Comparative Judgement for advancing research in applied linguistics

Abstract

Comparative judgement (CJ) is a data collection method in which judges are presented with two items, side-by-side, and asked to decide which is "better". By compiling the results of many such decisions, a scale can be developed to rank each item from best to worst. Though most commonly used for educational assessment, CJ is fundamentally a method for generating holistic, perceptually grounded measurements of hard-to-define constructs. This capability gives CJ broad potential in the field of applied linguistics, as it can address the need for more accurate measurement and definition of various applied linguistic constructs. In this tutorial, we provide a step-by-step guide on how to set up CJ studies and analyse the resulting data. We also discuss some of the method's strengths and weaknesses, and explore ways in which it might enhance and broaden the methodological toolkit of applied linguistic research.

Keywords

Comparative judgement, assessment, quantitative research methods, measurement

1. What is comparative judgement, and how can it help applied linguists?

Comparative Judgement (CJ) is a data collection method, with associated analytical procedures, which is typically used for measurement purposes. It was first described in the work of Louis Thurstone (1927, 1954), a pioneer in the field of psychometrics, who used it to measure hard-to-define psychological values like attitudes, beliefs and preferences. Later, CJ was introduced to the field of education and employed to assess a wide range of complex educational constructs, including mathematical reasoning, critical thinking, essay writing, and creative performance. The method holds significant potential for the field of applied linguistics as it can be used to address issues of construct definition and measurement through human perception. One area where it has already been used is in assessing linguistic competencies, including first and second language writing proficiency (Lesterhuis et al., 2022; Paquot et al., 2022; Şahin, 2021; Sims et al., 2020), translation (Han, 2021; Han et al., 2022), and sign language interpretation (Han & Xiao, 2022). Another area where the method has started to be used in AL is in defining and measuring complex constructs (e.g. Crossley et al., 2023; Zhang & Lu, 2024). Purpura et al. (2015) note that many applied linguistic constructs are multidimensional in nature and difficult to measure directly: syntactic complexity, lexical richness, fluency, second language proficiency. In many cases, such constructs are measured by dividing the construct into a limited number of very specific sub-components. For example, spoken L2 fluency is often assessed by combining statistical measures of the rate of speech and the frequency and duration of pauses (Kormos & Dénes, 2004). CJ provides a way for such constructs to be measured holistically. Further, because these measurements are derived from human judgements, they are grounded in an emic approach which emphasises the importance of human perception of (linguistic) phenomena over statistical or computational measures (Jarvis, 2017). Finally, several properties of CJ tasks offer potential advantages over alternate approaches to the measurement of linguistic phenomena, such as Likert scales.

In CJ, participants (who are referred to as *judges*) are shown pairs of items (often called *representations* or *performances*) and asked to compare them and decide which is "better". They are guided in this by a simple *task definition*. For example, in a study on the orthographic familiarity of English words, Bisson (2022) asked judges to decide "which [item] look[s] more similar to an English word". Many such comparisons are conducted by many judges, at the end of which process it is possible to generate a scale which ranks each item in the terms expressed in the task definition.

To date, most applied linguistic research on CJ has explored the method's potential for assessing proficiency. In such contexts, CJ is typically found to possess similar reliability and concurrent validity to more traditional approaches, while offering gains in efficiency. For example, Sims et al. (2020) asked two groups of judges – TESOL undergraduates and trained raters – to evaluate a set of L2 argumentative essays using both CJ and rubric-based methods. Many-faceted Rasch modelling was used to transform the rubric-based grades into rank-scales. The CJ and rubric-based scales were then compared. The authors reported that both sets of judges, using both assessment methods, produced highly reliable scales which correlated strongly with pre-existing rubric-based scores. Though the novel rubric-based results were marginally stronger than the CJ ones, CJ was more efficient, being on average 51 seconds per text faster than the rubric-based approach.

Other studies have reported similar findings when using CJ to assess other language competencies. For example, in a series of studies, Han (2021, 2022; Han & Xiao, 2022) has shown that CJ can yield accurate assessments of spoken- and sign-language interpreting. Correlations between CJ- and rubric-based assessments in these studies were around .85, even when judges possessed relatively little experience of assessing the target competence. Han (2022) also found that judges paid attention to construct-relevant features of each performance while conducting comparisons, providing evidence of the construct validity of the approach. These studies also illustrate the range of item types that can be evaluated using CJ: they involved asking judges to compare video recordings of learner performances rather than written texts.

Studies such as these use CJ to generate holistic assessments of proficiency which are rooted in the collective understandings of a target competence. They are therefore aligned with a constructivist epistemology which views proficiency not as a fixed construct but rather one shaped by collective perceptions and interpretations. Purpura et al. (2015) have argued that measurements with these properties should be prioritised by researchers seeking to define AL constructs: "given the centrality of measured constructs in the assertion of L2 research claims (…), it is imperative that the theoretical constructs we use in our work reflect collective understandings of the phenomena we wish to measure." (p. 38).

Several recent applied linguistic studies have drawn on CJ's capacity to generate measurements of this type. For example, Bisson et al. (2022) produced measures of orthographic (and phonological) familiarity by showing (or playing audio recordings of) pairs of Welsh words to lay judges and asking them to decide which looked (or sounded) more similar to an English word. In this way, the authors were able to produce scales ranking these words from most to least familiar. They then used these scales to explore the impact of familiarity on foreign language word learning. Crossley et al. (2023) used a similar approach to generate crowdsourced ratings of the lexical diversity found in pieces of L2 writing. The resulting scales reflect the type of collective, perceptual measures envisaged by Purpura et al., as well as by Jarvis (2017).

The aim of this paper is to show how applied linguists can make use of CJ to generate measurements with the above properties, whether to assess proficiency or to measure some other construct. The next section introduces the theory behind CJ. Following this, we provide a step-by-step tutorial on

how the method can be used in practice, before exploring some alternatives to CJ, its benefits and constraints, and some potential applications.

2. Theoretical and statistical underpinnings

2.1 Theoretical assumptions

Two key assumptions form the theoretical basis for CJ. The first is Thurstone's (1927) "law of comparative judgement", which can be expressed as the claim that comparing two items side-by-side is easier, and yields more reliable results, than evaluating items in isolation (termed *absolute* judgement). This claim is often described with reference to several vulnerabilities to which absolute forms of assessment (which include rubric-based assessment and the use of Likert scales) are prone. These include differences in judge severity and various types of rater bias, such as central/extreme tendencies. As Steedle and Ferrara (2016) explain, there is no possibility for such problems to occur in CJ because there is no need for judges to produce a rating for each item; all that is required is a decision as to which of two items is better.

