Original Manuscript

LANGUAGE
TESTING

Language Testing
1–28

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02655322251338605
journals.sagepub.com/home/ltj
Sage

Assessing intelligibility as conceptualised within the CEFR-companion volume (CV) framework using Adaptive Comparative Judgement

Jingwen Wang
University of Southampton, UK

Ying Zheng
University of Southampton, UK

Abstract

Two pivotal constructs, intelligibility (listeners' actual understanding) and comprehensibility (listeners' perceived difficulty of understanding), have dominated second language (L2) pronunciation research, marking a shift away from an emphasis on nativeness. The 2020 Companion Volume to the Common European Framework of Reference for Languages (CEFR-CV) presented a revised phonological competence scale, integrating both dimensions into a new definition of intelligibility. However, effective measurements to assess this refined construct are still lacking. This study explores the potential of Adaptive Comparative Judgement (ACJ) in measuring intelligibility as conceptualised by the CEFR-CV. ACJ employs judges who evaluate two stimuli based on a holistic criterion, selecting the better one. Through a collection of such binary decisions, judges' evaluations are statistically analysed, producing standardised estimates for each stimulus. Twelve judges assessed speech samples from 30 L1 (first-language)-Mandarin speakers of English performing four sentence repetition tasks. Incorporating Think Aloud Protocol (TAP) into the judgement process, the study combined quantitative and qualitative analyses, providing evidence for the efficacy of ACJ in measuring L2 speech. The findings, discussed in the context of existing literature on intelligibility and comprehensibility, unveil future research on the use of ACJ in L2 pronunciation assessment in further elucidating the intelligibility construct as defined by the CEFR-CV.

Keywords

Adaptive Comparative Judgement (ACJ), CEFR-CV, comprehensibility, intelligibility, pronunciation assessment, second language speech

Corresponding author:
Jingwen Wang, Languages, Cultures and Linguistics, University of Southampton, B65, Avenue Campus, Highfield Road, Southampton SO17 1BF, UK.
Email: Jingwen.Wang@soton.ac.uk

Introduction

Following Munro and Derwing (1995a, 1995b) and Derwing and Munro (1997), three related yet partially independent constructs have become dominant in second language (L2) pronunciation research: intelligibility, comprehensibility, and accentedness. In line with their definition, intelligibility refers to the extent to which a speaker's message is actually understood by a listener, comprehensibility is defined as the listener's perceived difficulty in understanding specific utterances, and accentedness is the listener's perception of the degree to which speech deviates from the native speaker norms. Munro and Derwing (1995a) found that intelligibility and comprehensibility are strongly and positively interrelated, while accentedness has a weak to moderate negative correlation with intelligibility, and a moderate to significant negative correlation with comprehensibility. This indicates that speakers with pronounced foreign accents may be intelligible but can remain less comprehensible due to the listening challenges they pose. Consequently, L2 pronunciation instruction now prioritises intelligibility as the primary goal and, to a lesser extent, comprehensibility, recognising that both aspects are crucial for successful communication (Thomson, 2017).

In the realm of L2 pronunciation, intelligibility measures can take various forms. Listeners might be asked to transcribe every word of utterances (Nagle et al., 2023), complete cloze exercises with omitted content words (Kennedy & Trofimovich, 2008), or identify specific segmentals within minimal pairs (Shehata, 2024). The focus of these assessments has been limited to what Kang et al. (2018) have referred to as the "local level" (p. 116), reflecting listeners' understanding of individual words rather than the ideas being conveyed. In contrast, Likert-type scales require listeners to evaluate their degree of comprehension at the global level on a continuum (e.g., from "hardly intelligible" to "perfectly intelligible"; Tsubota et al., 2004). Scalar ratings of intelligibility have been controversial because it is unclear how they differ from measures of comprehensibility (Kang et al., 2018). Comprehensibility has largely been measured following Munro and Derwing's (1995a) initial operationalisation, where listeners use a 9-point Likert-type scale to rate how difficult a stretch of speech is to understand, from 1 (*extremely easy*) to 9 (*extremely hard*). Likewise, the measure for assessing accentedness construct, such as 9-point scales, often mirror those used to evaluate comprehensibility.

The 2020 Companion Volume to the Common European Framework of Reference for Languages: Teaching, Learning, and Assessment (CEFR-CV) introduced significant revisions to its 2001 "Phonological Control" scale. It now conceptualises intelligibility as "accessibility of meaning for interlocutors, also covering the interlocutors' perceived difficulty in understanding (normally referred to as 'comprehensibility')" (Council of Europe, 2020, p. 133). For clear references throughout this paper and to distinguish it from the intelligibility and comprehensibility as termed by Munro and Derwing, we will henceforth refer to this definition as "CEFR-CV intelligibility." This integration of traditional intelligibility and comprehensibility concepts into a unified construct aligns more closely with the chief goal of the CEFR-CV, which is to prioritise communicative competence in L2 learning, emphasising both phonetic clarity and communicative ease. Moreover, it aims to better meet the practical demands of contemporary L2 pronunciation instruction and to standardise terminology across rubrics used in the various tests and assessments. However, this broader definition adds complexities, especially in developing effective measurements. Current approaches to evaluating either intelligibility or comprehensibility fall short of the comprehensiveness required for direct application or adaptation to this new framework. Meaningfully merging existing measures of both traditional constructs also proves challenging. Such adaptations or mergers might oversimplify the new construct and consequently, leading to overlooking its essential influential factors. Thus, there is an urgent need for a promising measure that can evaluate this expanded definition across both research and assessment contexts.

This study is a methodological exploration of the potential of Adaptive Comparative Judgement (ACJ) for measuring CEFR-CV intelligibility. Originating in the psychophysical domain, ACJ adapts Thurstone's (1927) "Law of Comparative Judgement" (CJ), initially used to scale the perceived magnitude of physical stimuli such as "loudness." In this approach, judges evaluate pairs of stimuli and make dichotomous decisions about their relative quality. The outcomes from these comparisons, made by a panel of judges, are statistically modelled to produce standardised quality estimates for each stimulus. These estimates are then used to position each stimulus along a continuum, resulting in either a logit-based scale, or a rank order that categorises all stimuli from best to worst performance. Acknowledging the favourable capability of CJ-based measures in evaluating complex constructs such as written or spoken production (Bisson et al., 2016; Kelly et al., 2022), we hypothesised that ACJ could be a promising approach to assessing CEFR-CV intelligibility. To the best of our knowledge, studies that have implemented ACJ in L2 pronunciation are limited. To ensure the integrity of this research, Pollitt's (2012) quality control standards for reliability of ACJ were applied. In addition, Weir's (2005) socio-cognitive validation framework was used to evaluate various facets of the validity of ACJ.

Literature review

Measurements of intelligibility and comprehensibility

As English continues to dominate as the global lingua franca, native speakers have become a minority (Eberhard et al., 2023). This shift has redirected L2 research on accentedness from the elusive goal of acquiring a native-like accent to achieving a high level of communicative proficiency, emphasising two pivotal constructs: intelligibility and comprehensibility. Intelligibility is often measured through orthographic transcription tasks, which can take on varied forms. For instance, scores can be calculated by the percentage of words correctly transcribed from full utterances (Chau et al., 2022; Nagle et al., 2023) or by counting only accurately transcribed content words (Kennedy & Trofimovich, 2008). The adequacy of transcription tasks in capturing the essence of intelligibility has been debated, as listening inherently involves a top-down process that assists in recognising mispronounced or unintelligible words by inferring their meanings from the overall gist. However, intelligibility essentially emphasises a bottom-up process in which comprehension is achieved by precisely identifying each word (Thomson, 2017). Alternatively, forced-choice identification tasks, which require listeners to distinguish between phonemes within minimal pairs (e.g., "cub" vs. "cup"), have been adopted

(Shehata, 2024). While this measure offers a more straightforward evaluation of intelligibility, it typically assesses only at the word level. More recently, Kang et al. (2022) introduced a hybrid approach to measure intelligibility using three tasks: a scalar rating from 1% to 100%, a sentence transcription task, and a phrase transcription task. The two transcription tasks were automatically graded using a fuzzy string-matching approach to compare the transcriptions to a gold standard transcript, with the final intelligibility score derived by averaging the results from all three tasks.

In addition, some researchers have measured intelligibility using Likert-type scales, where listeners rate their global understanding of speech on a scale from “hardly intelligible” to “perfectly intelligible” (Tsubota et al., 2004), as Isaacs (2008) suggests that understanding the gist of the message is often more important than recognising every word. However, scalar ratings of intelligibility are controversial because it is unclear how this differs from the operationalisation of comprehensibility (Kang et al., 2018). In contrast, comprehensibility is most often measured using a 9-point Likert-type scale (1 = *easy to understand*, 9 = *extremely difficult to understand*, Uchihara et al., 2023). However, variations exist, including reverse scales, giving lower points to speech that is difficult to comprehend (Crowther et al., 2016), five 7-point bipolar scales (Kang, 2010), and extended 100-point sliding scales (Nagle et al., 2022). Concerns have been raised about the accuracy of scalar ratings for measuring comprehensibility. Listeners may not accurately gauge the difficulty of understanding, as they often report a reduced perception of difficulty once they grasp the overall meaning (Thomson, 2017).

