Is CJ a valid, reliable form of L2 writing assessment when texts are long, homogeneous in proficiency, and feature heterogeneous prompts?

Abstract

Comparative judgement (CJ) is a method of assessment in which judges perform paired comparisons of pieces of student work and decide which one is "better". CJ has many potential benefits for the writing assessment community, including its reliability, flexibility, and efficiency. However, by reviewing the literature on CJ's application to L2 writing assessment, we find that while existing studies have established the plausibility of using CJ in this context, they provide little indication of the conditions under which the method is most likely to prove useful. In particular, by focusing on the assessment of relatively short texts, covering a wide proficiency range, and using a single essay prompt, they leave unresolved the question of how such textual factors affect CJ's reliability and validity. To address this, we conduct two studies exploring the reliability and validity of a community-driven form of CJ for evaluating L2 texts which were longer, featured a narrower proficiency range, and were more topically diverse than earlier studies. Our results suggest that CJ remains reliable under these conditions. In addition, comparison with rubric-based assessment using CEFR scales suggests that the CJ approach also has an acceptable level of validity.

Keywords

Comparative judgement, assessment, writing, learner corpora, crowdsourcing

1. Introduction

In comparative judgement (CJ), judges are shown pieces of student work side-by-side and asked to decide which of the two is "better". Many such paired comparisons are conducted by a team of judges, who are usually (but not always) people with experience and/or expertise in the target domain. A scale can then be constructed by inputting these judgements to the Bradley-Terry model (Bradley & Terry, 1952), which assigns a score to each item such that the texts which "win" the greatest number of comparisons receive the highest scores, and those that win the fewest receive the lowest.

In the research literature, CJ is frequently presented as an alternative to more traditional methods of assessment such as mark schemes or rubric-based marking, over which it is seen as having several advantages. Firstly, as Jones & Davies (2023) have argued, CJ is well-suited to the evaluation of competencies and constructs that are hard to pin down. For example, Jones & Inglis (2015) describe the use of CJ to assess students' mathematical problem-solving skills. This is a much more difficult thing to evaluate than traditional math tests, in which test-takers simply supply a correct answer, because of the wide variety of approaches that test-takers can take. Nonetheless, Jones & Inglis reported that CJ-based assessments of these skills were highly reliable, and correlated strongly both with scores of the same exam generated through rubric-based assessment, and with students' predicted GCSE scores.

Jones & Inglis' study also illustrates a second key benefit of CJ, which is that it removes the need both for the development of complex marking schemes or rubrics, and for the resource-intensive process of training examiners to use them. Rather than instructing judges on how to evaluate test items, CJ instead assumes that they already possess the necessary skills and knowledge. CJ tasks therefore typically include only very simple instructions on how decisions should be made, such as "Choose the

better translated version" for a translation assessment (Han et al., 2022), or "The better writing?" for an L1 writing assessment (Wheadon et al., 2020).

Some scholars have argued that CJ's lack of reliance on rubrics or mark schemes also provides benefits in terms of validity. For example, Han (2021, 2022) argues that CJ is not vulnerable to issues affecting rubric-based marking, such as halo effects (i.e. a tendency for a rater's judgement of one aspect of a text to influence their judgement of all others) or differences in rater severity (i.e. when some raters tend to give higher or lower scores than others), since all that is required of judges is to choose the better of two texts. Han also found that judges considered robustness against rater bias to be an advantage of CJ, compared to rubric-based marking. This is presumably because, in CJ, final grades are based on many judgements performed by many judges, reducing the influence of any single judge on any single text. This same process also helps to ensure that CJ-based evaluations refer to a broad range of aspects of text quality – since each judge brings their own knowledge, preferences, and experience, CJ can capture many perspectives on what constitutes a strong piece of work (Lesterhuis et al., 2022; van Daal et al., 2019). The absence of a rubric is central to this principle of allowing judges to draw on their own personal perspective, ensuring that they "are not tied to predefined criteria" (van Daal et al., 2019, p. 61) imposed by what Pinot de Moira et al. (2022) have called the "hyper-specificity" of some rubric-based approaches.

Much of the research providing evidence of CJ's reliability, validity, and efficiency has been conducted in the areas of mathematics and L1 writing assessment. Much less is known about how CJ performs in the assessment of second language writing. Below, we review research exploring CJ's efficacy for both L1 and L2 evaluation, and argue that L2 studies to date have explored too narrow a range of assessment conditions, leaving much to be assumed about how the method will perform in other, more challenging, contexts. We then describe two studies which seek to address this issue by using CJ to assess a set of texts with more complex characteristics.

1.1 CJ for L1 writing assessment

Studies on the efficacy of CJ for L1 writing assessment can be divided into two general types. The first focuses on CJ's reliability and concurrent validity. For example, a study by Wheadon et al. (2020), which described the CJ assessment of more than 55,000 English L1 texts written by UK primary school students, reported a mean reliability of between .85 and .91 across various year groups (measured using the Rasch-based measure of scale separation reliability (SSR), which can be considered comparable to Cronbach's alpha; Verhavert et al., 2018, and see Section 2.2, below). Concurrent validity is usually conceptualised in these studies as correlation with other measures of the same construct, such as rubric-based assessments of the same texts. Most studies report moderate to strong correlations. For example, Steedle & Ferrara (2016) reported an SSR of around .89 and correlations of around r = .66 with rubric-based assessments in a study of argumentative essays written by American high school students, and a series of studies by Heldsinger & Humphrey (2010, 2013; Humphry & Heldsinger, 2019; McGrane et al., 2018) reported reliability levels above .95 and correlations with rubric-based assessment of .90 or higher for the assessment of narrative, descriptive, and argumentative essays.

A second type of study into CJ for L1 writing assessment investigates CJ's construct validity. Such studies are designed to address the perceived opacity of CJ, relative to rubric-based marking, which stems from the fact that it is not possible to know what judges have taken into consideration while making judgements. In order to explore the extent to which judges consider a sufficiently broad range of context-relevant textual features, Lesterhuis et al. (2018) asked judges to make notes explaining their decisions during CJ. The authors reported that 93.5% of comments were "construct-relevant". While most of these comments pertained to text argumentation and organisation, all

areas of their target competence description were mentioned to some extent. Similarly, Chambers & Cunningham (2022) used think-aloud to track judge attention, and again found that judges mostly paid attention to construct-relevant text features but also noted some influence of construct irrelevant features such as handwriting and the presence of missing responses on judge decisions. Lastly, Lesterhuis et al. (2022) identified four different judging profiles (narrowly focused, broadly focused, source-focused, and language-focused), all of which involved a principal focus on argumentation and organisation, but differed in secondary focus. The authors argued that CJ is valid for the assessment of L1 writing, and that a diversity of judging approaches results in the highest validity, since "combining the judgments of different types of assessors into scores leads to more informed text scores, representing the full complexity of text quality" (p. 8).

1.2 CJ for L2 writing assessment

Research into CJ for L2 writing assessment has been slower to emerge. Most studies to date have sought to establish the basic reliability of the method, and tested its concurrent validity via comparisons with rubric-based grades, considered the gold standard in language assessment. For example, Sims et al. (2020) compared CJ with rubric-based marking for a set of 60 L2 English argumentative essays written by students in an intensive English program. They reported a reliability of around .90 for both novice judges (undergraduate TESOL students) and experts (trained raters) after around 11 comparisons per text. Correlations with rubric-based grades, produced by the same judges, were also strong ($r > .90$). Similar levels of reliability and validity were reported by Author et al. (2022) in a study evaluating the potential for CJ to be used to generate proficiency ratings for texts in learner corpora. They used a community-driven approach in which they recruited judges from the applied linguistic community, who provided judgements of 50 argumentative essays taken from the ETS corpus of TOEFL exam responses (Blanchard et al., 2014). They reported a reliability of .95 after 20 comparisons per text. A linear model showed a significant relationship between CJ and pre-existing rubric-based grades (adj. $R^2 = .51$, $p < .001$), with texts rubric-rated as "low" proficiency having a lower mean rank than those rated "medium", and the "medium" texts a lower rank than those rated as having "high" proficiency.

