Huang et al. Language Testing in Asia (2018) 8:14
https://doi.org/10.1186/s40468-018-0069-0

Language Testing in Asia

RESEARCH Open Access

# Evaluating CEFR rater performance through the analysis of spoken learner corpora

Lan-fen Huang[1*], Simon Kubelec[2], Nicole Keng[3] and Lung-hsun Hsu[1]

* Correspondence:
lanfen.huang@gmail.com
[1]Shih Chien University, 200 University Road, Nei-men, Kaohsiung 845, Taiwan
Full list of author information is available at the end of the article

## Abstract

**Background:** Although teachers of English are required to assess students' speaking proficiency in the Common European Framework of Reference for Languages (CEFR), their ability to rate is seldom evaluated. The application of descriptors in the assessment of English speaking on CEFR in the context of English as a foreign language has not often been investigated, either.

**Methods:** The present study first introduced a form of rater standardization training. Two trained raters then assessed the speaking proficiency of 100 learners by means of actual corpus data. The study then compared their rating results to evaluate inter-rater reliability. Next, ten samples of exact/adjacent agreement between Raters 1 and 2 were rated by six teachers of English in tertiary education. Two of them had attended rater standardization training with Raters 1 and 2, while the other four had not received any relevant training.

**Results:** The two raters agreed exactly in 44% of cases. The rating results between the two trained raters were closely correlated ($\rho = .893$). Cross-tabulation showed that in one third of the samples, Rater 2 scored higher than Rater 1 and they agreed more often at the higher levels. The better rating performance of Teachers 1 and 2 suggested that rater standardization training may have helped enhance their performance. The unsatisfactory proportion of correctly assigned levels in teachers' ratings overall was probably due to the high input of subjective judgment based on vague CEFR descriptors.

**Conclusions:** Regarding assessment, it is shown that the attendance of rater standardization training is of help in assessing learners' speaking proficiency in CEFR. This study provides a model for assessing data from spoken learner corpora, which adds an important attribute to future studies of learner corpora. The paper also raises doubts about teachers' ability to evaluate students' speaking proficiency in CEFR. As CEFR has been widely adopted in the relevant fields of English language teaching and assessment, it is suggested that the rating training framework established in this study, which uses learner corpus data, be offered to (prospective) teachers of English in tertiary education.

**Keywords:** Rater performance, CEFR, Teacher rater, Speaking assessment, Learner corpus, LINDSEI

## Background

The Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001, 2018) has commonly been adopted in language learning, teaching, and assessment throughout Europe and beyond. It has established common standards for formulating the objectives of language learning curricula and materials, as well as certifying learners' proficiency in language skills. However, research on the application of descriptors in CEFR in English speaking assessment has been lacking.

Ratings results using CEFR descriptors provide valuable information about learners' speaking proficiency, which will contribute to learner corpus studies in a significant way. Learners' proficiency levels, in some studies, were inferred on the basis of external criteria (e.g., institutional status (2nd- or 3rd-year English majors at university) in Götz (2013) and learners' scores on internationally recognized proficiency tests in Huang (2014)). However, they have been criticized for being unreliable indicators of learners' proficiency (Callies et al. 2014). The present study assesses learners' speaking proficiency by means of actual corpus data, and thus, the rating results contribute to learner corpus studies in investigating interlanguage across proficiency levels.

Another rationale behind this study is the need to group for English courses such as oral training and presentations arising in the English curricula, since oral communication skills are always vital in the global workplace. Teaching English as a foreign language at all levels, particularly in Asia, focuses very much on listening and reading skills, probably because it is easier to administer tests of listening and reading than of speaking and writing. When language teachers assess students' speaking performance, they usually choose from the scales used in academic institutions, such as the ACTFL Proficiency Guidelines (American Council on the Teaching of Foreign Languages 2012), and consult rubrics with numerical scales to cater for different speaking tasks. The most influential scale is CEFR, but one of the issues in adopting CEFR in curricula is whether teachers of English are able to assess spoken English on the CEFR scales.