Aspects influencing intelligibility and comprehensibility

Although the constructs of intelligibility and comprehensibility lack universally agreed-upon definitions, existing literature provides a foundation for considering features that contribute to speech quality. These aspects have been extensively used to define L2 speaking assessment rubrics. Research has consistently shown that both segmental (vowels and consonants) and suprasegmental (prosodic features such as intonation, rhythm, and pitch) features are important for the speech constructs (Huensch & Nagle, 2021; Kang et al., 2022; Nagle et al., 2023; Yenkimaleki & van Heuven, 2024). For example, Munro and Derwing (1995a) observed that phonemic and intonation features, along with grammatical scores, influence comprehensibility and accent ratings but do not correlate strongly with intelligibility. In a contrasting study, Huensch and Nagle (2021), who replicated and extended Munro and Derwing’s (1995a) work, concluded that phonemic errors, goodness of prosody, and grammatical errors are significant predictors of both intelligibility and comprehensibility, whereas foreign accents primarily associated to phonemic and grammatical errors.

These observations were partially supported by Trofimovich and Isaacs (2012) and Crowther et al. (2016). Trofimovich and Isaacs noted that accent is primarily linked to phonological features such as segmentals, word stress, and rhythm, while comprehensibility is more closely related to syntactic features like lexical richness and grammatical accuracy. Conversely, Crowther et al. (2016) argued that segmentals, word stress, and lexical and grammatical usage greatly affect both comprehensibility and accentedness. It should be noted that associations between identified features and constructs vary across studies, likely due to the different types of tasks used to collect speech samples. Trofimovich and Isaacs (2012) elicited non-spontaneous speech via a read-aloud task, whereas Crowther et al. (2016) used picture narratives to obtain more flexible, extemporaneous speech. Similar variations have been demonstrated by Trofimovich et al. (2009), who employed a sentence repetition task to measure comprehensibility and found that it is primarily influenced by pronunciation accuracy and fluency measures. In addition, the choice of measures used to assess these global pronunciation constructs also impacts the correlations between features and constructs. For instance, Kang et al. (2018) compared five intelligibility measures aiming to identify the most promising one. They found that phonological variables, such as segmental, prosodic, and temporal features, contribute differently to the outcome scores, showing varied associations with each intelligibility measure.

Moreover, Huensch and Nagle (2021) found that a faster speech rate has a positive correlation with both comprehensibility and accentedness. This findings was further confirmed by Choi and Kang (2023), who investigated the relationship among suprasegmentals features, fluency, and scores on a paired discussion task in Cambridge English Language Assessment (a high-stakes English speaking test). They claimed that speech rate metrics (syllables per second, articulation rate, phonation time ratio) and pitch features (pitch range, tone choices, pitch concord) as significant predictors of overall scores on the paired speaking tasks. Faster speech rates, a wide pitch range, and appropriate use of pitch to convey meaning were linked to successful task performances. Similarly, Kang (2010) and O’Brien (2014) associated slow speech rate, frequent pausing, and reduced pitch range with to decreased comprehensibility. Further clarifying this line of research, Trofimovich et al. (2022) concluded that a series of linguistic features affect comprehensibility, including pronunciation (segmental and suprasegmental), fluency (speech rate, pauses, and self-repair), and lexicogrammar (vocabulary diversity, grammatical accuracy, and complexity). In addition, Isaacs and Trofimovich (2012) and Tsunemoto and Trofimovich (2024) emphasised the importance of discourse-level features, stating that well-organised and structured discourse (e.g., coherence) is crucial for comprehensibility.

In summary, L2 pronunciation researchers have identified a series of linguistic features that contribute to the constructs of intelligibility and comprehensibility. These include segmentals, suprasegmentals, fluency, lexical-grammar, and discourse-level organisation. Moreover, listeners’ comprehension of L2 speech appears to be influenced by both their L1 background and exposure to specific L2 accents. Researchers have consistently found that the correlation between the degree of a foreign accent and intelligibility loss varies among listeners from different L1 backgrounds. For example, Pérez-Ramón et al. (2022) discovered that Spanish listeners experience less loss of intelligibility when hearing speech from English-Spanish bilinguals, compared to native English listeners. In contrast, the Czech cohort reported the greatest loss in intelligibility. However, regarding L1 familiarity, some studies report that a shared L1 background between speakers and listeners can enhance comprehension (Saito & Shintani, 2016), while others demonstrate no effect from shared L1 benefits (Major et al., 2002). These divergent findings can be attributed to variations in how constructs are operationalised, the composition of samples, and the characteristics of the speech used across studies.

Researchers maintain a neutral viewpoint, stating that both matched and mismatched interlanguage intelligibility benefits are present between L2 speakers (Miao & Kang, 2023). On the other hand, it is generally reported that native listeners understand more words and sentences spoken by familiar than unfamiliar accented speakers (Kennedy & Trofimovich, 2008). Moreover, accent familiarity does have a direct effect on comprehensibility (Miao, 2023); for instance, speakers with accents familiar to listeners might receive overall higher comprehensibility ratings (Carey & Szocs, 2024). However, differing positions, such as that of Munro et al. (2006), who suggest that accent familiarity does not correlate with either intelligibility or comprehensibility, also exist. Thus, the effects of shared L1 and accent familiarity remain unclear and are open to further discussion.

Challenges in measuring CEFR-CV intelligibility

In 2020, the CEFR substantially revised its phonological control scale with the release of its companion volume. The original 2001 single scale has now expanded into three scales: "Overall Phonological Control," "Sound Articulation," and "Prosodic Features." In addition, the 2020 CEFR-CV redefined intelligibility as the "accessibility of meaning for interlocutors, covering also the interlocutors' perceived difficulty in understanding (normally referred to as 'comprehensibility')" (Council of Europe, 2020, p. 133). This construct is now featured across the descriptors of these scales to distinguish proficiency levels. As noted, the redefined CEFR-CV intelligibility construct encompasses what Munro and Derwing term both intelligibility and comprehensibility. The reasons for expanding this definition are multifaceted. First, it aligns closely with the main objective of the CEFR-CV to improve communicative competence among L2 speakers, emphasising that both clarity of speech and ease of understanding are fundamental for communicative success. Second, although there is considerable overlap between the factors influencing both original dimensions, their theoretical separation, while valuable for research, may not meet the practical demands of L2 pronunciation instruction. The new definition thus aligns more closely with the real-world communicative demands of language learners. By focusing on the broader aspects, it might provide a more integrated framework for pronunciation teaching, encouraging L2 pronunciation teachers to focus not merely on the technical details of language production but also on practical communicative outcomes, which facilitates the development of teaching practices that address both balance linguistic accuracy and functional comprehensibility. Finally, it clarifies and standardises the terminology used and ensures consistency across different assessment scales. The terms intelligibility and comprehensibility are often used interchangeably and interpreted inconsistently or vaguely within L2 oral proficiency scales (Trofimovich et al., 2022). For instance, Band 8 of the IELTS speaking descriptors, as published by the British Council (2020), stated that L2 speech is "easy to understand throughout; L1 accent has minimal effect on intelligibility." This implies that what is actually being measured aligns more closely with Munro and Derwing's concept of comprehensibility (Trofimovich et al., 2022).

Although this refined definition offers several advantages, it also introduces additional complexities. Current measures of either intelligibility or comprehensibility in L2 pronunciation research lack the comprehensiveness required for direct application or adaptation within this new framework. For example, Huensch and Nagle (2023) suggested using comprehensibility as a proxy for intelligibility, drawing on Munro and Derwing's (1995a) finding that comprehensibility scores often serve as good predictors of intelligibility scores. They recommended employing scalar ratings to measure both constructs and commended this method for being quicker and easier to administer than traditional intelligibility measurements. However, since the dynamic interaction between intelligibility and comprehensibility remains unclear, this approach may risk misestimating the importance of factors that influence the variability in the strength of their interrelationship. Furthermore, finding a meaningful way to merge existing measures for gauging both constructs remains challenging. Kang et al.'s (2022) combined intelligibility measure could potentially be adapted for measuring CEFR-CV intelligibility by incorporating traditional comprehensibility scalar rating. However, the efficacy of its final result is uncertain, as such a merger might oversimplify or overlook crucial aspects of the new intelligibility construct. In addition, the current measures for both the intelligibility and comprehensibility constructs are implemented only in research settings; their applicability in broader contexts, such as in operational L2 speaking assessments, remains underexplored. Thus, there is a need for an effective measure for this expanded definition that can be adopted in broader practical contexts.

ACJ

Originating from Thurstone's (1927) "Law of Comparative Judgement," Comparative Judgement (CJ) was initially developed to quantify subjective properties such as the intensity of individual attitudes about societal phenomena. In this approach, judges compare randomly selected pairs (ideally covering all possible pairs) of stimuli based on a holistic criterion, making a series of binary decisions about which stimulus in each pair has better performance. These decisions are then statistically analysed using the Bradley-Terry-Luce model (Bradley & Terry, 1952; Luce, 1959), an adaption of the Rasch model, to produce standardised parameters and rankings for each stimulus. The foundational rationale of CJ is that humans are generally better at making comparative rather than absolute judgements, as isolated assessments tend to be less accurate and more susceptible to individual biases (Kelly et al., 2022). Recognising its potential, CJ has been widely adopted in the field of educational assessment since the 1990s, for example for evaluating complex work like primary students' writing (Pinot de Moira et al., 2022) and spoken language interpreting (Han, 2022). More recently, Sickinger et al. (2024) evaluated the effectiveness of standard holistic CJ, dimension-based CJ (their new criteria-based method), and traditional rubric marking for assessing young learners' L2 scripts. Their findings indicated that both CJ methods were not only reliable but also enabled judges to make decisions more quickly than with rubric marking. Furthermore, this method is considered especially relevant where predetermined criteria might not fully capture the multifaceted nature of these constructs and are prone to varied interpretations by different raters (Bisson et al., 2016; Kelly et al., 2022). In addition, CJ-based methods are able to distil the consensus among judges, thereby achieving a high level of reliability, a frequent challenge in traditional rubric marking approaches (Pollitt, 2012).

