Manuscript Running Header CROWDSOURCED COMPARATIVE JUDGEMENT

Article Type *Article*

Title Crowdsourced comparative judgement for evaluating learner texts: how reliable are judges recruited from an online crowdsourcing platform?

Author(s) [Peter Thwaites,[a] Nathan Vandeweerd,[b,] Magali Paquot [a, c]]

Author Affiliations [[a] Centre for English Corpus Linguistics, Institut Langage et Communication, UCLouvain, Belgium [b]Radboud University Nijmegen, Netherlands c Fonds de la Recherche Scientifique – FNRS, Belgium]

Author notes / acknowledgements (optional)

Correspondence concerning this article should be addressed to [Magali Paquot. UCLouvain. SSH/ILC. Collège Érasme, Place Cardinal Mercier, 1 - Louvain-la-Neuve 1348 - Belgique. E-mail: magali.paquot@uclouvain.be].

The authors wish to thank Ian Jones, San Verhavert, and Tanguy Dubois for their assistance at various stages of the preparation of this manuscript.

Biodata:

Peter Thwaites is a Postdoctoral Research Fellow at the Centre for English Corpus Linguistics, UCLouvain. His research focuses on second language teaching and assessment, with a particular focus on writing and vocabulary development. His most recent publications include 'Exploring the Impact of Lexical Context on Word Association Responses', published in the *International Journal of Corpus Linguistics* (2022) and 'Word Association Research and the L2 Lexicon' (with Tess Fitzpatrick, published in Language Teaching (2020).

Magali Paquot is Research Associate at the Centre for English Corpus Linguistics, UCLouvain. She specializes in using learner corpora to study key topics in SLA (complexity, phraseology, crosslinguistic influence, proficiency). She is co-editor in chief of the *International Journal of Learner Corpus Research* and a founding member of the Learner Corpus Research Association. Her research is published in journals including *Corpus Linguistics and Linguistic Theory, Language Learning, Language Testing* and *Studies in Second Language Acquisition*.

Nathan Vandeweerd is an Assistant professor in the Department of Language and Communication at Radboud University, Nijmegen. His research interests include second language acquisition, corpus linguistics and the effect of register on L2 production. His recent publications, on topics including L2 French lexical complexity and L2 writing

assessment, can be found in journals such as the *International Journal of Learner Corpus Research*, *Language Learning*, and *Studies in Second Language Acquisition.*

Crowdsourced comparative judgement for evaluating learner texts: how reliable are judges recruited from an online crowdsourcing platform?

Abstract

Recent studies of proficiency measurement and reporting practices in applied linguists have revealed widespread use of unsatisfactory practices such as the use of proxy measures of proficiency in place of explicit tests. Learner corpus research is one specific area affected by this problem: few learner corpora contain reliable, valid evaluations of text proficiency. This has led to calls for the development of new L2 writing proficiency measures for use in research contexts. Answering this call, a recent study by Paquot et al. (2022) generated assessments of learner corpus texts using a community-driven approach in which judges, recruited from the linguistic community, conducted assessments using comparative judgement. Although the approach generated reliable assessments, its practical use is limited because linguists are not always available to contribute to data collections. This paper therefore explores an alternative approach, in which judges are recruited through a crowdsourcing platform. We find that assessments generated in this way can reach near identical levels of reliability and concurrent validity to those produced by members of the linguistic community.

Keywords

Crowdsourcing, comparative judgement, learner corpora, writing assessment.

Introduction

Considerations of language proficiency are ubiquitous in second language research. Research has focused, for example, on how proficiency should be defined (Hulstijn 2011, 2015; Leclercq et al. 2014; Norris & Ortega 2003), what sub-components make up the construct (Jeon & In'nami 2022), and how proficiency can best be assessed (e.g. Bachman & Palmer 2010; Green 2020). Even in studies where proficiency is not an explicit research focus, it is frequently included as a grouping or control variable.

Several systematic reviews of proficiency measurement and reporting practices (e.g. Thomas 1994, 2006; Tremblay 2011) have argued that adequate proficiency measurement is essential to effective research design, since in the absence of such measures it is "unclear to what extent proficiency is intervening in the accurate interpretation of research findings" (Norris 2018 p. 9). A recent systematic review of this area (Park et al. 2022) painted a relatively pessimistic view of current proficiency measurement and reporting practices. The authors found that only 42% of studies in five key journals used an explicit test to measure the proficiency of their participants, and reported that institutional status – an indirect and often unreliable estimate of participant proficiency – was "by far the most frequently used means of assessment" in the studies they reviewed (p. 213). Park et al. conclude that "there has been relatively little aggregate change in conventions for assessing L2 proficiency" (p. 213) since the publication of earlier reviews, and suggest that this lack of progress is at odds with recent calls for greater methodological rigour in applied linguistic studies (e.g. Gass et al. 2021; Norris et al. 2015).

One area of applied linguistics with a particularly clear need for improvements in standards of proficiency measurement and reporting is learner corpus research. This is because in many learner corpus studies, proficiency is either the specific subject of investigation (e.g. Ionin & Díez-Bedmar 2021), or is a potential confound for the effect of other variables under investigation. For example,

learner corpora are sometimes used to explore the effects of L1 background on L2 productions (e.g. Werner et al. 2021); but doing so without controlling for the effects of proficiency can result in problems of interpretation. However, despite this, scholars have noted that few learner corpora contain adequate proficiency measures (Carlsen 2012; Paquot et al. 2022). This situation is well-exemplified by Park et al.'s (2022 p. 226) decision to exclude learner corpus research from their proficiency measurement review paper on the grounds that "information about L2 learners represented in corpora (e.g., information about their L2 proficiency) is, in many cases, not available".

According to Carlsen (2012), there are two specific issues. The first is that "most [learner corpus] constructors to date have […] used external criteria [such as] year of instruction or institutional status" as measures of text proficiency (p. 166). These are the same measures whose overuse in wider linguistic research is criticised by Park et al. (2022), along with numerous other scholars, on the grounds that they are unlikely to offer satisfactory reliability and validity.

Secondly, Carlsen (2012) argues that learner corpora are over-reliant on learner-centred proficiency measures, which seek to ascertain the proficiency of the learner who created each text rather than the proficiency demonstrated in the text itself. Learner-centred methods include direct proficiency evaluations such as placement tests in addition to indirect methods like institutional status. Carlsen argues that these approaches are problematic because they do not directly evaluate each text. A more satisfactory approach is to use text-centred methods, such as rubric-based evaluations of each text. Unlike learner-centred approaches, these methods are not vulnerable to fluctuations in learner proficiency "from one day to the next or from one test to another" (Carlsen 2012 p. 168), and are also uniquely focused on the target competence (e.g. writing). In contrast, learner-centred grades are often based on tests of general linguistic competence, or on indirectly related areas such as grammatical knowledge or receptive skills (Park et al. 2022).

The most well-established approach to text-centred L2 writing assessment, whether in learner corpus research or beyond, is rubric-based assessment by expert graders (Weigle 2002). However, while this approach is widely used in high stakes assessment contexts, it tends to be considered too expensive and time-consuming for evaluating proficiency in research contexts. This is evident from Carlsen's (2012) attempt to generate text-centred grades for the ASK corpus using a rubric-based approach. Carlsen had each text in the corpus graded by five expert graders, resulting in highly reliable but prohibitively time-consuming and costly evaluations. Therefore, while there are signs that proficiency measurement and reporting practices in learner corpora are improving (e.g. Boyd et al. 2014; Crossley et al. forthcoming; Gablasova et al. 2019), there remains a need for the development of new, text-centred approaches to the evaluation of L2 writing, which offer improvements in terms of affordability and time-efficiency without compromising on reliability and validity.