The second claim underpinning CJ is that it does not require judges to be provided with any assessment criteria beyond a simple task definition. Instead, so long as they collectively possess sufficient breadth and depth of expertise in the target domain, groups of judges will be able to make valid assessments simply by drawing on their own knowledge. Of course, this could raise concerns that individual judges might make decisions based on very different understandings of the target construct; but this problem is alleviated by the fact that in CJ, each item is evaluated by many judges, each of whose decisions are given equal weight. For example, in a writing assessment study by Paquot et al. (2022), each of 50 L2 English argumentative essays was evaluated by an average of twenty different judges. This has two effects. Firstly, it means that the influence of any single judge who might prioritise a given aspect of the target competence in their decisions is mitigated by the judging preferences of the rest of the group. By contrast, the reliability of absolute assessments is negatively affected when raters have different interpretations of target competences/constructs (Weigle, 2002), and for this reason substantial resources are spent in developing rubrics, training raters, and using double- or even triple-rating of items to minimise inter-rater variation. Secondly, the plurality of judges involved in CJ tasks allows a broad representation of the target construct to emerge, since each judge is assumed to bring a slightly different perspective to the task. In this sense, diversity of construct interpretation is actively welcomed in CJ, since broad understandings lead to broad construct representation. Indeed, Bisson et al. (2016, p. 143) have argued that CJ might offer broader construct representation than rubric-based assessment, since rubrics, in their effort to achieve high reliability, can often reflect "narrow and rigid [construct] definitions" which limit raters' ability to draw on their own expertise.

2.2 Mathematical basis

The mathematical underpinnings of CJ also derive from Thurstone's law of comparative judgement. In its original form, the law was expressed as a series of mathematical expressions allowing the calculation of scale values reflecting the likelihood of a given item "winning" a paired comparison against other items. Revisions by Bradley and Terry (1952) and Andrich (1978) resulted in the mathematical framework currently used for deriving scale scores (and corresponding rank-orders), and for the calculation of various properties of the scale as a whole. A procedure for calculating these scales is described in "Analysing CJ data", below. CJ's mathematical framework is very similar to the dichotomous Rasch model, which is used to model responses to tests with binary answers such as True/False questions. It has the properties of being *iterative*, in that each time a comparison occurs, the results are computed and each item's scale score updated; and *probabilistic*, in that the

probability of any given item "winning" a comparison against any other can be directly estimated from the two items' scale scores. For a detailed discussion of the mathematics behind CJ, readers are directed to Bramley (2007).

2.3 Validity

Naturally, researchers have a responsibility to provide evidence that these assumptions are valid. The most frequently-used approach to testing CJ's validity is to compare CJ-derived rating scales with scales generated through other forms of assessment (i.e. tests of concurrent validity). Sims et al.'s (2020) study, described above, is one example; the authors reported correlations of around $r = .90$ between CJ- and rubric-based rank scales. Other studies have shown that CJ scores also correlate with measures of less closely related constructs. For example, Marshall et al. (2020) found that CJ-based assessments of secondary students' performances on an English writing task and a statistical analysis task correlated with students' English and Mathematics Grade Point Averages, respectively.

Other studies have explored the extent to which CJ judges consider various aspects of their target construct when making decisions (i.e. tests of construct validity). For example, in a study exploring CJ's application to L1 writing assessment, Lesterhuis et al. (2018) found that 93.5% of judge comments referred to "construct-relevant" textual features. Most of these related to argumentation and organisation; a smaller set of comments referred to linguistic style and convention. A similar approach was taken by Han (2022), who conducted post-hoc interviews to explore what judges paid attention to while conducting comparisons of recordings of spoken language interpretation. Han found that judges mentioned a similar range and type of assessment criteria to those found in existing rubrics and definitions of the same construct; the author took this as evidence of the approach's validity.

3. Running a CJ study

In this section, we describe a five-step framework for the design, administration, and analysis of a CJ study. This is based on Lesterhuis et al.'s (2017) framework for running CJ studies in educational contexts, but modifies it slightly so that (a) all issues relating to sample size are contained within the same category, and (b) it is clear that tool selection should follow this process. The framework's components will be described in turn:

1. Defining the task
2. Determining sample size: number of items, judges, and comparisons
3. Choosing a tool
4. Administering a study
5. Analysing CJ data

To make the following sections easier to follow, we offer a hypothetical example of the type of data that can be studied using CJ. The example involves using the method to generate a scale rating 100 English pseudowords in terms of their plausibility (i.e. which are most similar to real English words). This example is referred to in each section below, and is also the subject of three videos and an *R* script, available in the supplementary materials, demonstrating the process of running a CJ study.

3.1 Defining the task

The first step in running any CJ study is to create the task definition which guides judges' decisions. This is a short question or statement which tells judges how to choose the "winner" of each comparison. Task definitions must align with the construct being measured, and should also be

written in language comprehensible to judges. Most definitions also avoid naming specific aspects of the target construct, since doing so could inhibit judges' ability to use the full range of their expertise in making each comparison.

In our pseudoword example study, the goal is to generate a rank scale which orders the pseudowords in terms of their plausibly as genuine English words. One possible task definition, suitable if our judges are applied linguists, is simply "Which is the more plausible pseudoword?". However, if we have decided to use laypeople as judges, a less technical definition may be preferrable – for example, "Which of the following artificial words looks most like it could be a real English word?".

Examples of task definitions in published studies include "select which [item] look[s] more similar to an English word", in Bisson et al.'s (2022) study of the similarity of Welsh and English words; and "Choose the best translated version" in Han et al.'s (2022) study of Chinese-English and English-Chinese written translation.

Lastly, some studies supplement their task definition with further information, such as a summary of the target construct (e.g. Landrieu et al., 2022) or even a complete mark scheme (e.g. Chambers & Cunningham, 2022; Gijsen et al., 2021). Such information can help to clarify the task for judges, and may also serve to counter the criticism that without them, CJ results can be opaque, making it "challenging to explain the basis for awarding a particular mark or grade" (Kelly et al., 2022). Nonetheless, the provision of such materials remains uncommon, since they risk contradicting the principle of allowing judges to make use of their own expertise.

3.2 Sample size: number of items, comparisons, and judges

The second step in conducting a CJ study is to determine parameters relating to sample size. Researchers should begin by choosing the number of items to include in a study. Most studies have used item sets of below 100 – for example, Han and Xiao's (2022) study of Chinese Sign Language interpreting used 36 items. However, there are examples of studies using far larger item sets. The largest study included in Verhavert et al.'s meta-analysis contained 1089 items, for example, while Wheadon et al. (2020) describe the development of a rating scale containing more than 50,000 items; this was made possible by conducting individual CJ tasks in around 85 different schools, then using an anchoring procedure to link the datasets. However, researchers aiming to use CJ to rate datasets of this size should be aware that little research is currently available on CJ's reliability at this scale (see below).

