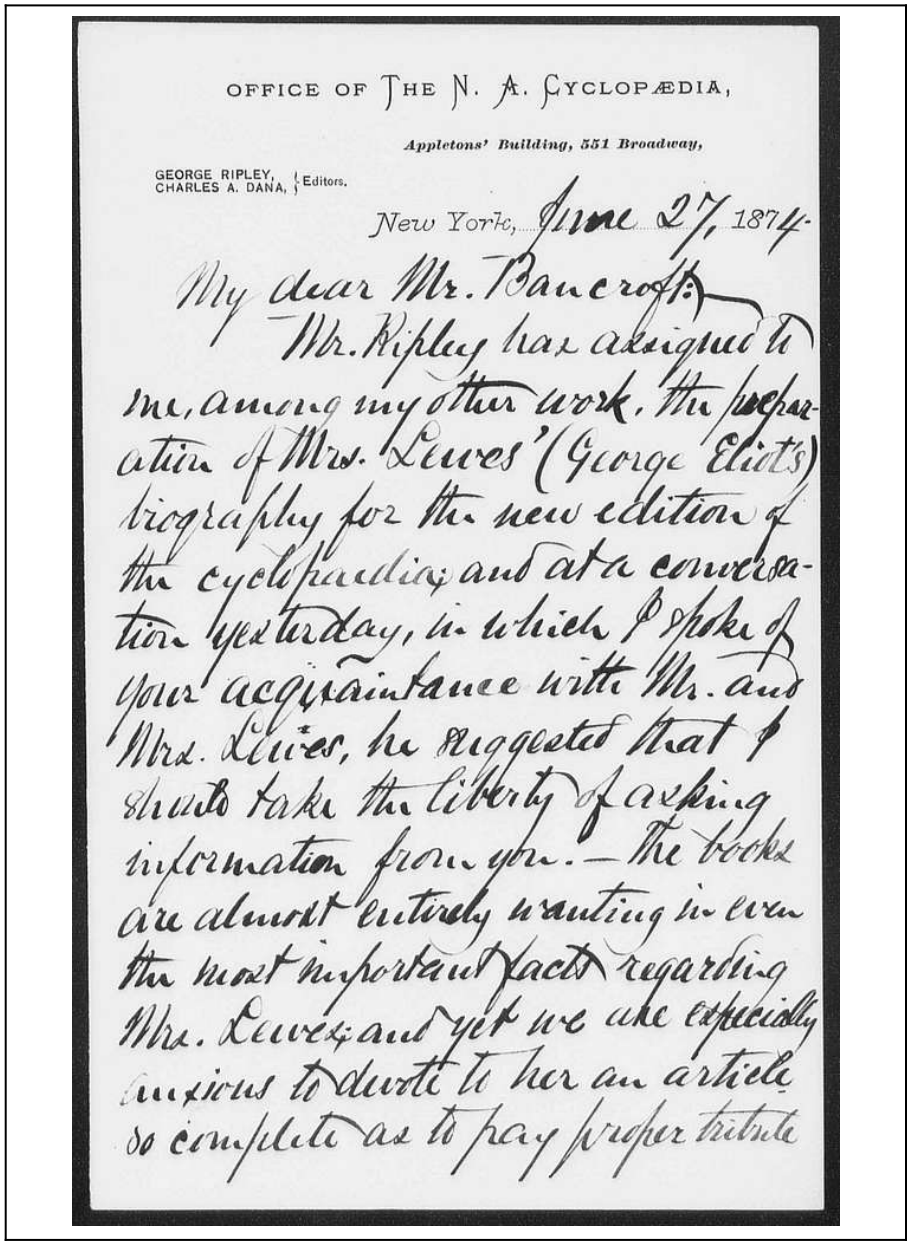
What is comparative judgement, and how can it help you?

