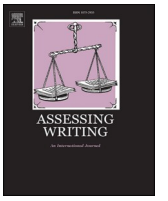
Assessing Writing 66 (2025) 100986

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw



Exploring the scoring validity of holistic and dimension-based Comparative Judgements of young learners' EFL writing

Rebecca Sickinger [1], John Pill [*,2], Tineke Brunfaut [3]

Lancaster University, Linguistics and English Language, Lancaster LA1 4YL, United Kingdom

ARTICLE INFO

Keywords:
Comparative Judgement
Construct relevance
Dimension-based judging
Holistic judging
Pairwise comparison
Scoring validity
Testing young learners

ABSTRACT

Comparative Judgement (CJ) is a pairwise comparison evaluation method, typically conducted online. Multiple judges each compare the quality of a series of paired performances and, from their decisions, a rank order is constructed and scores calculated. Research across different educational contexts supports CJ's reliability for evaluating written performances, permitting more precise scoring of scripts and for dimension-focused evaluation. However, scant insights are available about the basis of judges' evaluations. This issue is important because argument-based approaches to validation (common in the field of language testing and adopted in this study) require evidence to support claims about how scores are appropriate for test purpose. Therefore, we investigate the scoring validity of CJ, both when used holistically (the standard application of CJ) and when evaluating scripts by individual criteria (termed dimensions in the research context). Twenty-seven judges evaluated 300 scripts addressing two writing task types in a national English as a Foreign Language examination for young learners in Austria. Judges reported via questionnaires what they had focused on while judging. Subsequently, eight judges provided think-aloud data while evaluating 157 scripts, offering further insight into the writing features they considered and their decision-making during CJ. Findings showed that while most judges adapted a decision-making process similar to traditional rating methods, some adapted their method to accommodate the nature of CJ evaluation. Furthermore, results indicated that the judges considered construct-relevant criteria when using CJ, both holistically and by dimension, thus offering support to an argument for the appropriateness of using CJ in this context.

1. Introduction

The assessment of second language (L2) writing, the focus of this article, has conventionally involved solo raters evaluating written performances after extensive training (Shaw & Weir, 2007). Comparative Judgement (CJ) is an innovative evaluation method in which multiple judges, potentially without specific training, work individually to assess performances in a series of paired comparisons (a CJ session). It is often conducted remotely online using specialised software; judges see two performances side by side onscreen with a question that directs them to decide which is the "better" performance (see, e.g., Sims et al., 2020). Judges have been found capable of

* Corresponding author.
E-mail addresses: englishemailrs@gmail.com (R. Sickinger), j.pill@lancaster.ac.uk (J. Pill), t.brunfaut@lancaster.ac.uk (T. Brunfaut).
[1] ORCID: https://orcid.org/0009–0001-7160–136X
[2] ORCID: https://orcid.org/0000–0002-8799–0267
[3] ORCID: https://orcid.org/0000–0001-8018–8004

https://doi.org/10.1016/j.asw.2025.100986

Available online 15 October 2025

using CJ successfully guided by such a question alone (e.g., Paquot et al., 2022); however, that question can also be supported by supplementary material, for example, a description of the relevant competence (Sickinger et al., 2025; van Daal et al., 2019). Once the judge has selected the "better" performance, another pair of performances is displayed. Judges continue to make these pairwise comparisons until they have completed the CJ session. As judges make their decisions, a statistical model rank-orders the performances and assigns scaled scores. The final scaled score for each performance is formed from the decisions made by all the judges in the session (for more details on the CJ method, see Sickinger et al., 2025).

There is a considerable body of CJ research on assessments across a wide range of subject areas (e.g., design and technology, science, mathematics) and focused particularly on CJ's reliability (Bartholomew & Jones, 2022). Jones and Davies (2024) state CJ's benefits as its (a) efficiency – it is generally a faster process than traditional rating particularly as training needs (if any) are minimal; (b) reliability – for example, as judges only ever compare two scripts at a time, the problem of rater effects is eradicated; (c) variety – the ability to evaluate open performances is a key feature; and (d) precision – scripts can be awarded a score using the full scale stipulated by the test administrator.

Research into the use of CJ for L2 writing assessment (L2 script evaluation) is comparatively limited but increasing. Most studies have investigated and established CJ's reliability in this context, and have done this for its conventional holistic judging format, which requires one overall judgement on each L2 performance pair (e.g., Sims et al., 2020; Thwaites et al., 2024). Recently, Sickinger et al. (2025) have complemented this with reliability research on an adapted form of CJ, in which L2 scripts are evaluated by individual criteria (termed *dimensions* in the research context, and similar to rating criteria in analytic rating). The latter format, CJ by dimensions, requires judges to assess each performance pair in terms of a particular aspect (e.g., vocabulary) and offers greater potential for providing more detailed scoring feedback on performances.

As CJ's reliability has now been demonstrated, including in L2 contexts, the validity of the method is becoming a focus (Buckley et al., 2022). To date, validity remains under-researched for L2 CJ assessments (Thwaites et al., 2024, 2025), although the validation of assessment processes is fundamental to much contemporary research in the broader field of L2 proficiency testing. For example, argument-based approaches to validation – common in this field and adopted in the current study – require the provision of evidence to support claims (a logical argument) that demonstrate the adequacy of scores on a test as a representation of its purpose (Chapelle, 2021). These claims should account for the context of test use and the expectations of test users. Taking as an example the test in the current study (introduced in Section 3), contextual expectations might include a common understanding within the education system of the components of L2 English writing skills, the need to provide thorough feedback on test-taker performance, and the use of a scoring method that is interpretable by the general public. To show the appropriateness of a new scoring approach for this context, evidence is required to support claims made for the approach and to demonstrate that relevant issues have been investigated. The current study therefore seeks to provide evidence of CJ's scoring validity that pertains to its context of use, where clarity in the marking process is valued.

Typically, CJ is conducted without training or supporting information, which can result in a perceived lack of transparency in the judging process, potentially negatively influencing claims about scoring validity (Bramley, 2022; Chambers & Cunningham, 2022; Holmes et al., 2017; Walland, 2022). Other aspects of CJ that can affect views on scoring validity include the lack of clarity regarding construct-relevant criteria, and the need to identify, and possibly filter out, less reliable judges (Kelly et al., 2022). However, a benefit of CJ is its capacity to encompass a range of judges' subjective opinions and its inherent multiple-marking (Thwaites & Paquot, 2024). Validity is thought to be positively influenced by the range of experts evaluating performances (notwithstanding the attendant issue of defining that expertise), as it "increases the probability that the multidimensionality of text quality is represented in the text scores" (Lesterhuis et al., 2022, p. 8).

To contribute to the emerging validity research on CJ, the current study examines the scoring validity of CJ in an L2 context and expands it to CJ in both its holistic and dimension-based formats.

2. CJ validity

As Thwaites et al. (2025) note, "Little evidence exists of the construct validity of using CJ for L2 writing assessment" (p. 1). In the field of language testing, various frameworks (e.g., Shaw & Weir, 2007) have been conceptualised for considering the validity of L2 writing assessments. An argument-based validity approach has gained favour in recent years, requiring the provision of evidence to support claims made about features of a test and its score meanings (see Chapelle, 2021, for a full description and worked examples of an argument-based validation chain). Knoch and Chapelle's (2018) argument-based validation framework offers a guide to integrating a new scoring process into a test structure (and is therefore relevant to the current study). Their framework, based on analysis of studies into rating processes, sets out a chain of six inferences – evaluation, generalization, explanation, extrapolation, decision, and consequence – and the claims and evidence associated with each. It provides a systematic way to relate the scoring processes of a test to its validation, set out the claims of a validity argument and present related evidence. This exposition shows how far-reaching the impact of changing the scoring process of a test is. The framework has been used previously in relation to CJ (Sims et al., 2020), and an aim of the current study is to expand on this work to integrate the use of CJ into a validity argument (see Section 6.1). While some proponents of CJ argue that the validity of this method does not necessarily require an understanding of the basis for judges' decisions (e.g., Jones & Davies, 2024), such a view remains contested in the field of language testing.