These studies suggest that CJ holds potential as a method for L2 writing assessment. However, a rather weaker set of findings was reported in a study by Şahin (2021) on the CJ-based assessment of 35 L2 English writing samples. After around 20 comparisons per text, conducted by English instructors, Şahin reported a reliability of .72 and correlations with pre-existing rubric-based scores of .51. These levels of reliability and concurrent validity provide some counterpoint to the more positive findings reported above, and suggest that a better understanding is required of the factors that contribute to CJ's reliability and validity in the context of L2 assessment.

One potential source of variance in CJ's efficacy is the characteristics of the texts used for comparison. In seeking to establish baseline measures of the efficacy of CJ for L2 writing assessment, studies to date have tended toward relatively simple sets of texts, with broad proficiency spans, mean text lengths of 250 words or fewer, and relating to a single essay prompt. Each of these variables could affect the reliability and validity of a CJ task. Beginning with variance in word count, judges are more likely to suffer from fatigue when assessing longer texts than shorter ones. Tired judges may be more likely to make inattentive or inconsistent decisions than less fatigued ones, potentially having an impact on reliability and validity (Verhavert et al., 2019, 2022).

Regarding proficiency span, conducting comparisons is easiest where items differ widely in proficiency/quality, since it is easiest to decide which of two texts is "better" when they occupy different locations on the proficiency spectrum. As a result, item sets with broad proficiency spans will lead to greater agreement between judges and, in consequence, higher reliability (Crompvoets et

al., 2021, 2022). Judge statements tend to support this – for example one judge in a CJ study by Han (2021, p. 352) reported that "It was difficult to make CJ decisions when two renditions were of similar quality".

Lastly, CJ's efficacy may be affected when item sets include responses to several different prompts, rather than only one. In general, CJ is viewed as being robust to heterogeneity in item types (Jones et al., 2014, 2016; Jones & Davies, 2023). Nevertheless, a substantial body of research testifies to the potency of prompt effects in L2 assessment. For example, in a synthesis study In'nami and Koizumi (2016) reported that across 17 L2 writing assessment studies, task effects and task-examinee interactions accounted for substantial variance in L2 writing scores. In addition, there is evidence to suggest that raters perform differently in assessing different essay prompts. For example, Weigle (1999) reported that raters were significantly more severe in their assessments of some prompts than others.

In summary, while there is some evidence of CJ's potential in the area of L2 writing assessment, little is known about why studies to date have varied in the level of reliability and validity they have reported. As such, further research is required to clarify the conditions under which CJ can be expected to perform effectively. In particular, further research is needed on how CJ performs in the evaluation of complex texts. Such research would put L2 writing assessors in a better position to decide whether CJ might be suitable in their context.

1.3 The current study

This paper reports on two studies in which CJ is used to assess relatively complex sets of L2 texts. The studies are part of a wider project aiming to assess CJ's potential to enrich learner corpora with accurate, reliable assessments of text proficiency. As Carlsen (2012) and Park (2022) have both noted, few first-generation learner corpora contain satisfactory proficiency measurements, limiting their usefulness for studies investigating proficiency-related aspects of textual variation. While rubric-based rating has been used for this purpose, it tends to be considered too inefficient (Carlsen, 2012). This has led to calls for more alternative methods. CJ's flexibility and efficiency is well-suited to this role, but must be shown capable of yielding reliable grades for the often complex texts stored in learner corpora.

In the first study, we test the reliability and concurrent validity of CJ for evaluating texts with a narrow proficiency range (B1+ to C1 on the Common European Framework of References (CEFR)), and with long word counts (around 570 words). In the second study, we additionally explore the influence of topical diversity on CJ's efficacy by collecting judgements of responses to five different essay prompts, rather than only one.

For each of the two studies, there were two research questions:

1. Does comparative judgement of this set of texts result in a reliable scale?
2. How valid are these assessments in terms of their overlap with scores generated through gold standard writing assessment methods (i.e., their concurrent validity)?

Existing research allows us to make two hypotheses regarding these studies. Firstly, studies on the assessment of long, complex items, and those with relatively narrow proficiency spans/differences in quality, suggests that a larger number of comparisons will be required to reach satisfactory reliability. Since the L2 assessment literature provides no reason to assume that written L2 texts will perform any differently in this regard to any other type of item, our first hypothesis is as follows:

1. Both studies will result in acceptable reliability (>.80) and concurrent validity (correlations with rubric-based scores >.60); but reaching these levels will require more comparisons per item than in studies (e.g. Sims et al., 2020) featuring shorter, more heterogeneous texts.

Secondly, given evidence of prompt-rater interactions in L2 writing (e.g. Weigle, 1999), we hypothesise that:

2. Study 2 will yield lower levels of reliability and concurrent validity to Study 1, or require a larger number of comparisons to reach the same levels.

2. Method

2.1 Study design

The first stage in developing these two studies was to select two sets of texts of the desired length, proficiency range, and topical diversity. Given the wider project goal of L2 learner corpus enrichment, all texts were taken from the International Corpus of Learner English (ICLE; Granger et al., 2020). This resource contains around 9000 L2 English argumentative essays of a wide range of lengths (range = 69-4239 words, median = 548), written by language learners from 25 different L1 groups, and constituting responses to a wide range of prompts.

From this corpus, an initial selection of 500 texts was made. These texts formed the basis of a sub-project to generate triple-marked rubric-based grades for a subsection of ICLE texts which were (a) representative of the ICLE corpus in terms of length, (b) consisted of twenty responses for each of the corpus' 25 L1 groups, and (c) also contained a varied set of the most popular essay prompts in the corpus. The process for the selection and grading of these texts is described in detail in Author (in preparation) and will be only briefly described here. The following three steps were performed:

1. Filter the corpus by text length, leaving only texts whose length is within one standard deviation of the mean (i.e. 381-815 words);
2. Filter by essay prompt, leaving only texts whose prompt occurred at least 50 times and contained responses from at least 3 L1 groups;
3. Select exactly 20 texts for each of the 25 L1 groups from this filtered selection. Ideally, there would be four texts on each of five prompts per L1 group; in practice, however, this was not always possible.

Each text in this sub-corpus was graded by three raters using CEFR writing rubric C4 (Council of Europe, 2009, p. 187). Grading was conducted online and was preceded by two hours of CEFR descriptor refamiliarization, led by the second author. All raters had a minimum of 5 years of experience teaching or evaluating learners at CEFR levels B1-C2, and at least an undergraduate degree in TESOL or Applied Linguistics. To analyse the data, the many-Faceted Rasch measurement (MFRM) model (Linacre, 1989), an extension of the Rasch model (Rasch, 1960), was used. The MFRM model allows for additional facets such as rater severity and task difficulty to be part of the analysis model. Judges will vary in systematic ways and MFRM allows for the modelling of these differences (McNamara, 1996). 67 texts had to be dropped from the analysis as they either had a negative point measure correlation and/or were not in accordance with the MFRM model (i.e., misfitting). Note that none of these texts showed unusual properties, such as being classified as outliers or misfitting items, in the later CJ stage. The inter-rater reliability of the rubric-based assessments (overall Rasch-Kappa = -.03, judge Rasch-Kappa from -.13 to -.09) indicated that judge agreement was within expected ranges. Separation reliability – comparable to CJ's SSR measure – was .94.

The rubric-based scores were used to test the concurrent validity of the CJ-derived scores. Despite its shortcomings (see e.g. Panadero & Jonsson, 2020, for a summary), rubric-based assessment is well

established as the benchmark against which other forms of L2 writing assessment are tested, including CJ (e.g. Sims et al., 2020) and automated essay scoring methods (Powers et al., 2015; Vajjala, 2018).

After compiling the larger sub-corpus, the texts for CJ were selected. For both studies, we aimed to sample around 50 texts according to three criteria: 1) that they had a similar mean length to that of the ICLE as a whole; (2) that their writers came from a range of L1 backgrounds; and (3) that no L1 group provided a disproportionately high number of essays.

Study 1 concerns the comparative judgement of texts on a single topic. We chose the essay prompt "Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value". This is the most common prompt in the ICLE and features responses by 20 of the 25 L1 groups. Four essays on this topic were chosen for each of twelve L1 groups, resulting in a selection of 48 texts.