The next step is to decide how many comparisons should be conducted per item. A useful way to approach this decision is to first identify a target level of reliability. In CJ, reliability increases as a function of the number of comparisons conducted per item (though see below). Further, there is a well-developed literature on how many comparisons are likely to be needed to reach a given reliability level. This means that once researchers have decided on a target reliability level, they can immediately identify an approximate range for the number of comparisons they will need.

A particularly useful resource on reliability is Verhavert et al.'s (2019) meta-analysis of 49 CJ studies from a range of subject areas, including several studies from areas of applied linguistics. Verhavert et al.'s results suggested that a reliability of .70 (measured as scale separation reliability, or SSR, which is described in Section 3.5, below) is a suitable target for "low-stakes or formative assessments" (p. 542), while .90 is a more suitable target for high-stakes tests. They also report that, based on datasets with a mean of 84 items, and a range of 6-1089, an SSR of .70 can typically be reached in 10-14 comparisons per item, .80 requires around 20 comparisons, and .90 needs 26-37. Researchers

within applied linguistics appear to generally aim for reliability levels of around .80, and typically conduct 20-30 comparisons per item to achieve this. For example, Bisson et al.'s (2022) study of orthographic and phonological similarity reported SSR = .90 and .83 from a total of 32 comparisons per item; Han and Xiao's (2022) study of Chinese Sign Language interpreting reported reliability of .86 after around 20 comparisons per item, and Thwaites et al.'s (2024) study of L2 English argumentative essay assessment reported reliability of around .82 after 26 comparisons per item. For our pseudoword study, then, we might also target SSR >= .80, and estimate 20-30 comparisons per item to achieve this.

Once a general range has been identified for the target number of comparisons per item, we must consider the difficulty of making each comparison and the level of expertise possessed by judges in order to arrive at a final number. There are no strict guidelines here: researchers will simply have to make decisions about whether to aim for the upper or lower end of their identified range. Beginning with item difficulty, if the items selected for a study differ widely in quality/proficiency, judges will find it relatively easy to make decisions and a satisfactory level of reliability will emerge from relatively few comparisons. In contrast, if all items are of similar quality, each decision becomes more difficult, judges will differ in their decisions, and high reliability will therefore require more comparisons. Similar problems also appear to occur when judges lack expertise in a target construct (e.g. Jones & Alcock, 2014; Jones & Wheadon, 2015), meaning that lay judges require a larger number of comparisons per item to reach the same level of reliability as more expert ones. For example, Thwaites (Submitted) found that judges recruited from a crowdsourcing platform (who generally lacked experience or expertise in the target construct), were able to evaluate L2 argumentative essays to a similar level of reliability (i.e. SSR = .81), and concurrent validity (i.e. correlations with pre-existing rubric-based grades of the same texts r = .68), as a group of linguists recruited through a community-driven approach (SSR = .82, r = .68), but required more comparisons to do so (28 per item, compared with 24 for the linguists). Verhavert et al.'s (2019) meta-analysis reported similar findings.

Applying these ideas to our pseudoword study, there is little available information on how difficult judges might find the task, so we should err on the side of caution and choose a final number of comparisons at the upper end of the 20-30 range previously mentioned.

To calculate the total number of comparisons required, we can multiply the number of items (e.g. our 100 pseudowords) by the number of comparisons desired per item, then divide by two (since each comparison involves two items). For the pseudoword example, this would equate to (100 items * 30 comparisons) / 2 = total 1500 comparisons.

The final sampling consideration is to decide on judge numbers and demographics. Researchers again have a good deal of freedom here, but should consider four main points:

- Having a large number of judges helps to provide the diversity of experience and expertise required to ensure that the final rating scale reflects broad coverage of the target construct (Bisson et al., 2016);
- However, larger judging groups means fewer comparisons per judge. Few studies have been conducted on the minimum number of comparisons each judge should make, but (2019) reports that larger numbers of comparisons per judge can increase the chances that a scale's reliability will reach asymptote (i.e. the point at which an increase in reliability would require a lot more comparisons);

- The length and complexity of each comparison should be considered alongside the number of comparisons each judge is asked to complete. Longer or more complex items take more time to assess and are likely to lead to greater fatigue, which may reduce reliability.
- Expert judges are likely to yield high reliability more quickly than less expert ones, but in some contexts there may be benefits to gathering lay judgements of constructs.

Again applying these guidelines to the pseudoword study, either lay or expert judges might be sought, depending on whether the researcher desires a lay or expert definition of the construct of pseudoword plausibility. Whatever the decision, a reasonable approach to judge numbers might be to split the 1500 comparisons between 25 individuals, leaving each judge 60 comparisons. This allows a relatively large judging group to contribute to the underlying construct representation, while ensuring that each judge conducts enough comparisons to be reliable. It also takes into consideration the fact that each comparison is very brief, requiring only two pseudowords to be compared.

3.3 Selecting a tool

Numerous CJ platforms are currently available, all of which offer the basic functionality required to run CJ studies. However, each platform differs in their specifics. The most popular platform among researchers is *No More Marking* (*NMM*). This is a proprietary tool, like *Comproved* and *RM Compare.* An open-source alternative is *ComPAIR*. Web links, example studies, and key comparisons for each of these platforms are presented in Appendix 1.

One important area of variance between tools is the algorithm used to determine which items are paired together in each trial. There are two types: adaptive and pseudo-random. Adaptive algorithms, described by Pollitt (2012), first gather information on each item during initial rounds of comparisons, then use that information to pair items of similar quality in later rounds. This saves time by eliminating trials in which the outcome is too predictable (i.e. those comparing very strong to very weak items), cutting the time required for reliable scales to emerge. Unfortunately, several studies have suggested that adaptive algorithms are susceptible to inflation of reliability levels (Bramley, 2015; Bramley & Vitello, 2019; Crompvoets et al., 2022). For this reason, adaptive algorithms are only recommended in contexts where reliability does not need to be measured. One example is CJ's application to peer evaluation, in which learners are asked to conduct comparisons between their own productions and those of their peers. Since here the focus is on students' learning outcomes rather than the resulting rating scale, adaptive tools may be appropriate.

For most research purposes, however, pseudo-random algorithms are to be preferred (Bramley & Vitello, 2019). These algorithms work by first selecting the item with the fewest total comparisons (to ensures that each item receives an approximately equal number of comparisons), then choosing a partner for that item randomly. Pseudo-random algorithms therefore do not eliminate trials with very predictable outcomes, and for this reason require more comparisons than adaptive algorithms to reach high levels of reliability. However, the information provided by these predictable trials seems to contribute to the stability of the scale as a whole, leading to pseudo-random CJ's more trustworthy statistical properties.

There are numerous other sources of variance, relating to each platform's flexibility (i.e. whether it allows studies to be modified after comparisons have commenced), their costs, their need for specialist skills or web hosting requirements, and the range of item formats they can handle. The table in Appendix 1 summarises each platform's approach to these issues.