Typically, CJ studies have varied in their methods to demonstrate aspects of validity for CJ, including comparisons with final exam scores in science-based subjects (Hughes et al., 2016), predicted grades in maths (Jones & Inglis, 2015), automated scores in English language arts tests (Steedle & Ferrara, 2016), and relevant national demographics in L1 primary school writing (Wheadon et al., 2020). Correlation with expert scores has been used across educational contexts, including writing in L1 (McGrane et al., 2018) and L2 (Sims

et al., 2020). The word count of scripts, while sometimes classified as a construct-irrelevant feature of L1 writing (Landrieu et al., 2022; Whitehouse, 2013), has also been operationalised as a proxy for script quality and, thus, a validity measure (Sims et al., 2020). Additionally, the appropriateness of the criteria on which judges base their decisions has been used to support validity claims in writing in L1 (Lesterhuis et al., 2022; van Daal et al., 2019) and, recently, in L2 (Thwaites et al., 2025). For example, studies in L1 writing that investigated what judges consider when making CJ decisions (in CJ's conventional holistic format) have asked judges to note down what they considered when judging after each pairwise comparison (Whitehouse, 2013), presented judges with manipulated scripts to focus on chosen aspects of performances, and used think-aloud protocols (TAPs), where judges articulate their thoughts while conducting the judgements (Chambers & Cunningham, 2022). In the context of L2 writing assessment, Thwaites et al. (2025) asked judges to add comments directly to the judging platform whilst judging L2 scripts at Common European Framework of Reference for Languages (CEFR) levels B1–C1.

Research to date suggests that judges generally consider relevant features (e.g., Lesterhuis et al., 2018; Thwaites et al., 2025; van Daal et al., 2019). However, instances of judges using features perceived as construct-irrelevant, such as handwriting, have also been noted (Chambers & Cunningham, 2022). Such instances – "hidden in a holistic judgement and, therefore, untraceable" (Chambers & Cunningham, 2022, p. 8) – are concerning. Nevertheless, the use of CJ in its recently proposed adapted format, in which written performances are judged by dimensions/separate criteria (Sickinger et al., 2025), could potentially provide more transparency regarding the construct being considered.

2.1. The decision-making process

Applying CJ has been seen as a "paradigm shift" (Oates, 2022, p. 4) in the assessment context, allowing a wide range of performances to be judged without the need for rating scales, intensive rater training or post-test adjustments (Pollitt, 2012a, 2012b). Instead, Laming's (2004) statement that "All judgments are comparisons of one thing with another" (p. 9) coupled with Thurstone's (1927/1994) Law of Comparative Judgement lie at the heart of the CJ process. A key benefit of this comparison approach coupled with CJ's intrinsic multiple marking is the avoidance of long-recognised rater effects, such as inaccuracies, biased and inconsistent ratings (Pollitt, 2012a, 2012b). However, as noted by Kelly et al. (2022) and Thwaites et al. (2025), CJ currently lacks a comprehensive theoretical framework.

Research has considered how judges make these comparative judgements, although, as van Daal (2020) notes, much of this is centred on non-educational contexts. For example, the extent to which judges use intuition in their decision-making has been investigated for performances involving L1 English and statistics (Marshall et al., 2020), with judges of the English responses, for instance, reporting using the assessment standards in an almost equal ratio to intuition. In addition, these judges reported using intuition more than the judges for the statistics responses did (Marshall et al., 2020). The use of intuition by judges has also been reported in other studies (e.g., Curcin et al., 2019; Rotaru, 2022) and viewed as a benefit of CJ by some judges: "It was a liberating experience to use gut-reaction and professional judgement, rather than becoming bogged down in an overly complex mark scheme" (Walland, 2022, pp. 51–52). In Paquot et al. (2022) and Thwaites et al. (2025), where crowdsourced judges have reliably evaluated essays from a standardised English-language test (TOEFL iBT) without prior training or guiding instructions, it can arguably be presumed that a form of intuition was also used during the decision-making process.

In contexts where transparency is valued, an apparent over-reliance on intuition is likely to be viewed as a threat to scoring validity and limit the acceptance of the CJ method. An alternative approach in which performances are judged by different aspects (e.g., criteria/dimensions for written performances) could offer a solution, providing greater direction for judges and alleviating difficulties for judges inherent in the decision-making process, as well as offering more detailed feedback to test-takers. McGrane et al. (2018) investigated this possibility for L1 writing with mixed results. The method – where each performance was compared on two dimensions with universal exemplars, rather than with another, paired, performance – was found to be "cognitively challenging" by some judges (p. 308), and the authors recommended providing clearer descriptors in future research. The present study takes this recommendation forward.

3. The present study

The current study aims to provide further insight into the validity of CJ for evaluating L2 performances, particularly the criteria considered by judges. It aims to do so, firstly, by incorporating such insights into a validity argument using Knoch and Chapelle's (2018) validity framework. Secondly, in addition to investigating the validity of CJ in its conventional, holistic format, it also aims to do so for a newly introduced CJ method, *CJ-Dimensions*, where, unusually for CJ, judges evaluate the scripts by individual dimensions/criteria (Sickinger et al., 2025). Thirdly, while CJ L2 writing research has so far concentrated on adult learner contexts, this study aims to shed light on the under-researched context of L2 young learners.

The following research questions are addressed:

RQ1. To what extent do the Comparative Judgement methods of script evaluation (holistic CJ; CJ by dimensions) provide scores which are based on appropriate criteria?

RQ2. What decision-making approaches are evident when judges are evaluating scripts using Comparative Judgement (holistic CJ; CJ by dimensions)?

The research reported here was part of a larger project (Sickinger, 2023) conducted in the context of national, standardised testing

for monitoring standards in English as a Foreign Language (EFL) in lower-secondary schools (Year 8) in Austria. At the time of preparing this manuscript, the most recent test administrations were from 2013 and 2019, overseen by IQS (Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen), a subordinate department of the Ministry of Education, Science and Research. The writing subtest comprised two tasks: a shorter, A2-level task and a slightly longer, B1-level task (see Fig. 1; Siller et al., 2019). At both administrations, writing performances had been rated analytically by a group of English teachers from across Austria.

4. Method

4.1. Research design

The larger project (Sickinger, 2023) providing data for this article consisted of two phases (Fig. 2). Phase 1 primarily focused on quantitative, score-related data to investigate the reliability of CJ and the impact of judge characteristics; the findings of this are reported in Sickinger et al. (2025). However, additionally, Phase 1 also generated qualitative questionnaire responses, which form one part of the current study. Phase 2, which constitutes the other part and main focus of the current study, was conducted a year after Phase 1 and focused on qualitative data from think-aloud protocols performed by a subset of participants.

4.2. Participants

Phase 1 – including the questionnaires analysed here – comprised 27 participants (74 % female; M_{age}=44.07y, SD_{age}=11.95y). These participants were grouped according to their EFL teaching (experienced ≥4 years; Freeman, 2001) and their rating experience:

a) four inexperienced teachers who also were inexperienced in rating; henceforth labelled *inexperienced*
b) five experienced teachers who did not have rating experience on the Year 8 writing exam; henceforth labelled *t-experienced*
c) 18 experienced teachers who had previously trained as raters for the Year 8 writing exam; henceforth labelled *raters*.

For Phase 2, as the four *inexperienced* teacher participants had proven to be less reliable in Phase 1, only *t-experienced* and *raters* (n = 23) were invited. Eight participants (75 % female; M_{age}=47.13y, SD_{age}=9.22y), comprising four *t-experienced* and four *raters*, volunteered and evaluated the scripts whilst conducting concurrent, non-mediated think-alouds (Green, 1998).

4.3. Script evaluation

For the purpose of the project, IQS provided a set of scripts from the 2013 test (the 2019 scripts were unavailable for research purposes at the time). IQS used the official exam raters' scores to select performances from their database to represent all eight scoring bands for each of the four dimensions of the analytic rating scale. The test and the accompanying four-dimension rating scale[1] had been developed using an adapted version of Shaw and Weir's (2007) conceptualisation of writing evaluation and reflected the communicative language competences described in the CEFR (Gassner et al., 2011; Siller et al., 2019).

The Phase 1 questionnaire responses followed the participants' rating and judging of the 300 scripts provided: 150 *short scripts* from the short writing task and 150 *long scripts* from the longer task. The Phase 2 think-alouds resulted from new judgments a year later on 157 scripts – 80 short and 77 long scripts (specifically, 40 new short scripts and 37 new long scripts, 10 short and 10 long scripts from the Phase 1 pilot study, and, to ensure a range of scripts from low to high scoring, every fifth script from the Phase 1 *CJ-Holistic* rank orders, making 30 short and 30 long scripts). These scripts were evaluated by our participants using two different CJ methods:

a) *CJ-Holistic*, where a script was judged in its entirety against the entirety of a second script – the common CJ method;
b) *CJ-Dimensions*, where each script was judged in one of four dimensions against the same dimension in a second script. The dimensions reflected the criteria used for the 2013 exam's analytic rating: Task Achievement, Coherence & Cohesion, Grammar, and Vocabulary.

Scripts, presented to the judges onscreen in pairs, were randomly allocated by the CJ platform. In Phase 2, the researcher stopped the CJ session after the time allocated for each think-aloud was reached (see Table 1). In this phase, judges' evaluative decisions were not used for CJ score calculation, as the focus was the process; judges were not aware of this.

