Is comparative judgement just a quick form of multiple marking?

Tom Benton Research Division and **Tom Gallacher** GL Assessment

(The study was completed when the second author was based in the Research Division at Cambridge Assessment)

Introduction

For many years now, comparative judgement (CJ) has been proposed as an alternative method to traditional marking for summative assessment (see, for example, Pollitt, 2004). Traditional marking relies upon an examiner reviewing a piece of work (perhaps an essay) and then making an absolute judgement regarding the mark or grade that it deserves on a pre-determined fixed scale. However, according to proponents of CJ "… humans are very bad at making such absolute judgements" (Christodoulou, 2018). As such, they say it is better to avoid absolute judgements entirely and, instead, assign scores by repeatedly getting examiners to decide which of a pair of scripts (e.g., essays) is superior and then using a statistical model (the Bradley-Terry model) to convert the set of judgements into a single score for each script[1]. In particular, this avoids issues with ensuring that all examiners interpret the marking or grade scale in a consistent manner. For example, whereas absolute judgements may be influenced by the leniency or severity of individual examiners, comparative judgements automatically avoid such influences because in CJ, examiners never assign a mark or grade, and make only relative judgements between scripts.

The aim of this article is not to question the claim that CJ can produce valid achievement scores. Rather, the aim is to investigate why it is that CJ works as well as it does, and what would be required for traditional marking to perform equally well. Specifically, this article provides evidence supporting the following central conjecture:

> *Comparative judgement is just a form of multiple marking, with a very simple mark scheme and using 'fancy statistics'.*

As underlined above, there are three elements to this conjecture. Let us unpack them a little further:

- **Multiple marking:** The crucial term here is 'multiple'. Although judges in a CJ exercise do not give marks to scripts in the formal sense, multiple judgements are made regarding each piece of work. Usually (although not always) the judgements of several judges are combined to create an overall score for each script.
- **Simple mark scheme:** Rather than having a detailed mark scheme telling judges how to map particular performances to specific scores, judges are generally asked to do something cognitively much simpler – to simply say which of two scripts is superior. This is not to say that CJ exercises do not pay careful attention to defining the skills they are trying to measure. Most CJ exercises include some form of training to help judges know which aspects of performance to focus on. However, at the point of making judgements, the task that judges are required to undertake is clearly simpler than in traditional marking.
- **Fancy statistics:** In traditional marking, the score assigned by a marker is usually exactly the same as the one that is communicated to a candidate. In contrast, judgements from CJ are processed through a fairly complex statistical model to create scores. This model is a crucial part of the machinery for ensuring that the scores ultimately assigned to scripts are all upon a single comparable scale.

The point of the conjecture is that, although the three outlined elements are taken as a given in any CJ exercise, they are rarely applied to traditional marking *even though they could be*. For example, whereas CJ judgements are routinely processed through a statistical model, the same is not true of traditional marks. This is true despite the fact that, in modern on-screen marking systems, scripts are randomly assigned to markers, meaning that it is very simple to build statistical models that would adjust scores accounting for the different ways in which markers have used the mark scale. For example, it would be easy to adjust marks to account for the relative leniency or severity of different markers. To take another example, whereas every CJ exercise requires that multiple judgements are made about each script, in traditional marking each script is usually marked by only one examiner. Interestingly, studies of the effect of multiple marking on reliability (e.g., The Office of Qualifications and Examinations Regulation [Ofqual], 2014a) tend to be fairly negative about its impact. One possible reason for this negativity, compared with the positive reviews of CJ, is that research on the impact of multiple marking rarely considers the possibility of processing the resulting marks using a statistical model.

This article will show that, if the numbers of judgements for each script and the statistical models from CJ are applied to data from traditional marking, then we can produce scores of equal (or perhaps better) quality without the need for judges to actually make comparative judgements. More specifically, it will show via a pseudo-CJ approach that, if we recreate paired comparisons data such as that used in CJ exercises but based on data from a multiple marking study, we can produce scores that are just as useful as those from a real CJ study. In other words, it is not necessarily the quality of judgements themselves that improve when using CJ; rather it is simply that there are more of them and that statistical models are used to iron out differences in the leniency and severity of different judges. Given that it is in fact the ability to quickly produce judgements for each script that is at the heart of CJ's strength, the question of whether mark schemes can be simplified to facilitate a greater marking speed naturally arises.