For our pseudoword study, we might select *No More Marking*, because:

- It offers a pseudo-random algorithm, which is preferred to an adaptive option because the reliability of the data is critical to the reporting of the study;
- Though NMM only supports items in pdf and mp3 formats, pdf is sufficient for displaying pseudowords;
- Inviting judges is very simple on NMM - it requires only a link to be shared;
- NMM is free for researchers to use, lowering study costs;
- NMM's backend system allows easy monitoring of study progress (important during the following stage).

The supplementary videos illustrate study setup on this platform.

3.4 Running a CJ study

After setting a CJ study up, judges can be invited and will subsequently begin making comparisons. As they do so, the rating scale will begin to emerge. All platforms allow in-progress monitoring of this scale, though some make it easier than others (see "In-progress data monitoring" in Appendix 1). At this stage, the researcher's main task is to monitor the incoming data to ensure that judges are able to complete the task as requested, and that the target reliability can be achieved.

Researchers will need to pay particular attention to two concerns. Firstly, some judges may not have begun or finished their tasks (visible via counts of completed comparisons), while others might not be performing the task as requested. Signs of this latter problem might include a high judge infit score (see Section 3.5, below, for more on fit statistics) or very fast or slow decision times. Researchers should consider contacting judges whose data seems unusual. Second, researchers can also use in-progress data to consider changes to study parameters. For example, they may find that their target reliability level has been reached earlier than expected, and thus decide to stop collecting comparisons; alternatively, they might find that reliability is lower than anticipated, and on this basis assign more comparisons to each judge. Researchers can (if their platform permits – see "Flexibility" in Appendix 1) also choose to expand their study by adding more items, or broaden the judging base by adding new judges.

3.5 Analysing CJ data

CJ data is most commonly analysed using the *sirt* package (Robitzsch, 2022) of *R* (R Core Team, 2023). The supplementary materials contain a link to a video and an example *R* script showing the analysis of a simulated version of our pseudoword study using these tools.

CJ data should be analysed in three steps: production of an initial model, checking and correcting, then analysis of the final model. All of these steps require raw data to be downloaded from the chosen CJ platform. The data can take various formats, two of which are shown in Figure 1. In each example, rows correspond to a single comparison while columns specify (at minimum) the identity of the two items compared, the identity of the judge making the comparison, and the winner of the comparison. Platforms may also provide additional data, such as the time taken to make the decision ("timeTaken", Figure 1 left).

judge_id	timeTaken	chosen	notChosen
judge_017	190071	H7R5F9	TQR299
judge_017	228822	3W66KC	FNFJ7A
judge_017	216370	VDQXPR	7ME6E7
judge_032	56554	NYYCVV	FDMQXC
judge_032	23319	2NH6JQ	YUWZCX
judge_032	20567	Q5QMC2	FDMQXC
judge_032	11717	KH2WZ5	HWEQTL
judge_032	55019	MYNGCC	CRF6HZ
judge_032	51211	MNZA7Y	3W66KC
judge_032	31860	JU7ZLF	FQVYRQ

	id1	id2	result	judge
	<int>	<int>	<dbl>	<int>
1	11	83	1	1
2	81	96	1	1
3	26	28	1	1
4	17	86	1	1
5	72	95	1	1
6	28	53	1	1
7	3	76	0	1
8	58	61	0	1
9	44	69	1	1
10	58	76	0	1

Figure 1: Two formats for raw CJ data, from No More Marking *(left) and generated by the* btm_sim() function of R*'s* sirt *package (right). timeTaken indicates decision time in milliseconds.*

From this data, *sirt* generates a model containing various values. The core of the model is a rating scale which contains a scale score (denoted by *sirt* as "theta") for each item, indicating its quality relative to the rest of the dataset. The scale has a mean value of 0 (meaning that an item of absolutely average quality would also have a score of 0). Item values typically fall between around 6 (for the strongest items) and -6 (for the weakest). This value is also the basis of each item's rank.

This initial model must be cleaned before it can be considered final. This involves removing judges whose data reveals signs of inattentive or unprincipled decision-making. Several measurements can be used to detect this. One is judge infit, which shows the extent to which each judge's decisions deviated from what would be expected given decisions made by all others. Judges are considered to misfit the model if their infit is two or more standard deviations above the mean (Pollitt, 2012). However, researchers should treat misfit with caution because it does not necessarily imply low-quality decision making. High infit values can occur, for example, because a judge took a different perspective on some texts to that of other judges. Arguably, judges should not be removed on this basis only, since (as described above) CJ benefits from diverse judge perspectives.

Therefore, further steps are required to identify aberrant judge behaviour. Two approaches have been suggested in the literature. The first involves using two other measures – judges' median time per decision and their proportion of left (or right) clicks – to further investigate misfitting judges. All major CJ platforms automatically record decision times, while left click percentages can be manually calculated from data like that on the right side of Figure 1 if a platform does not provide them automatically. Very short decision times (relative to other judges) or very strong tendencies to select items from only one side of the screen suggest that judges may not have taken due care in their decisions. As with infit, neither of these measures alone should be taken as definitive evidence of inattentive judging, but taken together may represent sufficient grounds for judge removal. This

should not be done *ad hoc*; researchers should define conditions for judge removal based on a combination of these three measurements prior to study commencement. For example, Thwaites et al. (2024) developed a system in which judges who triggered two of three red flags – statistical misfit, median decision time of less than 5 seconds per item, and left click percentage of < 11% or < 89% – would be removed.

An alternative approach, suggested by Jones and Davies (2023), is to test the effect of misfitting judges on the overall model, removing them if they had a large impact on scale values. This can be done by first removing misfitting judges, then building a new version of the model and calculating correlations between the scale scores for the two models. A low correlation would indicate that the misfitting judge(s) significantly affected the scale, and therefore would justify their removal (although further research is needed into exactly what level of correlation justifies judge removal).

After investigating judge fit in one or both of these ways, researchers can either retain the original model (if no judges were removed) or proceed with the model without misfitting judges. They can then explore the scale's reliability. The most frequently reported measurement in CJ studies is scale separation reliability (SSR), which is considered analogous to Cronbach's alpha (Verhavert et al., 2018). This is automatically calculated by the *sirt* package; the *R* code in the supplementary materials shows how to extract it. An alternative approach to measuring CJ's reliability is the split-halves approach (Bisson et al., 2016; Jones & Davies, 2023). This involves randomly assigning judges to one of two groups, creating a rating scale for each, and then calculating correlations between the scale scores. The procedure is then re-run around 100 times, with the mean correlation serving as the reliability measure. This approach is somewhat more transparent than SSR, but has the disadvantage of requiring double the number of comparisons to what would normally be expected.