4.4. Comparative Judgement

4.4.1. Platform

For the purposes of this study, the scripts were uploaded to the *No More Marking* CJ platform (https://www.nomoremarking.com) and ten CJ sessions were created (Table 1). This platform was selected because it is freely available for researchers and has been used in previous CJ L2 research (Thwaites & Paquot, 2024).

4.4.2. Instructions

During a CJ session, judges' decisions are directed by a short question or instruction. For the two *CJ-Holistic* sessions, one evaluating the short scripts and one evaluating the long scripts, judges were told to "*Choose the better response*".

Short task

Read the instructions carefully and then write your text on the next page.

Time: **10 minutes**

Text: **40–70 words**

In your text, try **not** to use language from the task below.

You had a class project on "how to save the planet". This is the picture of your project. Now you write an **email** to your penfriend in England and tell him/her about it.

Say

- what exactly you did.
- who you did it with.
- if and why you liked/didn't like it.
- what your friend could do to save the planet.

E8W001

Long task

Read the instructions carefully and then write your text on the next page.

Time: **20 minutes**

Text: **120–180 words**

Use paragraphs.

In your text, try **not** to use language from the task below.

A youth-magazine for learners of the English language has started a writing-contest about teenage problems and experiences. Your English teacher has asked you to send in an exciting **story** about *'The first time I stayed at home all alone.'*

Write about a real story or a story you imagine.

In your **story,**

say
- when it was.

explain
- why your parents weren't in the house.

describe
- your feelings.
- a problem you had then.

explain
- how you solved the problem.

say
- what you have learned from this experience.

E8W006

(caption on next page)

Fig. 1. Short and Long Writing Tasks. *Note:* From IQS Bist-UE_Freigegebene_Items_E8_2013 (https://www.iqs.gv.at/downloads/archiv-des-bifie/bildungsstandardueberpruefungen/freigegebene-items). Copyright 2013 by BIFIE. Reprinted with permission.

For *CJ-Dimensions*, each judge conducted eight CJ sessions (see Table 1). For each of the four dimensions, judges conducted two CJ sessions, one for the short scripts and one for the long scripts. The directing questions for the dimensions of Task Achievement: *Which text best fulfils the criteria for Task Achievement?* and Coherence & Cohesion: *Which text best fulfils the criteria for Coherence & Cohesion?* encouraged judges to refer to CJ supporting materials (see below). The directing questions for Grammar: *Which text demonstrates the best range and accuracy of Grammar?* and Vocabulary: *Which text demonstrates the best range and accuracy of Vocabulary?* reflected the current EFL teaching philosophy in Austria of prioritising range over accuracy in marking (Siller et al., 2019). For all CJ sessions, judges were instructed to use the supporting materials during the decision-making process.

4.4.3. Supporting materials

Training for CJ is often minimal or deemed unnecessary (Thwaites & Paquot, 2024). Judges have also been found to be reliable when using CJ without supporting materials (e.g., Thwaites et al., 2024). However, as this study included some teachers who had not been rater trained, Supplementary Materials was provided in the form of two crib sheets: one for *CJ-Holistic* and one for *CJ-Dimensions* (see Supplementary Materials). The crib sheets presented key criteria reflecting the four dimensions of the original analytic rating scale (Task Achievement, Coherence & Cohesion, Grammar, and Vocabulary). They stated aspects of writing *expected* for each dimension and, for judgements involving more similar performance pairs (a potentially more complex process), additional guidance was provided with aspects of writing that *could be considered* for each dimension. The two crib sheets presented similar guidance; however, the instructions at the top of the sheets differed: for *CJ-Holistic*, judges were told to consider the scripts overall; for *CJ-Dimensions*, they were told to focus on only the individual dimension currently being judged.

4.5. Perception questionnaires

The Phase 1 judges completed short perception questionnaires upon completion of the *CJ-Holistic* sessions and, separately, the *CJ-*

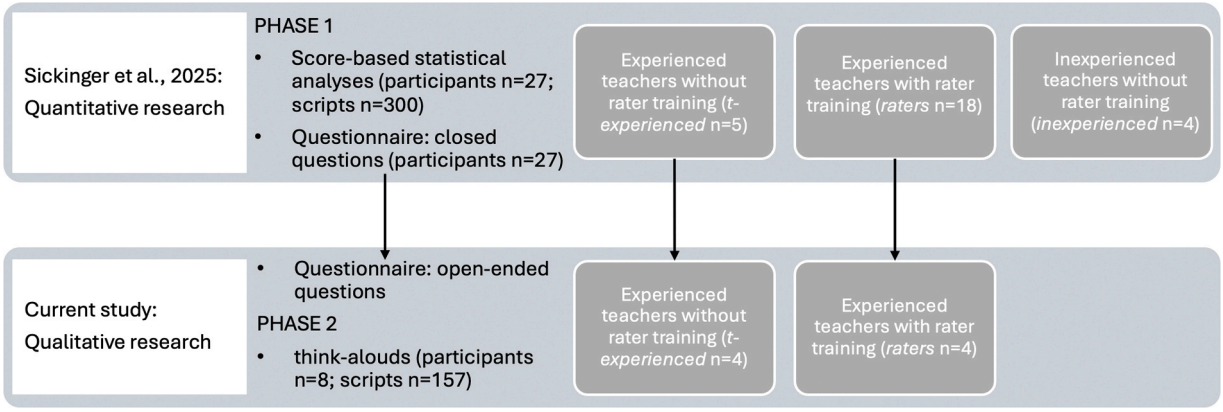


Fig. 2. Overview of Phase 1 and Phase 2. *Note:* One t-experienced judge in Phase 2 participated as a pilot judge in Phase 1 and did not receive reliability data.

Table 1
Think-Aloud Meetings.

Duration	CJ session
Meeting 1	
1 (15 min)	short *CJ-Holistic*
2 (20 min)	long *CJ-Holistic*
Meeting 2	
3 (10 min)	short *CJ-Dimensions* Task Achievement
4 (10 min)	short *CJ-Dimensions* Coherence & Cohesion
5 (10 min)	short *CJ-Dimensions* Grammar
6 (10 min)	short *CJ-Dimensions* Vocabulary
Meeting 3	
7 (15 min)	long *CJ-Dimensions* Task Achievement
8 (15 min)	long *CJ-Dimensions* Coherence & Cohesion
9 (15 min)	long *CJ-Dimensions* Grammar
10 (15 min)	long *CJ-Dimensions* Vocabulary

Dimensions sessions. In the questions relevant to the present study, for *CJ-Holistic*, judges were asked *What aspects of the [script] did you consider when judging?* and for *CJ-Dimensions*, judges were asked the same question addressing each dimension, e.g., *What aspects of the [script] did you consider when judging Task Achievement?* The results from these questionnaires provided insight from a larger number of participants and complemented the think-aloud data from Phase 2.

4.6. Think-aloud

Over three separate meetings lasting up to 60 min (Table 1), each judge in Phase 2 separately conducted ten CJ sessions whilst performing think-alouds using MS Teams screen-sharing with audio-recording. The three think-aloud meetings (1, 2, 3) took place on different days at least two days apart. *CJ-Holistic* was timetabled first so that these judgements would not be influenced by the more detailed considerations of the individual dimensions necessary for *CJ-Dimensions.* The four dimensions were covered in the same order by each participant for short and long scripts.[2] CJ sessions for the long scripts were allocated more time to allow judges more reading time and to accommodate longer verbalisations. Prior to judging, judges were provided with the two writing prompts (Fig. 1) and the two crib sheets. The resulting recordings were automatically transcribed, subsequently checked and manually amended where needed.

4.7. Analyses

The Phase 1 questionnaire data on criteria considered by judges when judging scripts were coded, following the descriptors shown on the analytic marking scale[1], and analysed. Criteria reported by only one judge were coded as "other" and are not included in the results below for reasons of space (a fuller account is given in Sickinger, 2023). Intra-coder reliability involved re-coding by the first author of the data, because IQS confidentiality regulations meant that only the first author was permitted data access. The resulting coefficients showed that the coding reliability was "substantial" or "better" (Landis & Koch, 1977) for all instances for *CJ-Holistic* ($\kappa > .76$) and *CJ-Dimensions* ($\kappa > .63$), except for three *CJ-Dimensions* Grammar codes (*tenses accuracy; general structures; general accuracy*), which showed "moderate" and "fair" agreement. In these cases, the original questionnaire data were re-examined to decide the final coding.

To analyse the Phase 2 think-alouds, a narrative approach with a focus on patterns was used (Gu, 2014).[3] The TAPs were annotated to indicate when the judge was commenting on, reading from, or paraphrasing the script; referring to the prompt; and/or referring to the crib sheet. In this article, judges' spoken text is italicised, references to the prompts (Fig. 1) are in bold font, and text read from the script is underlined. A slash (/) indicates non-sequential text. Explanatory comments are enclosed in square brackets. To ensure test-taker anonymity, names occurring in the TAPs have been replaced with initial letters.