Although CJ offers significant benefits, it has faced criticism for its reliance on judge expertise, scalability issues, and its time-consuming nature. Its process builds a broadly based consensus only guided by a holistic criterion, heavily depending on judges' interpretations based on their knowledge. While an "explicit" theoretical framework that links judge expertise with CJ requirements is currently lacking (Kelly et al., 2022), studies employing "qualified judges," with relevant field expertise and experience have generated satisfactory results across studies (Whitehouse & Pollitt, 2012). Moreover, research consistently shows that CJ results from both experienced and novice judges are nearly equally accurate (Han, 2022). In addition, CJ typically requires minimal training (Jones & Davies, 2024). Advances in technology have enabled CJ implementation via web-based platforms, automating the statistical scaling process and resolving previous difficulties in scaling. Moreover, studies have demonstrated that CJ is as fast or faster than traditional marking methods (Marshall et al., 2020; Sahin, 2021). In fact, a satisfactory level of Scale Separation Reliability (SSR) and parameter estimation can be achieved with relatively few pairings (Verhavert et al., 2019), greatly enhancing efficiency in real-world CJ applications.

Pollitt (2012) extended traditional CJ by introducing its adaptive variant, which refines the stimulus pairing process. ACJ further improves judgement efficiency by pairing stimuli of similar estimated quality, thereby maximising the value of each comparison. Pollitt noted that comparisons between stimuli of similar quality generate more "information" than those which differ greatly. Employing an algorithm, ACJ first estimates the value of each stimulus after initial judgements, then adaptively selects similar pairs for further comparison. Pollitt emphasised that ACJ retains all the advantages of CJ, including high reliability, validity, effective reduction of biases among judges, and suitability for assessing complex constructs or performances. Despite its strengths, the reliability of ACJ, specifically its SSR, has been questioned for potential "inflation." SSR is a type of reliability index specifically used in CJ-based contexts, and is comparable to reliability coefficients in classical test theory to demonstrate interrater consistency (equivalent to Cronbach's alpha, see Wright & Masters, 1982). It represents the ratio of true variance to observed variance (Pollitt, 2012). Bramley and Vitello (2019) investigated the reliability differences between ACJ and CJ methods and found that adaptivity indeed inflates reliability, with SSR reducing under non-adaptive and all-play-all conditions (where judges evaluate all possible pairwise comparisons). Furthermore, they noted that adaptivity does not directly impact the validity calculations of ACJ. Addressing these debates, Rangel-Smith and Lynch (2018) proposed a solution that moderates adaptivity levels while considering the Standard Deviation (*SD*) of the assessed items and the number of judgement rounds. As reported by Kimbell (2021), this approach received Bramley's agreement for its effectiveness in eliminating bias and solidifying reliability, which led to its implementation in RM Compare (https://compare.rm.com, Digital Assess, n.d.), a widely adopted ACJ online platform.

Compared to CJ, ACJ has had limited application in educational assessment research, particularly in language testing and assessment, where it is primarily used for evaluating textual rather than spoken outputs. Smith (2020) revealed the advantages of ACJ in assessing General Certificate of Secondary Education (GCSE) for English creative writing, noting that it enabled teachers to make reliable judgements more swiftly than traditional methods, thereby significantly enhancing their efficiency. Similarly, Paquot et al. (2022) evaluated the reliability and efficiency of crowdsourced ACJ in assessing text-based L2 proficiency assessments. They found strong reliability and a medium correlation with the original proficiency levels of the essays. This robustness was mirrored by Sherman et al. (2022), who utilised ACJ to assess students' work integrating design thinking into an English composition course, exploring progresses in students' rhetorical awareness. Although research on spoken stimuli is scarce, the potential of ACJ in evaluating spoken products has been demonstrated. Newhouse and Cooper (2013) compared the efficacy of three assessment methods: analytical rubric marking, ACJ, and individual teacher assessments in Italian language production evaluations. Their findings confirmed the high reliability of ACJ (mean $r=0.89$) and strong correlation with analytical marking, showing its effective alignment with other established assessment approaches.

There are several reasons for this study to favour ACJ over traditional CJ. First, while ACJ shares all the strengths of CJ, particularly its suitability for measuring complex constructs, it remains underexplored, especially in its capacity to assess spoken production. Second, among various CJ-based operationalisation platforms (e.g., "NoMoreMarking"), RM Compare stands out because it uniquely supports audio/video uploads. This capability makes it more suitable for examining pronunciation production, resulting in ACJ being incorporated in this study. Finally, since concerns about the inflated SSR of ACJ have been addressed in the RM Compare system, there should be no reluctance to use and explore ACJ. Given the previous intense debates over its reliability, we enhanced our reliability evidence by incorporating split-half reliability, calculated Pearson Product-Moment correlations for results from two independent groups of judges evaluating the same set of student work. Jones and Davies (2024) advocated reporting both SSR and split-half reliability, especially when employing CJ-based measures in novel settings.

Regarding validity evidence, most existing studies confirm the validity of ACJ by correlating its results with those obtained from traditional rubric scoring (i.e., convergent validity or criterion-related validity). This study expands this approach by adopting Weir's (2005) socio-cognitive validation framework, which also examines scoring and perceived validity. This framework, previously applied in research by Han (2022) and Han and Xiao (2022), assesses whether judges rely on irrelevant criteria during evaluations and evaluates how they perceive the usefulness of ACJ based on their experiences.

Therefore, this study aims to pilot and evaluate the efficacy of ACJ in the L2 pronunciation domain, addressing a critical research gap and potentially expanding the scope of intelligibility measurement techniques within the CEFR-CV framework. Specifically, this study explores two research questions:

> *Research Question 1 (RQ1).* How reliable is ACJ in assessing the CEFR-CV intelligibility of L2 speech?
> *Research Question 2 (RQ2).* How valid is ACJ in assessing CEFR-CV intelligibility, specifically in terms of its criterion-related validity, scoring validity, and the judges' perceived validity?

Method

Participants

Speakers. Thirty Mandarin-L1 English-L2 learners (10 males, 20 females; *M* = 24.53 years, *SD* = 1.52) were recruited to provide speech samples for this study. All had begun learning English in primary school, accumulating over a decade of English language learning experience (*M* = 12.57 years, *SD* = 1.83). To ensure homogeneity in proficiency, each speaker was required to have achieved an exact overall score of 6.5 on IELTS Academic, based on their self-report. In addition, all had been engaged in postgraduate-level studies at a UK university for at least 6 months at the time of data collection (*M* = 7.50 months, *SD* = 2.31).

ACJ judges. Twelve Mandarin-L1 speakers, proficient in English, were invited to evaluate speech samples using ACJ. To control for potential confounding variables of shared L1 and accent familiarity effects on listening comprehension, we deliberately recruited judges who shared the same L1 as speakers to mitigate systematic biases arising from varying degrees of accent familiarity. This homogeneous composition of judges aimed to maintain a consistent baseline in accent recognition, thereby enhancing the internal validity of the study. While such control may limit the reflection of real-world assessment practices, it is crucial for this initial exploration of the efficacy of ACJ in the current context. By focusing solely on the impact of pronunciation features and the effectiveness of the methodology itself, this approach allows the study to specifically attribute any differences in CEFR-CV intelligibility outcomes to the variables under investigation.

The selected judges were divided into two groups of six based on their experience in L2 language teaching and assessment: an Experienced Judge (EJ) group and a Novice Judge (NJ) group. The EJ group included three university lecturers and three postgraduate research students (anonymised as EJ01–EJ06, four females and two males, *M* = 32.75 years, *SD* = 3.42). All members of this group either held or were pursuing a Ph.D. in Applied Linguistics and had a minimum of 5 years' experience teaching and assessing Chinese English L2 students (*M* = 7.50 years, *SD* = 1.87). They also had knowledge of articulatory phonetics as part of their university training. Three members of the EJ group had previous experience using ACJ, having participated in a related study (part of our series project that includes the current one) that assessed the Read-Aloud task.

In contrast, the NJ group consisted of six female master's students in Applied Linguistics (anonymised as NJ01–NJ06, *M* = 23.15 years, *SD* = 1.14). These students were enrolled in an ELT/TESOL studies programme at a UK university and had completed two semesters of coursework at the time of the study. Four of the six NJ group members had completed a module entitled "Assessment of Language Proficiency," which provided them with a theoretical foundation in L2 assessment principles and practices. Despite a lack of practical assessment experience, their academic training equipped them with a basic field understanding. None had had prior exposure to or experience with ACJ.

Rubric raters. We enlisted two highly experienced male native English speaking raters to assess speech samples using traditional rubric marking. The raters were faculty members at the same UK university with doctoral degrees in Applied Linguistics, over three decades of teaching and research experience in the field, and expertise in L2 pronunciation and speaking assessment.

Data collection

Recording session. Data collection took place in a quiet office, where speakers individually completed a sentence repetition task. Recognising that the linguistic features influencing pronunciation constructs differ across task types, we initiated an ongoing project to further investigate these variations. As mentioned earlier, a prior study showed the potential of ACJ in assessing CEFR-CV intelligibility at the word level through a controlled Read-Aloud task. Future planned work will investigate the effectiveness of ACJ in evaluating CEFR-CV intelligibility at utterance level through open-ended questions, an uncontrolled task. The task in this study, was considered semi-controlled, requiring listeners to decode and then re-encode speakers' sentential-level utterances. Moreover, it allows for the analysis of learners' morphosyntax and phonology, yielding results that correlate with other measures of L2 performance (Yan et al., 2016).

