Research Matters / 33

A Cambridge University Press & Assessment publication

Editorial – the CJ landscape

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Tom Bramley Director, Research Division

References

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

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

Bramley, T. (2012). The effect of manipulating features of examinees' scripts on their perceived quality. *Research Matters: A Cambridge Assessment publication*, *13*, 18–26.

Bramley, T., & Oates, T. (2011). Rank ordering and paired comparisons - the way Cambridge Assessment is using them in operational and experimental work. *Research Matters: A Cambridge Assessment publication, 11*, 32-35.

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

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