One recent attempt at developing such an approach was reported in a study by Paquot, Rubin, and Vandeweerd (2022). The authors used crowdsourcing to recruit members of the linguistic community, who provided evaluations of texts from the ETS corpus (Blanchard et al. 2014) using the method of comparative judgement (CJ). Judges were shown two learner texts side-by-side and asked to choose which of them "was written by a more advanced speaker". An adaptive algorithm was used to manage the pairing of texts, such that as the algorithm begins to "learn" the quality of each text, it pairs items of similar quality in order to provide the maximum amount of information to the emerging model. A total of 318 comparisons, submitted by 43 judges, were collected in this manner. The scale separation reliability (a measurement of the stability of the rank assigned to each text, comparable to Cronbach's alpha; Verhavert et al. 2018) in the study reached .95, suggesting that this method was highly reliable. In addition, a linear model which tested the fit between the CJ-derived

rank order of texts and rubric-based grades for each text explained a moderate amount of variance (adjusted R^2 = .50, p < .001), and showed that the median crowdsourced ranking of texts rated as 'low' was significantly lower than those rated as 'medium', while the median crowdsourced ranking of texts rated as 'medium' was significantly lower than those rated as 'high'. In other words, the CJ-based assessment produced similar results to traditional rubric-based assessment, lending support to the use of CJ for assessing L2 writing

A significant feature of Paquot et al.'s (2022) study was its use of crowdsourcing to recruit judges. Typically, crowdsourcing studies in applied linguistics make use of platforms such as Amazon Mechanical Turk or Prolific to recruit judges with a wide range of backgrounds (Keuleers & Balota 2015). Paquot et al. (2022), however, took a more targeted approach, recruiting participants from the Applied Linguistics community using mailing lists and personal contacts. As Marsden and Morgan-Short (2023 p. 27) have noted, the study therefore represents a "community-driven" approach to learner corpus enrichment, in which the linguistic knowledge and experience of the "language aficionados" (Paquot et al. 2022 p. 10) who served as judges contributed to the study's validity.

One practical limitation of this approach is the availability of expert linguists to provide comparisons. Paquot et al.'s study required a period of around three months to generate grades for 50 texts. To contextualise this, the International Corpus of Learner English (ICLE; Granger et al. 2020), which lacks reliable proficiency information, contains more than 9000 texts. The time required to generate grades for such a large corpus (not to mention the demands placed on the applied linguistic community) would be substantial. An important next step is therefore to ask whether a more traditional crowdsourcing approach, in which judges are recruited using an online platform and may or may not possess relevant linguistic experience or expertise, could provide grades of similar reliability and validity.

Judge expertise is an ongoing area of investigation in CJ research. Judges tend to be categorised in one of three ways. Firstly, *expert* judges are defined as those with substantial knowledge, experience, and/or training in the evaluation of the target competence. Secondly, *peers* are those who "have performed the task under assessment or had experience with the task" (Paquot et al., 2022). Lastly, *novices* are those with no specific training in or experience with the target task or competence. To date, only one L2 study (Sims et al. 2020) has explored the influence of judge expertise on CJ for L2 writing assessment. The study compared the reliability and concurrent validity of CJ and rubric-based assessment approaches. The judges/raters were two groups of eight participants, each containing four novices (undergraduate students minoring in TESOL) and four experts (trained raters with 2-7 years of experience evaluating learner texts). Each group was asked to provide CJ- and rubric-based essay assessments of a set of 60 essays. Prior to the experiment, each text was also judged by a different set of expert judges using rubric-based methods. These prior grades were used to test the concurrent validity of the newly created CJ and rubric-based scores. The authors reported a high level of reliability for the CJ judgements overall. The expert raters were slightly more reliable than the novices (novices SSR = .89-.91, experts .92-.94), and showed slightly stronger concurrent validity (novices r = .90-.92, experts r = .94). These findings are supported by those of Verhavert et al. (2019), who conducted a meta-analysis of 49 CJ studies from a wide range of academic fields and found that novice judges were no less likely than peers or experts to generate reliable rating scales, but required a larger number of comparisons per text to reach the same reliability level as peer or expert judges.

While these studies provide a general background to what might be expected of judges with different backgrounds, it is important to note that the judge groups used in Paquot et al.'s original

(2022) study do not neatly fit into predefined categories. That study's linguists were a mix of experts and peers, since some reported significant experience of grading learner texts while others were students in English-language academic programs. In the replication of Paquot et al. which is presented in this paper, most judges are novices in the sense that they possessed neither experience nor training in (L2) writing assessment; but a small number of participants *did* report having these skills (see *Method*, below). As such, this study should be seen as contrasting two different crowdsourced populations, rather than peer and expert judges.

One issue not addressed above is the construct validity of CJ for writing assessment. A typical approach in CJ is to provide judges with only a minimal task description to guide their judgements (e.g. Wheadon et al. 2020, simply asked judges to choose 'the better text?'). This approach is based on the assumption that each judge brings slightly different perspectives to the task. The sum of these perspectives is a broad-based, emergent, collective understanding of the target construct, integrating the diverse perspectives of the judging team. The approach is not uncontentious: Bisson (2016 p. 143) has noted that to those used to rubric-based approaches, which specify exactly what judges should pay attention to, CJ can appear "opaque and under-defined". However, she goes on to argue that the relative lack of guidance given to judges in CJ is "a key strength", since it avoids the "narrow and rigid definitions" of target constructs in rubrics. A similar point is made by Pinot de Moira et al. (2022 p. 27), who argue that in seeking to maximise consistency, writing rubrics can become "hyper-specific", thereby potentially missing important aspects of target constructs and depriving judges of the ability to make full use of their expertise.

Nevertheless, it is important to avoid any assumption that CJ judges necessarily consider an acceptably broad range of construct-relevant textual features when making decisions. To date, no L2 studies have investigated what CJ judges pay attention to when comparing texts; this is clearly an area in need of dedicated research. However, a small number of studies have examined how judges make decisions when judging L1 writing. Four main findings emerge from these studies. Firstly, judges only rarely consider construct-irrelevant textual features such as the presence of crossings out, handwriting or font clarity, and layout (Chambers & Cunningham 2022; van Daal et al. 2019a; Lesterhuis et al. 2018). For example, Lesterhuis et al. (2018) reported that 93.5% of comments made by judges pertained to construct-relevant text features (e.g. strength of argumentation, organisation, and linguistic conventions such as spelling and grammar). Secondly, judges appear to pay principal attention to discourse-level features such as organisation and argumentation (Lesterhuis et al. 2018, 2022). Third, judges differ in their secondary focus; some pay more attention to linguistic features than to sources and referencing, for example (Lesterhuis et al. 2022). Lastly, optimal CJ assessments are achieved when judges are diverse in terms of focus (Lesterhuis et al. 2022). Though these findings cannot be assumed to automatically apply to CJ for L2 writing assessment, they do provide some support for the assumptions underlying CJ's construct validity.

The current study

In the present study we explore the basic reliability and concurrent validity of CJ assessment using traditionally crowdsourced judges. The study forms part of a wider project aiming to develop and report on best practices for using CJ for learner corpus enrichment. To this end, other studies within the project look at further aspects of CJ's suitability to L2 writing assessment, including its robustness to variance in the characteristics of the texts being assessed (e.g. text length, proficiency range, and topical variety; Thwaites et al. 2024) and the justifications that judges give for their decisions (Thwaites et al, in preparation).

Following the nomenclature suggested by Marsden et al. (2018), the study is a partial replication of Paquot et al. (2022), in that it makes only one significant change to the methods used in the original.

That change is the replacement of Paquot et al.'s crowdsourcing approach, in which members of the applied linguistic community were targeted, with a more traditional crowdsourcing approach using an online crowdsourcing platform.

The study addresses the following research questions:

1. Can a group of judges recruited using an online platform produce reliable evaluations of texts using comparative judgement?
2. What is the concurrent validity of these evaluations, measured against pre-existing rubric-based grades?
3. To what extent does the ranking scale produced by these judges agree with the scale produced by the linguists used in the original study?

Methods

Judge selection

To select judges, we first created a profile of the desired participants. Our priority was to recruit a set of judges reflective of what practitioners or researchers would be likely to find if they adopted a traditionally crowdsourced approach. As such, the only demographic decision we made was to specify that participants' L1 should be English (overwhelmingly the most common L1 on the platform), since doing so avoided adding a layer of linguistic variability to the study design. As Table 1 shows, this resulted in the admission of a minority of participants who possessed experience and/or training in grading L2 writing. We did not filter out these participants, since they are a natural part of the Prolific crowd, and because we assume that practitioners aiming to use crowdsourced CJ as a real-world assessment tool would not wish to exclude them.