However, this task may present the limitation of speakers simply mimicking the utterance, including its phonological features. Mimicking is easier for shorter utterances, which are recalled verbatim from short-term memory, compared to medium (8 to 15 syllables) and long (16 syllables or longer) sentences that require reconstruction for production (Yan et al., 2016). Thus, to reduce the likelihood of direct mimicry in this task, medium to long sentences of varying complexity were used. Specifically, the task included four sentences (Items 01 to 04), selected from the item bank of a high-stakes L2 speaking test (EnglishScore) at B2 level, ranging from 9 to 12 words (with 11 to 18 syllables). Sentences for items 03 and 04 were longer and more complex than those for 01 and 02. Speakers viewed a video clip of a person speaking a sentence, which they could watch up to twice before recording their repetition. They also had the option to re-record their attempt (without listening to the first attempt) before proceeding to the next sentence. Each speaker typically completed the recording session in six to eight minutes. Throughout the task, verbal communication between the researcher and the participants was prohibited. Speech samples were recorded at 44 kHz using a digital voice recorder, initially saved in .m4a format and later converted to .mp3 for compatibility with the ACJ platform.

ACJ and TAP session. The ACJ was operationalised using the RM Compare system. Each judge received a unique web link, granting access to their individual evaluation interface on the platform. In the judgement sessions, judges compared pairs of audio recordings side by side, making a selection based on a holistic criterion from the CEFR-CV intelligibility definition. That is, which of the two recordings presented was more intelligible and easier to process. Preferences were indicated by clicking an "A" for the left or "B" for the right sample, with the RM Compare system automatically recording decision history. Before the formal sessions, a 20-minute training was introduced to judges who were unfamiliar with ACJ methodology and the RM Compare system. We also clarified the holistic criterion, emphasising that decisions reflecting the judges' personal interpretations of intelligibility are acceptable. This session ensured that judges understood and could correctly apply the CEFR-CV intelligibility definition, allowing them to flexibly use their expertise to evaluate the speech samples.

With 30 speakers, the total number of possible pairwise comparisons was calculated as 30 × (30 − 1)/2 = 435 for each item, resulting in 1740 comparisons across four items by six judges or 290 pairs per judge. Theoretically, CJ-based measures require numerous comparisons to ensure high reliability, as increased data improve the accuracy of estimates in the Bradley-Terry-Luce model. Furthermore, as judges perform more comparisons, they become more familiar with the tasks and materials, reducing the novelty effect and simplifying the task complexity. This familiarity leads to more consistent and reliable judgements (Liu & Li, 2012).[1] However, the process can be tedious and time-consuming, risking decreased performance due to cognitive fatigue (Bramley et al., 1998). To maintain reliability while managing judge fatigue, Verhavert et al. (2019) thus suggest limiting comparisons to the minimum necessary. Jones and Davies (2024) recommend that for CJ, multiplying the number of scripts by 10 would yield a proper number of pairwise comparisons for each item. Accordingly, we aimed for 300 pairwise comparisons per item, assigning each judge to 50 judgements. Given its adaptivity, ACJ may need fewer comparisons than traditional CJ. This allocation ensured that each recording was evaluated around 20 times.

To verify scoring validity, that is, to assess whether judges' criteria align well with the CEFR-CV intelligibility construct, Think Aloud Protocols (TAPs) were implemented with 10 of the 12 judges. Due to the unavailability of one judge from the EJ group and to maintain balance, one from the NJ group was randomly selected to be excluded. The TAP requires participants to articulate their thoughts while completing the task, providing researchers with insights into participants' actions, reactions, and thoughts (Baxter et al., 2015). Judges were asked to verbalise the reasoning behind their decision-making reporting factors that made one recording more intelligible than the other. The first two judges conducted think-alouds across all 200 pairwise comparisons for four items, each session lasting about 5 hours. This process, however, revealed that the criteria referenced by these judges did not yield any new insights and their responses tended to become repetitive or simplified after approximately 15 pairs of comparisons per item. In addition, engaging in a think-aloud protocol can take resources away from the primary task, thus diminishing the cognitive capacity available for the task itself (Baxter et al., 2015). Consequently, we modified the approach for the eight subsequent judges, limiting them to think-aloud only during the first 15 comparisons per item (60 pairs in total) to maximise information gathering and ensure the judgement quality. Furthermore, researchers only prompted judges with questions such as, "Which sample do you find more understandable, and why?" to facilitate continuous verbalisation; no other eliciting questions were permitted. Judges then independently completed the remaining comparisons without introspective reporting, submitting their results within 2 weeks.

Follow-up interview session. Following the completion of the TAP sessions, brief semi-structured interviews with each judge were held to investigate their experiences of using ACJ to measure CEFR-CV intelligibility. The interviews featured three open-ended questions: the first explored judges' overall experiences with ACJ, the second evaluated their perceived benefits and drawbacks of ACJ, and the third compared ACJ with traditional scoring methods they had previously used. These discussions provided insights into the practical implementation and effectiveness of ACJ.

Rubric rating session. To examine the criterion-related validity of ACJ, two raters assessed speech samples using the rubric of EnglishScore Speaking Test, the source of our current task items. This rubric was specifically chosen for its appropriateness in assessing the types of speech elicited by these items and its close alignment with the CEFR-CV's phonological scale. Given that no rubric specifically designed for assessing CEFR-CV intelligibility currently exists, we opted for this closely available option that could provide a reasonable basis for comparison with the ACJ results. The rubric consisted of three 6-point subscales: (a) pronunciation, which included traditional intelligibility, which primarily focused on word recognition and prosody; (b) fluency, which included speech rate and pausing; and (c) task fulfilment, which assessed the completeness of the sentence repetition. Raters scored on each subscale, and the sub-scores were totalled to produce overall scores.

The raters specifically focused on Items 03 and 04 due to their greater complexity compared to the other two items, which enhances the discernment of variations in judgements, essential for assessing the validity of the ACJ method. The correlation between the two scoring approaches on these items can provide informative results. In addition, concentrating on these items helps manage the raters' workload and avoid the need to rate all 120 recordings. The first rater evaluated both items, whereas the second rater assessed only Item 04. This approach allowed us to compare rubric-based scores to ensure consistency between the raters.

Data analysis

ACJ data. Judges' binary decisions were automatically analysed in the RM Compare system using its built-in Bradley-Terry-Luce model. However, due to its limited adoption in prior ACJ studies and the minimal verification of its printouts against other statistical software, we conducted additional analyses using FACETS 4.1.6 to validate the logit estimates from RM Compare. Following Linacre's (2023) guidelines for analysing paired data in FACETS, three facets were modelled: two recordings for pairwise comparison and judges as a dummy facet for fit analysis. Each observation, whether a "win = 1" or "lose = 0" decision, was weighted by 0.5 and entered twice to increase computational stability as Linacre recommended. To examine the reliability of ACJ (RQ1), four types of statistical evidence were provided: (a) item infit statistics to measure the consistency of the evaluation of each item; (b) judge infit statistics to gauge each judge's internal self-consistency relative to others in their group; (c) SSRs to determine how closely a typical judge's rankings align with the consensus of the judging panel (Pollitt, 2012); and (d) split-half reliability to assess the consistency across different judge groups. We applied Pollitt's (2012) threshold for (item and judge) infit statistics, specifically established for ACJ, accepting values within the mean infit value plus two SDs as satisfactory. For the SSR, reliability levels above 0.70 are generally deemed adequate for low-stakes tests, while a minimum of 0.90 is necessary for high-stakes assessments (Nunnally, 1978). These standards also apply to the assessment of split-half reliability.

Rubric scores. Criterion-related validity is generally addressed by correlating the results of the new measurement with those from a measure already established as valid. In this study, such evidence was associated with the correlation coefficients between ACJ and rubric scoring results. Thus, to address the first part of RQ2 on criterion-related validity, the scores from the two rubric raters for Item 04 were first correlated to examine their accuracy, which would serve as a benchmark for ACJ results. Subsequently, the rank order derived from the rubric scores that the rubric raters provided for Items 03 and 04 was compared with rankings from both judge groups.

TAPs. To address the second part of RQ2, scoring validity, the criteria judges verbalised in the ACJ processes were examined. Understanding these criteria is important; if judges incorporate irrelevant characteristics, such as non-linguistic features, into their decisions, ACJ may not accurately measure CEFR-CV intelligibility. Given the ambiguity in defining the CEFR-CV intelligibility construct, the study relied on features contributing to listeners' understanding as identified in existing literature. To analyse these criteria, manifest content analysis of the TAPs was employed. This approach, a subset of content analysis, is favoured for its ability to count and categorise the visible, explicit elements of content, offering a more surface-level and quantitative focus on what is directly observable (i.e., frequency of the criteria judges referred to). Initially, the first author reviewed all protocols to develop tentative themes. After several reviews and discussions with the second author, an initial coding scheme was established, closely aligning with terms used in the literature. The initial coder revisited the data 1 month later to refine the codes, achieving a 92% exact agreement index between the two coding sessions. Discrepancies were resolved through further discussion.

Follow-up interview data. To address the third part of RQ2, perceived validity, a content analysis of the interview data was conducted, focusing on judges' perceptions of using ACJ. This analysis involved coding the data to identify and categorise judges' reported experiences with the judgement process.