4. Comparison with other methods

The key attribute of CJ is its *comparative* nature: it differs from other measurement tools in that it presents judges with items side-by-side, providing a clear context for each decision. The natural alternative is to use *absolute* forms of assessment, which involve judgement of items in isolation. These include rubric-based marking (for proficiency assessment) and Likert scales (to measure linguistic constructs like acceptability).

Section 2.1 compared CJ and rubric-based assessment, arguing that CJ does not suffer from problems like rater bias or differences in rater severity. In defence of a rubric-based approach, some of these shortcomings can be statistically controlled using many-faceted Rasch modelling (McNamara et al., 2019). Rasch analysis cannot, however, avoid the financial and temporal costs of developing rubrics and training raters. Nor can it overcome the challenge of developing rubrics that are both broad enough to provide comprehensive coverage of the target construct while also being specific enough to minimize rater variation (which would result in low inter-rater reliability). The group-based nature of CJ avoids this problem completely by allowing the construct to emerge from the collected expertise of the judges.

CJ has also been used to measure hard-to-define linguistic constructs such as similarity (Bisson, 2022), diversity (Crossley et al., 2023), and acceptability (Stadthagen-González et al., 2019). The latter study used a specific type of CJ called the two-alternative forced choice task (2AFC), which is identical to CJ except that it specifies that each judge should compare all possible item pairings within a given (sub)set, to study the acceptability of various types of codeswitching. The authors reported that this method was an effective way of testing explicit hypotheses regarding the types of codeswitch considered most and least acceptable. The thoroughness of its design makes 2AFC best

suited to contexts in which small numbers of simple items, such as single words or sentences, need to be compared to each other.

Absolute approaches to measuring a construct such as the acceptability of codeswitches would typically use tasks such as Likert scales (Schütze & Sprouse, 2013), and magnitude estimation (Bard et al., 1996), each of which involves showing participants individual items and asking them to provide a binary or numerical judgement (for more details on collecting applied linguistic data using these methods, see Spinner & Gass, 2019). While few studies have directly compared CJ to these approaches, Sprouse and Almeida (2017) found 2AFC to be a more sensitive task, capable of higher rates of statistical detection of "theoretically interesting contrasts between different sentence types" (p. 1) than any of the three absolute alternatives. One reason for this may be that comparison tasks are, as Thurstone's law of comparative judgement task suggests, easier and more transparent than those involving absolute judgement.

A final CJ-adjacent research method is item ranking. Here, judges are presented with sets of three or more items (not pairs, as in CJ), and asked to place them in order of quality. In theory, this allows items to be judged more efficiently: ranking items A, B, and C together should be quicker than completing three paired comparisons A-B, A-C, and B-C. However, little research to date has systematically investigated the efficiency, reliability, or validity of the ranking method; the few existing studies have differed in the number of items to be ranked in each trial (for example, Attali et al. (2014) asked judges to rank sets of five written tasks, while Bramley and Black (2008) have used sets of 10), making findings hard to interpret.

5. Benefits and constraints

CJ's principal strengths are its efficiency and reliability. These are well described in a recent study by Pinot de Moira et al. (2022) which compared CJ's classification accuracy, reliability, and efficiency with rubric-based grades of increasing sophistication – the simplest given by a single rater, and the most sophisticated being the aggregate of grades produced by four raters. They found that CJ was as fast as rubric-based grading using two raters, while its reliability and classification accuracy (defined as agreement with "definitive" grades provided by a senior moderator) were equal to triple rubric-based grading. This efficiency-to-reliability payoff is particularly impressive given that CJ also provides efficiency *prior* to data collection: it does not require materials such as rubrics or survey instruments to be developed, nor participants to be trained in their use. CJ can therefore be considered a useful tool for any applied linguistic contexts in which proficiency data needs to be collected; a specific context is described in Section 6, below.

Another advantage is CJ's compatibility with crowdsourcing approaches to data collection. This is illustrated in Crossley et al. (2023), who used Amazon Mechanical Turk to recruit participants for a lexical diversity judgement task, and in Paquot et al. (2022; Thwaites et al., Submitted), who used both community-driven and traditional crowdsourcing approaches to generate assessments of essays from learner corpora. Combining CJ with crowdsourcing in this way can facilitate the collection of data from diverse demographics and allow communities lacking access to expertise to recruit qualified participants.

CJ is also valued for its ability to compare items with different characteristics. For example, Jones et al. (2016) used CJ to explore the standards of mathematical ability required by various historical math exams. To do this, judges were asked to compare learner performances on tests of mathematics produced up to 50 years apart. In addition to their findings showing that required standards dropped slightly between the 1960's and the 1990's, the study also demonstrated CJ's robustness for comparing heterogeneous tests of the same underlying construct. A similar

robustness to heterogeneous items has also been demonstrated in an applied linguistic context: Thwaites et al. (2024) showed that CJ tasks featuring responses to multiple essay prompts were assessed just as quickly and reliably as those responding to only one prompt.

In other areas, CJ's advantages must be considered alongside concomitant drawbacks. One example is the scale score given to each item in a CJ task. These scores have the benefit of being much more precise than categorical grades, such as the IELTS bands or CEFR levels which result from absolute assessment. They also have a straightforward interpretation: they reflect the likelihood of any one item being considered "better" than any other. One the other hand, CJ scores are essentially norm-referenced (i.e. interpretable only in relation to items included in the same task) without additional steps being taken to align them with an external standard. Two potential solutions to this problem are to include previously graded items representing grade boundaries among a CJ set, allowing CJ scores to be anchored to these items (Marshall et al., 2020); and using a standard-setting exercise in which experts determine cut-off points for each level of the target standard (Fleckenstein et al., 2020).

Another area of uncertainty is CJ's validity. As discussed in Section 2.2, most CJ studies report strong concurrent validity, while dedicated studies have provided support for CJ's construct validity. Nevertheless, criticisms remain. Kelly et al. (2022) provide a helpful discussion of these. They argue that current uses of CJ diverge significantly from those described in Thurstone's early research, in which the method was intended for use "only with stimuli whose values could be evaluated relatively instantaneously" (Kelly et al., 2022, p. 5). This raises questions as to whether CJ is suitable for assessing the more complex items used in recent AL research, such as essays or video recordings. Secondly, they note CJ's perceived opacity – i.e. the difficulty of knowing what judges consider while making decisions. They suggest that variance in the expertise of judges used in CJ studies adds to this opacity, since in some studies judges lack expertise in the target construct. They highlight the need for "a clear theoretical framework linking the choice of [judge] expertise with the needs of the comparative judgment process" (Kelly et al., 2022, p. 8), and advocate further research into the variables affecting the construct validity of CJ scales. Jones and Inglis (2023) respond to these criticisms by citing substantial evidence of CJ's validity. Nevertheless, Kelly et al.'s work reflects a reluctance to accept that evidence of a method's validity in one field of research can be seen as sufficient evidence of its validity for another. With this in mind, it is likely that CJ's acceptance in the field of applied linguistics will require future studies providing evidence of CJ's validity for each specific application.