### Aims of the study

The aims of this study are to assess learners' speaking proficiency on the CEFR scales and to evaluate the rating performance of both experienced raters and teachers of English. It is hypothesized that the experienced raters' performance will be positively related and their inter-rater agreement will be high in rating 100 extracts from the Czech and Taiwanese sub-corpora of the Louvain International Database of Spoken English Interlanguage (LINDSEI-CZ and LINDSEI-TW) (Gilquin et al. 2010). The proficiency levels judged by two expert raters are used as a benchmark. Ten extracts with perfect/adjacent agreement between the two expert raters are re-assessed by six teachers of English. Their rating performance is evaluated to explore the possibility of adopting CEFR in assessing speaking, which is a crucial task regularly undertaken by teachers in higher education. The hypothesis is that the rating performance of teachers of English with prior training is better than those without any training. The comparison will show how far teachers of English have the same understanding of the CEFR scales and descriptors. Their feedback on CEFR is also reported.

### Assessing spoken learner English using CEFR

The drawbacks of the level descriptions have been discussed before in the literature; for example, Alderson (2007) noted that the descriptors are not clearly enough defined to be useful. He found that CEFR is "action-oriented"; thus, the scale itself evaluated language ability rather than the proficiency level. The descriptors in the latest version of CEFR (Council of Europe 2018) remain unchanged, but the section on phonology has been expanded. He further argued that there was a lack of empirical research to show the credibility of CEFR and pointed out the need for a large European learner corpus to contribute to empirical research on CEFR levels. The lack of empirical evidence taken from second language (L2) learner data and the need for corpus research were also critically discussed in Hulstijn (2007).

Over the years, however, there has been increasing use of learner corpora to gather empirical data on the CEFR levels. Granger et al. (2015) provided an overview in their book about learner corpus research, and there has also been a growing number of CEFR-linked learner corpora (e.g., the Cambridge Learner Corpus (CLC) (Harrison and Barker 2015); the Norwegian ASK corpus (Carlsen 2013)). Hawkins and Filipovic (2012) examined the "criterial features," how indicative they are of L2 proficiency, and how they are assumed to influence raters' decisions.

In a study of 27 French-speaking foreign language teachers from secondary education, Gilquin et al. (2016) found a difference in ratings between those given by English native and non-native raters. For instance, non-native teachers rated fluency lower than did the native raters (Gilquin et al. 2016; Rose 2017). Some studies also showed that native speakers seemed to be more tolerant than non-native speakers when judging accent (Koster and Koet 1993; Wester and Mayo 2014). Winke et al. (2012) also found that raters whose L2 is the test takers' L1 tended to be lenient in scoring. Gilquin et al. (2016) investigated the impact of rating factors, such as the frequency and type of discourse markers, and found similar results to those in Hyland and Anan (2006). Xi (2017) explains that often raters are not good at analyzing fine-grained linguistic features, and when using holistic scoring rubrics (e.g., CEFR), it is easy to prioritize the overall communicative effectiveness. In other words, even though the objective measures of linguistic features have an essential role in scoring, raters' evaluations tend to be influenced by their judgment of the overall effect of communication.

Wisniewski (2017) reviewing empirical work on learner language and CEFR noted that language tests based on CEFR levels are often for high stakes and impact on livelihoods. CEFR ability levels were not designed to map onto a development continuum along which language proficiency might be placed, and the levels were arrived at without the benefit of analyzing learner language. To allow more meaningful and accurate rating, more accessible CEFR-related learner corpora coming from transparent and reliable sources are urgently needed, and so are "more LC [learner corpora] for spoken learner language and for lower CEFR levels" (Wisniewski 2017, p. 246).

Learner corpus data can play a key role in speaking assessment, by providing rich data for level calibration and ironing out inconsistencies between raters. Yan (2014) also comments that rater alignment is particularly difficult at lower score levels and recommends training that focuses on rater disagreement.