Results

Validation results of RM compare system

Rasch analyses were conducted with FACETS on the dichotomous decisions made by both groups of judges across all four items to validate the results obtained from the RM Compare system. First, we calculated the correlations between the rank orders produced by both software systems across different judge types and items, resulting in eight exceptionally high coefficients, ranging from 0.99 to 1.00. These indices confirmed the validity of the results from RM Compare. However, FACETS uses the Rasch model as its built-in framework, while RM Compare employs the Bradley-Terry-Luce model. Despite their commonalities, slight differences in their algorithms lead to variations in reporting infit statistics and SSR figures. These differences will be further discussed in the reliability section.

Table 1. Distribution of item infit statistics.

	EJ group ($n=6$)			NJ group ($n=6$)		
	Item infit threshold	Item infit range	Percentage of misfitting recordings	Item infit threshold	Item infit range	Percentage of misfitting recordings
Item 01	⩽1.93	0.51–2.52	3.33% ($n=1$)	⩽1.70	0.54–1.87	3.33% ($n=1$)
Item 02	⩽1.79	0.01–1.75	0.00% ($n=1$)	⩽1.85	0.50–2.03	3.33% ($n=1$)
Item 03	⩽1.94	0.05–2.15	3.33% ($n=1$)	⩽1.82	0.06–2.08	3.33% ($n=1$)
Item 04	⩽2.08	0.02–2.30	3.33% ($n=1$)	⩽1.97	0.02–2.20	6.67% ($n=2$)

RQ1: Reliability evidence of ACJ in assessing CEFR-CV intelligibility

Infit statistics for items and judges, SSR indices, and split-half reliability between judge groups were analysed to confirm the reliability of ACJ. Table 1 displays the percentage of misfitting recordings for each item as identified by the RM Compare system, demonstrating high consistency across both judge types and items. Each of the four items included 30 recordings, totalling 120 recordings per judge group and 240 overall. Of these, only 8 recordings, or 3.33%, were judged inconsistent. Discrepancies in item infit values between different software systems were noted, likely due to the differing algorithms each employs. For instance, applying the same misfit cutoffs (no more than mean plus 2 SDs), RM Compare showed complete consistency for Item 02 judged by the EJ group, whereas FACETS illustrated inconsistent judgements for recordings of Speakers 06 and 18. In a review of all recordings judged inconsistent, as reported by both systems, FACETS recorded a slightly higher number of inconsistencies ($n=9$). Nevertheless, following this stringent standard, only 3.75% of the recordings were judged inconsistent. Overall, the item infit statistics from both software systems demonstrate a high convergence in the judgements by both groups.

Table 2 presents the distribution of judge infit statistics from the RM Compare system, demonstrating that all judges' infit values fall within the established consistency threshold, indicative of alignment with group consensus. Notably, NJ05 from the NJ group recorded an infit value of 1.82 for Item 03, approaching the upper limit. According to Pollitt (2012), although NJ05's performance is on the edge, it remains within the bounds of the threshold. To further confirm NJ05's alignment with other judges in this group, we revisited the FACETS results, which showed NJ05's infit value at 1.61, comfortably within its acceptable threshold of no more than 1.66. Consequently, NJ05's decisions were retained.

The SSR averaged 0.94 for the EJ group and 0.91 for the NJ group, indicating extremely high interrater reliability of the scales produced by ACJ, as detailed in Table 3. The variations in SSR indices reported by both tools were negligible, with differences amounting to less than 0.01. In summary, the comparison of infit statistics and SSR indices between the two software systems demonstrated the overall validity of the RM Compare results. However, the discrepancies observed warrant careful consideration.

Table 2. Distribution of judge infit statistics.

	EJ group ($n = 6$)		NJ group ($n = 6$)	
	Judge infit threshold	Judge infit range	Item infit threshold	Item infit range
Item 01	⩽1.53	0.82–1.33	⩽1.63	0.70–1.51
Item 02	⩽1.88	0.39–1.57	⩽1.69	0.49–1.37
Item 03	⩽1.64	0.48–1.52	⩽1.82	0.72–**1.82**
Item 04	⩽1.47	0.83–1.45	⩽1.63	0.56–1.48

Table 3. SSR across items and judge groups.

	EJ group ($n = 6$)	NJ group ($n = 6$)	Average across judge groups
Item 01	0.86	0.87	0.87
Item 02	0.98	0.87	0.93
Item 03	0.97	0.95	0.96
Item 04	0.95	0.93	0.94
Average across items	0.94	0.91	n/a

Following Jones and Davies' (2024) recommendation, we further explored split-half reliability by analysing correlations between the ACJ rankings from both judge groups. As shown in Figure 1, Spearman's rho (ρ) (a nonparametric measure of rank correlation) ranged from 0.87 to 0.93, demonstrating strong convergence between the judge groups' pairwise decision-making. Furthermore, the analysis indicates that NJs' achieved results are comparable to those of their more experienced counterparts.

RQ2: Validity evidence of ACJ in assessing CEFR-CV intelligibility

Criterion-related validity evidence. As noted, criterion-related validity was established through the correlation between the ACJ results and rubric scores. Prior to conducting this analysis, we calculated Pearson's correlation between the two rubric raters for Item 04, which was found to be 0.86 ($p < .05$), indicating high consistency between their scores. The ACJ rankings from both judge groups showed a strong correlation with the rank order (converted from the rubric scores) provided by both rubric raters for Items 03 and 04. Spearman's ρ varied from 0.82 to 0.90 (see Figure 2). This consistency indicates that the two scoring methods similarly assessed the speakers' performance on the focal construct.

Scoring validity evidence. To understand the criteria judges referred to in making their decisions, TAPs were analysed across 600 pairwise comparisons involving 10 judges. Our content analysis initially revealed 1526 codings related to 73 initial criteria, which were then refined and consolidated into 22 micro criteria across eight macro criteria. These criteria align with features contributing to listeners' intelligibility and

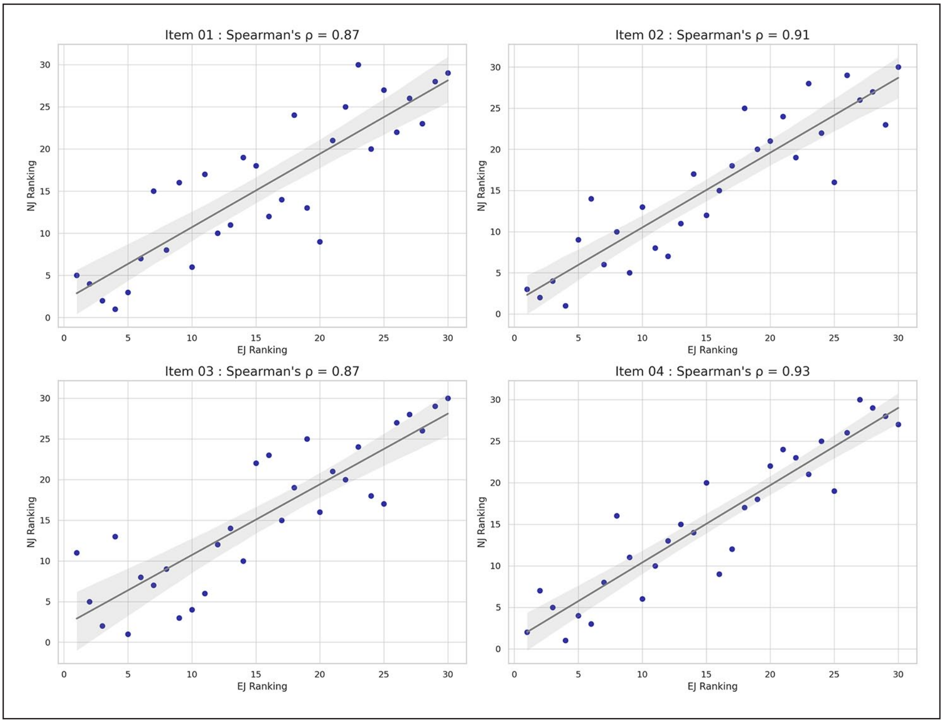

Figure 1. Correlations of ACJ results between judge groups.

comprehensibility as identified in the literature review (see Table 4). On average, each judge employed 2.54 assessment criteria per decision. Segmentals and grammar accounted for the largest proportion of the codes (22.61% and 22.45%, respectively), followed by suprasegmentals (21.89%), fluency (15.14%), and discourse (13.63%). Codes for acceptability, accentedness, and shared L1 benefits totalled less than 5%. Notably, judges tended to rely heavily on a select few micro criteria. Using an arbitrary threshold of 75 coded units (approximately 5% of all coded units), six significant micro criteria emerged: (a) articulation accuracy (*n*=261, 17.10%), (b) repetition accuracy (*n*=177, 11.60%), (c) grammatical accuracy (*n*=121, 7.92%), (d) fluency (*n*=118, 7.73%), (e) information accuracy (*n*=94, 6.16%), and (f) articulation clarity (*n*=84, 5.50%). The distribution of criteria usage varied slightly between EJ and NJ groups, though it was generally balanced. Specifically, the EJ group focused more on information content (*n*=43, 5.89%), while the NJ group paid more attention to intonation (*n*=52, 6.53%) and pausing patterns (*n*=43, 5.40%). These findings suggest that although there is overlap in criteria usage between the two groups, distinct focuses also exist. Furthermore, despite its low frequency (3.41%), the inclusion of the acceptability criterion (the degree of annoyance and irritability experienced by listeners when encountering listening difficulties), alongside other frequently mentioned macro criteria suggests that the judgements address segmental, temporal, and syntactic aspects of word recognition, as

