Assessing L2 writing using comparative judgement: what works?

Dear readers,

Last month on this blog, I wrote a guest post about comparative judgement (CJ) – a method of assessment in which pieces of student work are compared with each other rather than being evaluated in isolation. I argued there that CJ has several advantages over more widespread forms of assessment, such as rubric-based marking: it is generally less time-consuming to perform, also reduces the time and costs associated with training raters to use a rubric, and offers a broader, more inclusive approach to construct definition.

In this second post, I look at research that explores the suitability of comparative judgement for the assessment of second-language writing. But before I get into this research, it's worth briefly asking why it is necessary to put forward novel methods for L2 assessment in the first place.

When we assess our students' writing, there are several things that we would hope to be able to demonstrate. One is that our assessments are reliable – that is, given the same piece of student work being judged on the same criteria, the same score would be given regardless of what day it is, who is doing the evaluation, or any number of other potentially intruding factors. Another is that the evaluation should be valid – it should genuinely assess the aspects of L2 writing that it was intended to. It should also be efficient in terms of time and financial costs.

The problem that we tend to encounter is that the methods which give us the highest reliability and validity are also generally the most expensive and time-consuming. Analytic rubric-based methods, such as those used in the IELTS exam, are a case in point. While the careful rubric development rubric process and the extensive training given to raters make this a gold-standard approach in terms of reliability and validity, both processes make enormous demands on time and money. This may be feasible (indeed, necessary) for a high-stakes exam such as IELTS, but there are many testing scenarios in which it is less achievable. As a result, there will always be a need for assessment methods which offer gains in reliability, validity, or efficiency, so long as they don't sacrifice too much from any other area. The key questions pertaining to CJ (or any other novel method of assessment) is, how far can the method take us in any one of these directions, and what (if anything) is sacrificed to get there?

Regarding CJ, it appears on the surface that quite a lot of progress has already been made on these questions, despite the small number of studies conducted. For example, one study by a team of researchers team at Brigham Young University (Sims et al., 2020) compared CJ with more traditional holistic rubric-based assessment, and found similar levels of reliability and validity for the assessment of a set of 60 L2 academic essays. However, the CJ approach was more efficient, requiring 52 seconds less per essay than the rubric-based approach. Another study by our team at the Centre for English Corpus Linguistics, UCLouvain (Paquot et al., 2022), found that a CJ approach to grading 50 TOEFL writing exam scripts yielded very high levels of reliability and significant overlap with earlier rubric-based assessments (a measure of validity). The study therefore supports the view that CJ can deliver strong reliability and validity, although it does not comment on efficiency.

These results are encouraging. But if we dig beneath the surface, we find that many details remain unexplored. Here are some questions that still need to be answered:

- What proficiency ranges can CJ be used effectively with? Research in other fields has suggested that CJ is most effective where texts are diverse – after all, it is easier to differentiate an A2 from a C1 text than it is to tell apart texts at B1 vs B2 levels. Both the Sims et al. and Paquot et al. studies above worked with broad proficiency spans. Can CJ still work effectively with narrower ranges?

- Can CJ still prove reliable with longer texts? No study to date has reported high reliability with L2 texts longer than around 250 words. Is this the upper limit?
- Do the texts under evaluation all need to be answers to the same question, or can texts on different topics be compared with each other?
- How much training and experience do judges need? Studies have found that CJ can be used effectively with novice graders, but that they reach acceptable levels of reliability more slowly than experts (Verhavert et al., 2019). Sims et al.'s study supported this, finding that novice judges (who were undergraduate students minoring in TESOL) were slightly less reliable than experts (experienced writing teachers and examiners) after the same number of rounds of CJ assessment. But can this finding be replicated? And is this basic TESOL experience necessary?

These are among the questions that our research team is seeking to answer. Our approach to date has been to use crowdsourcing to recruit judges, who then complete paired comparisons of texts. We've used two distinct methods – our "expert" judges have been recruited through mailing lists and social media (and our blog posts!) related to applied linguistics, while our "novices" are users of the crowdsourcing platform Prolific. The table below sets out the experiments we've conducted so far.