	ICLE argumentative	**Experiment 1**	**Experiment 2**	**Paquot et al. (2022)**
n.	8965	48	50	50
Source	ICLE	ICLE	ICLE	ETS corpus
Text length mean	598	569 words	610 words	272 words
Text length range	69-4239	400-771	399-813	71-421
n. Essay topics	Approx. 1500	1	5 (10 essays each)	1
Text proficiency	Unknown	Approx. B1-C2	Approx. B1-C2	"Low" to "High"; roughly A1-C2
n. L1 groups	25	12	22	1
Essays per L1 group	1-982	4	1-4	50

Table 1: Details of the texts used in each experiment; ICLE and Paquot et al. (2022) data included for comparison.

For Study 2, ten essays were chosen for each of five essay prompts. No L1 was represented more than once in each group of ten essays, nor more than four times overall. Table 1 presents the characteristics of the texts in both studies, alongside data on the texts in the ICLE as a whole; details of the more easily evaluated texts used in Author et al. (2022) are provided for comparison. Table 2 provides details of the proficiency levels generated by the rubric-based MFRM model. Further descriptive data for the essay sets in the two studies is available in the supplementary materials.

	Experiment 1 texts	Experiment 2 texts	All rubric-rated texts
	N = 48	N = 50	N = 500
CEFR level			
A2	0 (0%)	0 (0%)	1 (0.2%)
A2+	0 (0%)	0 (0%)	1 (0.2%)
B1	0 (0%)	0 (0%)	14 (3.2%)
B1+	6 (16%)	7 (15%)	46 (11%)
B2	15 (39%)	8 (17%)	102 (24%)
B2+	11 (29%)	13 (28%)	104 (24%)
C1	6 (16%)	16 (35%)	125 (29%)
C2	0 (0%)	2 (4.3%)	40 (9.2%)
Removed	10	4	67

Table 2: Grades derived from MFRM model of rubric-based assessment

Judges for both CJ studies were recruited using a community-driven approach based on Author et al. (2022). Members of the linguistic community were invited to participate via applied linguistic mailing lists, on social media, and through personal invitation. For the second study, participants also included a group of fourth-year undergraduate students in TESOL. The two sets of judges were very similar on all measured variables (see Table 3).

For both studies, participants first completed a brief demographic survey and consent forms. They were then directed to the *No More Marking* (NMM) CJ platform. This platform was selected over alternatives because it allows crowdsourced participants to begin submitting comparisons immediately upon learning of the study.

In the CJ task itself, judges were shown two essays side-by-side and asked to decide "Which of the two texts is a more advanced piece of second-language writing?" (see the screenshot in Figure 1). This wording reflects a text-centred approach to proficiency – i.e., it asks judges to consider the text itself rather than the proficiency of the learner who produced it (Carlsen, 2012). It also explicitly mentions the target construct – i.e., "second-language writing", and in this sense is more specific than the wordings chosen by Sims et al. ('Choose the better response'; 2020, p. 35) and Author et al. ('Which text was written by a more advanced speaker?'; 2022, p. 11). Nevertheless, the task definition is minimal by comparison with rubric-based forms of writing assessment. This is in line with the view that CJ's validity is increased when judges are allowed to make use of their own experience and understanding of the target construct, rather than being constrained by excessive guidance. (Lesterhuis et al., 2022).

	Experiment 1	**Experiment 2**
	N = 69	**N = 65**
Education Level		
Doctorate	33 (48%)	21 (33%)
Master's degree	31 (45%)	27 (43%)
Bachelor's degree	5 (7.2%)	8 (13%)
High School	0 (0%)	7 (11%)
English proficiency		
C2	49 (71%)	44 (68%)
C1	16 (23%)	16 (25%)
B2	3 (4.3%)	4 (6.2%)
B1	1 (1.4%)	1 (1.5%)
Experience as L2 examiner		
Frequently	24 (35%)	26 (40%)
Occasionally	13 (19%)	12 (18%)
Rarely	8 (12%)	5 (7.7%)
Never	24 (35%)	22 (34%)
Experience as L2 instructor		
Frequently	36 (52%)	37 (57%)
Occasionally	14 (20%)	9 (14%)
Rarely	7 (10%)	6 (9.2%)
Never	12 (17%)	13 (20%)
Experience as content instructor		
Frequently	30 (43%)	21 (32%)
Occasionally	14 (20%)	13 (20%)
Rarely	7 (10%)	11 (17%)
Never	18 (26%)	20 (31%)
Training in writing evaluation	45 (65%)	48 (74%)

Table 3: CJ judge demographics

Each judge was asked to complete a minimum of five comparisons, and permitted to complete a maximum of 10. Judges who failed to complete the minimum number of comparisons were excluded from further analysis. The number of judges meeting this criterion was 70 in Study 1 and 66 in Study 2.

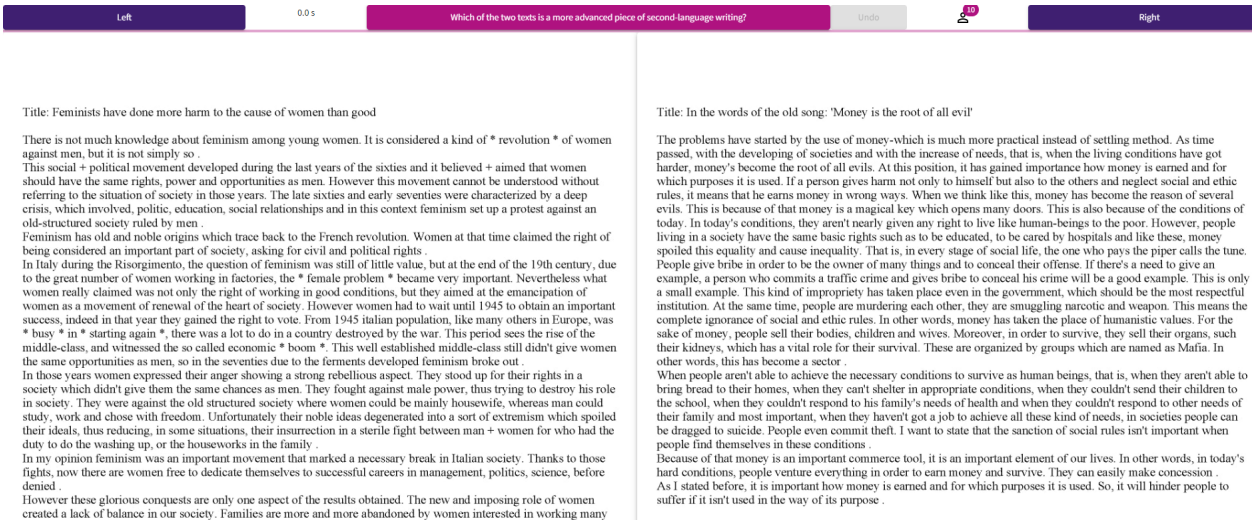

Figure 1: A screenshot of a text comparison from Experiment 2, in the No More Marking platform (nomoremarking.com).

Pairs of texts were chosen for comparison using NMM's pseudo-random algorithm, which first selects whichever text in the dataset has the fewest comparisons, and then randomly chooses a second text. Prior to the commencement of the data collection, we selected stopping conditions and judge inclusion criteria for the two studies. We decided to end each data collection upon the completion of 600 total comparisons. This equates to a total of 25 comparisons per text for the 48 texts in Study 1 (since each individual comparison constitutes a decision about each of two texts; (600 x 2)/48 = 25); and ((600 x 2)/50 =) 24 decisions per text for Study 2. This was based on Verhavert et al.'s (2019) finding that acceptable reliability (i.e. >= .80) can generally be achieved after around 20 comparisons per text. Data collection for Study 1 took two months to complete, while Study 2 took six months. The longer period required for the second data collection was because we wanted to avoid exhausting the goodwill of the applied linguistic community by publicising the study too heavily.

2.2 Analysis

Once all data had been collected, it was transformed into a rating scale in R (R Core Team, 2022 v.4.2.2), using the *btm* function of *sirt* package (v3.12-66; Robitzsch, 2022) with additional functions by Jones (2022). This package fits the judgement data to the Bradley-Terry model (Bradley & Terry, 1952), which yields a scale containing a score for each item which reflects the probability of that item "winning" a comparison against another item. Across the scale as a whole, these scores have a mean value of 0 and a standard deviation of around 2.5 (Bramley, 2007). The scale is then used as the basis for further statistical tests.