1. For further technical details of the way the scoring scale is defined, see Bramley and Vitello (2018).

Data

This article makes use of exactly the same set of data as an earlier study by Bramley and Vitello (2018). The data is drawn from 150 essays from a Higher Tier English Language General Certificate of Education (GCSE) paper sat in summer 2014. Candidates were asked to write a diary entry or blog on the subject of "…and it made me change my mind" and the resulting essays were originally marked out of 40. On average the essays were around 500-words long comprising between 1 and 7 pages of writing. The study by Bramley and Vitello explored the relative reliabilities of *adaptive comparative judgement* (ACJ) and comparative judgements where the pairs being compared were assigned at random (*random comparative judgement* or RCJ). The two CJ studies were undertaken by separate sets of examiners and scores for all 150 scripts were derived from each method.

Crucially for this research, the same set of scripts was also used in a study by Child, Munro and Benton (2015). The aim of their study was to evaluate the impact of some fairly cosmetic changes to a mark scheme on marking accuracy. As part of this study, all 150 scripts were each marked by 17 examiners who had not been involved in the original marking of this GCSE paper. In addition, all 150 scripts were also marked by the Principal Examiner (PE) for the paper.

As can be seen from our description, for the 150 essays being studied, 4 methods of scoring have been trialled in previous research. Table 1 provides a comparison of the types and number of judges involved in each of the four methods. It also provides details of the number of judgements used to create scores for the two CJ methods. For further details of the design of the CJ studies, see Bramley and Vitello (2018). All the CJ tasks were run online using the *No More Marking*™ website (www.nomoremarking.com).

Table 1: A comparison of the features of different methods of generating scores

Method	*Adaptive comparative judgement (ACJ)*	*Random comparative judgement (RCJ)*	*Traditional marking*	*Principal Examiner marking*
Reference	Bramley and Vitello (2018)		Child, Munro and Benton (2015)	
Format	On-screen judgements		Paper-based marking	
No. of judges involved in total	18	16	17	1
Who the judges were	Experienced markers of the given paper	Experienced markers of the given paper	Markers who had never marked the paper in question before	The Principal Examiner
Average number of judgements on each essay to produce scores for each script in current study	14.4	13.4	Typically 1 (but data also allows analysis of multiple marking)	1

Analyses

Predictive value

The research in this article compares the predictive value of scores derived in different ways (including each of those shown in Table 1). By predictive value, we mean the correlation between the scores assigned to the 150 essays and other external measures of student achievement. Specifically, the focus was upon how well scores derived in each of the above ways correlated with achievements in other tests of English Language and English Literature taken during the same examination session. This measure might be called *predictive validity* (we are seeing if the scores from the English Language examination can help us 'predict' something else that is true about the candidates), although, given that the various assessments being correlated were all taken during the same examination session, others might prefer the term *concurrent validity*. Certain experts might also take issue with any use of the term *validity* at all, on the grounds that validity can only be attached to various uses of test scores, rather than the test scores themselves. For this reason, we shall largely avoid the term validity throughout the article and instead use *predictive value* or *correlation* to describe the results.

The focus is upon predictive value rather than reliability for two reasons. Firstly, for ACJ, it is almost certain that the reliability coefficients routinely produced by this method are biased and give values that are far too high (Bramley & Vitello, 2018; Bramley, 2015). Secondly, even when restricting our attention to the other methods, we note that the values of reliability coefficients are dependent upon the scale used to report scores. As such, because traditional marking and CJ report scores on different scales, it is not clear that, for example, a scale separation reliability coefficient from CJ can be straightforwardly compared to (say) an inter-marker correlation coefficient from traditional marking. Focussing on the predictive value of scores avoids this issue.