In tertiary education in Taiwan, rating training for language teachers is not commonly provided, although they are most often the primary raters. This study aims to determine the validity and reliability of the use of CEFR in speaking assessment in the higher education context. It compares the rating performance of two groups of teachers: one of which attended a rater standardization session beforehand, whereas the other without training rated according to the guidelines provided.

## Data and participants

Most of the previous studies of oral proficiency were based on data collected in test-taking contexts, in which the test-takers would have been assessed by the grading criteria. It is argued that oral production in interviews is a better option (e.g., Magnan (1988) and Iwashita et al. (2008)) when assessing learners' speaking proficiency in CEFR. The spoken data for rating were extracted from one of the largest learner corpora of spoken English—LINDSEI (Gilquin et al. 2010). The learner language was elicited in interviews, which were more natural than testing contexts. The tasks that each learner did were similar to those in an English speaking course, in which teacher raters were themselves involved in the workplace. The two sub-sections below describe the data under investigation in more detail and introduce the nine raters.

### Audio extracts of spoken English

One hundred extracts of interviews with Czech (Gráf 2015) and Taiwanese learners (Huang 2014) were rated on the CEFR scale. The learner data were collected by twenty national research teams as a contribution to LINDSEI (Gilquin 2018). Each sub-corpus contained at least 50 interviews, involving three tasks. In Task 1, the learner spoke for about 5 min on his/her choice from three set topics. Task 2 was an approximately 7-min discussion including some follow-up questions about Task 1 and some general topics such as student life, hobbies, study and travel experiences, future plans, etc. Task 3 asked the learner to reconstruct a story based on a sequence of four pictures (Gilquin et al. 2010).

Task 1 in each of the 100 interview recordings was extracted using Audacity (2013 members of the Audacity development team 2013). This task was used for rating because it was based on self-selected topic and candidates were given a few minutes to prepare, making it more comfortable for them to produce English that represented their proficiency levels. The 5-min length of Task 1 allowed a rater to judge a candidate's CEFR level. The accompanying transcripts were not made available to raters. They performed a post hoc aural assessment.

### Raters' profiles

The trainer, acting as the third rater, and the two raters participating in the rating task were chosen because of their experience as Cambridge IELTS examiners. They had experience of examining in spoken English ranging from 6 months to 17 years and held recognized qualifications in teaching English, such as the Certificate in Teaching English to Speakers of Other Languages (CELTA), Diploma in Teaching English to Speakers of Other Languages (DELTA), and the MA in Teaching English to Speakers of Other Languages (TESOL). Their educational background and teaching experience are listed in Table 1. The three raters were native speakers of

**Table 1** Educational background and teaching experience of the trained raters

| | Rater 1 | Rater 2 | Rater 3 |
|---|---|---|---|
| First language | English | English | English |
| Current academic position | Lecturer | Lecturer | IELTS examiner trainer |
| Recognized qualification in teaching English | CELTA | MA in TESOL and TEFL certificate | CELTA, DELTA, and MA in TESOL |
| Years of teaching experience at university | 0 | 2.5 in Taiwan | Various in two countries |
| Years of teaching experience elsewhere | 14 in Taiwan | n/a | n/a |
| Familiarity with non-English accents | Asian (Korean/Chinese) | Chinese, French, European | Chinese |
| Past experience of rating speaking | Yes IELTS | Yes Research project | Yes KET, PET, and IELTS |
| Familiarity with CEFR before the rater standardization training | Somewhat | Basic | n/a |
| Past experience using the CEFR scales in marking spoken English | Yes Placement tests | No | Yes Placement tests |

English. Their teaching experience, mostly in Taiwan, ranged from 2.5 to 14 years. They had probably familiarized themselves with Chinese accents of English. Raters 1 and 2 reported low levels of familiarity with the CEFR and had applied it on language assessments such as placement tests.