The model was run twice for each study. The first run was used to identify and remove judges whose data suggested that their judgements may be unreliable. This is typically done using infit scores, which are calculated alongside the rating scale and indicate the extent to which a judge's decisions were consistent with those of other judges. Larger infit scores indicate greater inconsistency. In existing research, an infit score of two or more standard deviations above the mean is often taken as evidence of misfit – i.e. an unacceptable level of consistency with other judges. However, high infit

does not always indicate unprincipled judge behaviour but can instead stem from a judge taking a different (but valid) perspective on certain texts. As such, removing judges based on high infit alone risks losing the diverse judge perspectives which Lesterhuis et al. (2022, p. 8) argue are key to the validity of CJ. We therefore employed a cautious approach to judge removal which focused on gathering sufficient evidence of inexpert performance. We specified three "red flags" of inexpert performance. Judges who met two of the three were removed. Infit scores of two standard deviations above the mean were the first red flag. The other two were:

1. An average decision time of less than twenty seconds per comparison;
2. A left-text selection rate of 10% or less, or 90% or more (this measure indicates a strong preference for texts on any one side of the presentation).

For each of the two studies, the *btm* function was re-run after removing judges on this basis, resulting in the final models used to answer the two research questions. To answer RQ1, pertaining to the reliability of the judgements in the two studies, the scale separation reliability (SSR) was extracted from these models. This measure provides an indication of the extent to which each item is separated from other items on the rating scale. An alternative to SSR is split halves reliability (Bisson et al., 2016), which involves assigning all judges to one of two groups, calculating a rating scale for each group, and then calculating the correlation between them. The process is then repeated around 100 times, with the median correlation coefficient being taken as the final reliability score (Jones & Davies, 2023). This method is more transparent than SSR, but requires twice as many comparisons to be conducted. Since SSR provides the greatest comparability with existing L2 CJ studies, and due to the practical difficulty of recruiting so many judges through a community-driven approach, SSR was preferred for this study.

To answer RQ2, pertaining to the concurrent validity of the two CJ-derived rating scales, we used two methods. Firstly, to test the fit of the CJ rank order with the categorical CEFR grades derived from the MFRM analysis of our rubric-based assessments, we fit a linear model with CJ rank order as the dependent variable and CEFR grade as the independent variable. Since the rubric-based graders were allowed to use "plus" levels to assist them in the evaluation of borderline texts (e.g., B1+ grades could be assigned to texts on the border between B1 and B2 standards), these plus levels were included in the linear model. Secondly, since MFRM analysis returns a rank-order of each text in the same way that CJ does, we ran correlation tests to compare the rank order of each text from the CJ and rubric-based methods. Spearman tests were used due to their compatibility with rank-ordered data.

Finally, because it was important for us to understand the extent of the difference between CEFR grades and CJ ranks, we identified as an outlier any texts whose CJ rank order was 1.5 times higher or lower than the median rank of texts of the rubric-rated as belonging to the same CEFR band. This allows us to identify and explore texts which generated significantly different responses from the CJ and rubric-based judges.

3. Results

3.1 Study 1

3.1.1 RQ1 Reliability of CJ for texts on a single topic:

For the set of essays on a single prompt, a total of 659 comparisons were collected from 70 judges. Of these, 649 decisions from 69 judges (mean decisions per judge = 9.4) passed the acceptability checks described in Section 2.2. These decisions were entered into the btm model, as described above. The SSR value for the resulting model was .82. A reliability of .7 was definitively achieved after

18 comparisons per text, and the .8 reliability threshold was crossed at 25 comparisons. The final value of .82 represents an asymptote, since the reliability level increased by less than .01 between rounds 27 and 29.

3.1.1 RQ2 Concurrent validity for texts on a single topic:

The MFRM model derived from rubric-based assessment showed that the texts in Study 1 all occupied the proficiency range from B1+ to C1 (see Table 4). Note that no rubric-based grades are available for ten of the texts used in this study due to the data loss which occurred during the MFRM analysis, described in Section 2.1. A linear model was built to test the fit between the CEFR levels of the 38 remaining texts and their mean rank order from the CJ evaluation. No model assumptions were violated. The model showed significant fit between the two assessments (F (3, 34) = 14.23, p < .001); the adjusted R^2 value was .52. Paired comparisons of the mean rank of texts at each CEFR level are presented in Table 4. The difference in mean rank of texts at B1+ and B2 levels was non-significant (t = .241, p = 1), as was that between texts at B2+ and C1 levels (t = -1.36, p = 1). All other comparisons were statistically significant.

				vs. B2		**vs. B2+**		**Vs C1**	
CEFR grade	***n.***	***EMM***	**95% CI**	***t***	***p***	***t***	***p***	***t***	***p***
B1+	6	14.83	6.98-22.68	0.241	1	-3.46	<.01	-4.24	.001
B2	15	13.73	8.77-18.70			-4.72	<.001	-5.31	<.001
B2+	11	31.45	25.66-37.25					-1.36	1
C1	6	38	30.15-45.85						

Table 4: Summary of linear model comparing CJ rank order to rubric-based CEFR grades for Experiment 1. T-*tests based on 41 degrees of freedom. Bonferroni correction has been applied to p values. EMM = Estimated Marginal Means.*

Figure 2 plots each text according to its CJ rank order and its CEFR level. Exact and adjacent agreement between CJ-rank and rubric-based grades were calculated by comparing each text's rubric-based CEFR level and CJ-based rank with the number of other texts rubric-rated at the same CEFR level. For example, six texts were rubric-rated at B1+ level in the current study. Therefore, CJ judgements of these texts were judged to show exact agreement if their CJ rank fell within the first six texts. Adjacent agreement was calculated by summing the number of texts at adjacent CEFR levels before testing the CJ rank in the same way. For example, the six B1+ texts were summed with 15 B2 texts to get 21; CJ judgements of these texts were therefore considered to show adjacent agreement with rubric-based grades if their CJ rank fell within the span 1-21. This is illustrated in Figure 2 by horizontal lines corresponding to the number of texts rubric-rated at each CEFR level. If CJ rank perfectly corresponded to CEFR level, all B1+ texts (for example) would appear below the corresponding horizontal line. In reality, each CEFR level contains texts with quite a wide range of CJ ranks. Exact agreement was 39%, while adjacent agreement was 95%. Only two texts showed neither exact nor adjacent agreement between CJ rank and CEFR grade. These were Text Q081, which was the highest ranked text in the B1+ level, and the outlier Q238, CEFR graded as B2 but CJ-ranked as the fourth strongest text. Outliers are briefly discussed below.