The predictive value of different methods of generating scores (e.g., traditional marking and CJ) was evaluated via their Spearman rank order correlation with:

- the raw marks achieved by each candidate in the controlled assessment element of their English Language GCSE. The controlled assessment task was taken internally within schools, marked by teachers (and then moderated), and tested their skills in understanding extended literary texts and imaginative writing. Scores from this element of the English Language GCSE were available for 128 out of the 150 candidates included in the analysis.
- the Uniform Mark Scale (UMS) marks achieved by candidates across their entire English Literature GCSE. UMS marks summarise achievement across a number of different internal and external assessments within this GCSE and were available for 121 out of the 150 candidates.

Spearman correlations were chosen as they are invariant to the reporting scale used for scores.

Scores from traditional marking

As can be seen, for the *Traditional marking* column in Table 1, we can choose how many judgements of each essay we combine to produce scores. Several values were trialled for this current research study.

Specifically scores were produced via:

- **Single marking.** For each script in turn we randomly selected an examiner from the 17 available and used their mark for this script as the score.
- **Double marking.** Each script was assigned a score by taking the mean mark across two randomly chosen examiners.
- **Combining marks from all 17 markers (17-fold marking).** The average mark across all 17 examiners was assigned to each script. Both the median and the mean were trialled.

The use of double marking and 17-fold marking allows us to look at the effect of including one standard feature of CJ (multiple marking) within traditional marking.

In addition, the impact of applying fancy statistics to each of single, double, and 17-fold marking was also trialled. One way of doing this was to standardise each examiner's marks by subtracting their overall mean and dividing by their standard deviation. The predictive value of scores based upon single and double marking using these scaled marks could then be analysed. Alternatively, statistical processing could be done by treating each marker's scores as if they were scores from separate items and using a Rasch model[2] to calibrate the 'difficulty' of getting each particular mark from each given marker. Then for whichever marker (or combination of markers) was chosen within single, double, or multiple marking, a score could be assigned by combining the mark(s) assigned to the script with these calibrated difficulties. These steps were undertaken using the R package mirt (Chalmers, 2012).

Pseudo-CJ study

The central question in this article is whether any improved predictive value of CJ is achieved because of the judgemental process itself (i.e., improving the quality of judgements by making them relative judgements), or whether it is due to the increase in the number of judgements that are made about each script and the way these are analysed. In order to answer this question directly, a final method of creating scores out of traditional marks was trialled – pseudo-CJ. For the purposes of this method, all of the paired comparisons that were completed within the random CJ study were recreated based upon marks. For example, suppose that one row of data within this study indicated that the first judge in the RCJ study has compared *Script 131* and *Script 20*. We then look up the marks assigned to *Script 131* and *Script 20* by *the first marker* (not the same individual as the first judge because the judges used in different studies are entirely distinct and are randomly ordered in both cases). Whichever script was awarded the higher mark is treated as if they were judged as superior in a paired comparison study. If both scripts were awarded the same mark by the relevant marker, then the superior script in the pair is chosen completely at random. Suppose that the next row of RCJ data indicates that *Judge 10* compared *Script 108* and *Script 20*. We replace this data using information on which of these scripts was awarded the higher mark by *Marker 10*. This process is continued until the entire data set from the RCJ study has been replaced by paired comparisons based upon marks. The various paired 'comparisons' were then converted into scores in the same way as would be done for a real CJ study using the R package sirt (Robitzsch, 2018).

2. A more complex approach based upon the graded response model (GRM) was also trialled. However, it was not found to have any noticeable impact upon results and so, for brevity, the results from this method are omitted from this article.

Note that, the RCJ study was designed so that each judge only viewed each script once and, on average, each script was judged by 13.4 of the judges. As such, in the pseudo-CJ data, on average each script has data from paired comparisons of their marks against 13.4 other scripts based on the marking from 13.4 markers. It is very important for this element of the study that it is based on genuine instances of multiple marking. Whilst it would be possible to convert a single marker's marking into a set of paired comparisons, it would be pointless and would ultimately result in the same rank order of candidates as the original marks. For this method, it is the fact that the pseudo-CJ method combines the judgements of multiple markers that leads to the expectation that it will have increased predictive value relative to single marking.