Dear readers,

I'd like to introduce you to an assessment method that I think you might find interesting. It's called Comparative Judgement (CJ). In this post I'm going to explain how it works, and then in a later post I'll share with you some findings from our research at UCLouvain on CJ for second-language assessment. But before I explain how comparative judgement works, I want to tell you how it *doesn't* work.

Imagine being shown a sample of someone's handwriting, like the example below. Your task is to give me a grade reflecting how "good" it is. How would you go about this task?

Perhaps a suitable first step would be to make sure you could understand the handwriting. You might give it a score on this quality – legibility – perhaps using a five-point scale. The handwriting below probably scores quite poorly on this criterion, requiring some effort to decipher. But it is nevertheless consistent, well-proportioned, and rather elegant. These things should probably count towards its final score too, so maybe we need another criterion. But which one? And should it have equal weighting with legibility? What other features need to be considered?



OFFICE OF THE N. A. CYCLOPÆDIA,

Appletons' Building, 551 Broadway,

GEORGE RIPLEY, CHARLES A. DANA, } Editors.

New York, June 27, 1874.

My dear Mr. Bancroft:

Mr. Ripley has assigned to me, among my other work, the preparation of Mrs. Lewes' (George Eliot's) biography for the new edition of the cyclopædia; and at a conversation yesterday, in which I spoke of your acquaintance with Mr. and Mrs. Lewes, he suggested that I should take the liberty of asking information from you. — The books are almost entirely wanting in even the most important facts regarding Mrs. Lewes; and yet we are especially anxious to devote to her an article so complete as to pay proper tribute

Making evaluations of abstract constructs like "quality of handwriting" is not as easy as it first appears. Doing it well requires numerous steps - construct definition, rubric development, rater training – all of which can be time-consuming and expensive. These steps take on additional importance in high-stakes contexts, where there is pressure to deliver reliable and consistent results across many different test items.

Comparative judgement emerged as a response to the difficulty of making evaluations like this. But it was not the issue of construct definition that led its developer, Louis Thurston, to come up with it. It was, instead, the simple

insight that it is often easier to judge things in comparison to other things than it is to evaluate them in isolation: comparison facilitates decision-making.

The first paper on comparative judgement was published by Thurston in 1927. It involved presenting judges with test items, such as handwriting samples, side-by-side, and asking these experts simply to decide which of them was "better". There was no rubric: judges were expected to simply consult their own intuitions - what CJ researchers Ian Jones and Ben Davies have more recently described as their "'know it when you see it" knowledge – to make decisions. Through these studies, Thurston demonstrated that it was possible to generate reliable measurements of concepts like "quality of handwriting", for which no objective scale (such as weight in grams or height in centimetres) could exist.

But this blog is not about handwriting, it's about assessing English for Academic Purposes. So is it really possible to apply this simple, rubric-free method of comparative judgement to a task as complex and multi-dimensional as, for example, the evaluation of argumentative essays?

In fact, CJ is already being used for educational assessment in quite a wide range of contexts, and at increasingly large scale. A 2020 paper by Chris Wheadon and colleagues, for example, reported on a project involving the comparative judgement of more than 50,000 pieces of L1 writing across 85 UK primary schools. The study suggested numerous benefits of CJ in this context, including strong reliability and validity, increased efficiency compared to more traditional methods, and the potential to reduce judge bias by ensuring that every student's work is evaluated by several judges, rather than only by their class teacher.

There are currently few studies which look at CJ for L2 assessment. However, those that exist have generally reported similar findings. For example, our research team at UCLouvain last year published a pilot study involving the comparative judgement of TOEFL essays by members of the linguistic community, who we recruited using crowdsourcing. We found that the grading scale generated by these participants had a reliability of .95, and overlapped to a significant degree with ratings derived from more traditional, rubric-based methods (Paquot et al., 2022). This early evidence therefore suggests some promise for CJ as a way to assess L2 writing.