The next step was the selection of a crowdsourcing platform. Two recent studies (Douglas et al. 2023; Peer et al. 2022) have compared the quality of data produced on a range of current platforms. Data quality, in these studies, is evaluated using tests of the proportion of participants who pass comprehension, attention, and honesty checks, as well as by measuring the internal consistency of their responses. Both studies suggest that two platforms, Prolific and CloudResearch, provide higher quality data than alternative platforms such as Amazon Mechanical Turk. From these two, Prolific was chosen on the grounds that (a) Peer et al. (2022) reported that its participants scored slightly higher for honesty; (b) Douglas et al. (2023) reported a slightly lower cost per high quality participant. We also found Prolific's data security policies to be compatible with local regulations.

Next, we recruited 400 participants to take a survey on their experience and attitudes towards various aspects of language. The survey was developed with the intention of collecting data for an exploratory analysis of the relationship between participants' declared attitudes towards language (such as their tolerance for spelling or grammar errors) and the decisions they make in CJ tasks. A secondary purpose of the survey was to conduct data quality checks. Following Oppenheimer et al. (2009), we used instructional manipulation checks to check whether participants were paying due attention to questions. These included both "overt" (e.g. 'This item is a check of your attention and should be answered with "I don't understand"') and the "covert" ('Well-written texts contain no words') questions. Participants who failed to provide the expected response to any of these questions were not invited to participate in the CJ section of the study. As another check on data quality, participants were removed if they answered "strongly agree" to more than 80% of the survey's 42 Likert scale items. 10 participants were removed in total. The survey can be found in the supplementary materials.

The remaining 390 participants formed a potential judge pool for the CJ stage of the experiment. 80 participants were randomly sampled and offered the opportunity to take part in the CJ stage; the first 40 participants to take a slot became the final judge set. This number was chosen because it closely accorded to the number of participants used in Paquot et al.'s (2022) original study, and because that study suggested that this number of participants would be sufficient to reach an acceptable level of reliability. Participants first completed a Google form which was used (a) as an additional data quality screener, in which we cross-checked participants' L1 (three participants who declared their L1 to be a language other than English at this stage were removed from analysis), and (b) to allow us to provide participants with login details for the CJ platform. The Google form can be found in the supplementary materials. Table 1 presents a comparison of the profile of the judges in the original study and in this replication.

[TABLE 1 HERE]

Comparison with Paquot et al., 2022

All other consequential aspects of study design and procedure were the same as in the original study. These aspects included:

- The 50 TOEFL essays under comparison, taken from the ETS corpus (Blanchard et al. 2014);
- Their associated rubric-based prior scores, which graded each text as low, medium, or high proficiency;
- The same adapted version of the ComPAIR CJ platform (Potter et al. 2016), except that minor modifications were made to allow greater flexibility in the setup of new experiments (images of various ComPAIR screens can be found in the supplementary materials);
- The same adaptive text pairing algorithm, which ensures that (a) each essay receives approximately the same number of comparisons, and (b) each essay is compared with texts of a relatively similar quality;
- The mean number of comparisons performed by each participant (n = 10);
- The instructions given to participants – i.e. they were asked to read each pair of essays and decide "Which text was written by the more advanced speaker of English?".

The minimal instructions given in both studies reflects the assumption, discussed above, that CJ should allow judges to draw on their own, holistic understanding of the target construct, and that through the pooling of the collective understandings of the team of judges, a valid construct will emerge.

Lastly, as in the original study, the raw data produced by the ComPAIR platform was processed in R (v4.2.2; R Core Team 2023) using the *sirt* package (v3.12-66; Robitzsch 2022) with additional functions by Jones (2022). The *btm* function of this package compiles the result of each paired comparison according to the Bradley-Terry model (Bradley & Terry 1952). This process assigns a numeric score to each text, which increases when the text "wins" a comparison and decreases when it "loses". The scores have a mean value of 0, a standard deviation of around 2.5 (Bramley 2007), and indicate the "quality" of each item relative to all others. They can be interpreted as indicating the probability that each item will "win" a comparison with any other. The scores form the basis of a rank-ordered scale, from which reliability and fit metrics can be extracted. Reliability is measured using scale separation reliability (SSR). This takes a value between 0 and 1, with higher values indicating greater reliability, and reflects the extent to which the items occupy distinct locations on the scale. Fit statistics are calculated for each text and each judge, and take a value between 0 and infinity, where a score of 1 indicates that a text/judge performed in a manner consistent with that of other judges/texts. Fit scores greater than the mean fit score plus two times

the standard deviation are considered to misfit the overall model. This might indicate a text which has proven difficult for judges to evaluate, or a judge whose decisions differed from the rest of the judging team. Following Paquot et al. (2022) and Jones & Davies (2023), we assume that misfitting texts and judges may contribute useful information to the model and therefore do not remove them; instead, we briefly discuss them below.

Analysis

A total of 370 paired comparisons were collected from the 37 judges (slightly more than the 316 comparisons collected in the original study). From these comparisons, a rating scale was generated using the *btm* model described above. To answer RQ1, we extracted the SSR value (along with judge and text fit scores) from this model. We also ran the *btm* function on successive iterations of the data, beginning at round 1 and adding one further round of comparison in each iteration until all rounds were included. The SSR was recalculated at each round. This allows both visualisation of the evolution of reliability at each successive round, and comparison of the points at which the two scales reached an asymptote, which is the point at which adding further comparisons adds little additional reliability, and is defined as the point at which there was a total change in SSR of less than .01 for three consecutive rounds.

To investigate the extent of agreement of the CJ scale with the prior rubric-based scores for each text (RQ2), we first created a rank-ordering of the items, such that the weakest items had low rank values and the stronger items higher ones. We then fit a linear model with CJ rank order as the dependent variable and the rubric-based score (low, medium, or high proficiency) as the independent variable. An initial test using the *gvlma* package (v1.0.0.3; Pena & Slate 2019) showed that none of the assumptions for a linear model were violated. We then used the *lm* function to build the model, the *MASS* package (v7.3-60; Venables & Ripley 2002) to set contrasts between adjacent ETS grades, the *emmeans* package (v1.8.6; Lenth 2023) to compute the estimated marginal means, and the *effects* package (v4.2-2; Fox & Hong 2009) to visualize the model predictions. Together, these processes allowed us to explore whether the texts with low, medium, and high rubric-based scores had correspondingly low, medium, and high CJ rank. Lastly, to explore the extent of agreement on text quality between the judges in the original study and the replication (RQ3), we ran a Spearman correlation test comparing the rank order of each text across the two studies.

The study was pre-registered using the OSF Registrations service prior to the commencement of data collection. Study materials, R scripts, and data are available at https://osf.io/9y84n/.

Results

RQ1: Can a group of crowdsourced judges, recruited using an online platform, produce reliable evaluations of texts using comparative judgement?

At the end of the final round of comparisons, the reliability level for the Prolific participants' CJ-derived scale stood at SSR = .944. By comparison, the SSR level reported in the original study was .937 after the same number of comparisons. SSR values are comparable to Cronbach's alpha; in both experiments the SSR values can be interpreted as indicating very high reliability. Figure 1 shows the evolution of SSR values for the two studies on a round-by-round basis. Paquot et al. (2022) reported that in the original study, an asymptote was reached at round 14, since the total SSR increase from rounds 14-16 was lower than .01. In the present study, asymptote was also reached at round 14.

The generation of SSR values also permits investigation of judge fit data. From this information, we found that two judges misfit the overall model. Survey data collected prior to the CJ

part of the study suggested that these two judges were relatively typical in terms of their experience of teaching and assessing writing. However, data on their attitudes towards different aspects of writing showed that both appeared to consider discourse-level features to be slightly more important, and grammatical accuracy slightly less important, than did other participants.

[FIGURE 1 HERE]

RQ2: What is the concurrent validity of these evaluations, measured against rubric-based grades?

[TABLE 2 HERE]

A linear model predicting CJ rank order on the basis of the ETS rubric-based grades was statistically better than a baseline model without the rubric-based grades ($F(2, 47) = 28.58$, $p < .001$), indicating that adding the rubric-based assessment to the model significantly contributed to predicting the rank-score of a text. Specifically, the ETS scores explained a moderate amount of variance in the CJ-based rank order (adj. $R^2 = 0.53$). This figure is very similar to the variance explained by the corresponding model in the original study (adj. $R^2 = 0.50$). As shown in Table 2 and also in line with the results of Paquot et al., the rank for texts rated as "low" proficiency was significantly lower (*EMM* = 11.33, *95%CI* = 5.53-17.14) than that of texts rated as "medium" (*EMM* = 21.79, *95%CI* = 17.18-26.40), and the median rank for those rated "medium" was significantly lower than those rated as "high" proficiency (*EMM* = 38.16, *95%CI* = 33.53-42.77). This is also illustrated in Figure 2, and demonstrates the comparability of the two sets of assessments.