Among the six teachers of English (see Table 2), Teacher 2 had a distinct profile in terms of first language and teaching experience. He was a native English speaker on a PhD programme in Asian Cultural Studies in Taiwan. The other five differed slightly in their formal education and work experience. All had been teaching in the tertiary sector for at least 7 years and were all familiar with Chinese/Taiwanese non-English accents. As regards the recognized qualification in teaching English, five teachers had either received a Master's degree in TESOL/TEFL or a Certificate in the Teaching Knowledge Test (TKT, by Cambridge ESOL). Teachers 2 and 4 had experience of rating speaking, but none had experience of rating on the CEFR scales. Overall, these teachers of English had limited understanding of CEFR and had not applied it when assessing oral skills.

## Methods

This section first sets out the rating procedure and evaluation of the rating performance. Then, it outlines the rater standardization training and some measures to maintain the reliability of the scoring.

### Rating procedure and rating performance evaluation

One senior rater, two experienced raters, and six teachers of English were recruited to participate in this rating task. The rating procedure followed instructions provided by the Centre for English Corpus Linguistics at the Catholic University of Louvain, Belgium. It had been used in rating a random sample of five learners from each of the first 11 sub-corpora of LINDSEI (Gilquin et al. 2010).

**Table 2** Educational background and teaching experience of the teachers of English

| | Teacher 1 | Teacher 2 | Teacher 3 | Teacher 4 | Teacher 5 | Teacher 6 |
|---|---|---|---|---|---|---|
| First language | Mandarin Chinese | English | Mandarin Chinese | Mandarin Chinese | Mandarin Chinese | Mandarin Chinese |
| Current academic position | Lecturer | PhD student and lecturer | Lecturer | Lecturer | Lecturer | Lecturer |
| Recognized qualification in teaching English | TKT Module 1, 2, 3, and Young Learners, MA in Applied Linguistics (TEFL) | TEFL/ TESOL, Oxford Cambridge Phonics, BA in Education | MA in TESOL | MA in TESOL | MA in English Literature | MA in TEFL |
| Years of teaching experience at university | 9 in Taiwan | 12 in the USA, China, Korea | 14 | 7 | 10 in Taiwan | 12 in Taiwan |
| Years of teaching experience elsewhere | 0.5 | 12 (elementary/ secondary) in the USA, China, Korea | 3 | 8 | 0 | 4 (preschool/ young learners) in Taiwan |
| Familiarity with non-English accents | Chinese | Korean, Japanese, Chinese, Hong Kong, Taiwanese, and Indonesian | Chinese | Chinese, Japanese, Singaporean, Malaysian, and Indonesian | Chinese, Japanese | Chinese, Japanese, Spanish, French, Indian, and German |
| Past experience of rating speaking | No | Yes | No | Yes | No | No |
| Familiarity with CEFR before the rater standardization training | Basic | None | Basic | None | Limited | Limited |
| Past experience in using the CEFR scales for marking spoken English | None | None | None | None | None | None |

The senior rater served as a trainer and third rater for final assessment. The two experienced raters (R1 and R2) and two teachers (T1 and T2) attended 4 h of training in rater standardization (details in the next section). After the training session, R1 and R2 rated the speaking proficiency of 100 learners on the CEFR scales.

To compare the rating performance of the two raters, the proportion of exact and adjacent agreement and discrepancies of two and three sub-bands was calculated. Spearman's rank order correlation was also used to express inter-rater reliability. Values ranging from − 1 to 1 and those close to 1 indicate a strong positive relationship between raters. Values between 0.5 and 1 are considered to show very close correlation; those between 0.3 and 0.49 show moderately close correlation; while those between 0.1 and 0.29 show only slight correlation (Cohen 1988).

R1 and R2 completed their rating in 10 days. One week after the submission of rating, ten samples of exact/adjacent agreement between R1 and R2 were sent to six teachers of English. Teachers 1 (T1) and 2 (T2) attended the rater standardization training while the other four (T3 to T6) had no experience of examining proficiency tests and received no relevant training on rating on the CEFR scales. The rating performance of these six teachers of English was evaluated by its degree of correspondence with the standard CEFR levels set by R1 and R2. The proficiency levels assigned by teachers were also compared with the benchmark, given by Raters 1 and 2.