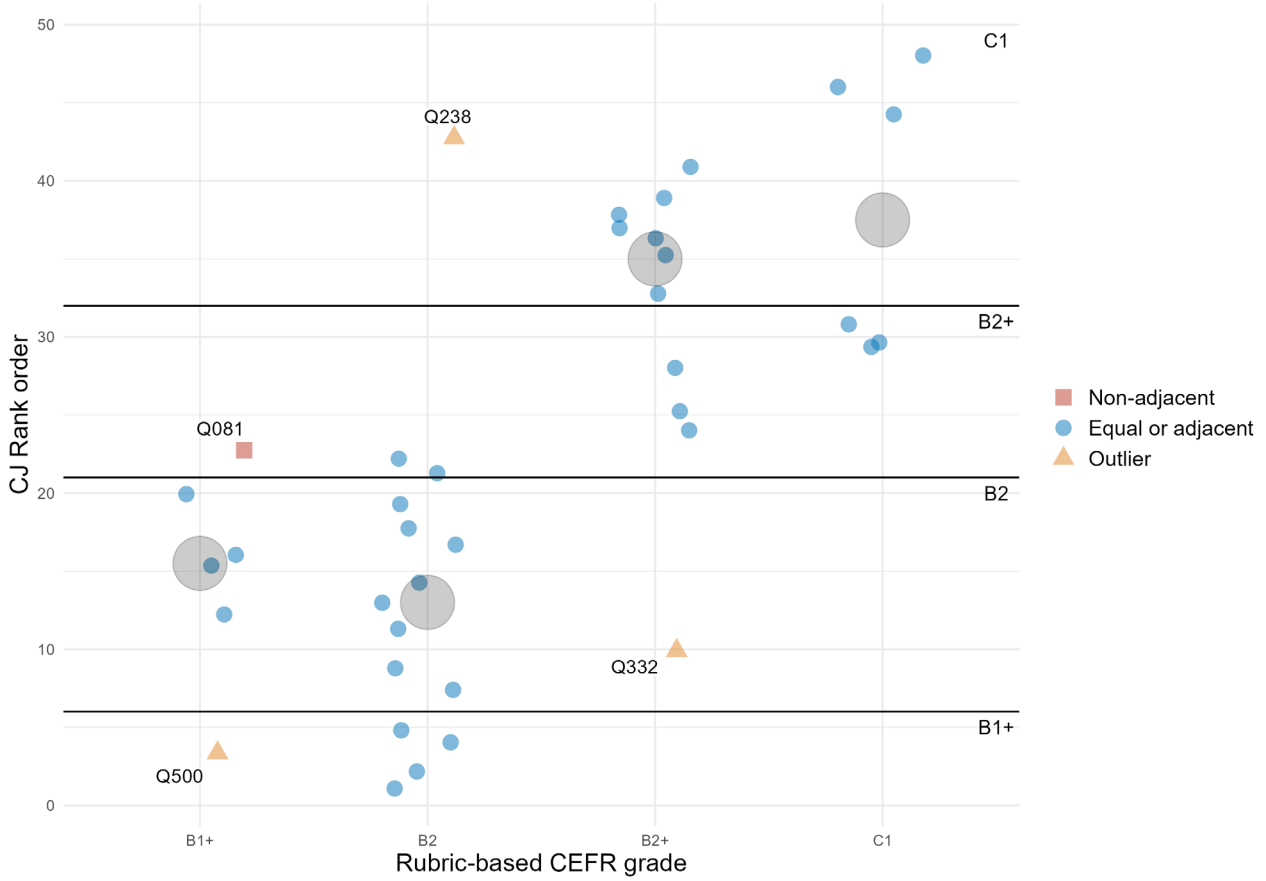

Figure 2: Scatterplot showing rank order of texts by CEFR level for Experiment 1. Small circles are texts; large circles show the mean CJ rank of texts at each CEFR level. Horizontal lines represent the number of texts at each CEFR level, according to MFRM rubric-based grades. "Equal or adjacent" texts are those whose CJ rank is within expected bounds – i.e. no higher or lower than texts either at the same CEFR level, or an adjacent CEFR level. "Non-adjacent" texts have unexpectedly high or low CJ grades, comparable to texts two CEFR bands higher or lower. Outliers are texts whose CJ rank order was 1.5 times higher or lower than the median rank of texts of the same CEFR level. Non-adjacent texts and outliers are labelled with their Text ID.

To explore the relationship between the CJ- and MFRM rubric-based datasets with greater granularity, a Spearman correlation showed that the CJ and MFRM rank-orders were significantly correlated (r = .668, p < .001). Levshina (2015) suggests that correlations between .3 and .7 indicate moderate strength. The correlation is plotted in Figure 3.

3.2 Study 2

3.2.1 RQ1 Reliability of CJ for texts on multiple topics:

For the set of texts comprising responses to five different essay prompts rather than only one, 610 comparisons were collected from 66 judges. Of these, one judge was removed, leaving 602 valid comparisons (mean = 9.3 decisions per judge). The btm model created from these comparisons had an SSR of .82. Though the final reliability of the two studies is almost identical, threshold values were reached slightly more quickly in Study 2: a level of .7 was passed at 13 comparisons per text, and .8 at 20 comparisons. An asymptote was reached at round 26.

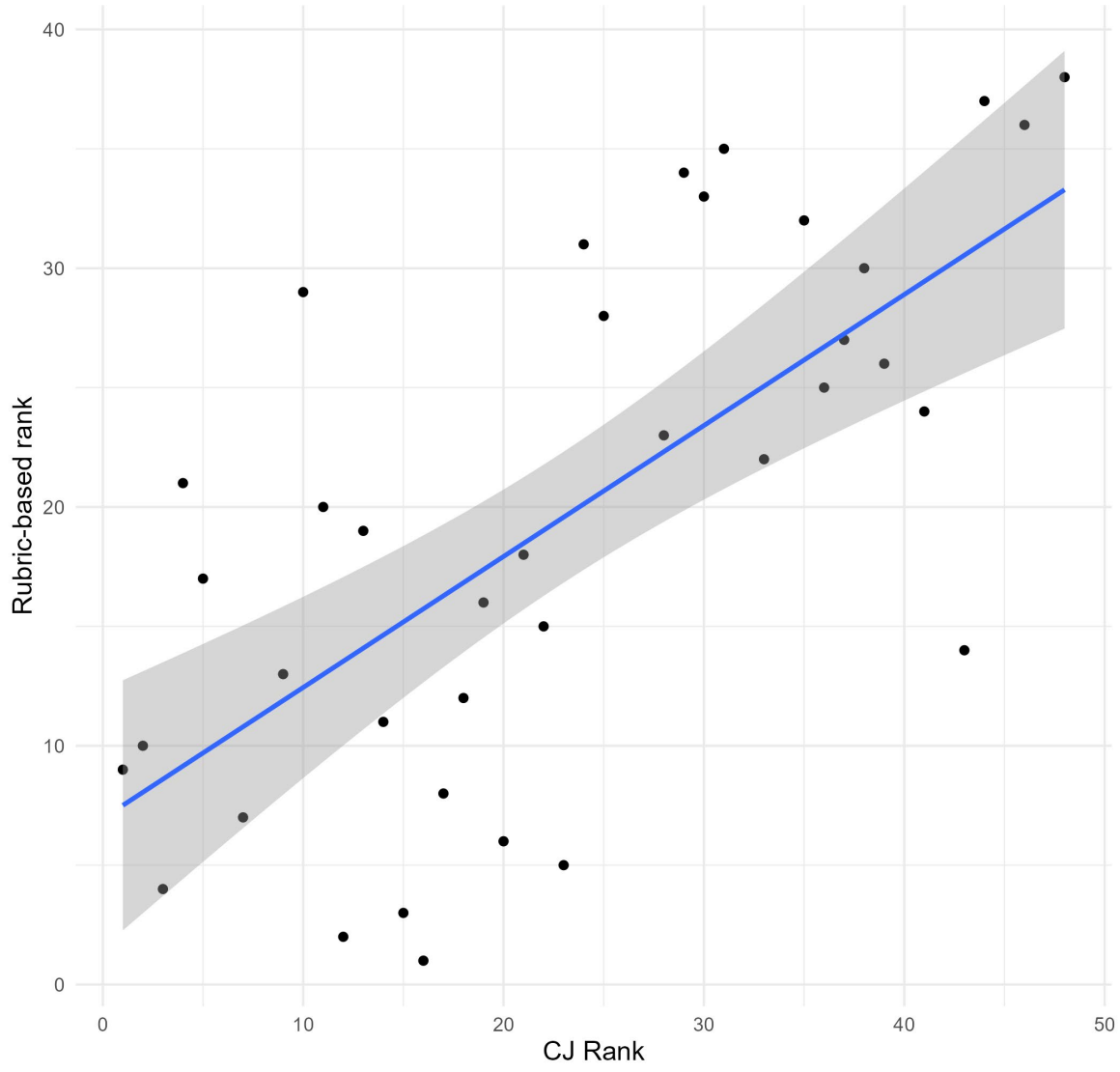

Figure 3: Correlation plot showing relationship between CJ- and MFRM rubric-based rank orders for Experiment 1

3.2.2 RQ2 Concurrent validity for texts on multiple topics:

The MFRM rubric-based model revealed that the texts in Study 2 covered a slightly broader proficiency range than those in Study 1, with two texts at C2 level in addition to those at levels B1+ to

				vs. B2		vs. B2+		Vs C1		Vs C2	
CEFR grade	***n.***	***EMM***	**95% CI**	***t***	***p***	***t***	***p***	***t***	***p***	***t***	***p***
B1+	7	11.43	3.00-19.86	-.43	1	-2.96	.05	-4.74	<.001	-3.68	<.01
B2	8	13.88	5.99-21.77			-2.60	.13	-4.45	<.001	-3.45	<.05
B2+	13	26.77	20.58-32.96					-2.04	.48	-2.05	.47
C1	16	35.19	29.61-40.77							-1.06	1
C2	2	44	28.22-59.78								