[FIGURE 2 HERE]

To investigate the rank order distribution of the texts rated at each proficiency level in more detail, Figure 3 shows the rank order of all texts as a function of their ETS proficiency grade. The horizontal lines correspond to the number of texts in the dataset which were assessed as "low" (n = 12) and "medium" (n = 31) proficiency. If the CJ rank order perfectly corresponded to the ETS grades, all low-proficiency texts would appear on or below the first line, all medium proficiency texts between the two lines, and all high proficiency texts above the second line. In reality, several texts in each ETS proficiency group were ranked one level higher than expected, while one text (424) was assessed as being low proficiency by the ETS rubric-based graders, but had a CJ rank corresponding to texts rated as high proficiency.

In the original study, texts were defined as outliers if their rank was "greater than the median plus 1.5 times the interquartile range (IQR) or […] less than the median minus 1.5 times the IQR" (Paquot et al., 2022, p. 17). Two texts in the original study (texts 411 and 424) were labelled as outliers according to this definition. Interestingly, both texts were also judged to be outliers in this replication. That is to say, the participants of both the original study and the replication judged these texts to be of different quality than was expected from their ETS proficiency grades. In addition, two further texts (400 & 402) were identified as outliers in this study but not in the original.

[FIGURE 3 HERE]

[FIGURE 4 HERE]

RQ3: To what extent does the ranking scale produced by Prolific judges agree with the scale produced by the linguists in the original study?

The third research question was tested using a Spearman correlation, which revealed that the rank orders from the two studies were strongly correlated ($r = .90$, $p < .001$; see Figure 4). This indicates a high level of agreement between the two sets of judges. Nevertheless, the rank-orders were not

perfectly aligned. To explore this, the absolute difference in each text's rank order between the two studies was calculated by subtracting the rank in the replication from the rank in the original study, then converting negative values to positive. On average, the difference in rank was five places, with a minimum of zero and a maximum of 15. Eight texts showed rank differences two or more standard deviations (i.e. ten or more places) from the mean. Overall, these findings suggest that while there was a high level of agreement between the two sets of judges in general, it was not unusual for some texts to be ranked quite differently.

Discussion

The foregoing analyses suggest that the set of judges recruited via Prolific crowdsourcing was able to generate a highly reliable rating scale, with an SSR level near-identical to that of the members of the Applied Linguistic community used in the original study (RQ1). In addition, these grades were significantly associated with the rubric-based assessments of the texts provided in the ETS corpus (RQ2), and correlated strongly with the scale produced by the more experienced set of judges in the original study (RQ3), demonstrating their concurrent validity. Together, these findings suggest that the Prolific crowd were no less able to distinguish stronger from weaker texts than the judges in the original study. This is remarkable given the differences in experience and training between the two groups: in the original study, 70% of judges had some experience of grading learner writing and more than 50% had received training as writing judges; in the replication, the corresponding figures were 27.5% and 7.5% (see Table 1).

Importantly, the process of generating this data was efficient in terms of time and cost, compared with other attempts to generate grades for learner corpus texts. The 370 comparisons took around 12 hours to collect, at a total cost of £336. By comparison, Paquot et al.'s (2022) original, community-based CJ approach allowed data to be collected at zero cost, but required several months for the submission of sufficient judgements. In addition, the study's reliance on the voluntary participation of members of the linguistic community is unlikely to be sustainable as a long-term data collection method. A second point of comparison is a recent project (Kanistra & Kollias, in preparation) to gather triple-graded, rubric-based evaluations of texts from the International Corpus of Learner English. The cost of triple-grading 50 essays in this manner was €450, which is only slightly higher than the costs of crowdsourced CJ. However, the rubric-rating process was much slower – the entire process of recruiting graders, training them to use CEFR writing rubrics, and allowing them time to complete their grades required around 3 months. The crowdsourced CJ approach therefore appears to represent the most favourable option in terms of efficiency.

One question arising from the above results concerns how a group of largely untrained and inexperienced judges could have generated evaluations of such apparent reliability and accuracy. After all, in their review of reliability scores in CJ studies, Verhavert et al. (2019) observed that novice judges tend to require a larger number of comparisons to reach a high reliability level than do more experienced judges. Why then, in the present study, were Prolific judges able to produce assessments of apparently high quality no less quickly than the linguists in Paquot et al. (2022)?

One potentially influential factor is the relative ease of the CJ task. The texts used in the study covered almost the entire L2 proficiency spectrum "from beginner to advanced" (Paquot et al. 2022 p. 19). This means that some decisions would have required judges only to distinguish, for example, texts at approximately A2 level from those at C1. This is likely to have contributed to the high levels of reliability demonstrated in this study, since a broad proficiency range is likely to lead to much less disagreement between judges, and thus more consistent data, than would be the case in a study containing texts of more homogeneous proficiency. It is noteworthy that one other study

reporting very high reliability for CJ assessments of L2 writing (Sims et al. 2020) also featured texts with a very broad proficiency range.

Another possible influence on the high reliability achieved in the present study is the use of an adaptive algorithm for text pairing, which works by pairing texts of a pre-determined level of similarity in quality (i.e. similar, but not too similar). On the surface, it is logical to use this method since it ensures that each comparison provides useful information to the emerging model. However, several studies have shown that adaptive algorithms can result in inflation of SSR values (Bramley & Vitello 2019; Rangel-Smith & Lynch 2018), and this has led to some scholars advocating a move away from adaptive tools (Jones & Davies 2023). Given this issue, it is likely that the SSR values reported above (and those in Paquot et al., 2022) are somewhat inflated, relative to studies using random text pairing.

These points suggest that further research, preferably using non-adaptive text pairing methods, will be needed to test whether judges recruited through a crowdsourcing platform can produce acceptable assessments of texts covering narrower proficiency spans. Two further studies within the same wider project as the present paper have explored this, with results suggesting that CJ is significantly more difficult when texts are more similar in proficiency (Thwaites et al. 2024), and that while crowdsourced judges are able to generate assessments of similar concurrent validity to linguists, a larger number of comparisons must be conducted to allow them to reach satisfactory validity (Thwaites et al., Submitted). Together with the findings of the present study, these results suggest that judges largely lacking specialist training and linguistic expertise can produce reliable assessments of texts, although their reliability and efficiency begins to decline as the proficiency span represented in the texts narrows.

Although the above reveals some limits to Prolific users' ability to reliably evaluate L2 writing, crowdsourced CJ appears overall to be a surprisingly effective assessment tool. One way of making sense of this is that judges whose L1 is English already possess enough knowledge of what constitutes a successful piece of writing to allow them to make comparative judgements of reasonable accuracy when texts are sufficiently distinct in quality. It may be that these participants had sufficient exposure to relevant texts to allow them to judge what constitutes a successful piece of writing, and to therefore be able to make comparative judgements with reasonable accuracy. Usage-based theories of language development might support this account, given that they would argue that exposure to language forms the basis of linguistic competence (e.g. Ellis 2009; Tomasello 2003).

Given the above, it might be tempting to argue that the Prolific judges' L1 English proficiency allowed them to compensate for their comparative lack of linguistic knowledge and experience, next to Paquot's et al.'s linguist judges, some of whom were L2 users of English. However, we would caution against this explanation. As Hulstijn (2011, 2015) has argued, L1 status does not necessarily provide the "higher language cognition" – i.e. the ability to use and understand "low frequency lexical items or uncommon morphosyntactic structures [pertaining to] topics addressed in schools and colleges" (Hulstijn 2011 p. 231) - likely to be necessary for effective assessment of argumentative essays. This type of linguistic competence is likely to develop, in both L1 and L2 users of a given language, through education. As such, judges' level of education is likely to be as important an influence on CJ performance as L1 status. Unfortunately, in the present study we did not collect data on judges' level of education. Analysis of the subsection of all Prolific users whose L1 is English reveals that around 60% have at least a bachelor's degree, suggesting that the Prolific crowd are in general relatively well-educated. However, given that we cannot assess the influence of judges' level of education using the present data, we suggest that future research investigate the influence of

judges' expertise, L1 background/target language proficiency, and level of education on CJ performance.