### Rater standardization training

Raters' standardization training is commonly provided to ensure the reliability of scoring by human raters (Luoma 2004; Alderson et al. 2001, 1995) and is considered effective (Shohamy et al. 1992; Weigle 1994; Davis 2015). Two experienced raters and two teachers of English participated together in a training session provided by a trainer, who was an IELTS Examiner Trainer, and also serves as Rater 3 and the Rater Monitor for Raters 1 and 2. The purpose of this training session was to train the two experienced raters, in preparation for rating 100 samples extracted from LINDSEI-CZ and LINDSEI-TW on the CEFR scales.

At the beginning of the training, all participants signed a confidentiality agreement and filled in a rater's profile detailing their experience in English language teaching and testing. The training session followed a sequence of seven steps:

(a). The Principal Investigator opened the session and introduced the background of the rating task.

(b). The trainer and participants established a shared understanding of the five criteria on the CEFR scales: range, accuracy, fluency, phonological control, and coherence (Council of Europe 2001).[1] Phonological control had been adopted to replace interaction because interaction with the interviewer had not been required in the recordings for the rating (see Appendix).

(c). The Two Band Fit (or Match) (2BF) assessment system was introduced, and the IELTS Speaking assessment criteria (band descriptors–public version) (British Council 2015) were discussed. A 2BF is a technique taught to speaking test examiners, so that they can make an informed accurate assessment of a candidate's performance across the criteria of fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation. Each criterion outlines a list of descriptors of linguistic performance features ranging from Band A1– (very low) to C2+ (very high). In the course of the interview, examiners make a progressive assessment of a candidate's performance and equate this with two band descriptors of the four criteria described above, e.g., is the candidate's performance sample closer to the descriptor for A2, or that for B1? Further, the candidate's performance may be inconsistent across the assessment criteria: use of sophisticated lexemes, yet extended pausing and hesitation. In such cases, a candidate is awarded a marked profile to indicate the inconsistency, e.g.,
B1– or B2+. This part was particularly relevant for the two trained raters, who were also qualified IELTS examiners, activating their professional knowledge and further facilitating their rating performance on the CEFR scales.

(d). In selecting a score, the IELTS Speaking Test 2BF assessment system was applied. This asked the raters to match a speaker's production performance to the criteria and descriptors outlined in the IELTS Speaking Test Public Band Descriptors (British Council 2015), and then decide whether the performance more closely matched the descriptors of the higher or lower band outline in the IELTS Public Band Descriptors. The score was then cross-referenced to a table detailing a match between the IELTS Public Band Descriptor scores and the CEFR ratings.

(e). The CEFR levels were illustrated through six audio-taped samples[2] that had been rated by the trainer before the training. The participants reported their scores aloud and discussed reasons for the consensus score. The discrepancies were also discussed leading to a shared understanding and reducing the ambiguity of the rating scales.
(f). Five more recorded performances were rated by the participants, each followed by participants' justification and trainer's commentaries.
(g). The Principal Investigator explained the rating procedure,[3] and then the two experienced raters received an audio CD with 100 extracts and electronic rating forms. They independently rated the samples at home and returned their rating results on an Excel table within 10 days.

In addition to rating procedures for raters to follow, two measures were taken to maintain scoring reliability. First, the 100 audio-taped performances of Czech and Taiwanese learners were offered in random order at the intervals between five learners of each L1. CZ001 to CZ005, take the codes for R1, for example, were renamed 001 to 005; TW001 to TW005 corresponded to 006 to 010. The remaining files were coded in this way. This arrangement made raters listen to different accents as well as proficiency levels, which may have made it harder to be inconsistent and reduced potential measurement errors. As Bachman (2004) argues, the order of rated materials affects raters' decisions; for instance, an essay of average quality followed by some very poor essays might be rated higher than it deserves. Second, it was strongly suggested that the raters should spend no more than 2 h at a time on rating to prevent raters from being affected by fatigue.