	Paquot et al, 2022	Paquot replication	New study 1	New study 2
Judges	Experts	Novices	Experts	Experts
Text length	250 words	250 words	550 words	600 words
Proficiency range	A1-C2	A1-C2	B1-C2	B1-C2
n. essay prompts	1	1	1	5
Reliability	.95	.95	.81	.82

Four studies in the CLAP CJ project.

There are three main findings to share. Firstly, we found that our novice judges were able to generate grades of very similar reliability and validity to our experts. Moreover, the use of Prolific meant that our novices – most of whom had zero experience of teaching or evaluating L2 writing – were able to complete almost 400 comparisons in less than 12 hours, and at a cost lower than rubric-based assessment.

Secondly, we found that expert judges were able to generate reliable evaluations of texts which were longer (around 550 words) and reflected a narrower proficiency range (B1-C2), than those used in earlier studies. The statistical level of reliability of around 0.8 was lower than the 0.95 reported in studies using shorter, less homogeneous texts, but still comparable to the reliability of many rubric-based assessments (the difference in reliability might also be related to differences in the comparison methods – those used in the earlier studies may have slightly inflated reliability levels).

Thirdly, we found essentially the same level of reliability in a study using experts to judge texts of the same length and proficiency range as those above, but which comprised answers to five different essay prompts instead of just one.

Of these preliminary findings, it is the first which tends to catch people's attention. Can it really be that judge expertise is not needed? In fact, there are two caveats to this finding which strictly limit their interpretation. One is that they concern the set of texts with wide variance in proficiency: we don't know what would happen if we asked novices to evaluate the more homogeneous texts. The other is our focus on reliability: we don't know how novices and experts differ in terms of construct validity. In other words, do novice judges pay due attention to the same breadth and depth of textual features that experts do? It may be, for example, that novices consider only a few obvious features, such as grammatical accuracy or

task achievement, but miss features that an expert would spot. Over the coming months, we will investigate these and other issues in greater detail.

Grading is much discussed these days. In the UK, workload-related pressures have turned marking into a flashpoint, while continuing developments in AI-based Automated Essay Scoring (AES) pose questions as to how long humans might continue to evaluate student work. With the landscape changing so fast, it's essential to start talking about what kind of a future we want regarding who grades student work, and how they grade it. Do we wish to insist upon human grading? What is the role of our expertise? How can we make the grading process more efficient without compromising on our principles?

CJ can contribute to these discussions. It appear to offer a reliable and efficient approach to grading, able to be adapted to make use of as much expertise as is considered necessary. It therefore offers a distinct new assessment option, for example, for teams of teachers who want to keep grading in the human domain but nevertheless needs gains in efficiency. But it can also, as the above experiments show, be used to compare the reliability, validity, and efficiency of various groups of judges – perhaps including those powered by AI. We hope that our studies will help to stimulate such research, helping us to make the best possible decisions about the future of grading.

Meanwhile, we are almost at the end of our "expert" judge data collection period, having collected more than 1200 comparative judgements from more than 150 judges. Thanks to everyone who has taken part! If you'd like to give CJ a try but haven't yet had a chance to do so, there is still time. Please visit www.tiny.cc/L2assessment to get involved.

Thanks for reading these blog posts! If you'd like to stay in touch with developments on this research project, please follow our Twitter @cecl_UCL.

Peter Thwaites

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

References

Paquot, M., Rubin, R., & Vandeweerd, N. (2022). Crowdsourced Adaptive Comparative Judgment: A Community-Based Solution for Proficiency Rating. *Language Learning*, *72*(3), 853–885. https://doi.org/10.1111/lang.12498

Sims, M. E., Cox, T. L., Eckstein, G. T., Hartshorn, K. J., Wilcox, M. P., & Hart, J. M. (2020). Rubric Rating with MFRM versus Randomly Distributed Comparative Judgment: A Comparison of Two Approaches to Second-Language Writing Assessment. *Educational Measurement: Issues and Practice*, *39*(4), 30–40. https://doi.org/10.1111/emip.12329

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *26*(5), 541–562. https://doi.org/10.1080/0969594X.2019.1602027