What are the mechanics behind the apparent efficacy of CJ? One of the keys is a basis in collective, rather than individual, decision-making. In CJ, each item in a test set is not judged only once (or perhaps twice, in high stakes contexts) as in rubric-based assessment. Instead, it is compared to many other items – sometimes as many as twenty or thirty. These comparisons are spread between many judges – for example, all the teachers of a given subject across a school district, or all the EAP teachers in a given university. Each judge is assumed to bring a slightly different understanding of the target construct, or slightly different preferences regarding the weight of each aspect, to the mix. For example, in the context of L2 writing assessment, one judge might place the highest value on grammatical accuracy, while another might pay more attention to the overall coherence of the text.

With each decision therefore being based on slightly different intuitions, many factors are taken into consideration and a broad, collective conception of the target construct is allowed to emerge. Studies on the validity of CJ provide evidence of this. Typically, these studies involve judges being asked to provide comments on the rationale for each comparative judgement they make. These comments are then analysed to test the extent to which the judging team, as a whole and as individuals, consider the full extent of the target construct. In the context of L1 writing assessment, several such studies have suggested that while few individual judges mention all aspects of a construct, across an entire judging team there is good construct coverage (Van Daal et al., 2019), with very few comments concerning construct-irrelevant features (Chambers & Cunningham, 2022).

The evidence suggests, then, that CJ can provide reliable and valid assessments even of complex constructs like writing quality. But it's difficult to be swayed by research findings alone, so perhaps you'd like to try CJ for yourself? Our ongoing projects make it possible for you to do that. By

following the link below, you'll be taken to a platform where you can provide 5-10 comparative judgements of argumentative essays from the ICLE corpus (preceded by a short survey and consent form). This will give you a taste of how the CJ process works, while at the same time helping us to answer research questions about CJ for L2 assessment. If you *do* decide to try it out, would be hugely grateful if you could try to provide at least 5 judgements, since this is the threshold for being able to use your data scientifically.

The link is: www.tiny.cc/L2assessment.

We hope you've enjoyed reading this first post on CJ! In a second post to follow, we will share some initial results from our research exploring the reliability of CJ for L2 writing assessment, and ask whether expert graders are really any better at judging L2 writing than laypeople are. We hope you'll join us again to find out!

Peter Thwaites

Centre for English Corpus Linguistics
Institute of Language and Communication
Université Catholique de Louvain, Belgium

References

Chambers, Lucy, and Euan Cunningham. 'Exploring the Validity of Comparative Judgement: Do Judges Attend to Construct-Irrelevant Features?' Frontiers in Education 7 (2022). https://www.frontiersin.org/articles/10.3389/feduc.2022.802392.

Daal, Tine van, Marije Lesterhuis, Liesje Coertjens, Vincent Donche, and Sven De Maeyer. 'Validity of Comparative Judgement to Assess Academic Writing: Examining Implications of Its Holistic Character and Building on a Shared Consensus'. Assessment in Education: Principles, Policy & Practice 26, no. 1 (2 January 2019): 59–74. https://doi.org/10.1080/0969594X.2016.1253542.

Jones, Ian, and Ben Davies. 'Comparative Judgement in Education Research'. International Journal for Research and Method in Education, under review.

Paquot, Magali, Rachel Rubin, and Nathan Vandeweerd. 'Crowdsourced Adaptive Comparative Judgment: A Community-Based Solution for Proficiency Rating'. Language Learning 72, no. 3 (2022): 853–85. https://doi.org/10.1111/lang.12498.

Thurstone, L. L. 'A Law of Comparative Judgment'. Psychological Review 34, no. 4 (1927): 273–86. https://doi.org/10.1037/h0070288.

Wheadon, Christopher, Patrick Barmby, Daisy Christodoulou, and Brian Henderson. 'A Comparative Judgement Approach to the Large-Scale Assessment of Primary Writing in England'. *Assessment in Education: Principles, Policy & Practice* 27, no. 1 (2 January 2020): 46–64. https://doi.org/10.1080/0969594X.2019.1700212.