The amount of examiner time required to produce scores is an important consideration in any research into different methods of marking. After all, regardless of how reliable it may be, there is little point in suggesting a method that would require hours of examiner time for every single essay. So, in order to provide context for the analyses we have described, we also examined the amount of time taken to complete traditional marking for each script and how this compared to the length of time required to complete each individual paired comparison. In addition, the analysis calculated the correlation between CJ scores and traditional marks in order to investigate whether the two approaches appeared to be generally measuring the same skills. Finally, in order to test whether scores from CJ might be more strongly influenced by superficial features of essays, the analysis investigated whether the association between essay length and scores differed depending upon whether CJ or traditional marking was used.

Results

Speed of marking

To begin with, since it affects our interpretation of results regarding predictive value, we present some information regarding the speeds of CJ, and of traditional marking. As part of the software used to undertake the CJ exercises, the time taken to make each individual judgement was recorded. The mean amount of time for each judgement was recorded as 3.5 minutes. Given that, as mentioned earlier, the essays were each around 500-words long (so each pair of essays consisted of about 1,000 words), this indicates that judgements took about 1 minute for every 300 words of written text. According to results published at http://www.readingsoft.com/, this is close to a typical on-screen reading speed for a good reader[3].

As part of the study by Child et al. (2015), examiners were asked to complete a questionnaire which included a question on how long they felt it took them to mark each script. It is important to note that markers were not only asked to mark the essays being studied in this article (Question 4 in the original exam paper) but also another slightly shorter question (Question 2). The results in Table 2 reflect the amount of time examiners felt that it took them to mark *both* of these questions. As can

3. A fairly recent blog has suggested that paired comparisons of GCSE English essays can be done within as little as 23 seconds on average (https://blog.nomoremarking.com/judging-gcse-english-efficiency-and-reliability-9a8df9b80096). With essays of the length of those in our research, this would mean judges were reading at almost 3,000 words per minute – a speed that is not credible (equivalent to reading this entire footnote in 2.2 seconds). It is possible that the research in the blog is based upon much shorter essays than our research. Alternatively, it may be that, in contrast to our research, judges saw the same essay many times and so could work from memory rather than re-reading each essay in full every time.

be seen, most examiners responded either that it took "6–10 minutes" or that it took "11–20" minutes to mark both of these questions. Given that there are slightly more respondents in the "11–20" category, we might estimate that it took markers around 11 minutes to mark each script on average. Table 2 gives information on the amount of time it took markers to mark two questions in one script: Question 2 (an essay marked out of 14) and Question 4 (an essay marked out of 40). If we assume that the amount of time taken to mark each of these questions is roughly proportional to the available number of marks, we can derive a rough estimate that it took about 7 minutes on average for an individual examiner to complete traditional marking for one essay.

Table 2: Examiners' responses regarding how long it took them to complete traditional marking

Typically, how long did it take you to mark one script (Questions 2 and 4)?	*Number of examiners*
3–5 minutes	2
6–10 minutes	6
11–20 minutes	8
Over 20 minutes	1
Total	**17**

Recall that each paired comparison in the CJ exercise took 3.5 minutes to complete, whereas we estimated that traditional marking for a single one of these essays took about 7 minutes (i.e., twice as long). Also note that each paired comparison provides a judgement for two essays, whereas marking only deals with one essay at a time. Putting these facts together indicates that CJ produces judgements for essays at about four times the speed of traditional marking.

Predictive value

The Spearman correlations of each method of producing scores for the English essays with scores on the two external achievement variables are shown in Table 3. To enable some of the main patterns to be seen more clearly, the same information is presented visually in Figure 1.

To begin with, we note that the predictive value of RCJ was higher than that from ACJ for both the English Language and English Literature assessments. This was despite ACJ having a reported (scale separation) reliability coefficient of 0.97 compared to 0.72 for the RCJ method. This confirms the conclusions from Bramley and Vitello (2018) that reliability coefficients from ACJ are biased upwards. It is simply not credible that ACJ could display improved reliability to this extent without it translating into any meaningful improvement in predictive value.