5. Results

5.1. Features of writing reported by judges

To address RQ1 (Are CJ scores based on appropriate criteria?), 26 Phase 1 judges self-reported the aspects of writing they considered during CJ. (One participant did not respond to the questions.) In their questionnaire responses following *CJ-Holistic*, judges typically reported considering broad categories, e.g., Grammar. As Fig. 3 shows, most judges reported considering Coherence & Cohesion, Vocabulary, and Task Achievement or key aspects of that dimension such as the content/bullet points for inclusion (see Fig. 1). Just under half listed Grammar. However, the criteria "range" and "accuracy" referred to by many judges also implicitly referenced Grammar and/or Vocabulary. There were more self-reports from judges regarding range than for accuracy, reflecting Austrian marking policy of prioritising range over accuracy, which was emphasised in the crib sheets. Three participants specifically referenced the crib sheets which directed judges to consider the separate dimensions and to prioritise range over accuracy.

All aspects of writing reported were arguably construct-relevant, reflecting elements of the construct operationalised in the Year 8 exam's original analytic rating scale. For example, a *rater* judge noted "naturalness of text", which could be linked to the analytic rating scale's descriptor "fairly fluent text", and another noted "would an English speaking person understand the text", which could again be linked to the rating scale text "clear and coherent text" with only a "few inaccuracies which do not impair communication" (Siller et al.,

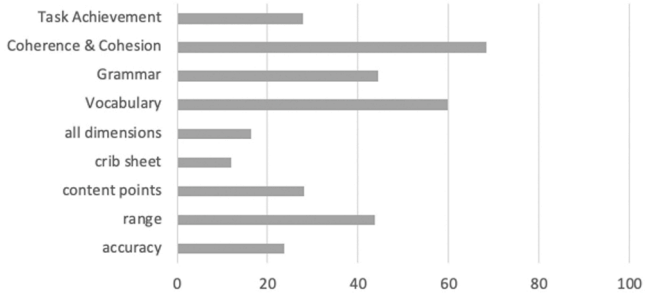


Fig. 3. CJ-Holistic: Percentage of Judges Mentioning Aspects of Writing (n = 26).

2019, p. 23).

As shown in Fig. 4, judges reported considering a wider range of criteria during *CJ-Dimensions*. For Task Achievement, generally the focus was on whether the script addressed the prompt through the test-takers including and/or elaborating the content points. For Coherence & Cohesion, paragraphing and linking words/connectives were commonly noted along with other cohesive devices, and some judges referenced overall fluency and clarity. For Grammar, the use of structures generally, and specifically tenses in the script, were key features for the judges, as were range and accuracy, with range being mentioned more frequently. Finally, for Vocabulary, range and accuracy were mentioned, with range mentioned by almost twice as many judges as accuracy. Apart from very few exceptions (e.g., "nice"), responses were related to the construct (as defined by the analytic rating scale).

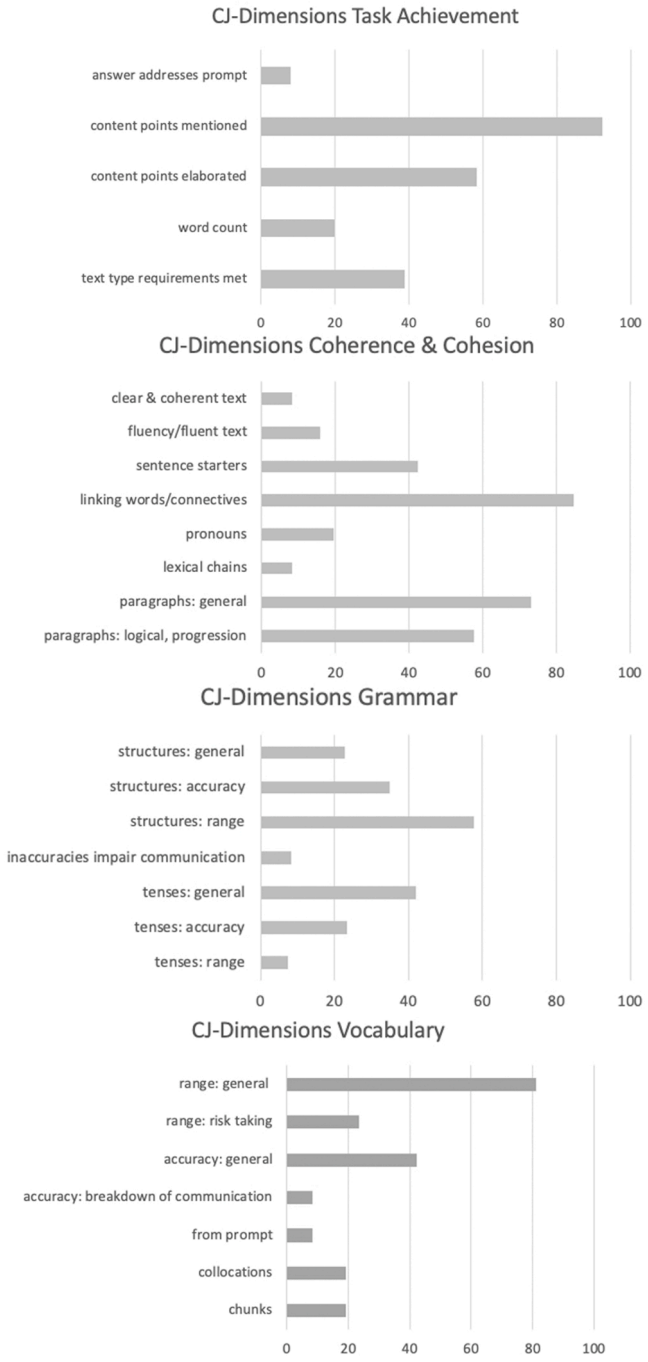


Fig. 4. CJ-Dimensions: Percentage of Judges Mentioning Aspects of Writing by Dimension (n = 26).

5.2. Judges' decision-making approaches

Phase 2 of the study was designed to provide more insight into the decision-making process as well as the criteria considered. Analysis of the eight Phase 2 judges' TAPs revealed their decision-making approaches when evaluating scripts using CJ (RQ2). Below, we first report the results for the conventional holistic CJ approach, and then for the novel dimension-based CJ approach.

5.2.1. Decision-making during CJ-Holistic

Two decision-making approaches were observed during *CJ-Holistic*. One approach, applied by six judges (three *t-experienced* and three *raters*), reflected the IQS analytic rating process, where raters were trained to read scripts carefully and then consider each dimension in turn. Typically, these judges initially read aloud from each script. Usually, they read the scripts carefully and completely, often considering each dimension in turn, before comparing the two scripts. We label this approach *read and compare*. In the second approach, used by two judges (one *t-experienced* and one *rater*), the judges scanned the scripts noting key textual aspects before making their decision regarding the better of the pair. We label this approach *comparative scan*.

5.2.1.1. Read and compare. Judges using the *read and compare* approach varied in their approach to *CJ-Holistic*: some read more than others, some made quicker decisions than others, and some verbalised their decision-making more extensively. However, generally, as summarised in Fig. 5, they looked for superficial features in both scripts first, and then read the first script. Next, they read the second script and compared it to the first as they read.

In the following example, which shows the complete process for a pair of long scripts, a *t-experienced* judge discovers during step 1 that both test-takers have employed incorrect text-types. The judge then moves on to step 2 and reads the left-hand script noting key features (step 2), before reading the second script on the right-hand (step 3), and finally making a decision (step 4). For reference, the task prompt for the long task is in Fig. 1.