6. Potential applications

CJ is a method in development. Though increasingly popular in educational assessment, researchers are only now beginning to explore its full potential. Here, we highlight two applications which applied linguists have begun to explore: the method's use as an off-the-shelf approach to testing linguistic proficiency in research contexts, and its potential for generating qualitative data which can contribute to discussions of linguistic construct definition and measurement.

Beginning with CJ's use for assessing language proficiency in research studies, a recent study by Park et al. (2022) called for the development of reliable, efficient measurement tools which researchers can use to assess the proficiency of texts or participants used in their studies. The need for such tools is due to the high prevalence of unsatisfactory proficiency measurement and reporting practices in applied linguistics. Park et al. surveyed five key SLA journals and reported that the majority of studies used indirect measures like institutional status to assess the proficiency of study participants. These measures have long been considered unreliable (e.g. Thomas, 1994). CJ is increasingly used for this

purpose. For example, Wengelin et al. (2024) used the method to assess the proficiency of a set of written texts used in a study exploring the impact of spelling difficulties on overall writing quality, while Paquot et al. (Paquot et al., 2022; Thwaites et al., In press) use crowdsourced and community-driven CJ to generate assessments of texts in learner corpora – a context similarly suffering from insufficient proficiency reporting – thereby facilitating research which uses learner corpus data to explore proficiency-related variation.

The use of CJ to contribute to linguistic construct definition arises from its use as a measurement tool. Several examples of this latter usage have already been given above (e.g. Bisson, 2022; Sprouse & Almeida, 2017; Stadthagen-González et al., 2019). Such approaches can be expanded to facilitate exploration of construct definition by requesting judges to leave comments explaining the decisions they make during their comparisons. This qualitative data can then be coded and analysed to explore what judges consider to make one item a better example of a given construct than another. This is the approach taken by numerous studies exploring CJ's validity for educational assessment (Lesterhuis et al., 2018, 2022), but the method could equally well be used to explore linguistic constructs. For example, the construct of linguistic complexity could be explored by asking judges to leave comments on why they considered one text to be more or less complex than another. All major CJ platforms allow judges to leave comments, making such experiments quite simple to conduct (see "Judge commenting" in Appendix 1 for information on how platforms differ in their implementation of this function).

Another way in which CJ might be used to explore construct definition would be to use multiple regression techniques to explore relationships between a CJ-derived scale containing items which represent a given construct, and the various theoretical components of that construct. A cluster of studies which uses CJ in this way can be found in the area of readability research. Typically, such studies first use CJ to generate measurements of the difficulty or readability of a set of texts (for example using crowdsourcing platforms (Crossley et al., 2017, 2019) or by targeting a specific demographic such as Chinese learners of English (Zhang & Lu, 2024)); and then entering the resulting scale as the dependent variable in a regression analysis in which various measurements assumed to contribute to readability, such as lexical richness or syntactic complexity, are used as explanatory variables.

Both of these approaches to construct definition can easily be set up to offer comparison of how different demographic groups conceptualise a construct. For example, in an ongoing L2 writing proficiency study by Thwaites et al. (in preparation), comparative judgements and accompanying comments on a set of L2 essays were collected from three distinct judging groups – laypeople recruited through crowdsourcing, linguists recruited through community-driven methods, and trained writing assessors. By coding and analysing the resulting comments, it is possible to explore how each of these judging groups conceptualised the construct of L2 writing proficiency. A similar approach could be used using regression-based methods, and could be applied to the study of many other linguistic constructs. For example, CJ could contribute to the study of perceived fluency (Suzuki et al., 2021) by exploring how participants with various L1 backgrounds differ in their perception of the fluency of language learners. The CJ-adjacent method of 2AFC could also be used in this way, to explore whether differences in judge profiles influence the outcomes of studies in which competing hypotheses (such as those pertaining to the types of codeswitch perceived to be most and least acceptable; Stadthagen-González et al., 2019) are tested through comparisons of linguistic items.

A study by Morton (2022) provides an interesting link between the use of CJ for construct definition and another potential usage, as a teacher training tool. Morton asked seven CLIL teachers – four with a background in content teaching and the other whose background was language teaching – to use

CJ to evaluate learner performances on a writing task. He then held group feedback sessions in which the teachers were encouraged to reflect on the resulting rank scale. This facilitated discussion of how the two groups of teachers differed in their conceptualisation of the quality of these texts. Morton presents evidence that these discussions "may have been a catalyst for building new understandings of the content-language relationship" (p18). The study therefore integrates CJ's applications to educational assessment and construct definition while also serving to create learning opportunities for teachers in training.

7. CJ potentially has many other applications. These are still being explored by researchers. For example, Bouwer et al. (2018) investigated CJ's potential as a peer assessment and feedback tool by asking a group of learners to use CJ to provide feedback on L1 texts written by their peers, and comparing the resulting comments with those generated by another group of learner using an analytic list of assessment criteria. The results of the study suggested that CJ encouraged learners to pay attention to "higher order" aspects of their peers' writing, such as their content and structure, while the learners using the analytic scale paid increased attention to lower-level aspects such as grammatical control and vocabulary use. While the authors noted that both types of feedback were useful, they suggested that "feedback on higher level aspects is generally associated with improved writing performance" (p.8). The authors therefore concluded by suggesting that the CJ could serve as a "powerful instructional tool" (p. 9). Moreover, CJ is still being employed to measure psychological constructs such as motivation or anxiety, providing an alternative to traditional Liker-type rating scales in the development of personality tests (e.g. Bürkner, 2022; Merk et al., 2017).Conclusion

Comparative judgement offers enormous potential to support applied linguistic research, particularly as an accurate, reliable, and highly efficient alternative to absolute approaches to educational assessment and the measurement and definition of complex, multidimensional constructs. This tutorial has sought to explain how researchers can begin to use CJ in their own work, as well as explaining some advantages and anticipating some of the challenges that they might encounter.

Although CJ has already begun to be adopted by applied linguists (as the studies cited in this article show), much work remains to be done to fully explore and elaborate its various potential applications. In particular, while an increasing number of studies use the method to measure and explore various linguistic constructs, there remains too little research on how CJ compares to alternative (i.e. absolute) approaches to the same tasks. More research is also needed in the field of language assessment, where questions remain regarding CJ's validity for proficiency assessment, and more needs to be done to identify methods for aligning CJ scores to L2 proficiency scales such as the CEFR. Nevertheless, we hope that this study will persuade applied linguists of the potential of this approach to data collection and measurement, and of the benefits of working towards these future research goals.