Before going too much further, it is worth noting that, as would be hoped by proponents of CJ, both the ACJ and RCJ methods provided greater predictive value than the raw mark values from traditional single marking. The difference between the predictive value of CJ and that of traditional marking was similar in both direction and scale to that reported by Steedle and Ferrara (2016). However, it is notable that all of the advantage of ACJ and almost half of the advantage of RCJ vanished once these marks were scaled either using the simple mean/*SD* method, or by using Rasch analysis. In other words, much of the apparent advantage of CJ can be explained by its use of fancy statistics. Just as statistical methods are used in CJ to ensure that scores are on a consistent scale, statistical methods can also be applied to traditional marks to ensure the same thing and, at least in this data set, doing this improved the predictive value of the marks.

Perhaps most importantly for our research question, we see that the pseudo-CJ method based upon marking yielded predictive values similar to the RCJ method. In other words, if the number of judgements for each essay and the method of analysis are held constant, it makes no difference whether the data comes from actual paired comparisons or from pseudo-comparisons derived from marks. This implies that there is nothing magical about placing two essays next to each other that enables humans to make better decisions regarding their quality. Rather, the benefit comes from the number of judgements that are made and the ways these are combined. Of course, it may well be that it is the simplicity of the paired comparison task, and the resultant increased speed of judgements, that facilitates collating this number of judgements.

The predictive value of RCJ is similar to that of double marking if marks are combined without any form of scaling. If double marking is combined with any form of scaling, its predictive value increases further. Table 3 also shows estimates of the amount of time taken by judges to produce the scores given by each method. As noted earlier, we expect that each marker took around 7 minutes to mark each essay meaning that double marking would take 14 minutes for each essay. We have also seen that each paired comparison took 3.5 minutes (or 1.75 minutes for each essay being compared) so that we would expect either CJ task to require more than 20 minutes of examiner time for each script on average. In other words, the time expended on each CJ method was roughly equivalent to the time required to complete triple marking. However, the predictive value of the method was only equivalent to that of double marking.

The predictive value of combining the marks from all markers was even greater. This is unsurprising as we would expect using a larger number of markers to improve the marking reliability of the resulting scores. Finally we note that, although the predictive value of marks from the PE was above that of double marking, the predictive value of their marks was not as high as that of the consensus mark derived across all markers.

Comparison of rank order of CJ score with combined mark from all markers

As noted in Table 3, the predictive value of RCJ scores was lower than that of marks derived from all markers combined – for example, using the mean of all awarded marks. With this in mind, it is of interest to explore whether this is because RCJ measures something fundamentally different to marking (perhaps rewarding different script features), or whether this can just be explained in terms of the relative reliability of the two methods. This question is explored in Figure 2 which shows the association between the two sets of scores. As can be seen, there was a relatively strong correlation between the two measures (0.80). To interpret this correlation, we note that the reported reliability of the RCJ method was 0.721. Thus, if the method of taking the mean of the 17 marks was perfectly reliable, and the two methods were measuring precisely the same thing, we would expect a correlation of $\sqrt{0.721} = 0.85$ between the two methods. That the actual correlation is not far off this value suggests that the RCJ method measures essentially the same skills as those rewarded in traditional marking.

Table 3: Different methods of scoring the essays and the Spearman correlations of the scores with marks achieved in other tests

Scoring method for English Language examination essays		*Average number of judgements on each essay*	*Estimated average time spent on each essay (minutes)*	*Correlation of score with…*	
				English Language Controlled Assessment (N=128)	*English Literature Whole Qualification (N=121)*
Comparative judgement	ACJ	14.4	25.2	0.512	0.578
	RCJ	13.4	23.5	0.572	0.611
Pseudo-CJ based on marks		13.4	93.8	0.562	0.650
Single marking	Raw marks	1	7.0	0.482	0.540
	Scaled marks (mean/*SD*)	1	7.0	0.512	0.576
	Rasch scaled marks	1	7.0	0.513	0.577
Double marking	Raw marks	2	14.0	0.556	0.620
	Scaled marks (mean/*SD*)	2	14.0	0.577	0.646
	Rasch scaled marks	2	14.0	0.577	0.646
All markers	Rasch model	17	119.0	0.661	0.738
	Median mark	17	119.0	0.646	0.732
	Mean mark	17	119.0	0.660	0.740
Principal Examiner's marks		1	Unknown	0.608	0.692