Table 5: Summary of linear model comparing CJ rank order to CEFR grades for Experiment 2. T*-tests based on 42 degrees of freedom. Bonferroni correction has been applied to p values. EMM = Estimated Marginal Means.*

C1 (see Table 5). Four texts were removed because they were dropped from the MFRM model. The linear model comparing the CJ rank order and CEFR level of the remaining 46 texts met all model assumptions, and was again significant (F (4, 41) = 9.555, p < .001); the adjusted R^2 value was .43. This is slightly lower than in Study 1. Paired contrasts showed that in most cases, rank orders were not significantly different between adjacent CEFR levels (e.g., B1+ vs B2 t = -2.96, p = 1; B2 vs B2+ t = -2.60, p = .13), but were significant for contrasts spanning two CEFR levels (e.g., B1+ vs B2+ t = -2.96, p = .05). The exact and adjacent agreement levels were 37% and 89%, respectively. Five texts received CJ ranks corresponding to a CEFR level two bands higher or lower than would be expected based on their actual CEFR level. Three of these texts are the outliers Q030, Q237 and Q437; the other two were B1+ texts whose CJ rank placed them alongside texts ranked at B2+ level (see Figure 4).

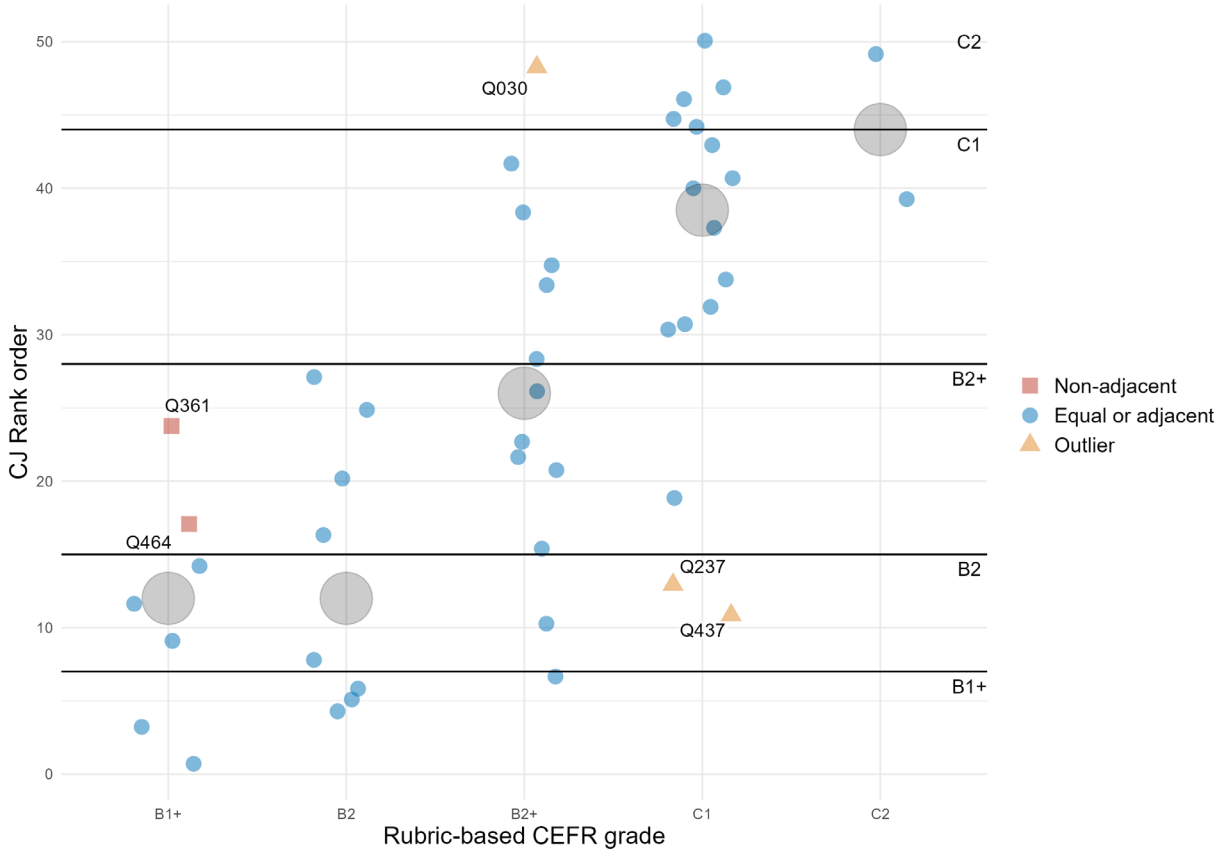

Figure 4: Scatterplot showing rank order of texts by CEFR level for Experiment 2. See Figure 1 for further information.

Lastly, we again ran a Spearman correlation test to identify the extent of overlap between the CJ and MFRM-based rank orders. The correlation was again of upper-moderate strength, almost identical to that of Study 1 (r = .677, p < .001). The correlation is shown in Figure 5.

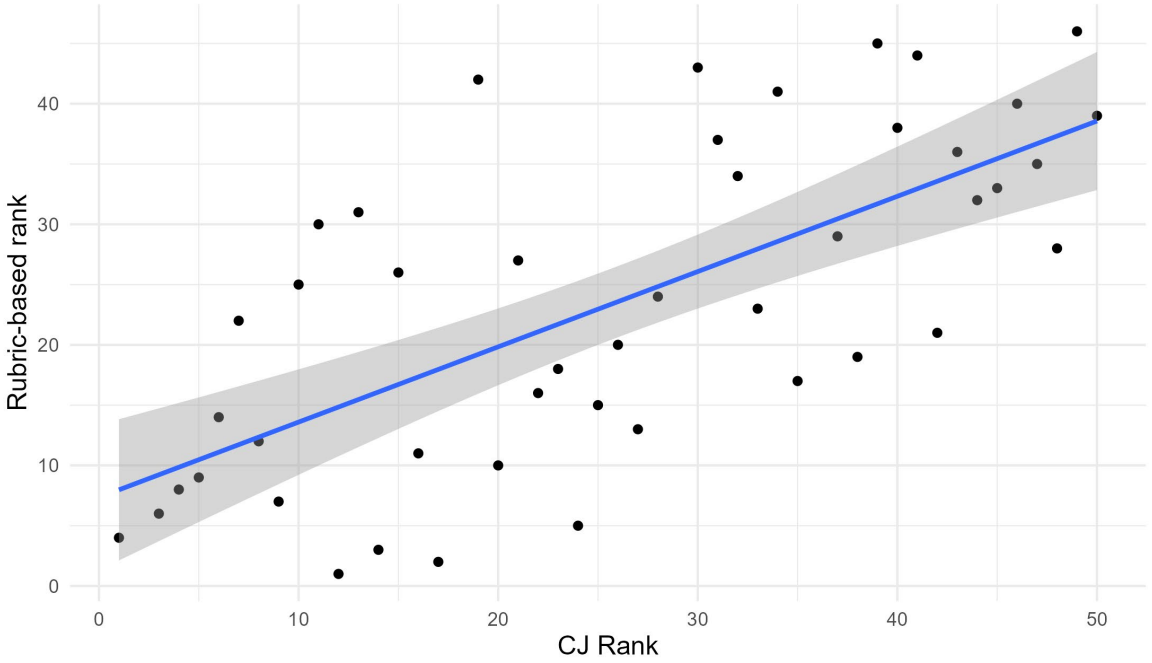

Figure 5: Correlation plot showing relationship between CJ- and MFRM rubric-based rank orders for Experiment 2.

4. Discussion

The two studies tested two hypotheses regarding the reliability and concurrent validity of CJ for assessing written L2 texts which were longer and covered a narrower proficiency span than those of existing studies (Study 1), and which included responses to a greater diversity of topics (Study 2). The first hypothesis, that reliability and validity levels for these more challenging texts would be lower than those in earlier studies which features simpler sets of texts, was supported. In both studies, reliability levels were around SSR = .81, much lower than the values above .90 in Sims et al. (2020) and Author (2022), and after more comparisons per item (around 25, compared with around 11 in Sims et al. and around 14 in Author). Similarly, the concurrent reliability levels were also lower than in the earlier studies – correlations with rubric-based assessments equalled around .67, compared with around .90 in Sims et al (2020). This is likely best explained by the complexity of the text characteristics in the present study.