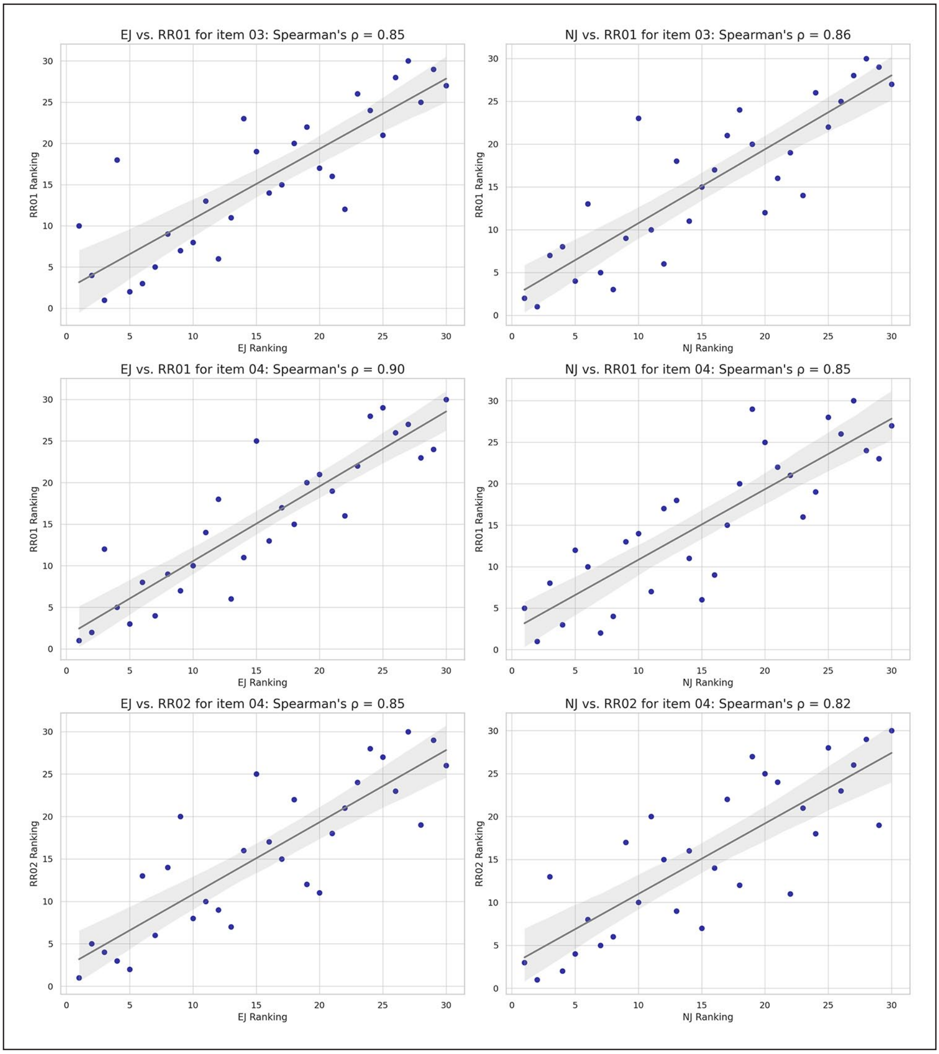

Figure 2. Correlations between ACJ results and rubric scores.

well as listeners' perceptions. These factors closely align with the scope of intelligibility targeted by the CEFR-CV framework, confirming that judges from both groups properly applied the CEFR-CV intelligibility definition in their judgement process.

Perceived validity evidence

Following the TAP session, judges provided five positive views (P) and three negative comments (N) regarding their experience with ACJ. Table 5 categorises each perspective, including descriptions and judges' pseudonyms. As demonstrated in Table 5, the

Table 4. The criteria judges referred to in assessing CEFR-CV intelligibility.

Macro criteria	Micro criteria	No. of codes		Total	%
		EJ	NJ		
Segmentals	• Articulation accuracy	146	115	261	22.61%
	• Articulation clarity	38	46	84	
	Total count	184	161	345	
Suprasegmentals	• Intonation	19	52	71	21.89%
	• Sentence stress	26	26	52	
	• Rhythm	33	12	45	
	• Voice confidence	9	20	29	
	• Tone	9	17	26	
	• Liaison	8	18	26	
	• Speech naturalness	13	18	31	
	• Timbre	14	6	20	
	• Word stress	12	4	16	
	• Loudness	5	9	14	
	• Pitch	0	4	4	
	Total count	148	186	334	
Fluency	• fluidity	51	67	118	15.14%
	• Pausing pattern	24	43	67	
	• Self-correction	13	11	24	
	• Speech rate	11	11	22	
	Total count	99	132	231	
Grammar	• Repetition accuracy	82	95	177	22.45%
	• Grammatical accuracy	60	61	121	
	• Omission(s)	15	14	29	
	• Addition(s)	11	5	16	
	Total count	168	175	343	
Discourse (information delivery)	• Information accuracy	41	53	94	13.63%
	• Information content	43	23	66	
	• Information completeness	18	30	48	
	Total count	102	106	208	
Acceptability	• Extra processing effort	15	16	31	3.41%
	• Distraction	8	10	18	
	• Irritation	0	3	3	
	Total count	23	29	52	
Accentedness	• Foreign accent	6	5	11	0.72%
Shared L1 Benefits	• Accent familiarity	0	2	2	0.13%
Total Count		730	796	1526	

Note: EJ = Experienced Judge; NJ = Novice Judge.

advantages of ACJ and the number of judges who favoured it somewhat outweighed its disadvantages and the number of judges who expressed concerns. Its reliability and comprehensiveness emerged as the most frequently cited benefits. However, these positive attributes were mainly raised by experienced rather than NJs, suggesting that

Table 5. Perceived advantages and disadvantages of ACJ.

Themes	Description	Judge ID
P1: reliable	ACJ ensures quality assessments through judge consensus.	EJ02, EJ04, EJ05, NJ03
P2: comprehensive	Judges can automatically evaluate a broad set of speaking constructs with ACJ.	EJ01, EJ02, EJ04, NJ01
P3: objective	ACJ minimises individual biases, thereby promoting fair evaluations.	EJ03, EJ04
P4: flexible	It is adaptable to various assessment criteria when used with ACJ.	EJ01, NJ01, NJ02
P5: engaging	ACJ renders the marking process funnier and more interesting.	NJ01, NJ05
N1: subjective	ACJ relies heavily on judges' preferences and expertise.	NJ05
N2: impractical	ACJ is more challenging to implement than traditional marking methods.	NJ02
N3: unfair	There are perceptions of potential inequity in outcomes.	NJ04

familiarity with rubric scoring might increase preference for ACJ compared to traditional making methods:

> P1: ". . .ACJ can synthesise the opinions of different people to form a general agreement." (from EJ05)

> P2: "I find this method is very meaningful. Because during the evaluation process, it allows me to naturally take into account many aspects [of speaking]. . ." (from NJ01)

Although some judges perceived ACJ as objective and flexible (i.e., P3 and P4), others found its implementation subjective and complex (i.e., N1 and N2). This divergence is evident in the following verbatim comments:

> P3: "In ACJ, we don't see who's work we're judging or how often it's been viewed, which really lets us focus on the quality itself, it feels fairer that way." (from EJ04)

> N1: ". . . sometimes I feel that a speaker's voice is very pleasant and charismatic, which makes me inclined to choose him, overlooking some more objective criteria. This could increase my subjectivity in making my decision." (from NJ05)

> P4: "I've found that as the difficulty of the items increases, the performance of the speakers becomes more varied, and my scoring criteria change accordingly. . . However, this change allows me to consider more aspects of the speech, which is more effective than being confined to a fixed set of standards." (from NJ01)

> N2: "It's really tough to do this! Especially when the two speakers are at similar proficiency, it's really hard to make the decision." (from NJ02)

Finally, one judge expressed that ACJ makes the evaluation process more interesting than the traditional scoring approach (P5). However, another alluded to its perceived unfairness (N3), as demonstrated in the following extracts:

> P5: "Working with ACJ makes the marking process a bit of fun. It's like taking a test, but with binary choices. It introduces a game-like element to evaluation that's surprisingly enjoyable." (from NJ01)

> N3: "I feel that in the comparison process, the better side always becomes my reference point, which doesn't seem quite fair. Even though the pairings are random and the number of pairings is limited, if one side in a comparison is worse, there are surely other students who are worse than the lesser side in that pair, but such pairings might not occur." (from NJ04)

NJ04's concern about the perceived unfairness of ACJ highlights potential issues with its adaptive algorithm: early, less favoured performances may not have the opportunity to be compared against better-performing stimuli later on. This may occur because the adaptive algorithm tends to pair stimuli with similar estimated qualities. Thus, performances consistently less favoured early on may be considered of lower quality and consequently, not paired with higher-quality performances later. However, this is not always the case, as the adaptive algorithm is designed to maximise the information collected from each comparison by actively seeking to pair stimuli that were not directly compared before. This strategy ensures that each stimulus is paired against a diverse set of others. In addition, the statistical models used in ACJ are robust to the order of comparisons, estimating the overall quality of each stimulus based on the entire set of comparisons. Therefore, while the adaptive nature of ACJ may occasionally lead to the concern described, the design of the algorithm and robustness of the statistical models somewhat mitigate them.