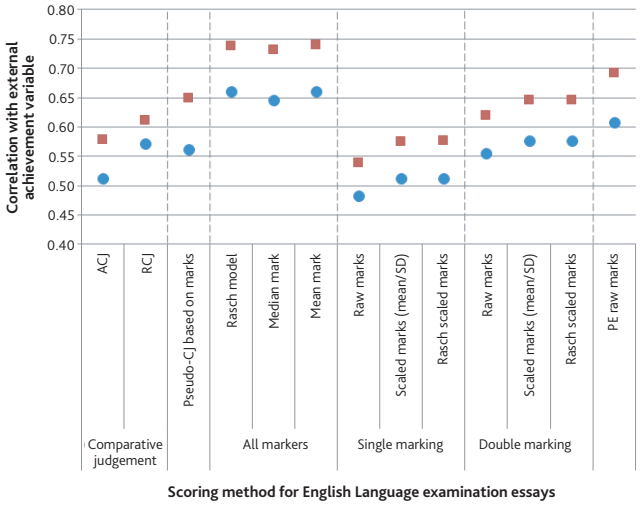

Figure 1: Spearman correlations of the scoring methods with external achievement variables

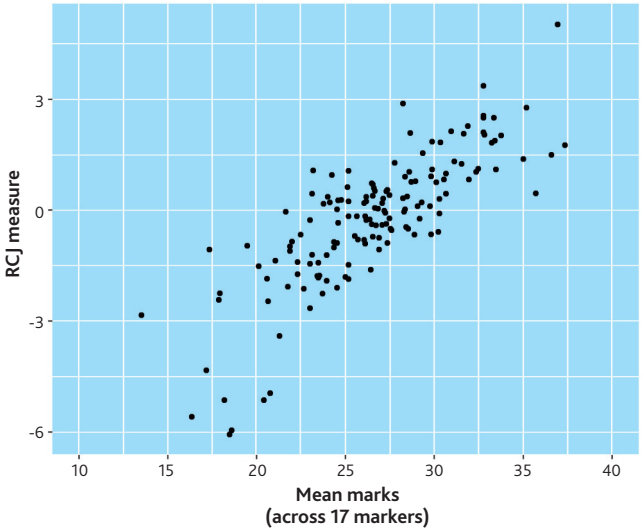

Figure 2: The association between scores from RCJ and the mean mark awarded to candidates across all 17 markers (i.e., not including the PE) (Note: Spearman correlation=0.80).

Relationship with length of responses

One concern over the use of CJ is that it is harder to check whether the features of responses that are used to make ratings match with those that are intended. For example, in the context of a study of comparative judgements of videos of an English speaking assessment, Pollitt and Murray (1996) mentioned that "It was no surprise to find evidence that the judges were influenced to some extent by the candidates' personalities, physical attractiveness, nationalities, and cultural backgrounds". In the case of our research, we might be concerned (as other authors have been) that judges taking part in a CJ exercise may be particularly susceptible to influence from extraneous factors, such as handwriting or the sheer length of responses (Bramley, 2007).

Although no measure of handwriting quality was available, one simple check that could be performed was whether the association of essay length with essay scores was consistent across the different methods of rating them. In particular, if judges within the CJ exercise were using essay length as a shortcut to make judgements, rather than fully considering the extent to which candidates have displayed the desired set of skills, then we might see a stronger association between essay lengths and scores from CJ than is the case with traditional marking. Of course, given the results shown in Figure 2, we have already stated that it appears that both CJ and marking are measuring the same set of skills. However, exploring the relationship with essay length provides another check.

The approximate word count of each essay was calculated using a method similar to the one described in Benton (2017). Figure 3 shows a scatter plot of these approximate word counts against standardised scores from three methods: RCJ, the mean mark from all markers (except the PE), and the PE's marks. Scores were standardised by subtracting the mean and dividing by the standardisation. Standardised scores are used in Figure 3 as it puts results from all of the different scoring methods on the same scale.

