significantly different results.

6 Similarly, Goldstein and Healy (1995) suggest that creating confidence intervals with estimates plus or minus 1.39 times the standard errors can ensure that the intervals for significantly different estimates will not overlap. This is equivalent to recommending the use of 84 per cent confidence intervals.

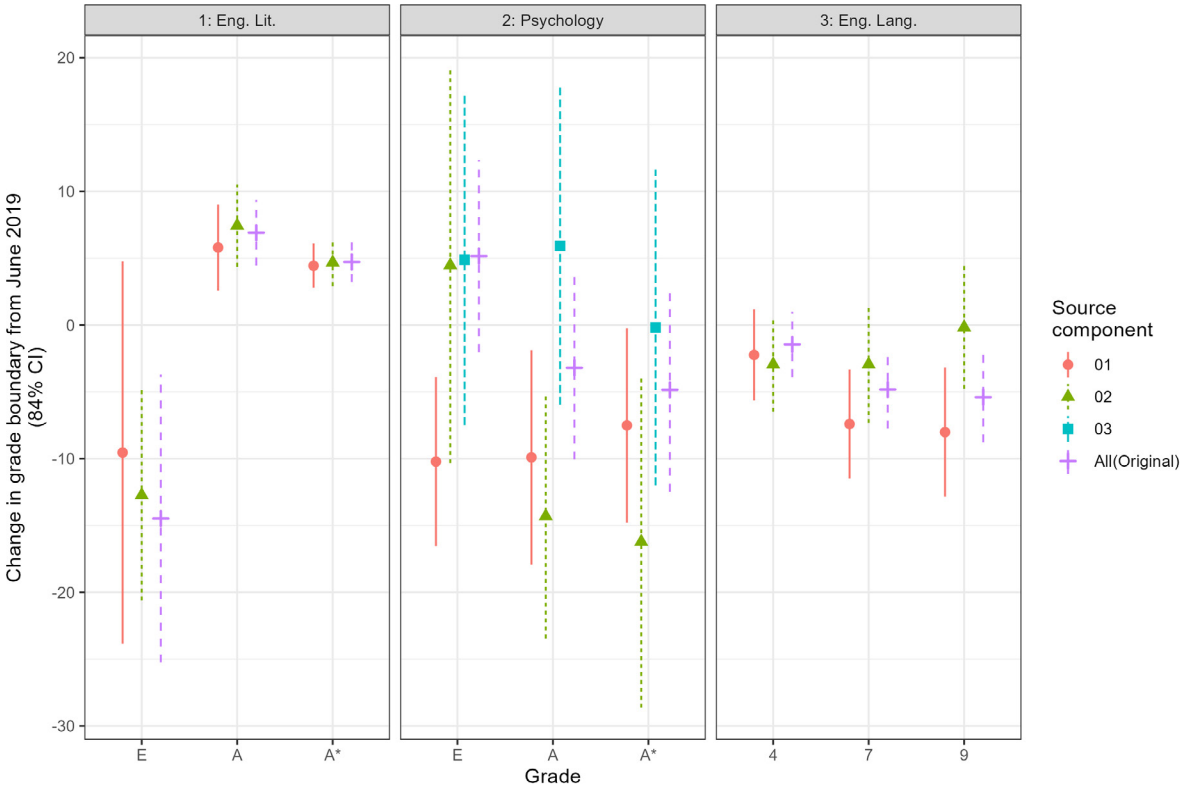

Figure 3: Changes in grade boundaries from June 2019 to autumn 2020 based on CJ evidence for each individual component alone and also based on all CJ evidence combined (with 84 per cent confidence intervals).

Aside from looking at the differences between estimated grade boundaries, it is also worth comparing the standard errors of grade boundaries from individual components to those based on the full set of CJ evidence. From Table 4 it can be seen that the confidence intervals of qualification grade boundaries derived from CJ evidence from a single component are mostly wider than those that combine all the CJ evidence. Specifically, the median width of confidence intervals from using all CJ evidence was plus or minus 4.5 marks. In contrast, the median width of confidence intervals using evidence from a single component was plus or minus 7 marks.

Conclusion

The aim of this research was to explore whether the use of CJ in awarding could be made more efficient by restricting judgements to a single component and then using the results to help infer grade boundaries for the qualification overall. Having compared the results of using CJ from a single component to using CJ exercises on all components, there are at least two reasons why we recommend that CJ in awarding should continue to incorporate separate exercises for each component:

- Relying on a single component leads to a noticeable decrease in precision. It should be noted that using CJ in awarding is already somewhat imprecise with analysis providing a range of scores that are consistent with judges' decisions rather than a single score.

- Relying on a single component effectively assumes that changes in the levels of candidate performance in one component are likely to be reflected in others. However, the analysis in this article has revealed that different components may suggest differing amounts of change in performance levels. This may be true even if our data indicates strong correlations between scores on different components. As such, if we accept that CJ results provide a realistic means of assessing changes in performance, we cannot assume that results from a single component are sufficient to infer how performance has changed on the qualification as a whole.

With regard to the issue of the loss of precision, it is possible that this might be addressed by increasing the size of the single CJ exercise, for example, by increasing the numbers of judges, scripts or comparisons included in the exercise. However, given the rate at which standard errors associated with CJ exercises fall with additional resources (see Benton, 2021) we would not expect this to provide a practical means to address this issue.

In terms of the evidence for using CJ in awarding at all, it would have been reassuring if we had found that the CJ results on every individual component in a qualification suggested the same level of change in performance among candidates. For example, this might have suggested that differences on all components were explicable in terms of a change in the general ability or prior attainment of the candidates entering a qualification. However, the fact that such consistency was not found for all three qualifications cannot be taken to imply a problem with the use of CJ for awarding. Particularly given the context in which the examinations were taken (a global pandemic), and the relatively small number of candidates entering autumn exams, it is genuinely plausible that changes in performance levels differ across components.

Further research might explore whether there are conditions where using a single component can be effective. Intuitively, we would expect the consistency of evidence from different papers to increase with greater overlap in the topics that they assess. Furthermore, it would be interesting to repeat the analysis in this article on data collected outside of the conditions of a global pandemic to see whether this leads to greater consistency of evidence. For example, we might speculate that the reason for the different patterns shown in Psychology paper 3 in our analysis is that the interruption to students' studies in 2020 meant they did not get to fully cover the topics in this paper.

Overall, this article provides little evidence as to whether the use of CJ in awarding is effective. Although this has been explored in various previous pieces of research (e.g., Benton et al., 2020; Benton et al., 2022; Curcin et al., 2019) it remains an open research question. However, this article does suggest that, if CJ is to be used in awarding activities, it is best if judges explicitly review performance on all of the different components. After all, we have seen that it is at least plausible that evidence from different components may lead to different results. Furthermore, relying on a single CJ exercise for a single component to grade an entire qualification decreases precision and makes additional assumptions that may or may not be correct.

Acknowledgement

With thanks to all those involved in the original data collection on which this research is based.

References

Benton, T. (2021). Comparative judgement for linking two existing scales. *Frontiers in Education, 6*, 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.

Benton, T., & Elliott, G. (2015). The reliability of setting grade boundaries using comparative judgement. *Research Papers in Education*, *31*(3), 352–376.

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.

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

Cumming, G. (2009). Inference by eye: Reading the overlap of independent confidence intervals. *Statistics in Medicine, 28*(2), 205–220.

Curcin, M., Howard, E., Sully, K., & Black, B. (2019). *Improving awarding: 2018/2019 pilots*. Ofqual report Ofqual/19/6575.

Goldstein, H., & Healy, M. J. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *158*(1), 175–177.

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