Discussion

This study explored the potential of ACJ for measuring CEFR-CV intelligibility through a semi-controlled task (sentence repetition), with promising initial results. In terms of reliability, the infit statistics for both items and judges suggested that over 95% of recordings were consistently judged, and all judges made convergent judgements. Notably, while using FACETS to validate RM Compare results, subtle discrepancies in infit statistics between the software were observed, likely due to differences in their built-in models. This point underscores the risks of relying on a single software, particularly in high-stakes assessment contexts, where misfitting scripts might require identification for re-judgement, or misfitting judges might need exclusion for further analysis. For instance, NJ05, identified in RM Compare as reaching the infit threshold for Item 03, was found to have a perfectly acceptable infit index in FACETS. Furthermore, various cutoffs used to standardise infit statistic values can lead to differing outcomes. For example, some judges' infit statistics in this study might fall outside the standard range of 0.5 to 1.5, as used in Han and Xiao (2022). Thus, while results from RM Compare are generally precise, adopting ACJ in high-stakes settings demands careful consideration. Moreover, it is important to use multiple methods to cross-validate ACJ results or to choose cutoffs appropriate for a certain context, ensuring the robustness and fairness of assessments.

In addition, the SSRs reached a relatively high level for both judge groups, exceeding the established lower limit of 0.90 for high-stakes assessments (Han & Xiao, 2022; Verhavert et al., 2019). Although the "inflated" reliability of ACJ was calibrated in RM Compare (Kimbell, 2021), without a comparison with traditional CJ results, whether there is inflation in the SSRs of this study remains unclear. Furthermore, no substantial differences were observed between the SSRs of the two judge types, with a finding in line with Han's (2022), who conducted research using two groups of judges (experienced and novice) to assess interpreting via CJ. In his study, the SSRs demonstrated significant consistency across both judge types. Moreover, the split-half reliability averaged 0.90 in this study, which implies that NJs were able to assess the CEFR-CV intelligibility of speakers' renditions similarly to the EJs. Given that other influential variables (e.g., shared L1 benefit) were controlled in this study, these results suggest that applying CJ-based measurement may not require extensive experience, expertise, or investment in rater training, backing Pollitt's (2012) and Jones and Davies' (2024) claims.

The correlations between the results from the rubric raters and both judge groups also demonstrated a high level of concurrence. These high correlations suggest that both ACJ and rubric scoring capture similar information about the L2 speech samples being evaluated. In other words, ACJ has shown potential for effectively assessing pronunciation with CEFR-CV intelligibility as its central construct. Moreover, ACJ provides judges with greater flexibility to assess various aspects of pronunciation, potentially offering more adjustable and comprehensive evaluation results than rubric ratings, which are bound by fixed standards. Although Pollitt (2012) drew similar conclusions, further research and evidence are required to support these findings. Furthermore, due to the limited sample size in this study (e.g., two raters), caution is warranted when interpreting these results. Further research with larger sample sizes is necessary to strengthen the evidence.

Content analysis of the TAPs revealed that judges reported decision-making criteria align with those features identified as contributing to L2 pronunciation constructs. We identified six micro criteria frequently mentioned by judges: (a) articulation accuracy (*n*=261, 17.10%), (b) repetition accuracy (*n*=177, 11.60%), (c) grammatical accuracy (*n*=121, 7.92%), (d) fluidity (*n*=118, 7.73%), (e) information accuracy (*n*=94, 6.16%), and (f) articulation clarity (*n*=84, 5.50%). These criteria span five macro assessment domains: segmentals, suprasegmentals, fluency, grammar, and discourse. This aligns with L2 studies on intelligibility and comprehensibility that have found that these constructs are associated with a wide range of linguistic features (e.g., Kang et al., 2020; Trofimovich et al., 2022). Nevertheless, the focus on syntactic and semantic aspects often depends on task type. Typically, this emphasis is evident in tasks that elicit spontaneous speech, providing a context for evaluating speakers' syntactic and semantic performance. The close alignment between our analysis and existing literature indicates a degree of scoring validity of ACJ in assessing CEFR-CV intelligibility. Furthermore, the criteria identified in the TAPs cover dimensions operationalised in CEFR-CV intelligibility, suggesting that judges appropriately applied this definition as a holistic criterion in their judgement process. It should be noted, however, that the discourse dimension in this study primarily measures the quantity and accuracy of the speakers' information delivery and does not fully explore speech coherence and content breadth due to task type limitations. Consequently, further research into more uncontrolled and extended task types is advised (Munro & Derwing, 1995a).

Regarding the perceived validity of ACJ, judges recognised its potential in L2 pronunciation assessment settings but expressed varied opinions about its practicality and fairness. Most judges, especially those experienced with traditional marking methods, viewed ACJ positively, noting its reliability and comprehensiveness. They believed that ACJ improves assessment accuracy by distilling a consensus among a group of judges (Kelly et al., 2022). They also appreciated the flexibility of the judgment process, which allows judges to adjust their criteria usage based on their expertise rather than adhering strictly to established standards. However, concerns were raised about the impractical and unfair implementation of ACJ. One novice judge noted difficulties in decision-making when faced with two closely matched stimuli, while another expressed concern that not presenting all possible pairwise comparisons could disadvantage some stimuli. These issues may stem from the adaptive nature of the pairing algorithm, which, while designed to present more informative pairs by matching stimuli with similar proficiency levels, may inadvertently increase judgment difficulty and create potential unfairness. In response, we suggest introducing a button for judges to indicate "similar proficiency" between stimuli, in addition to choosing a "winner" or "loser." This modification would require adjustments to the Bradley-Terry-Luce model to accommodate the new option. It is worth noting that introducing a third option could increase the complexity of the estimation procedure and the number of comparisons needed to achieve stable and reliable results. However, the exact impact of such an option on the required number of judgments would depend on various factors, such as the distribution of proficiency levels within the stimulus set, the frequency with which judges select this option, and the specific modifications made to the statistical model to accommodate the additional response category. We also advocate for refined algorithms that can more effectively account for the uncertainty of estimated qualities and pair stimuli that have not yet been directly compared.

Conclusion

In this exploratory study, we implemented ACJ to assess CEFR-CV intelligibility, evaluating its reliability, validity, and utility. The quantitative and qualitative data together suggest that ACJ offers a promising and valid measurement. Moving forward, we identify four potential research directions to expand understanding of ACJ. First, achieving a clearer and more detailed comprehension of the benefits and drawbacks associated with the use of ACJ is needed. Since this was not the primary focus of this study, our findings provide only an initial picture of adopting ACJ in assessing L2 pronunciation.

Second, to refine the pairing and modelling algorithm of ACJ, we propose the development of a more sophisticated algorithm to better pair stimuli that have not yet been compared while still ensuring that these pairs maximise information gain. Furthermore, we suggest integrating a "similar proficiency" option within the ACJ pairwise comparison process. This adjustment would necessitate transitioning from the current dichotomous model to a more nuanced tripartite model.

Third, despite overall consistency in judges' performances, subtle differences were observed in the criteria they prioritised during evaluations. These variations may arise from the methodological limitations inherent in TAP, where self-reporting may not always be natural or comprehensive; judges might only verbalise thoughts that they are consciously aware of and wish to share (Baxter et al., 2015). Therefore, even though the CEFR-CV intelligibility construct can potentially be developed based on TAP data, cross-validation remains essential for future studies.

Finally, it is necessary to explore a broader range of task types and to involve judges from diverse L1 backgrounds. While this study validated the effectiveness of ACJ in assessing semi-controlled tasks, its applicability to more extensive question types, such as picture narratives, requires further clarification. In addition, this study, by strictly controlling judges' L1 backgrounds to minimise the impact of shared L1 benefits, does not fully capture the authenticity of English as a lingua franca communication contexts. Given that both task type and listeners' L1 backgrounds influence the linguistic features that contribute to the pronunciation construct, examining criteria use across various tasks and judges with different L1s can deepen our understanding of the variability within the CEFR-CV intelligibility construct, thereby enriching the conceptual framework.

Acknowledgements

This manuscript is based on a thesis completed by the first author to fulfill the requirements for a Ph.D. degree at the University of Southampton, United Kingdom. We would like to thank Prof. Talia Isaacs for her insightful editorial suggestions and the anonymous peer reviewers whose assistance was invaluable in completing this manuscript. We also extend heartfelt thanks to all the participants for their crucial contributions to this research. Special thanks to *RM Compare* for providing an unlimited licence for their software, essential for data collection.

Author contributions

Jingwen Wang: Conceptualisation; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Validation; Visualisation; Writing—original draft.
Ying Zheng: Project administration; Supervision; Writing—review & editing.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Jingwen Wang https://orcid.org/0009-0000-7615-2674
Ying Zheng https://orcid.org/0000-0003-2574-0358

Note

[1] Note that the ACJ system dynamically updates parameter estimations in response to real-time decisions. This ensures that earlier judgments do not disproportionately affect later ones. Through this adaptive process, the system maintains consistent reliability of judgments throughout the assessment, effectively mitigating potential biases associated with the sequence of task presentations.

References

Baxter, K., Courage, C., & Caine, K. (2015). *Understanding your users: A practical guide to user research methods* (2nd ed.). Elsevier. https://doi.org/10.1016/C2013-0-13611-2

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*, 141–164. https://doi.org/10.1007/s40753-016-0024-3

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., Bell, J. F., & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone paired comparisons. *Education Research and Perspectives*, *25*, 1–24. https://search.informit.org/doi/epdf/10.3316/aeipt.99127

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. https://doi.org/10.1080/0969594X.2017.1418734

British Council. (2020). *How IELTS is assessed*. https://takeielts.britishcouncil.org/teach-ielts/test-information/assessment

Carey, M. D., & Szocs, S. (2024). Revisiting raters' accent familiarity in speaking tests: Evidence that presentation mode interacts with accent familiarity to variably affect comprehensibility ratings. *Language Testing*, *41*(2), 290–315. https://doi.org/10.1177/02655322231200808

Chau, T., Huensch, A., Hoang, Y. K., & Chau, H. T. (2022). The effects of L2 pronunciation instruction on EFL learners' intelligibility and fluency in spontaneous speech. *TESL-EJ*, *25*(4), 1–28. https://files.eric.ed.gov/fulltext/EJ1334064.pdf

Choi, S., & Kang, O. (2023). The roles of suprasegmental features in assessing paired speaking tasks in high-stakes language assessment. *System*, *119*, 1–12. https://doi.org/10.1016/j.system.2023.103183

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment—Companion volume*. Council of Europe Publishing. http://www.coe.int/lang-cefr.