One question that was briefly addressed by the authors of the original study concerned what differentiates CJ from rubric-based assessment. This was explored by analysing two texts whose CJ ranks were higher than would be expected from their rubric-based rating of "low", therefore potentially indicating differences between CJ and rubric-based assessment approaches. Paquot et al. (2022) noted that while the one of these texts (Text 411; original study rank = 17th; full text provided in the supplementary materials) was coherent, cohesive, and exemplified sophisticated lexical usage, there were still numerous spelling and grammatical errors. Text 411 also received a relatively high rank position in the current replication (27th), perhaps suggesting that the two sets of judges - the applied linguists in Paquot et al. (2022) and the Prolific crowd in this study – may have been sensitive to the same type of textual features. A similar pattern can be observed in Text 400 (see Example 1), which the judges in both studies ranked higher than would be expected from its "medium"-grade rubric-based assessment. Like Text 411, Text 400 made use of cohesive devices and included some sophisticated lexical units (e.g. *sustainable, acknowledge, tenacity; push + limits, take + risks, blind + risks, broad[en] + scope*), but also contained relatively frequent grammatical errors or unconventional phrasings (*The two concepts of success and risk are they related?*; *You appreciate more something when…*) and occasional spelling errors (*ennemy*, *suceed*), which perhaps contributed to the difficulty in ranking this text consistently. It is noteworthy that both sets of CJ judges regarded each of these texts as stronger than their rubric-based assessments would have suggested. However, these two texts provide only the most limited basis for understanding differences between CJ and rubric-based assessment: further research will be necessary in order to understand this area more clearly.

Example 1: Text 400
ETS rating = medium; replication rank = 8; median medium level rank = 21.8; (Blanchard et al., 2014)

The two concepts of success and risk are they related? Are we sure to achieve our goal when we're willing to take some risks? A statement says that 'Successful people try new things and take risks rather than only doing what they already know how to do well'. We could reply to this statement by another saying 'The best is the ennemy of the good'. It means that success is always appealing, but it over all needs to be sustainable. However, I think the fact that we take a risk to try to achieve our aim is exciting. On one hand if we managed, if we indeed suceed, the achievement feels even better. The dimension of risk makes it more unique. You appreciate more something when you have struggled for it. It also makes you move forward, push your limits backwards. On the other hand, if we fail to suceed because we have dared commit in the project as the risks were present, then it makes you revise your vision of the risk. Are the risks taken worth the final goal? To my opinion, I'm willing to take risks when I have a minimum of visibility. I mean, I do not take blind risks and I do not count only on my luck. Risks must be assessed. And it is important enough to be underlined, when we are taking risks, we shall not commit anyone else's responsability without he or she acknowledges and agrees with it. So, I do agree with the initial statement. Indeed, I think successful people are people willing to move forward, who makes prove of tenacity to succeed. If you never try anything new, you don't broad your scope.

Whatever the difference between CJ and rubric-based assessments, the comparison between the two approaches highlights how little is known about what judges pay attention to when conducting judgements on L2 writing. This information is crucial to establishing argument for CJ's

construct validity, which hinges on evidence of acceptable construct relevance (i.e. that judges have made decisions on the basis of relevant features) and construct representation (i.e. that judges have taken into account a sufficiently broad range of features, such as argumentation, organisation, and lexico-grammatical range and accuracy; Green 2020; Messick 1989). As discussed in the introduction, a limited number of L1 writing studies (Chambers & Cunningham 2022; van Daal et al. 2019b; Lesterhuis et al. 2018, 2022) have researched what CJ judges pay attention to as they make their decisions, with the results suggesting that judges attend to construct-relevant features and focus on a range of aspects of text quality, including organisation, argumentation, and mechanical features such as grammatical accuracy, spelling, and punctuation. However, these findings cannot be directly applied to L2 contexts because of differences in how L1 and L2 writing is assessed (Weigle 2002).

The challenge of establishing CJ's construct validity is, as Kelly et al. (2022) have argued, made more complex by the fact that CJ studies vary significantly in the type of judges used to make comparisons. Purely within the context of this replication and its original study, this variance includes not only the use of community-driven and crowdsourcing judge recruitment methods, but also the limited English proficiency of some judges (in the original study), the lack of experience or training in linguistics or writing assessment (in this replication), and judges' unknown level of education (this replication). Existing evidence from beyond the context of CJ suggests that differences of focus between such judges are likely. For example, Schoonen et al. (1997) found evidence that while lay judges appear to be as reliable as experts in the assessment of content-related aspects of L2 writing, they may be less reliable in assessing language usage. There is also evidence of training-related variance in CJ: Paquot et al. (2022) reported that judges with training in language assessment had significantly different fit scores to those without training. However, Paquot et al.'s study was not able to demonstrate specifically what differences in focus accounted for this finding, and there is besides very little research on how judge characteristics affect the features taken into account during CJ.

One immediate implication of this variance is that researchers must refrain from making broad claims about the validity of CJ in general, or even CJ for specific purposes such as L2 writing assessment. Instead, they should focus on identifying specifically with what type of judges, and in specifically what contexts, CJ is claimed to be a valid form of assessment.

A longer-term goal for CJ research will be for dedicated studies to compare the validity of CJ as conducted by different types of judges. This is the approach taken in an ongoing study by Thwaites et al. (in preparation), which explores comments made by three sets of judges – a crowdsourced group similar to that used in the present study, a community-driven group similar to that used in Paquot et al. (2022), and a group of trained L2 writing assessors – while conducting CJ. By analysing these comments in terms of their relevance, specificity, and breadth of construct representation, the study aims to offer a more detailed picture of CJ's validity for assessing L2 argumentative essays in general, and of which types of judge provide the most valid assessments more specifically. Initial results suggest that comments provided by trained assessors are more specific and cover a broader range of construct-relevant textual features than those of other groups, but that both community-driven and crowdsourced judges also demonstrate construct relevance and an acceptable level of construct representation. A second study exploring correspondences between what crowdsourced judges claim to value in L2 writing and what textual features actually account for variance in their CJ judgements (Vandeweerd et al., 2024) provides further evidence that these judges make decisions based on relevant criteria. Studies such as these will be essential to the task of developing a strong validity argument for CJ's use in L2 writing assessment.

Conclusion

There is a need for the development of reliable, efficient new approaches to L2 writing assessment to facilitate proficiency measurement in applied linguistic research in general, and in learner corpus research in particular. Following on from an initial study by Paquot et al. (2022), the present study suggests that a crowdsourced form of comparative judgement, in which judges are recruited through an online platform, offers a reliable approach to L2 writing assessment, with acceptable levels of concurrent validity, and might therefore be a suitable method for assessing L2 essays covering a broad proficiency range. In addition to being a direct, writing-focused, text-centred approach to evaluating learner texts, the method can produce results much more quickly than can rubric-based grading or community-driven CJ, and at slightly lower cost. This makes the method an attractive option for researchers who need to rapidly assess the proficiency of relatively simple L2 writing samples. However, further research will be need to focus on exploring the method's reliability for assessing more complex sets of texts, as well as its construct validity, before strong claims can be made regarding its general efficacy.

Supplementary Materials

Supplementary materials for this paper can be found at the project's Open Science Foundation page: https://osf.io/9y84n/.

Captions

Figure 1: Comparison of the evolution of SSR levels in the original study and this replication.

Figure 2:Median CJ rank of texts graded as low, medium, and high proficiency

Figure 3: Rank order distributions of the texts in each ETS proficiency band. Outliers are labelled with their text ID.

Figure 4: Correlation of rank orders in the original study and this replication

Bibliography

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Bisson, M.-J., Gilmore, C., Inglis, M., & Jones, I. (2016). 'Measuring Conceptual Understanding Using Comparative Judgement', *International Journal of Research in Undergraduate Mathematics Education*, 2/2: 141–64. DOI: 10.1007/s40753-016-0024-3

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2014). 'ETS Corpus of Non-Native Written English'. Linguistic Data Consortium. Retrieved from https://doi.org/10.35111/7ez0-x912.