In spite of these lower values, however, both studies returned acceptable levels of reliability and validity: SSR levels of .81 are between the thresholds of .70 and .90 for low- and high-stakes assessment, respectively, suggested by Verhavert et al. (2019), and beyond the Cronbach's alpha threshold of .80 suggested by Hoyt (2018) to indicate "good dependability of scores". Rank-order correlations in the upper-moderate range additionally suggested a close relationship between CJ and rubric-based evaluations.

The second hypothesis was that the reliability and concurrent validity of Study 2, which involved comparison of texts on five different prompts, would be lower than that of Study 1, in which all texts were responses to a single prompt. This hypothesis was not supported: the two studies had near-identical reliability and concurrent validity scores, suggesting that CJ was no more difficult when judges were required to evaluate texts on a range of topics than on just one. This supports the view that CJ is robust to heterogeneity in item sets (Jones et al., 2014, 2016; Jones & Davies, 2023) even in

the field of L2 writing assessment, where variance in writing tasks has been shows to affect rater performance (In'nami & Koizumi, 2016; Weigle, 1999). This is an encouraging result from the perspective of using CJ for learner corpus enrichment, since it suggests that reliable results can be achieved even where a corpus contains responses to a variety of prompts. One caveat is that although a variety of prompts were used in Study 2, all items were nevertheless the same textual genre (i.e. argumentative essays). Further research will be required to establish whether CJ remains reliable for comparing L2 texts of different genres (e.g. diagram description tasks and narrative passages).

Which specific textual characteristic(s) is the likeliest cause of the weaker reliability found in the present study? The data seem to rule out topical diversity, since Study 2 returned reliability and validity levels near-identical to those of Study 1. The influence of the two remaining factors – text length and proficiency span – cannot be definitively separated from the present data, since our study did not examine texts which were long but covered a broad proficiency span[1]. However, there is more empirical and theoretical support for an influence of proficiency span than for text length. As discussed in the introduction, several studies have identified an influence of proficiency span on CJ efficacy (Crompvoets et al., 2022; Han, 2021, 2022). Two such studies have even located a potential cognitive basis for this effect: Van Daal et al. (2017) suggest that judges experience greater complexity when comparing texts of similar quality than when comparing those which differ more clearly, while Gijsen et al. (2021) find evidence of this experienced complexity in judges' eye-tracking data, and demonstrate its association with reduced accuracy in decision-making. By contrast, few studies have systematically investigated the influence of text length on CJ efficacy. Intuitively, while it seems likely that longer texts would be harder to judge than shorter ones due to the amount of reading required, it is also possible that longer texts contain more evidence upon which judges can base their decisions, facilitating decision-making.

One question which clearly arises from the present data is what accounts for remaining differences between the CJ- and rubric-based assessments. One possibility is that judges attend to different aspects of text quality when using the two methods. Research on the content validity of CJ in L1 writing assessment suggests that CJ judges pay closer attention to higher-order aspects such as organisation and argumentation than to surface-level aspects such as observance of language conventions and proper use of referencing (Lesterhuis et al., 2018, 2022; van Daal et al., 2019). In their L2 CJ study, Author et al. (2022) examined several texts whose CJ rank was higher than would be expected based on their rubric-based grade, and found some evidence of a similar pattern. They noted that some of the texts which were CJ-ranked higher than their rubric-based grade would suggest were strong at the discourse-level (e.g. aspects such as argumentation and coherence), but also contained lower-level weaknesses such as basic spelling errors.

In both Studies 1 and 2, there were occasional texts whose CJ rank was markedly different from what would be expected based on their rubric-based grade (extreme examples of this are labelled outliers in Figures 2 and 4). However, examination of these texts does not offer clear support for the pattern suggested by Author et al. To illustrate, two examples of diverging texts are provided in the supplemental materials. Text Q030 was CJ-ranked much higher (48^{th} – higher ranks indicate stronger texts) than would be expected from its rubric-based grade (Q030 was graded as a B2+ text; the mean CJ rank for this grade was 26^{th}). Based on the above, we might expect this text to be strong at the level of discourse but weaker in terms of accuracy. This is not obviously the case: the text is arguably stronger in its lower-level characteristics (excellent spelling, very infrequent lexico-grammatical

[1] This was because the ICLE corpus does not contain sufficient texts at the A1-B1 levels.

errors) than at the level of discourse (it makes relevant points and uses clear linking expressions, but the author's position does not become clear until very late in the text). This can be compared with text Q437 (CJ rank = 11th, rubric-based grade C1 - the mean CJ rank for this grade was 38th) which is similarly accurate to Q030 at the level of lexico-grammar, and also contains a mix of strong and weak discourse-level aspects: it contains a clear thesis statement and a straightforward conclusion, but its body paragraphs are somewhat unfocused and not always clearly related to the topic. Ultimately, it is impossible to know exactly why these two texts were judged to differ so widely by CJ judges. It may be that judges prefer texts with one or other of the discourse-level features in the texts, but a systematic investigation into the decisions made during CJ and rubric-based assessment will be necessary before any conclusions can be drawn.

Lastly, another potential avenue for future research on the comparison of CJ and rubric-based assessment of L2 writing is to explore which of the two processes yielded the higher internal accuracy and consistency. In the present study, rubric-based grading with MFRM was treated as a gold-standard approach to L2 writing assessment, because this facilitated the evaluation of CJ's concurrent validity. However, even given the corrections that MFRM offers to the potential shortcomings of rubric-based assessment (McNamara et al., 2019), rubric-based assessment inevitably includes an element of measurement error. For this reason, we should be cautious about assuming that rubric-based scores are necessarily more accurate than CJ ones. A recent study by Pinot de Moira et al. (2022) suggests an alternative approach. The authors compared CJ and rubric-based assessment of L1 writing in terms of classification accuracy (i.e. how often each individual judgement matched what would be expected from each text's "definitive" score) and classification consistency (i.e. "a measure of the probability that a script will be given the same grade under repeated administrations of an assessment"; p32). The study found CJ to have higher accuracy than rubric-based assessment, and to be more consistent than single, double, and triple rubric-based assessment. Such an approach might offer a way to validate studies, such as the present one, that explore CJ's efficacy for L2 writing assessment.

5. Conclusion

This study sought to investigate whether CJ can be used to effectively assess sets of relatively long L2 texts which covered a relatively narrow proficiency range (B1+-C2), and which contained responses to a variety of essay prompts. The results suggest that these text characteristics do appear likely to affect reliability and validity levels (as well as the number of comparisons required to reach satisfactory reliability), but it nevertheless remains possible to achieve satisfactory evaluations.

These findings broaden the range of potential contexts in which CJ might effectively be used. Section 2.1 already discussed the idea of using CJ to enrich learner corpora with proficiency information. Similarly, the method could be used to assess and accurately reporting on the proficiency of the texts/learners involved in research studies. This would answer recent calls within applied linguistics for the development of off-the-shelf assessment methods for research purposes (e.g. Norris, 2018; Norris & Ortega, 2003; Park et al., 2022). CJ might also be used to assess texts in contexts such as schools or university departments, where texts of similar quality/proficiency frequently need to be assessed. One potential limitation is the essentially norm-referenced nature of CJ scores: if alignment with an external standard such as a CEFR band is required, additional steps will need to be taken. One possible approach is to include anchor items reflecting minimum standards for each band of the target proficiency scale (Jones & Davies, 2023).

The insights provided by this study regarding the reliability and concurrent validity of CJ for L2 assessment now need to be supplemented by research into construct validity. If CJ is to become an accepted tool for L2 writing assessment, it will be necessary to demonstrate that the decisions from

which CJ scores derive are being made on a principled basis, without excluding important aspects of text quality. We hope that this study will encourage such research.

Acknowledgements

The authors wish to thank all those who participated in the CLAP project as CJ judges. This research was supported by grant number 40008459 awarded by the Le Fonds de la Recherche Scientifique (FNRS) to the third author.

Bibliography

Author et al., (2022).