> *OK, should be a story.*
>
> *OK, there is on the left To A. The first day I stay/ at home here on the right it says, Hey, I'm K. Uh-huh, the left one is writing a letter or an email or something like that. It's not just, should it be a story. I'm sure, yes.* [appears to check prompt] *Good. /the first day I/* [inaudible reading of script] ***when it was****. Let's look at the* ***bullet points****. It was on a Friday yes it's here.*
>
> *My parents are/ so I find it is cool to be alone at home you can make what you will. OK, this is very confusing on the left. No*
>
> *er full stops.*
>
> */is great. But one problem is i must wash the clotes, whash the Kitchen and clean the house and the other problem is I must/ So grammar is quite bad. So I solved the problem, maybe I had a little/* ***What was the problem****? OK, the bullet points are here.*
>
> *They are done.* ***Describe feelings****.*
>
> *I learned/ so* ***bullet points*** *are done. Grammar and vocabulary? Vocabulary is OK, but not so good grammar is bad on the left.*
>
> *What have you got on the right?*

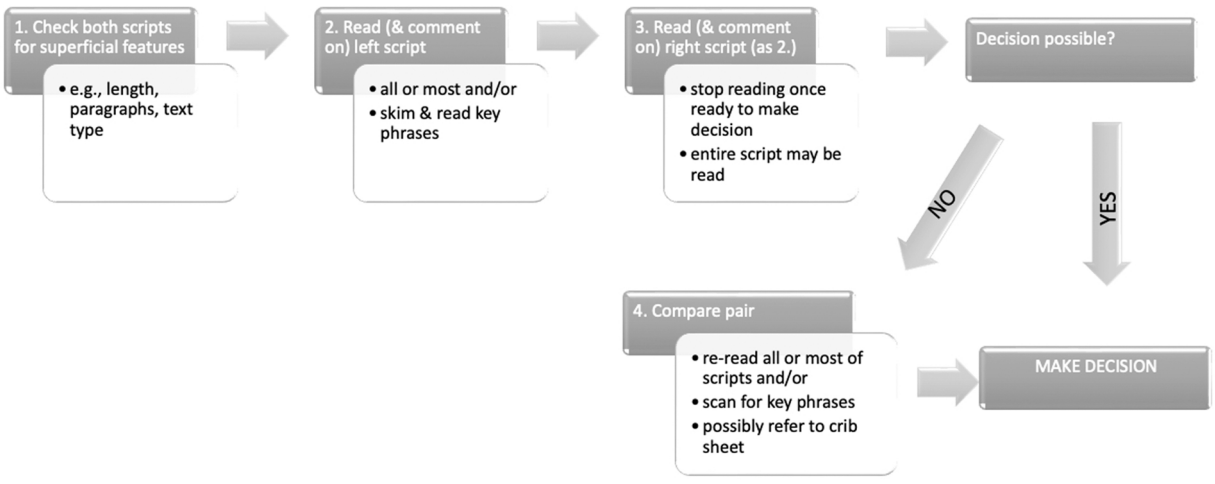


Fig. 5. Read and Compare Approach to CJ.

When I was ten I was the first time alone. It was in the evening because my parents went for a dinner away. It [judge stops and rereads as written] *At first I thought it was* [script = would be] *easy to be alone. It's yeah, it's not so confusing. The grammar is better. I don't know it exactly exactly/ electricity/ lights/ course it didn't work so the right one is better, but the grammar? /after my panik-attack I had an idea. So I think all the* ***bullet points*** *are done, right one is better.*

Some *read and compare* judges read all or the majority of a script but some skim-read. When seeing the same script again, these judges would not always re-read the script (steps 2 and 3). Many *read and compare* judges referred to the prompt (Fig. 1) to check the content points had been addressed (Task Achievement) while reading the text as part of step 2 and/or 3 before evaluating the other dimensions. Once the left-hand script had been evaluated (step 2), judges then evaluated the right-hand script (step 3). Sometimes, once the second, right-hand script appeared to be the better or worse, rather than reading all that script (step 3), the judge instead selected key textual aspects, relating to the prompt (Fig. 1) to support their judgement decision (step 4), as in the previous example.

Judges sometimes referred to the crib sheet for assistance when finalising their decision (step 4), looked for key difference(s) between the pair and, at times, as in the following example from a *t-experienced* judge, considered broader criteria: *OK, uhm, I think the text in the right has fewer mistakes and so on, but I'm going to go for the text on the left because, uhm, it's more daring.*

For *read and compare* judges, the features frequently mentioned during think-aloud for Coherence & Cohesion reflected the qualitative questionnaire data – mainly paragraphing, fluency, connectives, and sentence starters. However, the TAP data for Grammar and Vocabulary showed a narrower focus, predominantly the use of tenses and relevant words or phrases. Often, the features of writing being considered by the judge were discernible from the TAPs simply by the text the judge chose to read aloud and/or their manner of reading, e.g., their emphasis.

5.2.1.2. Comparative scan. In contrast to *read and compare* judges, judges using the *comparative scan* approach (Fig. 6) simply quickly compared key extracts from a pair of scripts (step 1). These judges tended to skim-read the scripts and make a quick decision often without verbalising the criteria, which made it difficult to ascertain exactly what they considered in their decision-making. For example, one *rater*'s entire verbalisation for a pair of short scripts was simply: *The next two: clearly the right one*. Sometimes, this *rater* judge verbalised more and read key text. Again, the entire verbalisation for a pair of short scripts is shown

I think that saving the planet is a project which is really hard to manage. Millions of people/ comfortable car/ save lots of electricity. Left one, the right one is no way better.

The *comparative scan* approach reflects the final decision stage (step 4) of the *read and compare* approach: comparing key extracts from the scripts. This example, showing the full process for two short scripts, highlights how the *t-experienced comparative scan* judge frequently selected key extracts from varying locations in and across the scripts (step 2). Here, paragraphs and sentence structure are the key features considered:

[reads salutation and first sentence from the left script] *Dear J, We had a project about "how to save the planet." It is terrible!* [salutation and first sentence from the right script] *Dear J, Last Friday, our class had the project how to save* [penultimate sentence from the right script] *you can save the world too. paragraphs, right paragraphs.*

Right is better.

The two *comparative scan* judges were typically fast; possibly, the ease of the task for these experienced teachers/raters made the process too fast for verbalisation (Brooks, 2012), resulting in a lack of transparency in the decision-making process (Barkaoui, 2011). However, when differences between the two scripts were less obvious, the judgment process took considerably longer, with these judges moving through some of or all the four dimensions in turn, typically with more text read aloud.

Fundamentally, in *comparative scan*, the decision between the two scripts was taken quickly and more detailed evaluation of the scripts was only necessary for more complex decisions (Fig. 6).

5.2.1.3. CJ-Holistic overall. Judges employing the *comparative scan* approach completed more pairwise comparisons within the allocated times than judges using the *read and compare* approach (Table 2). The *comparative scan* judges had also been among the fastest and most reliable in Phase 1 (Sickinger, 2023). The number of paired performances evaluated by *comparative scan* judges and the

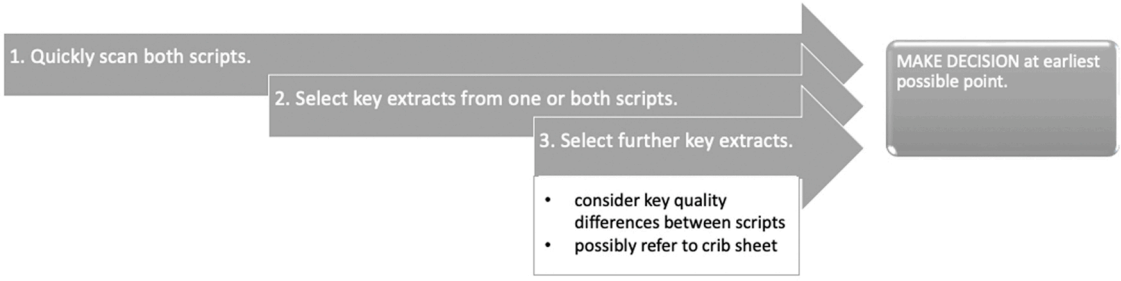


Fig. 6. Comparative Scan Approach to CJ.

reliability of their judgements seems to indicate that *comparative scan*, the decision-making approach adopted by these judges, was a successful adaptation to the CJ method.

While two decision-making approaches (RQ2) were clear from the *CJ-Holistic* TAPs, the criteria considered in the judgement decision (RQ1) were less clear from the TAPs. Some judges verbalised the criteria being considered, for example, from a *read and compare rater* judge: *In terms of coherence and cohesion it doesn't flow at all, but there are some sentence starters but basic ones*. However, other judges, particularly *comparative scan* judges, did not verbalise the criteria they were considering. This makes it harder to determine whether decisions were made on relevant criteria, although the text extracts judges selected to read provided some insight. Nevertheless, relevant verbalisations did reflect criteria reported in the questionnaire data from Phase 1.

5.2.2. Decision-making during CJ-Dimensions

Judges employed the same judging approaches in *CJ-Holistic* and *CJ-Dimensions*.

5.2.2.1. Task Achievement.

When evaluating Task Achievement, *read and compare* judges usually first checked that the correct text-type had been used and that the test-taker had answered the prompt's content points. For the short task (an email with appropriate salutation and sign-off), most *read and compare* judges considered the pair of scripts for these generic features before considering first one script then the other. However, as the long prompt required a story with no required key features, most *read and compare* judges started with step 2 for the long scripts (Fig. 5). Notably, some judges also read, rather than skimmed, more of the text for the long scripts.

Read and compare judges rarely mentioned other dimensions while focusing on Task Achievement and when they did, they tended to remark on doing so, highlighting the difficulty in separating the aspects of writing. For example, one *rater* judge said about a script: *I don't understand really. It's not that coherent right now. It's not coherence, but it's she's not really getting the message across if you ask me. She's not addressing the content points coherently*.