Declaration of Interest

This research was supported by grant number 40008459 awarded by the Le Fonds de la Recherche Scientifique (FNRS) to the corresponding author.

Supplementary materials

A short video demonstrating how to set up a CJ study on NoMoreMarking.com is available at: https://youtu.be/5coqml1fD_Q

A longer video with a fuller elaboration of study setup is here: https://youtu.be/JkfsiR_jrSc

A video explaining how to analyse CJ data in R is here: https://youtu.be/q-Cuk9gTnFU

An OSF page containing the same video, plus an R script and supporting documents for running CJ analyses is here: https://osf.io/mvk4j/files/osfstorage

References

Andrich, D. (1978). Relationships Between the Thurstone and Rasch Approaches to Item Scaling. *Applied Psychological Measurement*, *2*(3), 451–462. https://doi.org/10.1177/014662167800200319

Attali, Y. (2014). A Ranking Method for Evaluating Constructed Responses. *Educational and Psychological Measurement*, *74*(5), 795–808.

Badham, L., & Furlong, A. (2023). Summative assessments in a multilingual context: What comparative judgment reveals about comparability across different languages in Literature. *International Journal of Testing*, *23*(2), 111–134. https://doi.org/10.1080/15305058.2022.2149536

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude Estimation of Linguistic Acceptability. *Language*, *72*(1), 32–68. https://doi.org/10.2307/416793

Bartholomew, S. R., Zhang, L., Garcia Bravo, E., & Strimel, G. J. (2019). A Tool for Formative Assessment and Learning in a Graphics Design Course: Adaptive Comparative Judgement. *The Design Journal*, *22*(1), 73–95. https://doi.org/10.1080/14606925.2018.1560876

Bisson, M.-J. (2022). Learning words with unfamiliar orthography: The role of cognitive abilities. *Studies in Second Language Acquisition*, *45*(4), 838–852. https://doi.org/10.1017/S0272263122000390

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

Bouwer, R., Lesterhuis, M., Bonne, P., & De Maeyer, S. (2018). Applying Criteria to Examples or Learning by Comparison: Effects on Students' Evaluative Judgment and Performance in Writing. *Frontiers in Education*, *3*. https://www.frontiersin.org/articles/10.3389/feduc.2018.00086

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. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/487059/2007-comparability-exam-standards-i-chapter7.pdf

Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgment* [Cambridge Assessment Research Report]. Cambridge Assessment.

Bramley, T., & Black, B. (2008). *Maintaining performance standards: Aligning raw score scales on different tests via a latent trait created by rank-ordering examinees' work* [Cambridge Assessment Research Report]. Cambridge Assessment.

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.

Bürkner, PC. (2022). On the information obtainable from comparative judgments. *Psychometrika,* 87, 1439–1472. https://doi.org/10.1007/s11336-022-09843-z

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

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

Crossley, S. A., Cushing, S., Jarvis, S., & Kyle, K. (2023). Crowd-Sourcing Human Ratings of Linguistic Production. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*, 1515–1520. https://escholarship.org/uc/item/2zh6n03c

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, *42*(3–4), 541–561. https://doi.org/10.1111/1467-9817.12283

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas. *Discourse Processes*, *54*(5–6), 340–359. https://doi.org/10.1080/0163853X.2017.1296264

Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R. J., & Köller, O. (2020). Linking TOEFL iBT® writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assessing Writing*, *43*, 100420. https://doi.org/10.1016/j.asw.2019.100420

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

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

Han, C., & Xiao, X. (2022). A comparative judgment approach to assessing Chinese Sign Language interpreting. *Language Testing*, *39*(2), 289–312.

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, *34*(4), 537–553.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, *39*(10), 1774–1787. https://doi.org/10.1080/03075079.2013.821974

Jones, I., & Davies, B. (2023). Comparative judgement in education research. *International Journal of Research & Method in Education*, *47*(2), 170–181. https://doi.org/10.1080/1743727X.2023.2242273

Jones, I., & Inglis, M. (2023). The validity of comparative judgement: A comment on Kelly, Richardson and Isaacs. *Centre for Mathematical Cognition*. https://blog.lboro.ac.uk/cmc/2023/05/05/the-validity-of-comparative-judgement-a-comment-on-kelly-richardson-and-isaacs/

Jones, I., & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation*, *47*, 93–101. https://doi.org/10.1016/j.stueduc.2015.09.004

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*, *29*(6), 674–688. https://doi.org/10.1080/0969594X.2022.2147901

Kormos, J., & Dénes, Mariann. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145–164. https://doi.org/10.1016/j.system.2004.01.001

Landrieu, Y., De Smedt, F., Van Keer, H., & De Wever, B. (2022). Assessing the quality of argumentative texts: Examining the general agreement between different rating procedures and exploring inferences of (dis)agreement cases. *Frontiers in Education*, *7*, 1–16. https://doi.org/10.3389/feduc.2022.784261

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

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2017). Comparative judgement as a promising alternative to score competences. In E. Cano & G. Ion (Eds.), *Innovative Practices for Higher Education Assessment and Measurement* (pp. 119–138). IGI Global. https://doi.org/10.4018/978-1-5225-0531-0.ch007

Marshall, N., Shaw, K., Hunter, J., & Jones, I. (2020). Assessment by Comparative Judgement: An Application to Secondary Statistics and English in New Zealand. *New Zealand Journal of Educational Studies*, *55*(1), 49–71. https://doi.org/10.1007/s40841-020-00163-3

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

Merk, J., Schlotz, W., & Falter, T. (2017). The Motivational Value Systems Questionnaire (MVSQ) : Psychometric analysis using a forced choice Thurstonian IRT model. *Frontiers in Psychology*, 8. https://doi.org/10.3389/fpsyg.2017.01626

Morton, T. (2022). Using cognitive discourse functions and comparative judgement to build teachers' knowledge of content and language integration for assessment in a bilingual education program. *Journal of Immersion and Content-Based Language Education*, *10*(2), 302–322. https://doi.org/10.1075/jicb.21017.mor

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

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

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

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), 89–113. https://doi.org/10.20343/teachlearninqu.5.2.8

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the Validity of Quantitative Measures in Applied Linguistics Research1. *Language Learning*, *65*(S1), 37–75. https://doi.org/10.1111/lang.12112

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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*, 378–388.

