Comparative judgement
assess student production without absolute judgements

Josh Sumner[1]

Potential impact	medium
Timescale	medium term
Keywords	comparative judgement, assessment, reliability

What is it?

Comparative Judgement (CJ) has emerged as a technique that typically makes use of holistic judgement to assess difficult-to-specify constructs such as production (speaking and writing) in Modern Foreign Languages (MFL). In traditional approaches, markers assess candidates' work one-by-one in an absolute manner, assigning scores to different elements (analytic marking). In CJ, however, markers compare two pieces and consider the overall merits of each. They make one binary, holistic judgement as to which is better. This approach exploits humans' natural ability to compare; we find it easy, for example, to say which of two people is taller, but struggle to give precise estimates of height.

By using a collection of 'paired comparisons', in which items are judged several times, a rank order from 'worst' to 'best' is produced. Properties such as overall consistency of judgement can be evaluated, as can difficult-to-rate items or unreliable assessors.

Technology facilitates implementation of CJ: work is uploaded to web-based software. Multiple markers ('judges') make comparisons of two pieces of work presented side-by-side. Software using adaptive CJ, involving 'rounds'

1. Dean Close School, Cheltenham, United Kingdom; jma.sumner@gmail.com; https://orcid.org/0000-0002-5490-9266

of marking of work increasingly similar in quality, requires fewer comparisons but produces arguably equally reliable rank orders. CJ has proven reliable in assessment of first language, mathematical problem-solving, and written work in humanities. Findings include a higher level of inter- and intra-assessor reliability compared to traditional assessment, though research into application in MFL is limited; Pollitt and Murray's (1993) small-scale study concentrated on foreign language speaking, and there have been trials in some UK schools.

As research has found that 23% of students candidates receive the 'wrong' grade at General Certificate of Secondary Education (GCSE) in MFL using traditional techniques (Rhead, Black, & de Moira, 2018, p. 17), teachers, school leaders, and examination boards may consider eschewing analytic marking using criterion-based mark schemes in favour of holistic CJs.

Example

The MFL department at Sandringham Research School (2018) trialled CJ using the software *No More Marking* (www.nomoremarking.com) to assess writing in end-of-year exams. Teachers were presented with pieces of two anonymised students' work – both their own and others' – on screen side-by-side, and judged which was overall 'better'. The same piece of work was judged numerous times, by different teachers; through different comparisons, an algorithm brought together all judgements, providing a rank order. The department found a reliability metric of 0.89 and that student work was quicker to assess, though could not be used to give individual feedback.

In future, the introduction of pre-marked items into comparisons, 'anchor responses', could allow grades to be assigned using norm-referencing. This technique could be used by examination boards.

Benefits

With CJ, there is no change to the preparation or administration of tasks, only to assessment, but its benefits are numerous.

CJ saves time; judges make one judgement rather than numerous ones against different criteria. This replicates the natural process of reading and is faster. Higher reliability is achieved without needing time-consuming moderation.

Used across a department, as part of the process teachers see not only the work of their own class, but a range of student responses, without requiring judgement on the reliability of colleagues' marking as in a moderation. CJ thus has a formative perspective for teachers.

CJ does not require elaboration of mark schemes prior to a test, nor in a 'standardisation' process. 'Unpredictable' responses are more easily dealt with, and teachers may find students produce more novel, ambitious responses; traditional marking may stymie linguistic development and limit creativity as students are concerned with 'jumping through hoops'. CJ exploits teachers' expert knowledge and professional competency of 'good' production without demanding it be tightly defined.

CJ allows for a more accurate ranking order by avoiding markers using the middle of any level-based rubric, precluding the 'bunching' of marks due to reluctance to give zero or full marks. Determination of a rank order is more accurate than with criterion-based marking and inter-assessor reliability is higher due to repeated comparisons.

Potential issues

CJ is only suitable for summative assessment. Analytic scales provide feedback to students and teachers regarding relative strengths and weaknesses. A position in a ranking order, or a score, gives no information regarding learning, nor how to improve. Teachers wanting to use a task assessed through CJ formatively may need to mark work again analytically. However, subsequent instruction could be improved by teachers' knowledge of a cohort's performance. Examination boards may be reluctant to adopt CJ. The relativistic approach makes it difficult to appeal marks; the basis of assessment is a series of comparisons by numerous examiners, not transparent scores given by one.

Use of holistic judgement precludes weighting of elements of production (e.g. communication is weighted more highly than accuracy at GCSE). Markers may be swayed by salient features, such as inaccuracy of spelling. In addition, implementation of CJ for speaking is problematic, as it relies on memory; two audio files can only be subsequently rather than concurrently compared, unlike writing.

Furthermore, the absence of prescriptive mark schemes, hailed as a benefit, may only work for so long. Research into use of CJ in Geography found examiners used mark schemes implicitly due to knowledge of criteria of traditional approaches: a shared construct existed in an established community of practice through familiarity with extant methods.

Looking to the future

> The issues involved in using CJ to assess MFL production are much like those involved in assessing other complex constructs, and studies into its use for these have been positive. There is nothing, in my opinion, that makes MFL production, particularly writing, a wildly different construct. Consideration of implementation of CJ is crucial, lest we resign ourselves to the unreliability of current assessment. Examination boards could consider its use in high-stakes assessment, and schools could employ it to produce more reliable internal assessments which also save teachers time.

References

Pollitt, A., & Murray, N. L. (1993). What raters really pay attention to. *Studies in language testing, 3*, 74-91.

Rhead, S., Black, B., & de Moira, A. (2018). *Marking consistency metrics: an update*. Ofqual.

Sandringham Research School. (2018, December 7). Comparative judgement: a workshop exploring its use in MFL. *Sandringham Research School Blog*. https://researchschool. org.uk/sandringham/news/comparative-judgement-a-workshop-exploring-its-use-in-mfl/

Resources

Christodoulou, D. (2016). *Making good progress? The future of assessment for learning*. Oxford University Press.

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

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in education: principles, policy and practice*, 19(3), 281-300. https://doi.org/10.1080/096959 4x.2012.665354

Stock, P. (2017). Compare the marking: using comparative judgement to assess student progress at secondary level. *Impact, 1*, 12-15.

Oxford Education. (2019). *What is comparative judgement?* https://youtu.be/KLVtvBPXQ2U

If you want to try a simple comparative judgement experiment, you might enjoy doing this activity: https://www.nomoremarking.com/demo2

Innovative language pedagogy report
Edited by Tita Beaven and Fernando Rosell-Aguilar

Publication date: 2021/03/22