The two *comparative scan* judges also continued with their previous approach during *CJ-Dimensions*. For example, the *t-experienced* judge searched for evidence of the test-taker mentioning, and possibly elaborating, the content points by reading around and between scripts:

> *Task Achievement again.* [starts reading from the left-hand script] *Yours names names Last week when I have been in England we had got a project it's called "how to save the planet". We cleaned the Kindergarten* [moves on to read from the right-hand script] *I have a class project with my class "how to save the planet". I cleaning the streets and pick up the trash from the ground.* [laughs] That it really…*I liked it. Please help with it's important for the world* [moves back to the left-hand script and reads the last sentence] *It would be nice if you don't throw the trash away. Of course it's left. I really like that because elaboration.*

This judge's search for evidence to support decision-making was often verbalised as questions. For example, when reading a long script and searching for the responses to two content points (Fig. 1) this judge asked: *What's the problem?*; *Why were you alone?*

For *CJ-Dimensions* Task Achievement, the *comparative scan* approach was again faster than *read and compare*, particularly for the *rater* judge (Table 2).

5.2.2.2. Coherence & Cohesion.

The greatest variety of features of writing that was verbalised was given for Coherence & Cohesion, and arguably all features were construct-relevant. Sometimes, rather than explicitly mentioning criteria they were considering, some *read and compare* judges used emphasis whilst reading aloud to indicate which criteria were being considered.

Table 2
Number of Pair Judgements Made in Allocated Time by Judges.

	comparative scan judges (n = 2)		*read and compare* judges (n = 6)	
	faster judge	slower judge	fastest judge	slowest judge
CJ-Holistic				
short scripts (10 min)	28	24	12	6
long scripts (15 min)	17	16	11	7
CJ-Dimensions				
Task Achievement				
short scripts (10 min)	21	12	11	5
long scripts (15 min)	17	11	11	4
Coherence & Cohesion				
short scripts (10 min)	32	13	9	5
long scripts (15 min)	23	15	10	4
Grammar				
short scripts (10 min)	23	13	10	6
long scripts (15 min)	22	10	8	6
Vocabulary				
short scripts (10 min)	24	14	9	7
long scripts (15 min)	22	16	8	6

Note: All judges in Phase 1 were reliable; quantitative data were not available for one *read and compare* judge who took part in the Phase 1 pilot (Sickinger, 2023).

For Coherence & Cohesion, some judges considered paragraphing during step 1; however, most judges began their judging with step 2 (Fig. 5). Generally, *read and compare* judges tended to follow a structured approach, as shown by this *rater's* verbalisations which followed their reading of the second script (step 3):

> *The text quite flows gently. It's very nice. There are also the sentence connectors are there but, because.*
>
> *And the basics are there and very good* sentence *starters. There is: When, I was, After*
>
> *Uh, When, Then so it's clearly organised, not so much with paragraphs, but there's a logical progress of ideas and is clearly organised. I'd go for the right one.*

When one script was clearly better, some judges would not complete step 3 (reading the second script completely) before moving on to step 4 (the decision-making process). At this point most judges summarised their thoughts, as illustrated by one *t-experienced* judge: *the text is clearer on the one the left side and easier to read and, yeah, more logical I would say*. As demonstrated by another *t-experienced* judge, judges often clearly verbalised their knowledge of the appropriate criteria: *Although my feeling tells me I really like because it's funny the text on the right looking at Coherence and Cohesion: logical progression, the lexical chains, the connectives and sentence starters, I picked the text on the left.*

The *comparative scan* judges verbalised fewer features of writing; however, their text selection provided insights into the criteria they were considering. The two judges using the *comparative scan* approach also differed slightly in how they compared the scripts. For example, the *rater* sometimes just commented on the reasons for the selection of a script:

> *Left one has got more paragraph, right one hasn't got anything. Left one has got sentence starters. Looks more flu- fluent*
>
> *than the right one, so it's* gonna *be the left one.*

As they had done for Task Achievement, the *t-experienced comparative scan* judge would sometimes question the text for features of Coherence & Cohesion: *Ah, here is a paragraph; What about sentence starters here?* Sometimes these questions showed some frustration: *Why don't they make paragraphs?*, *Why don't they know how to write?*, *How can you compare these two texts?*

Again, the *comparative scan* judges were faster (and had been reliable judges in Phase 1) (Table 2).

5.2.2.3. Grammar . Distinct from the other dimensions, *read and compare* judges employed differing strategies for short and long scripts for this dimension, commenting more and reading less for the short scripts. Also, for this dimension, *read and compare* judges usually commenced with step 2 (Fig. 5), reading the left-hand script.

These judges tended to focus on the incorrect use of tenses. They also referenced mistakes more often, either by (a) commenting or (b) verbalising corrections. For example, (a) a *t-experienced* judge said: *She sayd wrong past and wrong…All mistakes in wrong, uh, tense… wrong verb forms and so on*, and (b) a *rater* said: *I write you uh, so I'm writing it should be.*

However, these judges continued to apply the philosophy of range over accuracy, as verbalised by two *t-experienced* judges: *I'm going for the right one because although it's not always, uh, correct, it is attempted, uh, the right or appropriate tenses are attempted in most cases which they're not on the left*; and, more succinctly, *range prioritised over accuracy*.

The *comparative scan* judges typically read out verbs, sometimes with comments:

> *I did it with/ I like it/ It's present I don't like it present, so half present, half past on the left* (*who?*); or without: [reading from the left script] *We cleaned the schoolpark. We put papers/ I liked it/* [reading from the right script] *it was great/ I liked most that/ The right.*

Unusually in the TAP data, this *comparative scan rater* judge noted expectations: *So they're supposed to have a lot written in past or present perfect at the end. We'll see*. This behaviour, coupled with the *comparative scan t-experienced* judge's interrogative style, suggests that these two judges were actively seeking out relevant content in the scripts. They also continued to prioritise range over accuracy for this dimension as the *t-experienced* judge noted: *At least she tries or he tries to form an if sentence*.

Once again, *comparative scan* was the quicker method (Table 2), and for Grammar these judges were among the most reliable (Sickinger, 2023). It seems reasonable to conclude that scanning scripts for key grammatical features is an appropriate strategy for evaluating grammar in *CJ-Dimensions*.

5.2.2.4. Vocabulary. When evaluating Vocabulary, as they did for Grammar, judges showed an appreciation of test-takers' range and their accuracy, noting mistakes, typically misspellings. Again, as in the other CJ sessions, when evaluating similar scripts, judges searched for features to differentiate the scripts, as exemplified by a *t-experienced* judge: *These texts are quite equal, but I would prefer the left because there are more different verbs used in a in a nicer way.*

For Vocabulary, as for Grammar, *read and compare* judges (Fig. 5) usually started their evaluation with step 2 (reading the left script); however, for this dimension, they tended to read more. Some judges also commented on the language. For example, a *t-experienced* judge carefully notes mistakes:

> *It gives an accident accident spelled correctly difficult word outside accros spelt wrong from my door and two peoples instead of people were ringing at my door. I was really scared and when he was coming inside I fightet with him instead of fought.*

Judges rarely referenced other dimensions whilst judging Vocabulary; however, sometimes, like this *rater* judge, they clearly felt they couldn't be ignored: *Some good words but not so, the grammar is atrocious.* However, it is unclear for examples such as this whether judges simply verbalised their awareness or whether such instances altered judges' decision-making.

The two *comparative scan* judges followed their usual approach (Fig. 6). The *rater* tended to read key words from each script in turn but occasionally moved between the two scripts. The *t-experienced* tended to focus on each script in turn for the long scripts, and move between the pair of scripts more for the short scripts, again reading out key words. On occasion, this judge also questioned the texts, as before: *Where are the good words*?

The difficulty of choosing between two similar texts is clear in the verbalisation of the *comparative scan t-experienced* judge: *I can't compare them.* In this instance a tiny difference, the spelling of "project" was used to make the decision: *with k and with c, maybe the only, only difference here.*

Once again, *comparative scan* was the quicker method (Table 2), and for Vocabulary these judges were very reliable (Sickinger, 2023). As with Grammar, it seems that scanning scripts for key language is an appropriate strategy for evaluating Vocabulary in *CJ-Dimensions*.

6. Discussion

This study aimed to provide insight into the validity of CJ for the evaluation of L2 writing performances by focusing on the decision-making approaches adopted by judges and the criteria they considered during that decision-making. The first phase – presented in detail in Sickinger et al. (2025) – primarily focused on quantitative data and the reliability of the method. Relevant to the current study, it also asked the 27 participants to respond to questionnaires which included items relating to the criteria they had considered when making their judgements. The second, qualitative phase – the main focus of this current study – asked eight participants to conduct think-alouds whilst judging performances using CJ.