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

Schütze, C. T., & Sprouse, J. (2013). Judgment data. In R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 27–50). https://books.google.com/books?hl=en&lr=&id=gU5kAgAAQBAJ&oi=fnd&pg=PA27&dq=Sch

%C3%BCtze,+C.+T.,+%26+Sprouse,+J.+(2013).+Judgment+Data.&ots=LT2n8zg7Mb&sig=TBE8X6R73VMrq_KfxvfO7DQ0GpE

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

Spinner, P., & Gass, S. M. (2019). *Using judgments in second language acquisition research*. Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9781315463377/using-judgments-second-language-acquisition-research-patti-spinner-susan-gass

Sprouse, J., & Almeida, D. (2017). Design sensitivity and statistical power in acceptability judgment experiments. *Glossa: A Journal of General Linguistics*, *2*(1), 14.

Stadthagen-González, H., Parafita Couto, M. C., Párraga, C. A., & Damian, M. F. (2019). Testing alternative theoretical accounts of code-switching: Insights from comparative judgments of adjective–noun order. *International Journal of Bilingualism*, *23*(1), 200–220. https://doi.org/10.1177/1367006917728390

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

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The Relationship Between Utterance and Perceived Fluency: A Meta-Analysis of Correlational Studies. *The Modern Language Journal*, *105*(2), 435–463. https://doi.org/10.1111/modl.12706

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44*(2), 307–336.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286. https://doi.org/10.1037/h0070288

Thurstone, L. L. (1954). The measurement of values. *Psychological Review*. 61(1):47-58.

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. https://doi.org/10.1016/j.asw.2024.100843

Thwaites, P., Kollias, C., & Paquot, M. (Submitted). *Testing crowdsourcing as a means of recruitment for the comparative judgement of L2 argumentative essays*.

Thwaites, P., Vandeweerd, N., & Paquot, M. (In press). Crowdsourced comparative judgement for evaluating learner texts: How reliable are judges recruited from an online crowdsourcing platform? *Applied Linguistics*.

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

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. https://books.google.com/books?hl=en&lr=&id=b6x_vQQR_-8C&oi=fnd&pg=PR10&dq=weigle+2002+assessing+writing&ots=YecaVlJQt7&sig=xAwXGcZoOtwecfZsfcfFQN2Wchg

Wengelin, Å., Kraft, S., Thurfjell, F., & Rack, J. (2024). What can writing-process data add to the assessment of spelling difficulties? *Reading and Writing*, *37*, 1635–1658. https://doi.org/10.1007/s11145-024-10524-9

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

Zhang, X., & Lu, X. (2024). Testing the Relationship of Linguistic Complexity to Second Language Learners' Comparative Judgment on Text Difficulty. *Language Learning*, 1–35. https://doi.org/10.1111/lang.12633

Appendix 1: A comparison chart of four CJ tools

	ComPAIR	Comproved	No More Marking	RM Compare
License	Open source - free, but requires own server.	Proprietary. Long-term and single-use licenses available. Research mode currently in development (also requiring license).	Proprietary, but free for researchers.	Proprietary - requires license. Various options available, including for researchers. Limited trial version also available.
Web address	Homepage: https://compair.open.ubc.ca/; Source code: https://github.com/ubc/compair	www.comproved.com	www.nomoremarking.com	Compare.rm.com
AL example studies	Potter et al. (2016) Paquot et al. (2022)	An earlier version of the tool, named D-PAC was used in: Badham & Furlong (2023) Lesterhuis et al. (2022)	Han & Xiao (2022) Bisson et al. (2022) Sims et al. (2020)	None, but see Bartholemew et al. (2019) for an example from the field of design assessment.
Documentation	Available via the project homepage; see also Potter et al. (2016). But note that little support for server-side installation and database management is available; expertise is required.	Downloadable manual available from https://comproved.com/en/getting-started-with-comproved/	Detailed information for researchers, including R code for processing results, at https://nmm.notion.site/No-More-Marking-for-researchers-70cb4eec46d547cd91c65ff2066d415f	Online help page at https://compare.rm.com/help-centre/
Support	compair.support@ubc.ca	Via contact form at https://comproved.com/en/contact/	Via chat function at www.nomoremarking.com	Via contact form at https://compare.rm.com/get-in-touch/
Additional requirements	Server space, server administration, database management.	None	None	None
Adding judges	Judge details uploaded via backend system.	Judge details uploaded/added via backend system. A second, anonymous option is in development, allowing judges to join via a link.	Judges join via a link. They subsequently receive an email allowing them to return later.	Judge details uploaded/added via backend system.

Adding learners (for peer feedback studies)	Learner details uploaded by researcher via backend system.	Learner details uploaded by researcher via backend system.	Learner details uploaded by researcher via backend system.	Learner details uploaded by researcher via backend system.
Adding items (learners upload own item)	Learners may upload a file, or type text into the interface.	Learners upload items into the interface.	Learners can be sent a link to type a text, or researchers can upload items.	Student judges upload items into the interface.
Adding items (researcher uploads items)	Items must be uploaded to database manually; expertise required.	Researcher uploads to backend.	Researcher uploads to backend.	Researcher uploads to backend.
Judge commenting	Possible, via a versatile text editor. Removing this comment box requires manipulation of source code.	Possible: comment boxes can be set to appear after each comparison in the task settings menu. The comment interface provides separate boxes for positive and negative comments about each item.	Possible, by clicking the task description while making comparisons. This location is not obvious and should be pointed out to judges if required. Cannot be turned on or off.	Possible: a simple comment box can be set to appear after each comparison.
File types supported	.pdf, .txt	.mp3, .mp4, .jpg, .jpeg, .png, .gif, .pdf; URL links	.pdf, mp3	.jpg, .png, .gif, .tif, .targa, .bmp, pdf; .mp4, .mp3, .avi, .mpg; Office documents (e.g. .docx, .xlsx, .ppt); URL links
Algorithms	Random & adaptive algorithms available. Adaptive algorithm uses Elo pair selection; likely vulnerable to reliability inflation.	Pseudo-random	Pseudo-random	Adaptive; algorithm developed in response to reliability inflation concerns (Rangel-Smith & Lynch, 2018), but independent verification needed.
Text capacity	Unlimited	Unlimited	Unlimited	Unlimited, but item volume contributes to fees
Flexibility	Both new judges and new items can be added after experiment begins, but the handling of the latter is uncertain, particularly if using an adaptive algorithm.	Judges can be added after experiment begins; new items cannot.	Both new judges and new items can be added after experiment begins. The algorithm will allow new texts to catch up with old ones.	Neither new judges nor new items can be added after comparisons have begun.

In-progress data monitoring	Only possible by periodically downloaded and analysing data.	Available	Available	Available
Branding (adding institutional logos etc.)	Can be added but require coding proficiency.	Can be added with institutional licenses.	Cannot be added.	Cannot be added.
Additional remarks	Substantial customisation is possible but requires coding proficiency (primarily in Python).	Forthcoming research-focused features promise to integrate survey tools to streamline demographic data collection, allow comparison of items on multiple criteria, and offer several other new features.	Currently the most popular tool among researchers.	A custom comparison mode allows micro-management of comparisons.