After the rater standardization training, the two trained raters were given a CD-ROM, which contained a rating form and the audio files of the extracts from 100 learners' interviews (50 Czech learners and 50 Taiwanese learners). Each extract produced on average 5-min discussion of one of the three set topics; it took each rater at least 8 h to listen to the audio files.

## Results and discussion

### Analysis of rating performance

The major analysis in this study concerns the rating performance of the two experienced raters, who participated the rater standardization training. Their scoring reliability needed to be high, with a strong positive correlation between their scores, in order to offer learners' speaking proficiency levels as another learner variable in LINDSEI as well as setting the benchmark for the six teacher raters.

#### *Inter-rater reliability between two experienced raters*

The rating of five analytic scores and one global score by the two experienced raters resulted in a detailed proficiency level on the CEFR scales for each learner, as illustrated in Table 3. The six levels in CEFR are A1, A2, B1, B2, C1, and C2. The rating task allowed raters to add borderline levels, which were indicated by plus (+) and minus (−) signs. This resulted in 18 awards—A1−, A1, A1+, A2−, A2, A2+, B1−, B1, B1+, B2−, B2,

**Table 3** Examples of rating results by Raters 1 and 2

| Competency | Range | | Accuracy | | Fluency | | Phonological control | | Coherence | | Holistic score | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Extract\rater | R1 | R2 | R1 | R2 | R1 | R2 | R1 | R2 | R1 | R2 | R1 | R2 |
| 1 | B2+ | C1 | C1− | C1 | C1− | C1− | B2+ | C1 | C1− | C1 | B2+ | C1 |
| 2 | C1− | C1+ | C1− | C1+ | B2+ | C1+ | C1− | C1+ | B2+ | C1+ | C1− | C1+ |
| 3 | Cl− | B2+ | B2+ | B2+ | C1− | C1− | B2+ | B2+ | C1− | C1− | C1− | B2+ |
| 4 | C1 | C1+ | C1+ | C1 | C2− | C1 | C1 | C1 | C1+ | C1 | C1+ | C1 |
| 5 | B2+ | C1+ | B2+ | C1+ | C1 | C1+ | B2+ | C1 | C1− | C1+ | C1− | C1+ |

B2+, C1−, C1, C1+, C2−, C2, and C2+. These ratings on the CEFR scales were converted to numerical equivalents from 1 to 18 for the statistical analysis using Excel and SPSS.

To estimate inter-rater reliability, the degree to which the two raters agreed with each other was first examined. For holistic scoring, exact agreement was reached for 44 of the 100 extracts. Adjacent agreement (by one sub-band score) accounted for 46, and the scores which differed from one other by two or three sub-band scores are represented by 9 or 1 respectively. In addition, Spearman's rho, 0.893, exceeds the threshold of 0.5 suggested by Cohen (1988). This indicates that the variables are positively linearly related, signifying very strong correlation between Raters 1 and 2. In other words, the inter-rater reliability is very high.

Table 4 below summarizes that the correlation coefficient for holistic scoring is highest. As reported in Shechtman's (1992) study, it seems easier to assign a holistic score than to rate each of the five competencies. The inter-rater agreement is higher in holistic scoring than in analytic scoring. Similar findings were obtained for writing assessments in Zhang et al. (2015).

The current study used adjudicated band scores to assign learners' speaking proficiency levels on the CEFR scales. The 22 jagged scores were transferred to a third rater for adjudication. These include 15 cases of adjacent agreement across levels (e.g., Extract code 3 in Table 5, R1 gave C1− (numerical score 13) and R2 gave B2+ (12)); six cases with two discrepancies in awards by two sub-bands (e.g., Extract code 1 in Table 5, B2+ (12) vs. C1 (14)), and one case with two awards discrepant by three sub-bands (i.e., rater 1 gave B2− (10) and rater 2 gave C1− (13)). Three cases of two sub-band differences on the same levels (e.g., Extract codes 2 and 5 below, C1− (13) vs. C1+ (15)) were considered standard in view of the fact that both raters awarded the same level to the learners; thus, they were grouped directly without rater 3's judgment.