For *CJ-Holistic*, the questionnaire data, though self-reported, suggest that most judges considered construct-relevant criteria with broad aspects of each dimension being mentioned by many participants. This consideration of construct-relevant criteria was also evident in the TAP data either explicitly through verbalisations or through the text selected to be read aloud. However, the specific criteria used to justify a decision between two scripts were not always clear from the TAPs, and further research into this aspect of CJ is indicated. Nevertheless, the *CJ-Holistic* think-aloud verbalisations often referred to key features of writing. For example, for Task Achievement, the focus was on the content points; for Coherence & Cohesion, the focus was on the overall structure of the script and ideas along with sentence connectors, sentence starters, and paragraphs; for Grammar, tenses were the most important feature; and for Vocabulary, the range of words and accuracy (particularly spelling) were key. The TAPs suggest these key features were usually sufficient to decide between two performances. However, it is not clear whether judges sought to give equal weighting to the four dimensions in their evaluations, as was initially intended by the test designers when planning the analytic rating scale (Gassner et al., 2011). Indeed, the *CJ-Holistic* transcripts showed judges taking the most time on the content points required for Task Achievement. It is unclear whether this dominance of aspects of Task Achievement in the transcripts simply reflects the time needed to check that the test-taker had mentioned and/or elaborated upon the content points or whether it indicates the judges giving more weight to Task Achievement.

For *CJ-Dimensions,* participants listed more detailed features of writing for each dimension when completing the self-reported questionnaires than they had for *CJ-Holistic*. Trained *raters* noted considering more varied aspects of writing than their untrained *experienced* peers did, probably reflecting the in-depth training they had received. Similarly, the *CJ-Dimensions* TAP data showed judges typically referencing dimension-specific features, with those trained as raters considering a broader range. These findings reflect the results of Thwaites et al.'s (2025) investigation that trained raters were "significantly more specific in their comments" (p. 10).

Judges also took advantage of CJ's open nature, often seen as a positive feature of the method (Lesterhuis et al., 2022; Pollitt, 2012a, 2012b; Wheadon et al., 2020), referring to scripts being exciting to read, for example. On occasion, some TAP transcripts (from both *read and compare* and *comparative scan* judges) suggested that judges were not always clearly separating the dimensions when judging *CJ-Dimensions*. This difficulty in clearly defining boundaries between dimensions was noted in a previous study involving similar scripts rated using the analytic rating scale (Pibal et al., 2018).

The TAP data clearly showed two different approaches to conducting CJ: *read and compare*, similar to rating, and *comparative scan*, seemingly more tailored to the CJ process. Both methods were reliable (Sickinger, 2023); however, *comparative scan* was considerably faster (Table 2). Arguably, *comparative scan* is a condensed version of *read and compare,* where only the last step 4 (compare pair) from *read and compare* is used. The differences between the two approaches appear to be the time spent searching for textual evidence of criteria, the amount of evidence found, and how the evidence is gathered (e.g., by reading the entirety of the scripts in *read and compare* or by skimming for key features in *comparative scan*). The TAPs suggest that once *read and compare* judges finished reading the scripts, they often began to consider the performances in a positive manner, seeking out similarities between scripts (e.g., correct tense usage) before considering differences (e.g., a greater range of tenses). The decision would then be that one script was better or that one lacked quality and was therefore the worse of the pair. The *comparative scan* judges followed a similar process but in a quicker, more direct way.

Generally, more complex decisions took longer for the judges. Previous studies have also shown that decisions between two similar performances are more difficult and take longer for judges than decisions between two performances of obviously different quality (Gijsen et al., 2021; van Daal et al., 2017). In addition, previous studies have noted that judges experience more difficulties when judging longer performances (Thwaites et al., 2024).

Although the judges spoke with little hesitation and in parallel to the decisions they were making, the TAPs are incomplete records (Barkaoui, 2011). Often, the process seemed quick and easy for the judges, and decisions were at times made without accompanying verbalisations (Brooks, 2012). Therefore, although two decision-making approaches could be clearly identified (RQ2), whether decisions were made on appropriate criteria is, at times, less clear from the TAPs, although supported by the questionnaire data (RQ1).

6.1. Validity

With this study, we argue that the application of Knoch and Chapelle's (2018) validity framework can be extended to CJ to support the introduction of CJ as an evaluation tool. In this framework, inferences made regarding test scoring are supported by claims which are evidenced by data. These are discussed in turn below.

Sickinger et al. (2025) provided data informing the *evaluation inference* that CJ scaled scores can differentiate between performances in a comparable manner to ratings and that judges can judge scripts reliably. The Phase 1 data supported this, as CJ scores from teachers experienced in classroom evaluation and/or rating were reliable and correlated significantly with expert raters' scores. The present study allows to build the argument further by informing the *generalization* and *explanation inferences*. The questionnaire data from Phase 1 and the Phase 2 TAPs presented above – particularly from *CJ-Dimensions*, which encourages a more detailed examination of the test-takers' language proficiency – provide support to the claim that judges make consistent decisions using construct-relevant criteria. Support for the *extrapolation inference* – claiming that the judgement criteria reflect the evaluation criteria – can be found both in the fact that CJ requires a large pool of judges, potentially allowing for a broader coverage of writing features, and in the present study's questionnaire and TAP data, which show that the judges as a group considered a wide definition of language quality. The *decision inference* requires the scaled scores from CJ to be suitable for reporting purposes and the decisions that rest upon the awarded scores. In support of this claim, the new *CJ-Dimensions* approach in particular allows more detailed feedback and encourages judges to focus on specific aspects of the tested language in turn. Finally, the *consequence inference* claims that the CJ process promotes the interests of the test users. Supporting this claim is CJ's potential to involve many teachers in the judging process, which could improve washback from the writing test to teaching and learning in the classroom, particularly with the more detailed feedback possible with *CJ-Dimensions*.

7. Conclusion

To our knowledge, this is the first empirical study investigating the scoring validity – a fundamental principle of test quality – of the Comparative Judgement method for L2 writing in the context of young learner assessment. Another innovative element of our study concerns the novel dimension-based approach. Our findings show that judges used construct-relevant criteria in their decision-making, thus supporting scoring validity in this context, where judges are familiar with rating and/or assessing classroom texts and are reminded of the target construct via crib sheets. Scoring validity is also supported by the possibility of providing feedback to test-takers on their performance. We also observed judges employing one of two strategies during decision-making: *read and compare*, which reflects the traditional rating method, and *comparative scan*, a fast and efficient approach adapted to CJ.

In drawing conclusions, we want to acknowledge this study's limitations. Further research is recommended given the TAP verbalisation limitations and the small scale of the study, in particular the limited number of judges who demonstrated the *comparative scan* decision-making approach. We also note the effect of similarity of the paired performances on the difficulty of a judge's decision. It may be the case that different criteria may be used by judges depending on their perception of the similarity or difference of the performances. This topic warrants further research.

This study used an argument-based validation approach, presenting evidence to support claims made about the context of use of the test and a possible new scoring approach, taking Knoch and Chapelle's (2018) framework as a model. The framework allowed the expectations from the test's context of use – particularly the expectation of clarity regarding the skills being judged – to inform an investigation of how CJ, usually applied holistically, can be adapted to consider separate dimensions of L2 writing performance. The study therefore fills a research gap regarding vital validity aspects of the CJ method. A key takeaway message for those considering CJ to evaluate L2 writing concerns the value of a dimension-based approach and the use of crib sheets to remind judges of the intended construct. Also, CJ merits a different decision-making approach from traditional rating, and *comparative scan* seems to be successful and therefore worth encouraging.

Endnotes

1. The analytic rating scale is available at https://www.iqs.gv.at/_Resources/Persistent/0d01b5451f07d44135bf4957f09c5f6d32ff9e7a/Assessment_Scale_E8_2019.pdf
2. In Phase 1 of the study (see Fig. 2), a counter-balanced design was employed for the quantitative examination of the three methods of evaluation: *CJ-Holistic*, *CJ-Dimensions*, and the traditional rating method. The order of evaluation method was found to have no effect on results (Sickinger, 2023). Because of this outcome, along with the involvement of only eight participants, a counter-balanced design was considered unnecessary for the Phase 2 study reported here.
3. Initial plans were for the TAP data to be coded for quantitative analysis. However, judges differed significantly not only in their approaches to CJ (as explained in Section 5.2, where two distinct styles of CJ decision-making processes are detailed) but also in their approach to think-aloud. During the think-aloud sessions, some judges read all text clearly out loud and verbalised extensively while others appeared to use emphasis or text selection to support their judgements, and yet others read silently and made minimal verbalisations. Therefore, to retain all useful insights, however verbalised, a narrative approach focusing on patterns was used.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Rebecca Sickinger: Writing – review & editing, Writing – original draft, Visualization, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **John Pill:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Tineke Brunfaut:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Rebecca Sickinger reports administrative support was provided by the *Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen*. Rebecca Sickinger reports a relationship with the *Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen* that includes: consulting or advisory. John Pill reports a relationship with the *Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen* that includes: consulting or advisory. Tineke Brunfaut reports a relationship with the *Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen* that includes: consulting or advisory.

