CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

How do judges in Comparative Judgement exercises make their judgements?

Conference Abstract

Tony Leech
Lucy Chambers

AEA-Europe conference, 10-12 November 2022

Author contact details:

Tony Leech & Lucy Chambers
Assessment Research and Development
Research Division
Shaftesbury Road
Cambridge
CB2 8EA
UK

anthony.leech@cambridge.org
lucy.chambers@ cambridge.org
https://www.cambridge.org/

As a department of the university, Cambridge University Press & 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 University Press & 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: Research Division

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:

Leech, T. & Chambers, L. (2022, 11 November). *How do judges in Comparative Judgement exercises make their judgements?* Open paper presentation, 23rd Annual Meeting of the Association for Educational Assessment – Europe, Dublin, Ireland.

Use of comparative judgement (CJ) in assessment contexts is increasing. It has two main applications: as an alternative to marking and for maintaining standards (whereby grade boundaries in an exam are decided such that it is no easier or more difficult for a candidate to get a grade in the current year as in previous years). This paper concerns the latter.

Currently, in England, standard maintaining involves the use of cohort level prior attainment data to suggest statistically recommended grade boundaries that retain the standard of performance at a grade from year to year. Judgement of actual scripts is involved chiefly as a check on these boundaries. This system has been criticised for not allowing for legitimate cohort level improvements in ability over time to be reflected. For this reason, Ofqual, the qualifications regulator in England, has taken an interest in methods that allow better use of judgemental evidence, such as CJ (Curcin et al, 2019).

Recent technological developments including routine scanning of student scripts mean assessment processes are progressively moving to being digitally facilitated – something partially fuelled by the Covid-19 pandemic. This, and the digital implementation of CJ, have allowed CJ standard maintaining methods to be investigated. Standard maintaining using CJ involves judges comparing packs of two or more candidate scripts, with each pack containing scripts from both the current year and a benchmark year, to decide which candidate responses are better. Statistical methods are then used to process the results to determine the overall difference in difficulty between the two years' papers (Benton, 2021). CJ has produced reliable and usable results as a standard maintaining tool (Benton et al, 2022).

However, two central issues in CJ, perhaps under-explored, are "what processes do judges use to make their decisions" and "what features do they focus on when making their decisions?" Curcin et al. briefly discussed these, finding that judges in their pilot CJ exercises mainly judged scripts question by question, gave questions with more marks a higher weighting in their overall judgement, used missing responses as a differentiator of quality, and based their judgements on mark scheme requirements. No judges explicitly suggested they were re-marking scripts. Subject-specific features of candidate responses were important to judges, while, pleasingly, superficial features were seldom mentioned.

We extend this discussion by reference to the outcomes of a series of OCR/Cambridge University Press & Assessment studies exploring the use of CJ for maintaining standards, conducted using in-house CJ software. Our contribution is to focus explicitly on what CJ judges are doing when judging, and what they attend to in their judgements. We therefore render more explicit some of the validity assumptions underlying both CJ, and standard maintaining more generally.

We report results of both a study into the processes used by judges when making CJ judgements, and the outcomes of surveys of judges who have used CJ. First, using insights from observations of judges undertaking a CJ exercise on a physical education exam paper and their being asked to think aloud while they judged, we highlight the variety of processes used when making their decisions, including, encouragingly, comparative reference of scripts to each other, but also re-marking and question-by-question evaluation. Then we explore the self-reported views of judges from varied CJ exercises. Issues discussed include whether they were able to make holistic judgements and take into account differences in difficulty between

papers from different years, and what parts of papers or types of questions they attended to the most in coming to their judgements.

From this, we develop a four-dimension model of the features that impact what judges attend to – the structure of the question paper, different elements of candidate responses, judges' own preferences and features of the CJ task itself – and explore how these interact in complex ways to contribute to judge decision-making.

We conclude by questioning, in light of these factors, whether the judgements made in CJ (or in the judgemental element of current standard maintaining procedures) are meaningfully holistic. Our insights, alongside those relating to the reliability and practicality of CJ, therefore challenge and complicate the issue of whether CJ should be chosen to replace the judgemental element of current standard maintaining procedures. This work therefore makes an important contribution to the discussion of the validity of this assessment process, made possible by emerging technology, and to discussions of how assessment is undertaken, and standards maintained, in the future.

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., 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.

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