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), 141–164. https://doi.org/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; https://doi.org/10.35111/7ez0-x912. https://catalog.ldc.upenn.edu/LDC2014T06

Bradley, R. A., & Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, *39*(3/4), 324–345. https://doi.org/10.2307/2334029

Bramley, T. (2007). Paired Comparison Methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), Techniques for monitoring the comparability of examination standards (pp. 246–300). Qualifications and Curriculum Authority.

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.

Carlsen, C. (2012). Proficiency Level—A Fuzzy Variable in Computer Learner Corpora. *Applied Linguistics*, *33*(2), 161–183. https://doi.org/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*. https://www.frontiersin.org/articles/10.3389/feduc.2022.802392

Council of Europe, N. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR): A Manual*. Council of Europe Language Policy Division.

Crompvoets, E. A. V., Béguin, A. A., & Sijtsma, K. (2020). Adaptive Pairwise Comparison for Educational Measurement. *Journal of Educational and Behavioral Statistics*, *45*(3), 316–338. https://doi.org/10.3102/1076998619890589

Crompvoets, E. A. V., Béguin, A. A., & Sijtsma, K. (2021). *Pairwise comparison using a Bayesian selection algorithm: Efficient holistic measurement*. PsyArXiv. https://doi.org/10.31234/osf.io/32nhp

Crompvoets, E. A. V., Béguin, A. A., & Sijtsma, K. (2022). On the Bias and Stability of the Results of Comparative Judgment. *Frontiers in Education*, *6*. https://www.frontiersin.org/articles/10.3389/feduc.2021.788202

Gijsen, M., van Daal, T., Lesterhuis, M., Gijbels, D., & De Maeyer, S. (2021). The Complexity of Comparative Judgments in Assessing Argumentative Writing: An Eye Tracking Study. *Frontiers in Education*, *5*. https://www.frontiersin.org/articles/10.3389/feduc.2020.582800

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

Han, C. (2021). Analytic rubric scoring versus comparative judgment: A comparison of two approaches to assessing spoken-language interpreting. *Meta : Journal Des Traducteurs / Meta: Translators' Journal*, *66*(2), 337–361. https://doi.org/10.7202/1083182ar

Han, C. (2022). Assessing spoken-language interpreting: The method of comparative judgement. *Interpreting*, *24*(1), 59–83.

Han, C., Hu, B., Fan, Q., Duan, J., & Li, X. (2022). Using computerised comparative judgement to assess translation. *Across Languages and Cultures*, *23*(1), 56–74. https://doi.org/10.1556/084.2022.00001

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, *37*(2), 1–19. https://doi.org/10.1007/BF03216919

Heldsinger, S., & Humphry, S. M. (2013). Using calibrated exemplars in the teacher-assessment of writing: An empirical study. *Educational Research*, *55*(3), 219–235. https://doi.org/10.1080/00131881.2013.825159

Hoyt, W. T. (2018). Interrater reliability and agreement. In *The reviewer's guide to quantitative methods in the social sciences* (pp. 132–144). Routledge.

Humphry, S. M., & Heldsinger, S. (2019). A Two-Stage Method for Classroom Assessments of Essay Writing. *Journal of Educational Measurement*, *56*(3), 505–520. https://doi.org/10.1111/jedm.12223

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing*, *33*(3), 341–366. https://doi.org/10.1177/0265532215587390

Jones, I. (2022). *Sirt functions* [Computer software]. https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fgithub.com%2FNoMoreMarking%2Fsirt%2Fblob%2Fmain%2FR%2Fsirt_functions.R&data=05%7C01%7Cpeter.thwaites%40uclouvain.be%7C63005d2103e44f64362108db792c2b22%7C7ab090d4fa2e4ecfbc7c4127b4d582ec%7C0%7C0%7C638237002831653403%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=yI2oKl7iGJmT3ncrc9Zmo5RPYnhublrngnHe9fH9LV0%3D&reserved=0

Jones, I., & Davies, B. (2023). Comparative judgement in education research. *International Journal of Research & Method in Education*, *0*(0), 1–12. https://doi.org/10.1080/1743727X.2023.2242273

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics*, *89*, 337–355.

Jones, I., Swan, M., & Pollitt, A. (2014). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, *13*(1), 151–177. https://doi.org/10.1007/s10763-013-9497-6

Jones, I., Wheadon, C., Humphries, S., & Inglis, M. (2016). Fifty years of A-level mathematics: Have standards changed? *British Educational Research Journal*, *42*(4), 543–560. https://doi.org/10.1002/berj.3224

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*, *0*(0), 1–15. https://doi.org/10.1080/0969594X.2022.2147901

Lenth, R. V. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://CRAN.R-project.org/package=emmeans

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*. https://www.frontiersin.org/articles/10.3389/feduc.2022.823895

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. https://doi.org/10.17239/L1ESLL-2018.18.01.02

Levshina, N. (2015). How to do linguistics with R. *Data Exploration and Statistical Analysis, Amsterdam-Philadelphia*.

Linacre, J. M. (1989). *Many-faceted Rasch measurement* [PhD Thesis, The University of Chicago]. https://search.proquest.com/openview/9947a2b1a43f10331c468abcca2ba3e6/1?pq-origsite=gscholar&cbl=18750&diss=y

McGrane, J. A., Humphry, S. M., & Heldsinger, S. (2018). Applying a Thurstonian, Two-Stage Method in the Standardized Assessment of Writing. *Applied Measurement in Education*, *31*(4), 297–311. https://doi.org/10.1080/08957347.2018.1495216

McNamara, T. (1996). *Measuring Second Language Performance* (First Edition). Pearson Educational.

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice and language assessment*. Oxford University Press.

Norris, J. M. (2018). *Developing C-tests for estimating proficiency in foreign language research*. Peter Lang.

Norris, J. M., & Ortega, L. (2003). Defining and measuring SLA. *The Handbook of Second Language Acquisition*, 716–761.

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. https://doi.org/10.1111/lang.12475

Pena, E. A., & Slate, E. H. (2019). *gvlma: Global Validation of Linear Models Assumptions*. https://CRAN.R-project.org/package=gvlma

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. https://doi.org/10.1177/00345237221118116

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice*, *19*(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354

Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating Automated Essay Scoring: A (Modest) Refinement of the “Gold Standard”. *Applied Measurement in Education*, *28*(2), 130–142. https://doi.org/10.1080/08957347.2014.1002920

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* https://psycnet.apa.org/record/1962-07791-000

Robitzsch, A. (2022). *sirt: Supplementary Item Response Theory Models*. https://CRAN.R-project.org/package=sirt

Şahin, A. (2021). Feasibility of using comparative judgement and student judges to assess writing performance of English language learners. *Journal of Pedagogical Research*, *5*(4), 140–154. https://doi.org/10.33902/JPR.2021474154

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

Steedle, J. T., & Ferrara, S. (2016). Evaluating Comparative Judgment as an Approach to Essay Scoring. *Applied Measurement in Education*, *29*(3), 211–223. https://doi.org/10.1080/08957347.2016.1171769

Vajjala, S. (2018). Automated Assessment of Non-Native Learner Essays: Investigating the Role of Linguistic Features. *International Journal of Artificial Intelligence in Education*, *28*(1), 79–105. https://doi.org/10.1007/s40593-017-0142-3

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2019). 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. https://doi.org/10.1080/0969594X.2016.1253542

van Daal, T., Lesterhuis, M., Coertjens, L., van de Kamp, M.-T., Donche, V., & De Maeyer, S. (2017). The Complexity of Assessing Student Work Using Comparative Judgment: The Moderating Role of Decision Accuracy. *Frontiers in Education*, *2*. https://www.frontiersin.org/articles/10.3389/feduc.2017.00044

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). Springer. https://www.stats.ox.ac.uk/pub/MASS4/

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

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), 428–445. https://doi.org/10.1177/0146621617748321

Verhavert, S., Furlong, A., & Bouwer, R. (2022). The Accuracy and Efficiency of a Reference-Based Adaptive Selection Algorithm for Comparative Judgment. *Frontiers in Education*, *6*. https://www.frontiersin.org/articles/10.3389/feduc.2021.785919

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2), 145–178. https://doi.org/10.1016/S1075-2935(00)00010-6

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. https://doi.org/10.1080/0969594X.2019.1700212