Crowther, D., Trofimovich, P., & Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility: Nonnative listeners' perspectives. *Journal of Second Language Pronunciation*, *2*(2), 160–182. https://doi.org/10.1075/jslp.2.2.02cro

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*(1), 1–16. https://doi.org/10.1017/S0272263197001010

Digital Assess. (n.d.). *RM compare* [Software]. https://www.rmcompare.com/

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2023). *Ethnologue: Languages of the world* (26th ed.). SIL International. https://www.ethnologue.com/language/eng

Han, C. (2022). Assessing spoken-language interpreting: The method of comparative judgement. *Interpreting*, *24*(1), 59–83. https://doi.org/10.1075/intp.00068.han

Han, C., & Xiao, X. (2022). A comparative judgment approach to assessing Chinese Sign Language interpreting. *Language Testing*, *39*(2), 289–312. https://doi.org/10.1177/02655322211038977

Huensch, A., & Nagle, C. (2021). The effect of speaker proficiency on intelligibility, comprehensibility, and accentedness in L2 Spanish: A conceptual replication and extension of Munro and Derwing (1995a). *Language Learning*, *71*(3), 626–668. https://doi.org/10.1111/lang.12451

Huensch, A., & Nagle, C. (2023). Revisiting the moderating effect of speaker proficiency on the relationships among intelligibility, comprehensibility, and accentedness in L2 Spanish. *Studies in Second Language Acquisition*, *45*(2), 571–585. https://doi.org/10.1017/S0272263122000213

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *Canadian Modern Language Review*, *64*(4), 555–580. https://doi.org/10.3138/cmlr.64.4.555

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34*(3), 475–505. https://doi.org/10.1017/S1366728912000168

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

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, *38*(2), 301–315. https://doi.org/10.1016/j.system.2010.01.005

Kang, O., Hirshi, K., Hansen, J., Looney, S., & Miao, Y. (2022). Using lexical stress, speech rate, rhythm, and pauses to characterize and normalize second language speech intelligibility. In *Proceedings of Meetings on Acoustics* (*volume 50*, pp. 1–9). AIP Publishing. https://doi.org/10.1121/2.0001790

Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, *68*(1), 115–146. https://doi.org/10.1111/lang.12270

Kang, O., Thomson, R. I., & Moran, M. (2020). Which features of accent affect understanding? Exploring the intelligibility threshold of diverse accent varieties. *Applied Linguistics*, *41*(4), 453–480. https://doi.org/10.1093/applin/amy053

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

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review*, *64*(3), 459–489. https://doi.org/10.3138/cmlr.64.3.459

Kimbell, R. (2021). Examining the reliability of adaptive comparative judgement (ACJ) as an assessment tool in educational settings. *International Journal of Technology and Design Education*, *32*, 1515–1529. https://doi.org/10.1007/s10798-021-09654-w

Linacre, M. J. (2023). *A user's guide to FACETS: Rasch-Model computer programs*. https://winsteps.com/a/Facets-Manual.pdf

Liu, P., & Li, Z. (2012). Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics*, *42*(6), 553–568. https://doi.org/10.1016/j.ergon.2012.09.001

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review*, *66*(2), 81–95. https://doi.org/10.1037/h0043178

Major, B., Quinton, W. J., & McCoy, S. K. (2002). Antecedents and consequences of attributions to discrimination: Theoretical and empirical advances. *Advances in Experimental Social Psychology*, *34*, 251–330. https://doi.org/10.1016/S0065-2601(02)80007-7

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

Miao, Y. (2023). The relationship among accent familiarity, shared L1, and comprehensibility: A path analysis perspective. *Language Testing*, *40*(3), 723–747. https://doi.org/10.1177/02655322231156105

Miao, Y., & Kang, O. (2023). An empirical approach to measuring accent familiarity: Phonological and correlational analyses. *System*, *116*, 1–14. https://doi.org/10.1016/j.system.2023.103089

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, *38*(3), 289–306. https://doi.org/10.1177/002383099503800303

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, *28*(1), 111–131. https://doi.org/10.1017/S0272263106060049

Nagle, C. L., Huensch, A., & Zárate-Sández, G. (2023). Exploring phonetic predictors of intelligibility, comprehensibility, and foreign accent in L2 Spanish speech. *The Modern Language Journal*, *107*(1), 202–221. https://doi.org/10.1111/modl.12827

Nagle, C. L., Trofimovich, P., O'Brien, M. G., & Kennedy, S. (2022). Comprehensible to whom? Examining rater, speaker, and interlocutor perspectives on comprehensibility in an interactive context. *The Modern Language Journal*, *106*(4), 675–693. https://doi.org/10.1111/modl.12809

Newhouse, C. P., & Cooper, M. (2013). Computer-based oral exams in Italian language studies. *ReCALL*, *25*(3), 321–339. https://doi.org/10.1017/S0958344013000141

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

O'Brien, M. G. (2014). L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning*, *64*(4), 715–748. https://doi.org/10.1111/lang.12082

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

Pérez-Ramón, R., Lecumberri, M. L. G., & Cooke, M. (2022). Foreign accent strength and intelligibility at the segmental level. *Speech Communication*, *137*, 70–76. https://doi.org/10.1016/j.specom.2022.01.005

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

Rangel-Smith, C., & Lynch, D. (2018, June). Addressing the issue of bias in the measurement of reliability in the method of Adaptive Comparative Judgment. In N. Seery, J. Buckley, D. Canty, J. Phelan (Eds.), *Research and practice in technology education: Perspectives on human capacity and development* (pp. 378–387). *36th Pupils' Attitudes Towards Technology Conference*.

Sahin, 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

Saito, K., & Shintani, N. (2016). Foreign accentedness revisited: Canadian and Singaporean raters' perception of Japanese-accented English. *Language Awareness*, *25*(4), 305–317. https://doi.org/10.1080/09658416.2016.1229784

Shehata, A. (2024). Arabic speech intelligibility: Perception of spoken Arabic by native and non-native speakers. *Language Teaching Research*. *0*(0), 1–16. https://doi.org/10.1177/13621688241231628

Sherman, D., Mentzer, N., Bartholomew, S., Chesley, A., Baniya, S., & Laux, D. (2022). Across the disciplines: Our gained knowledge in assessing a first-year integrated experience. *International Journal of Technology and Design Education*, *32*, 1369–1391. https://doi.org/10.1007/s10798-020-09650-6

Sickinger, R., Brunfaut, T., & Pill, J. (2024). Comparative Judgement for evaluating young learners' EFL writing performances: Reliability and teacher perceptions of holistic and dimension-based judgements. *Language Testing*. https://doi.org/10.1177/02655322241288847

Smith, M. (2020). Adaptive comparative judgement. In M. Gregson & P. Spedding (Eds.), *Practice-focused research in further adult and vocational education* (pp. 77–98). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-38994-9_5

Thomson, R. (2017). Measurement of accentedness, intelligibility, and comprehensibility. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 11–29). Routledge. https://doi.org/10.4324/9781315170756

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286. https://psycnet.apa.org/buy/1994-28135-001

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*(4), 905–916. https://doi.org/10.1017/S1366728912000168

Trofimovich, P., Isaacs, T., Kennedy, S., & Tsunemoto, A. (2022). Speech comprehensibility. In T. M. Derwing, M. J. Munro & R. I. Thomson (Eds.), *The Routledge handbook of second language acquisition and speaking* (pp. 174–187). Routledge. https://doi.org/10.4324/9781003022497

Trofimovich, P., Lightbown, P. M., Halter, R. H., & Song, H. (2009). Comprehension-based practice: The development of L2 pronunciation in a listening and reading program. *Studies in Second Language Acquisition*, *31*(4), 609–639. https://doi.org/10.1017/S0272263109990040

Tsubota, Y., Dantsuji, M., & Kawahara, T. (2004). An English pronunciation learning system for Japanese students based on diagnosis of critical pronunciation errors. *ReCALL*, *16*(1), 173–188. http://journals.cambridge.org/abstract_S0958344004001314

Tsunemoto, A., & Trofimovich, P. (2024). Coherence and comprehensibility in second language speakers' academic speaking performance. *Studies in Second Language Acquisition*, *46*, 795–817. https://doi.org/10.1017/S0272263124000305

Uchihara, T., Webb, S., Saito, K., & Trofimovich, P. (2023). Frequency of exposure influences accentedness and comprehensibility in learners' pronunciation of second language words. *Language Learning*, *73*(1), 84–125. https://doi.org/10.1111/lang.12517

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

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan. https://doi.org/10.1057/9780230514577

Whitehouse, C., & Pollitt, A. (2012). *Using Adaptive Comparative Judgement to obtain a highly reliable rank order in summative assessment*. Centre for Education Research and Policy. https://filestore.aqa.org.uk/content/research/CERP_RP_CW_20062012_2.pdf?download

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Mesa Press. https://research.acer.edu.au/measurement/2/

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, *33*(4), 497–528. https://doi.org/10.1177/0265532215594643

Yenkimaleki, M., & van Heuven, V. J. (2024). Developing interpreter trainees' speech comprehensibility: Does nativeness of the instructor matter? *TESL-EJ*, *27*(4), 1–29. https://doi.org/10.55593/ej.27108a8