The relationship between essay length and scores in Figure 3 is reminiscent of a similar chart shown in Benton (2017) for English Literature essays. As was seen in the earlier research, very short essays tended to be awarded lower marks than those of average length or longer. However, there is no evidence that very long essays were awarded higher scores than those of average length. More importantly for our research, we can see that the link between essay length and scores was very similar regardless of which method was used to generate them. In particular, there is no evidence of CJ being any more likely to favour long essays than the PE. This supports the earlier evidence that, in this study, both CJ and marking were measuring the same set of skills. Given that the judges involved in the CJ exercise were already experienced markers of the examination being studied, perhaps this similarity should come as no surprise.

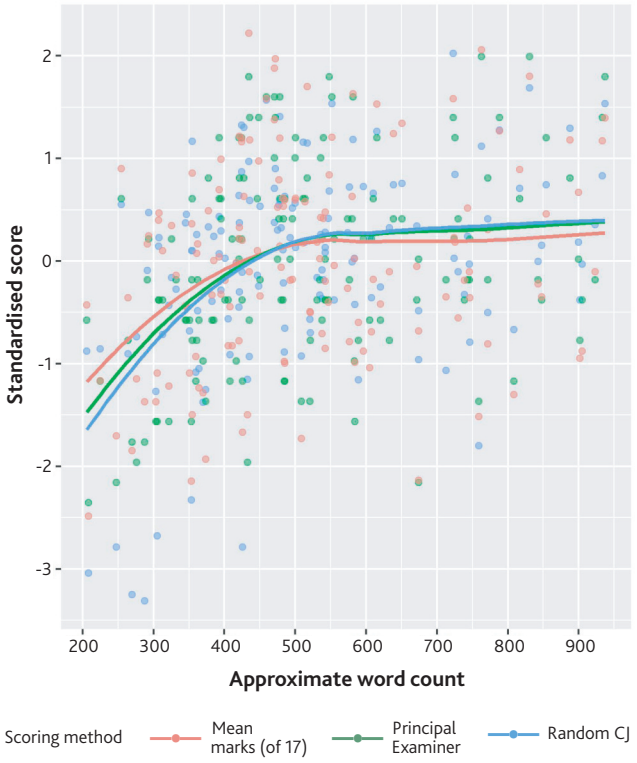

Figure 3: The association between approximate word count and essay scores derived via three methods

Conclusion

The central conjecture of this article was that CJ is just a form of multiple marking combined with a simple mark scheme and fancy statistics. The evidence from this study supports this conjecture. In particular, we have seen that if we recreate the paired comparisons in a CJ study using marks from a multiple marking study, the resulting scores from such a pseudo-CJ exercise have at least as much predictive value as scores from an actual CJ exercise. In other words, it is the number of judgements that are made about scripts and the way they are analysed within a CJ exercise that is important. The physical act of placing two essays next to each other and deciding which is better does not appear to produce judgements that, in themselves, have any more predictive value than getting the same individual to simply mark a set of essays.

Having seen that the strength of CJ lies in its use of multiple judgements and a strong statistical model, the evidence in this article suggests that these same techniques could be employed in traditional marking to improve the reliability and validity of scores. This need not be achieved by fully recreating CJ studies. For example, a statistical model can be applied to scores from traditional marking through much simpler techniques, such as marker scaling to address differences in the leniency and severity of different examiners.

The key question in whether CJ could replace marking is which of the two methods makes better use of the resources available to us. In this particular study, it appears that multiple marking was the more efficient. In particular, whilst the various CJ exercises took roughly as long as triple marking, they only achieved the predictive value of double marking. Thus, in this case, double marking (combined with statistical scaling) would appear to be the more efficient method.

It is important to note that, this is only a single study and has only considered CJ and marking for a particular task – marking GCSE English essays. Alterations to the design of either the CJ study or the marking study may lead to different results. For example, what would happen if judges in the CJ study were explicitly encouraged to make decisions more quickly? Could this be done without harming the reliability and validity of the resulting scores? Or in the case of marking, what if the mark scheme was simplified to encourage marking to be done more quickly with the express intention of subsequently using statistical scaling to iron out differences in leniency and severity between examiners? Indeed, given that our evidence shows that, provided they are analysed properly, absolute judgements can be just as useful as relative judgements, it is natural to ask how the process of producing absolute judgements can be

made quicker. For example, would getting examiners to give an intuitive score for each essay on a scale from 1–10, and then processing these scores using a Rasch model result in an equally useful set of student scores as traditional marking, but in a fraction of the time?