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., & Schone, K. (2014). 'The MERLIN corpus: Learner language and the CEFR'. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik: European Language Resources Association (ELRA).

Bradley, R. A., & Terry, M. E. (1952). 'Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons', *Biometrika*, 39/3/4: 324–45. [Oxford University Press, Biometrika Trust]. DOI: 10.2307/2334029

Bramley, T. (2007). 'Paired Comparison Methods'. Newton P., Baird J.-A., Goldstein H., Patrick H., & Tymms P. (eds) *Techniques for monitoring the comparability of examination standards*, pp. 246–300. Qualifications and Curriculum Authority: London.

Bramley, T., & Vitello, S. (2019). 'The effect of adaptivity on the reliability coefficient in adaptive comparative judgement', *Assessment in Education: Principles, Policy & Practice*, 26/1: 43–58. Taylor & Francis.

Carlsen, C. (2012). 'Proficiency Level—a Fuzzy Variable in Computer Learner Corpora', *Applied Linguistics*, 33/2: 161–83. DOI: 10.1093/applin/amr047

Chambers, L., & Cunningham, E. (2022). 'Exploring the Validity of Comparative Judgement: Do Judges Attend to Construct-Irrelevant Features?', *Frontiers in Education*, 7.

Crossley, S. A., Tian, Y., Baffour, P., Franklin, A., Kim, Y., Morris, W., Benner, M., et al. (forthcoming). 'The English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus', *International Journal of Learner Corpus Research*.

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2019a). '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*, 26/1: 59–74. Routledge. DOI: 10.1080/0969594X.2016.1253542

——. (2019b). '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*, 26/1: 59–74. Routledge. DOI: 10.1080/0969594X.2016.1253542

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). 'Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA', *PLOS ONE*, 18/3: e0279720. Public Library of Science. DOI: 10.1371/journal.pone.0279720

Ellis, N. C. (2009). 'Optimizing the Input: Frequency and Sampling in Usage-Based and Form-Focused Learning'. *The Handbook of Language Teaching*, pp. 139–58. John Wiley & Sons, Ltd. DOI: 10.1002/9781444315783.ch9

Fox, J., & Hong, J. (2009). 'Effect Displays in R for Multinomial and Proportional-Odds Logit Models: Extensions to the effects Package', *Journal of Statistical Software*, 32/1: 1–24. DOI: 10.18637/jss.v032.i01

Gablasova, D., Brezina, V., & McEnery, T. (2019). 'The Trinity Lancaster Corpus: Development, description and application', *International Journal of Learner Corpus Research*, 5/2: 126–58. John Benjamins. DOI: 10.1075/ijlcr.19001.gab

Gass, S., Loewen, S., & Plonsky, L. (2021). 'Coming of age: the past, present, and future of quantitative SLA research', *Language Teaching*, 54/2: 245–58. Cambridge University Press. DOI: 10.1017/S0261444819000430

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2020). *International corpus of learner English v3*., Vol. 2. Presses universitaires de Louvain Louvain-la-Neuve.

Green, A. (2020). *Exploring language assessment and testing: Language in action*. Routledge.

Hulstijn, J. H. (2011). 'Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment', *Language Assessment Quarterly*, 8/3: 229–49. DOI: 10.1080/15434303.2011.565844

——. (2015). *Language Proficiency in Native and Non-native Speakers: Theory and research*. John Benjamins. DOI: 10.1075/lllt.41

Ionin, T., & Díez-Bedmar, M. B. (2021). 'Article use in Russian and Spanish learner writing at CEFR B1 and B2 Levels: effects of proficiency, native language, and specificity', *Learner corpus research meets second language acquisition*, 10–38. Cambridge University Press.

Jeon, E. H., & In'nami, Y. (Eds). (2022). *Understanding L2 Proficiency: Theoretical and meta-analytic investigations*. Bilingual Processing and Acquisition, Vol. 13. Amsterdam: John Benjamins Publishing Company. DOI: 10.1075/bpa.13

Jones, I. (2022). 'sirt functions'.

Jones, I., & Davies, B. (2023). 'Comparative judgement in education research', *International Journal of Research & Method in Education*. Routledge. DOI: 10.1080/1743727X.2023.2242273

Kelly, K. T., Richardson, M., & Isaacs, T. (2022). 'Critiquing the rationales for using comparative judgement: a call for clarity', *Assessment in Education: Principles, Policy & Practice*, 1–15. Routledge. DOI: 10.1080/0969594X.2022.2147901

Keuleers, E., & Balota, D. A. (2015). 'Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments', *Quarterly Journal of Experimental Psychology*, 68/8: 1457–68. Taylor & Francis. DOI: 10.1080/17470218.2015.1051065

Kanistra, P. & Kollias, C. (2024). 'Aligning the ICLE 500 written scripts to the CEFR: a technical report'. Catholic University of Louvain.

Leclercq, P., Edmonds, A., & Hilton, H. (2014). *Measuring L2 proficiency: Perspectives from SLA*., Vol. 78. Multilingual Matters.

Lenth, R. V. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means*.

Lesterhuis, M., Bouwer, R., van Daal, T., Donche, V., & De Maeyer, S. (2022). 'Validity of Comparative Judgment Scores: How Assessors Evaluate Aspects of Text Quality When Comparing Argumentative Texts', *Frontiers in Education*, 7.

Lesterhuis, M., van Daal, T., Van Gasse, R., Coertjens, L., Donche, V., & De Maeyer, S. (2018). 'When teachers compare argumentative texts: Decisions informed by multiple complex aspects of text quality', *L1-Educational Studies in Language and Literature*, 18: 1–22. DOI: 10.17239/L1ESLL-2018.18.01.02

Marsden, E., & Morgan-Short, K. (2023). '(Why) Are Open Research Practices the Future for the Study of Language Learning?', *Language Learning*. DOI: 10.1111/lang.12568

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). 'Replication in Second Language Research: Narrative and Systematic Reviews and Recommendations for the Field', *Language Learning*, 68/2: 321–91. DOI: 10.1111/lang.12286

Messick, S. (1989). 'Validity'. *Educational measurement, 3rd ed*, The American Council on Education/Macmillan series on higher education, pp. 13–103. American Council on Education.

Norris, J. M. (2018). 'Developing and investigating C-tests in eight languages: Measuring proficiency for research purposes'. *Developing C-tests for estimating proficiency in foreign language research*, pp. 7–33. Peter Lang: Berlin.

Norris, J. M., & Ortega, L. (2003). 'Defining and measuring SLA', *The handbook of second language acquisition*, 716–61. Wiley Online Library.

Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). 'Guidelines for Reporting Quantitative Methods and Results in Primary Research', *Language Learning*, 65/2: 470–6. DOI: 10.1111/lang.12104

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). 'Instructional manipulation checks: Detecting satisficing to increase statistical power', *Journal of Experimental Social Psychology*, 45/4: 867–72. DOI: 10.1016/j.jesp.2009.03.009

Paquot, M., Rubin, R., & Vandeweerd, N. (2022). 'Crowdsourced Adaptive Comparative Judgment: A Community-Based Solution for Proficiency Rating', *Language Learning*, 72/3: 853–85. DOI: 10.1111/lang.12498

Park, H. I., Solon, M., Dehghan-Chaleshtori, M., & Ghanbar, H. (2022). 'Proficiency Reporting Practices in Research on Second Language Acquisition: Have We Made any Progress?', *Language Learning*, 72/1: 198–236. DOI: 10.1111/lang.12475

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). 'Data quality of platforms and panels for online behavioral research', *Behavior Research Methods*, 54/4: 1643–62. DOI: 10.3758/s13428-021-01694-3

Pena, E. A., & Slate, E. H. (2019). *gvlma: Global Validation of Linear Models Assumptions*.