Acknowledgements

We would like to thank the participants who took part in this study, and the *Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen* (IQS) for providing us with Year 8 EFL writing exam scripts and the exam's analytic rating scale.

Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.asw.2025.100986.

Data availability

The data that has been used is confidential.

References

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: an empirical study of their veridicality and reactivity. *Language Testing, 28*(1), 51–75. https://doi.org/10.1177/0265532210376379

Bartholomew, S. R., & Jones, M. D. (2022). A systematized review of research with adaptive comparative judgment (ACJ) in higher education. *International Journal of Technology and Design Education, 32*, 1159–1190. https://doi.org/10.1007/s10798-020-09642-6

Bramley, T. (2022). Editorial – the CJ landscape. *Research Matters, 33*, 5–9. ⟨https://www.cambridgeassessment.org.uk/Images/research-matters-33-a-summary-of-ocrs-pilots-of-the-use-of-comparative-judgement-in-setting-grade-boundaries.pdf⟩.

Brooks, V. (2012). Marking as judgment. *Research Papers in Education, 27*(1), 63–80. https://doi.org/10.1080/02671520903331008

Buckley, J., Seery, N., & Kimbell, R. (2022). A review of the valid methodological use of adaptive comparative judgment in technology education research. *Frontiers in Education, 7*, 1–6. https://doi.org/10.3389/feduc.2022.787926

Chambers, L., & Cunningham, E. (2022). Exploring the validity of comparative judgement: do judges attend to construct-irrelevant features? *Frontiers in Education, 7*, 1–14. https://doi.org/10.3389/feduc.2022.802392

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment.* SAGE.

Curcin, M., Howard, E., Sully, K., & Black, B. (2019). *Improving awarding: 2018/2019 pilots*. Ofqual. ⟨https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/851778/Improving_awarding_-_FINAL196575.pdf⟩.

Freeman, D. (2001). Second language teacher education. In R. Carter, & D. Nunan (Eds.), *The Cambridge guide to teaching English to speakers of other languages* (pp. 72–79). Cambridge University Press. ⟨https://ia600500.us.archive.org/8/items/ilhem_20150321_1654/%5BDavid_Nunan,_Ronald_Carter%5D_The_Cambridge_guide_t.pdf⟩.

Gassner, O., Mewald, C., Brock, R., Lackenbauer, F., & Siller, K. (2011). *Testing writing for the E8 Standards.* BIFIE.

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*, 1–11. https://doi.org/10.3389/feduc.2020.582800

Green, A. (1998). *Verbal protocol analysis in language testing research. A handbook*. Cambridge University Press.

Gu, Y. (2014). To code or not to code: dilemmas in analysing think-aloud protocols in learning strategies research. *System, 43*, 74–81. https://doi.org/10.1016/j.system.2013.12.011

Holmes, S., Black, B., & Morin, C. (2017). *Marking reliability studies 2017: Rank ordering versus marking – which is more reliable? Ofqual.* ⟨https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/859250/Marking_reliability_-_FINAL64494.pdf⟩.

Hughes, K., Hardy, J., Galloway, R. K., Rhind, S., McBride, K. L., & Donnelly, R. (2016). *Ask, answer, assess: peer learning from student-generated content.* ⟨https://www.advance-he.ac.uk/knowledge-hub/ask-answer-assess-peer-learning-student-generated-content⟩.

Jones, I., & Davies, B. (2024). 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. (2015). The problem of assessing problem solving: can comparative judgement help? *Educational Studies in Mathematics, 89*(3), 337–355. https://doi.org/10.1007/s10649-015-9607-1

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

Knoch, U., & Chapelle, C. (2018). Validation of rating processes within an argument-based framework. *Language Testing, 35*(4), 477–499. https://doi.org/10.1177/0265532217710049

Laming, D. (2004). Human judgment: the eye of the beholder. *Thomson Learning*.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310
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*, 1–10. https://doi.org/10.3389/feduc.2022.823895
Lesterhuis, M., van Daal, T., van Gasse, R., Coertjens, L., Donche, V., & 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), 1–22. https://doi.org/10.17239/L1ESLL-2018.18, 01.02.
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*, 49–71. https://doi.org/10.1007/s40841-020-00163-3
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
Oates, T. (2022). Foreword. *Research Matters, 33*(4). ⟨https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/issue-33-spring-2022/⟩.
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
Pibal, F., Sigott, G., & Cesnik, H. (2018). The role of error in assessing English writing in the Austrian Educational Standards Baseline Test. In G. Sigott (Ed.), *Language testing in Austria: Taking stock = Sprachtesten in Österreich: Eine Bestandsaufnahme* (pp. 421–444). Peter Lang.
Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education, 22*(2), 157–170. https://doi.org/10.1007/s10798-011-9189-x
Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354
Rotaru, V. (2022). Using adaptive comparative judgement for assessing GCSE History NEA responses: research report. *Qualifications Wales*. https://doi.org/10.13140/RG.2.2.30614.83524
Shaw, S. D., & Weir, C. J. (2007). *Examining writing: research and practice in assessing second language writing.* Cambridge University Press.
Sickinger, R. (2023). An exploration of comparative judgement for evaluating writing performances of the Austrian year 8 test for English as a Foreign Language. *[Doctoral dissertation, Lancaster University]*.
Sickinger, R., Brunfaut, T., & Pill, J. (2025). Comparative judgement for evaluating young learners' EFL writing performances: reliability and teacher perceptions of holistic and dimension-based judgements. *Language Testing, 42*(2), 137–166. https://doi.org/10.1177/02655322241288847
Siller, K., Lackenbauer, F., Berger, A., Sickinger, R., & Kulmhofer-Bommer, A. (2019). *Testing writing for the E8 Standards – Technical Report*.
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
Thurstone, L. (1927/1994). A law of comparative judgment. *Psychological Review, 101*(2), 266–270. https://doi.org/10.1037/0033-295X.101.2.266
Thwaites, P., Jadoulle, P., & Paquot, M. (2025). Comparative judgment in L2 writing assessment: reliability and validity across crowdsourced, community-driven, and trained rater groups of judges. *Assessing Writing, 65*, 1–15. https://doi.org/10.1016/j.asw.2025.100937
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*, 1–14. https://doi.org/10.1016/j.asw.2024.100843
Thwaites, P., & Paquot, M. (2024). Comparative judgement for advancing research in applied linguistics. *Research Methods in Applied Linguistics, 3*, 1–13. https://doi.org/10.1016/j.rmal.2024.100142
van Daal, T. (2020). Making a choice is not easy?! Unravelling the task difficulty of comparative judgement to assess student work [Unpublished doctoral dissertation]. *University of Antwerp*.
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 Kemp, 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*, 1–13. https://doi.org/10.3389/feduc.2017.00044
Walland, E. (2022). Judges' views on pairwise comparative judgement and rank ordering as alternatives to analytical essay marking. *Research Matters, 33*, 48–67. ⟨https://www.cambridgeassessment.org.uk/Images/research-matters-33-judges-views-on-pairwise-comparative-judgement-and-rank-ordering-as-alternatives-to-analytical-essay-marking.pdf⟩.
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
Whitehouse, C. (2013). Testing the validity of judgements about geography essays using the adaptive comparative judgement method. *Centre for Education Research and Policy*. ⟨https://filestore.aqa.org.uk/content/research/CERP_RP_CW_24102012_0.pdf?download=1⟩.

Rebecca Sickinger holds a PhD in Linguistics from Lancaster University (UK), with specialisation in language testing. Rebecca works in Austria and has taught in a range of educational contexts from early years to university with a focus on lower-secondary EFL teaching. She has also worked with the Austrian administrative body responsible for national lower-secondary testing of EFL standards.

John Pill is a lecturer at Lancaster University (UK), where he teaches and researches language testing. His research interests include testing language for specific purposes, with particular focus on healthcare and academic settings, the scope and definition of language constructs in different contexts, speaking assessment, language assessment literacy, test users' perspectives, and test impact.

Tineke Brunfaut is Professor of Applied Linguistics at Lancaster University (UK), where she specializes in language testing, second language listening, reading, and integrated skills. She has published widely on cognitive and affective factors and processes in language testing, methodology in language testing research, and language test development. She also regularly conducts language test training and consultancy work for professional and educational bodies around the world.