It is, of course, not necessarily true that the results shown here with respect to GCSE English essays would be repeated for other subjects. It is doubtless the case that certain types of student performances lend themselves more readily to CJ, whereas others are easier to mark. Nonetheless, the results here are important in understanding where the benefits of CJ derive from. Recognising that these are not solely caused by switching the way in which judgements are elicited, but also in the number of such judgements and how they are analysed, allows for a more nuanced comparison of the relative advantages of CJ and traditional marking. Failing to recognise these differences risks the two approaches never being compared on a like-for-like basis.

This article should not be taken as a criticism of the existing system for marking high-stakes examinations. Indeed, the examinations regulator in England, Ofqual, have themselves stated that "fundamentally, we believe this is a system that people can have confidence in" (Ofqual, 2014b, p.3). However, seeking for improvements in any system is worthwhile, and we hope that this article can provide a useful addition to existing research in terms of thinking about whether, and how, CJ might form a part of any improvements.

References

Benton, T. (2017). How much do I need to write to get top marks? *Research Matters: A Cambridge Assessment publication, 24*, 37–40.

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

Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgement*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Bramley, T., & Vitello, S. (2018). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, https://doi.org/10.1080/0969594X.2017.1418734.

Chalmers, R.P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1–29.

Child, S., Munro, J., & Benton, T. (2015). *An experimental investigation of the effects of mark scheme features on marking reliability*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Christodoulou, D. (2018). Comparative judgement – the next big revolution in assessment? *researchED*, *1*(1), 13–14.

Ofqual. (2014a). *Review of Double Marking Research*. Ofqual/14/5381. Coventry: Office of Qualifications and Examinations Regulation.

Ofqual. (2014b). *Review of Quality of Marking in Exams in A Levels, GCSEs and Other Academic Qualifications*. Ofqual/14/5379. Coventry: Office of Qualifications and Examinations Regulation.

Pollitt, A. (2004, June). *Let's stop marking exams*. Paper presented at the annual conference of the International Association for Educational Assessment, Philadelphia, USA.

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic, & N. Saville (Eds.), *Performance testing, cognition and assessment. Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (pp.74–91). Cambridge, UK: Cambridge University Press.

Robitzsch, A. (2018). *sirt: Supplementary item response theory models*. R package version 2.5–45. Retrieved March 19, 2018 from https://CRAN.R-project.org/package=sirt

Steedle, J. T., & Ferrara, S. (2016). Evaluating Comparative Judgment as an Approach to Essay Scoring. *Applied Measurement in Education, 29*(3), 211–223.

How have students and schools performed on the Progress 8 performance measure?

Tim Gill Research Division

Introduction

In October 2013, the Department for Education (DfE) announced that new 'headline' performance measures for schools would be introduced to replace the previous measure of the percentage of students achieving five or more grades A* to C at General Certificate of Secondary Education (GCSE) Level, including English and Mathematics. The new measures (known as *Attainment 8* and *Progress 8*) are based on performance in a student's best eight subjects at GCSE (or equivalent), although with some restrictions; students are required to take the English Baccalaureate (EBacc[1]) qualifications in English and Mathematics, as well as at least three other EBacc qualifications. The remaining three slots can be filled either by other EBacc qualifications or by other approved, non-EBacc qualifications.

One of the reasons for the introduction of the new measures was concern that the previous measures penalised schools with a low-attaining intake. As Progress 8 is a value-added measure, it already accounts for the prior attainment of the student and should, in theory, no longer penalise these schools. The following method is used to calculate school-level Progress 8 scores:

- Calculate the Attainment 8 score for each student. This is the total points score for their eight highest scoring eligible qualifications. Points are based on the grade achieved (e.g., for GCSEs, points are on a 1–8 scale; 1 = G, 8 = A*[2]).

1. EBacc is the *English Baccalaureate*, a school performance measure which shows the proportion of pupils studying the 'core' academic subjects at KS4. Only specific qualifications (mainly GCSEs) are eligible for inclusion in the EBacc.

2. This example is for 'old' GCSEs. The scores for new GCSEs (9–1 grading scale) are slightly different.