Pinot de Moira, A., Wheadon, C., & Christodoulou, D. (2022). 'The classification accuracy and consistency of comparative judgement of writing compared to rubric-based teacher assessment', *Research in Education*, 113/1: 25–40. SAGE Publications Ltd STM. DOI: 10.1177/00345237221118116

Potter, T., Englund, L., Charbonneau, J., MacLean, M. T., Newell, J., & Roll, I. (2016). 'ComPAIR: A New Online Tool Using Adaptive Comparative Judgement to Support Learning with Peer Feedback', *Teaching & Learning Inquiry*, 5/2. DOI: 10.20343/teachlearninqu.5.2.8

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rangel-Smith, C., & Lynch, D. (2018). 'Addressing the issue of bias in the measurement of reliability in the method of Adaptive Comparative Judgment'. *PATT36 international conference. Research & Practice in technology education: Perspectives on Human Capacity and Development*, pp. 378–88.

Robitzsch, A. (2022). *sirt: Supplementary Item Response Theory Models*.

Schoonen, R., Vergeer, M., & Eiting, M. (1997). 'The assessment of writing ability: expert readers versus lay readers', *Language Testing*, 14/2: 157–84. SAGE Publications Ltd. DOI: 10.1177/026553229701400203

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. DOI: 10.1111/emip.12329

Thomas, M. (1994). 'Assessment of L2 proficiency in second language acquisition research', *Language learning*, 44: 307–307.

——. (2006). 'Research synthesis and historiography', *Synthesizing research on language learning and teaching*, 13: 279. John Benjamins Publishing.

Thwaites, P., Kollias, C., & Paquot, M. (2024). 'Is CJ a valid, reliable form of L2 writing assessment when texts are long, homogeneous in proficiency, and feature heterogeneous prompts?', *Assessing Writing*, 60: 100843. DOI: 10.1016/j.asw.2024.100843

——. (Submitted) 'Testing crowdsourcing as a means of recruitment for the comparative judgement of L2 argumentative essays.'

Thwaites, P. & Paquot, M. (in preparation). 'Concurrent validity of crowdsourced, community-driven, and trained-rater approaches to using comparative judgement for L2 writing assessment.' (working title).

Tomasello, M. (2003). *Constructing a language: A usage-based approach to child language*. Cambridge, M.A.: Harvard University Press.

Tremblay, A. (2011). 'Proficiency assessment standards in second language acquisition research: "Clozing" the Gap', *Studies in Second Language Acquisition*, 33/3: 339–72. Cambridge University Press. DOI: 10.1017/S0272263111000015

Vandeweerd, N., Thwaites, P., & Paquot, M. (2024). 'Crowdsourcing L2 proficiency assessment using comparative judgement: What is really important to novice judges?'. Forthcoming presentation at the EuroSLA33 conference, Montpellier, France.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*., Fourth. New York: Springer.

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: 541–62. Routledge. DOI: 10.1080/0969594X.2019.1602027

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2018). 'Scale Separation Reliability: What Does It Mean in the Context of Comparative Judgment?', *Applied Psychological Measurement*, 42/6: 428–45. SAGE Publications Inc. DOI: 10.1177/0146621617748321

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Werner, V., Fuchs, R., & Götz, S. (2021). 'L1 Influence vs. Universal Mechanisms: An SLA-Driven Corpus Study on Temporal Expression'. Le Bruyn B. & Paquot M. (eds) *Learner Corpus Research Meets Second Language Acquisition*, Cambridge Applied Linguistics, pp. 39–66. Cambridge University Press: Cambridge. DOI: 10.1017/9781108674577.004

Wheadon, C., Barmby, P., Christodoulou, D., & Henderson, B. (2020). 'A comparative judgement approach to the large-scale assessment of primary writing in England', *Assessment in Education: Principles, Policy & Practice*, 27/1: 46–64. Routledge. DOI: 10.1080/0969594X.2019.1700212

Figures

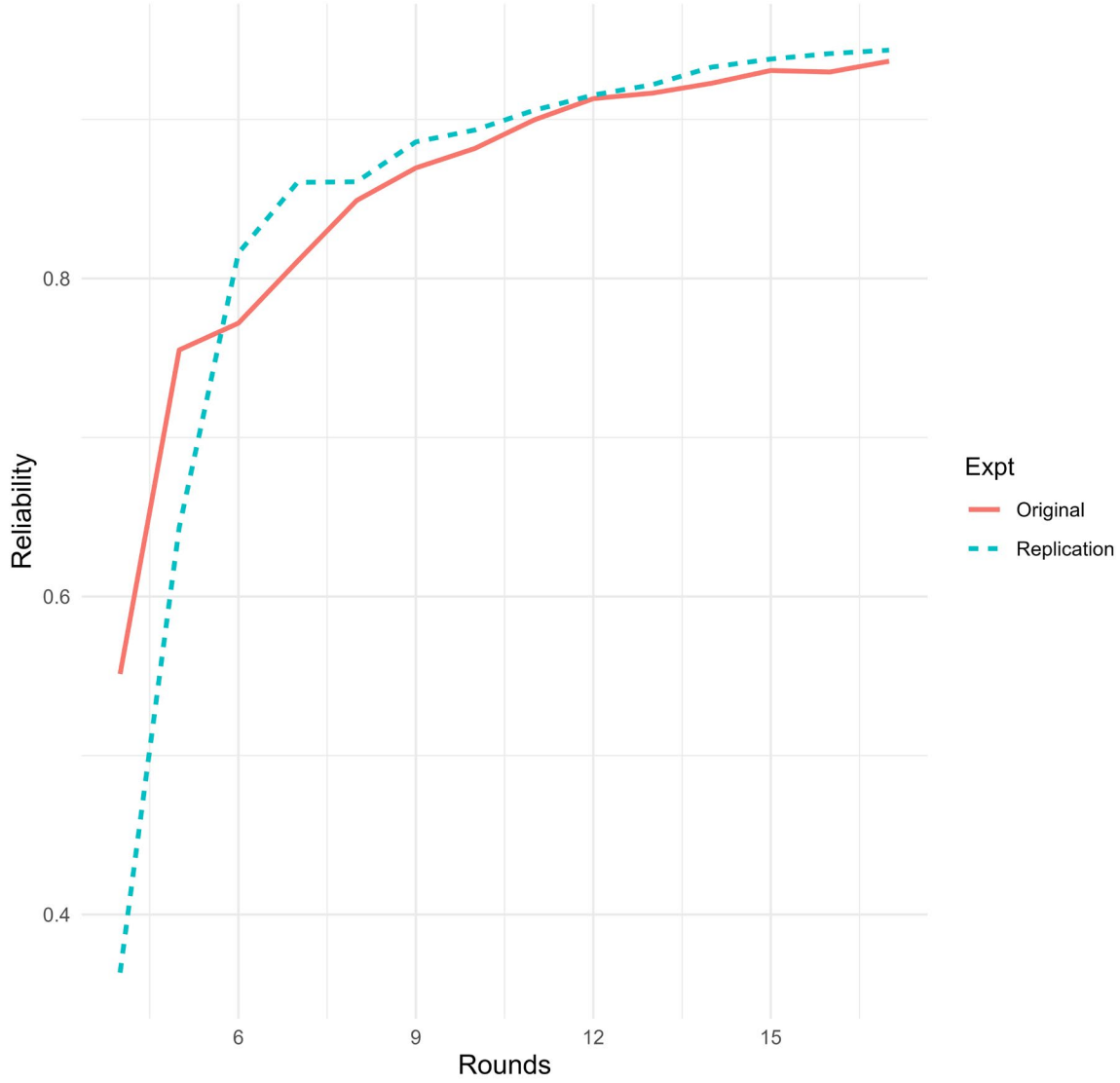

Figure 3: Comparison of the evolution of SSR levels in the original study and this replication.