**Table 4** Inter-rater agreement between Raters 1 and 2 in assessing speaking on the CEFR scales

| (Dis)agreement between R1 and R2 | Range | Accuracy | Fluency | Phonological control | Coherence | Holistic score | Mean |
|---|---|---|---|---|---|---|---|
| Exact agreement 0 | 38 | 34 | 46 | 35 | 34 | 44 | 39 |
| Adjacent agreement 1 | 51 | 53 | 42 | 50 | 52 | 46 | 49 |
| Discrepancy by two sub-bands 2 | 8 | 10 | 9 | 15 | 11 | 9 | 10 |
| Discrepancy by three sub-bands 3 | 3 | 3 | 3 | 0 | 3 | 1 | 2 |
| Total | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| Spearman's rho | 0.882 | 0.887 | 0.851 | 0.877 | 0.864 | 0.893 | 0.875 |
| Jagged score | 21 | 23 | 25 | 31 | 25 | 22 | 25 |

**Table 5** Examples of rating results with converted numerical points by raters 1 and 2

| Extract\rater | R1 | Holistic score | R2 | Holistic score | Holistic score disagreement |
|---|---|---|---|---|---|
| 1 | B2+ | 12 | C1 | 14 | 2 |
| 2 | C1− | 13 | C1+ | 15 | 2 |
| 3 | C1− | 13 | B2+ | 12 | 1 |
| 4 | C1+ | 15 | C1 | 14 | 1 |
| 5 | C1− | 13 | C1+ | 15 | 2 |

The speaking proficiency levels of 100 learners were determined using the CEFR scales. The distribution is shown in Table 6. Most of the learners in LINDSEI-CZ were at Level C1 and most Taiwanese learners were at B2. Slightly over half the total of 100 learners were at B2. The C1 level accounted for 38%. There were only nine and two learners at B1 and C2 respectively.

To further examine inter-rater agreement, the holistic scores awarded by Raters 1 and 2 were cross-tabulated (see Table 7). The format of this is adapted from Luoma (2004). The columns are designated for CEFR levels and their converted numerical scores as awarded by rater 1, and the rows are designated for the scores from Rater 2. The numbers in the table total 100. The cell at the intersection of Column C2− and Row C2− is 2, which means two cases of exact agreement at C2−; the cell at the intersection of Column C1+ and Row C1 reads 3, which refers to three cases of adjacent agreement. The only case of holistic scores with a difference of three sub-bands is listed in the cell at the intersection of Column B2− and Row C1−. No ratings are very far from the diagonal.

To ease the reading of the cross-tabulation, the diagonal containing the numbers with absolute agreement are shaded with dark gray and the cells with light gray denote adjacent agreement. The numbers above the diagonal indicate samples that were awarded one sub-band higher by Rater 1 than Rater 2 (22 cases in total), whereas the numbers below the diagonal indicate cases that were scored higher by Rater 2 than Rater 1 (34 altogether). This distribution suggests that in one third of the samples, Rater 2's scores were higher than rater 1's.

It can also be seen in Table 7 that there seems to be more agreement at the higher CEFR levels. Such tendency is further explored by comparing one rater's (dis)agreement with the other across CEFR levels, shown in Table 8. To compare Rater 1 with Rater 2, the proportion of exact agreement increases from 23% at B1, 38% at B2, and 52% at C1 to 100% at C2. This phenomenon is also found in rater 2's ratings in relation to R1's, with 27% at B1, 29% at B2, 66% at C1, and 100% at C2. This suggests that it is less challenging to rate speaking samples at higher CEFR levels. It can also be interpreted that the descriptors at levels B1 and B2 are not less helpful for the judgment of experienced raters.