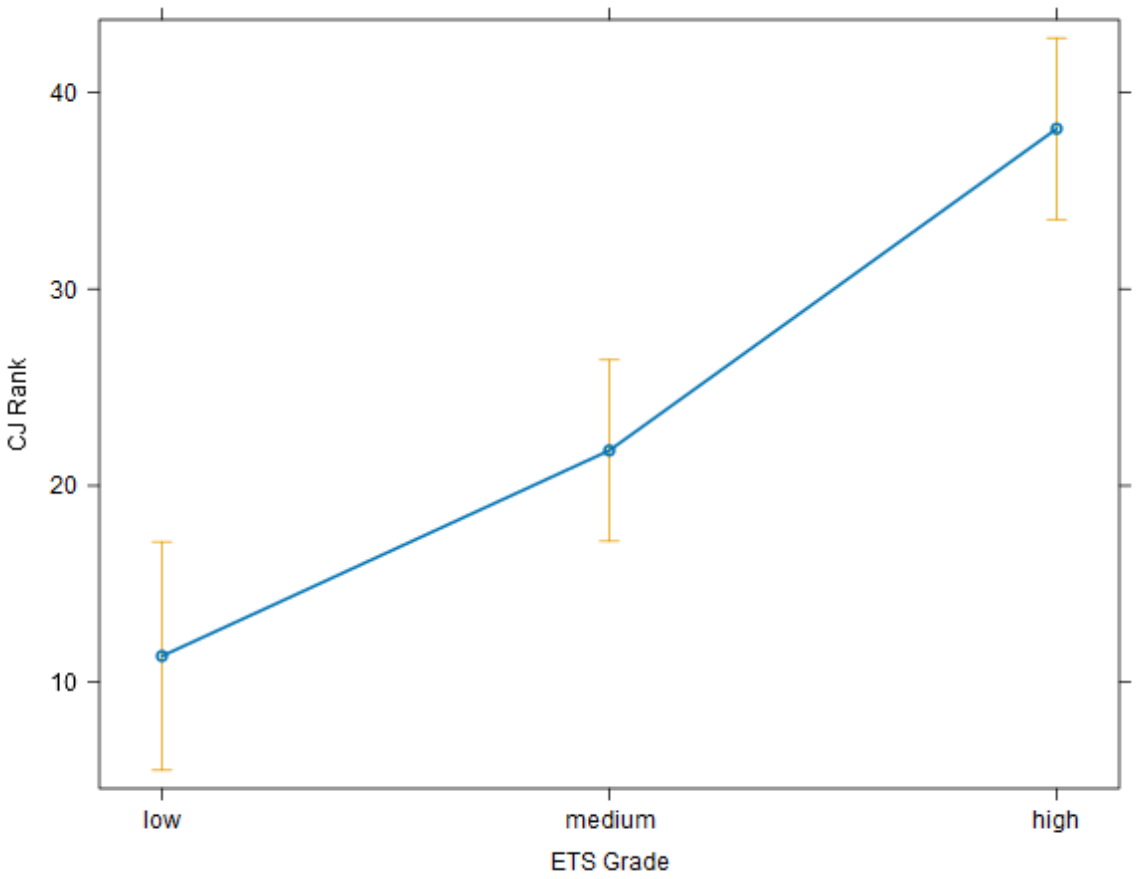

Figure 4:Median CJ rank of texts graded as low, medium, and high proficiency

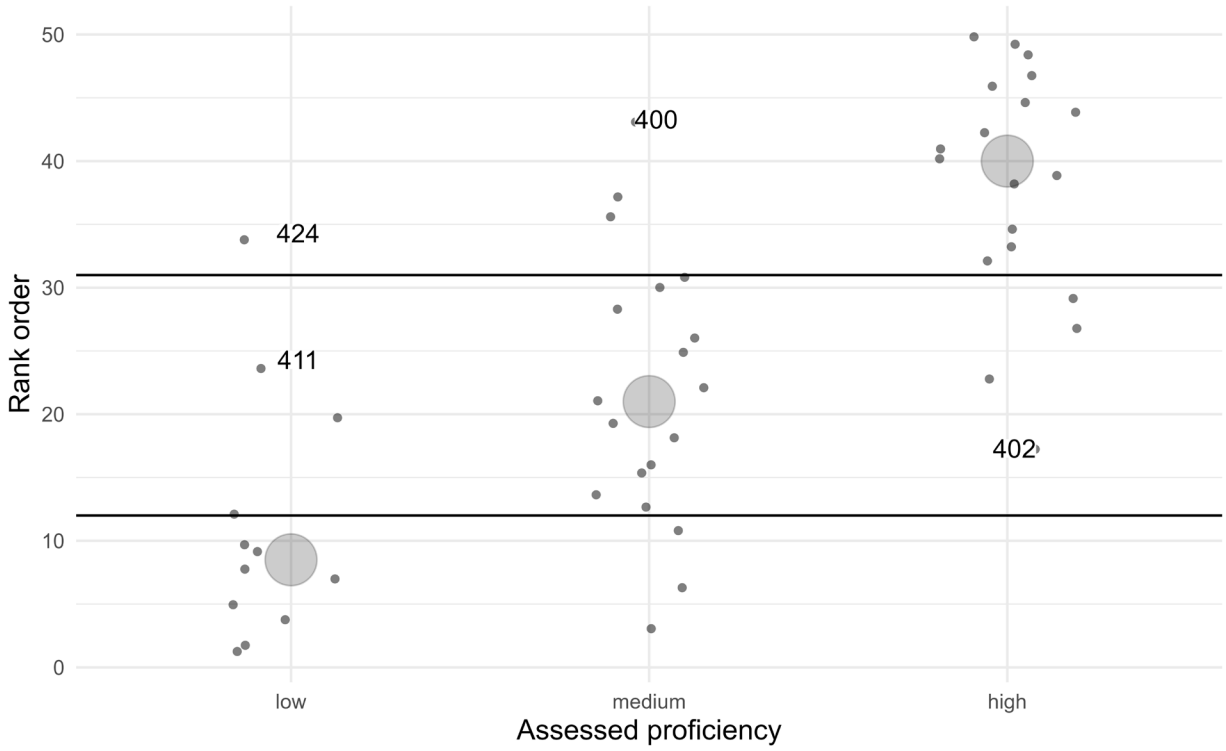

Figure 3: Rank order distributions of the texts in each ETS proficiency band. Outliers are labelled with their text ID.

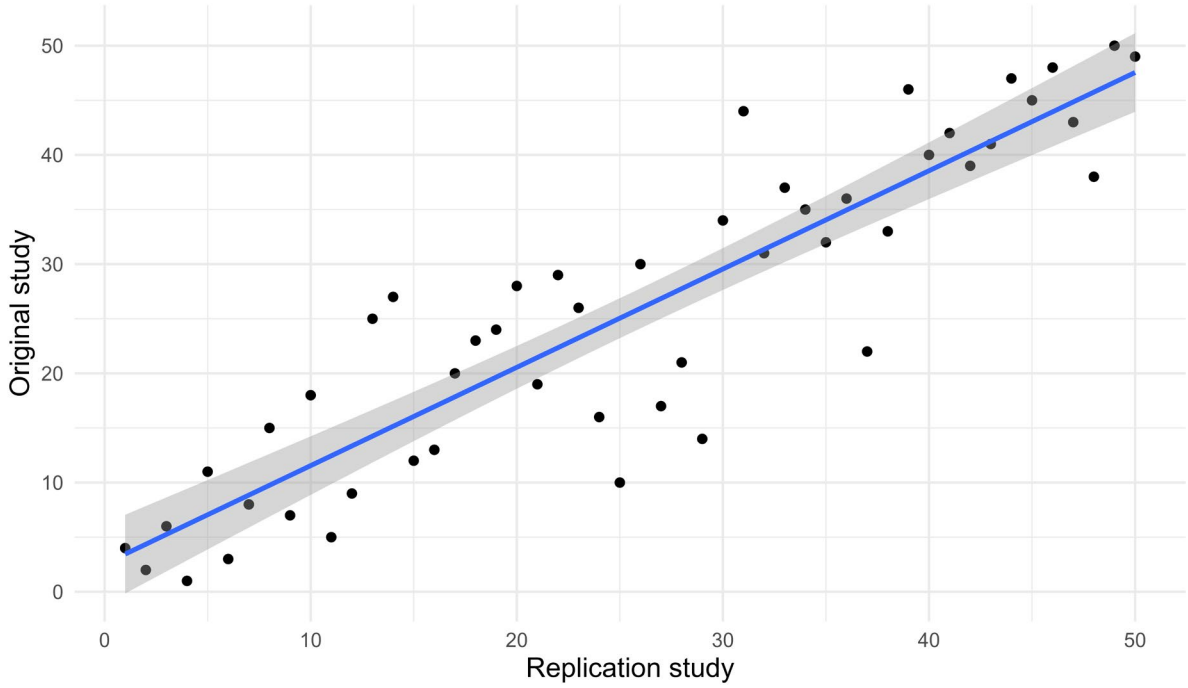

Figure 4: Correlation of rank orders in the original study and this replication