**Table 6** The distribution of learners' speaking proficiency on the CEFR scales

| CEFR levels | LINDSEI-CZ | LINDSEI-TW | Total |
|---|---|---|---|
| B1 | 0 | 9 | 9 |
| B2 | 12 | 39 | 51 |
| C1 | 36 | 2 | 38 |
| C2 | 2 | 0 | 2 |
| Total | 50 | 50 | 100 |

**Table 7** Cross-tabulation of holistic scores awarded by Raters 1 and 2

Cross-tabulation of holistic scores awarded by Raters 1 and 2

| **Raters** | **R1** | A1- - A2+ | B1- | B1 | B1+ | B2- | B2 | B2+ | C1- | C1 | C1+ | C2- | C2 | C2+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **R2** | Scores | 1-6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| A1-<br>\|<br>A2+ | 1<br>\|<br>6 | | | | | | | | | | | | | |
| B1- | 7 | | | 1 | | | | | | | | | | |
| B1 | 8 | | | | 1 | | | | | | | | | |
| B1+ | 9 | | | 3 | 3 | 4 | 1 | | | | | | | |
| B2- | 10 | | | | 2 | 5 | 4 | 1 | | | | | | |
| B2 | 11 | | | | 1 | 4 | 7 | 4 | | | | | | |
| B2+ | 12 | | | | | | 5 | 2 | 2 | | | | | |
| C1- | 13 | | | | | 1 | | 7 | 5 | 1 | | | | |
| C1 | 14 | | | | | | | 4 | 5 | 17 | 3 | | | |
| C1+ | 15 | | | | | | | | 2 | | 3 | | | |
| C2- | 16 | | | | | | | | | | | 2 | | |
| C2 | 17 | | | | | | | | | | | | | |
| C2+ | 18 | | | | | | | | | | | | | |

***Rating results of ten samples by six teachers of English***

Table 9 lists ten samples that were sent to the six teachers of English for independent rating. R1 and R2 awarded the same CEFR levels, except for item 3 (Extract code 003) with a slight difference of one sub-band. Teachers (T) 1 and 2 who were trained together with R1 and R2 performed equally well; each had six samples at the same level as R1 and R2 and the remaining four at an adjacent level. T3, T4, and T6 without any training in scoring on the CEFR scales diverged more from R1 and R2 with five, seven, and five samples respectively at an adjacent level. Among the jagged scores, the cells shaded in light gray are lower by one level, while those in dark gray are higher. Among these six teachers of English, T5's rating performance is the more accurate. All the four jagged cases marked by T1 are lower than R1 and R2. In contrast, T4 tended to give higher marks. Interestingly, in contrast to the tendency reported by Koster and Koet

**Table 8** Distribution of (dis)agreement between Raters 1 and 2 across CEFR levels

| Rater 1's (dis)agreement with Rater 2 across CEFR levels | | | | | | Rater 2's (dis)agreement with Rater 1 across CEFR levels | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CEFR Levels | Samples | Discrepancy (sub-bands) | | | Total | CEFR Levels | Samples | Discrepancy (sub-bands) | | | Total |
| | | 0 | 1 | 2 or 3 | | | | 0 | 1 | 2 or 3 | |
| B1 | 13 | 3 | 9 | 1 | 13 | B1 | 11 | 3 | 7 | 1 | 11 |
| | | 23% | 69% | 8% | 100% | | | 27% | 64% | 9% | 100% |
| B2 | 37 | 14 | 21 | 2 | 37 | B2 | 49 | 14 | 28 | 7 | 49 |
| | | 38% | 57% | 5% | 100% | | | 29% | 57% | 14% | 100% |
| C1 | 48 | 25 | 16 | 7 | 48 | C1 | 38 | 25 | 11 | 2 | 38 |
| | | 52% | 33% | 15% | 100% | | | 66% | 29% | 5% | 100% |
| C2 | 2 | 2 | 0 | 0 | 2 | C2 | 2 | 2 | 0 | 0 | 2 |
| | | 100% | 0% | 0% | 100% | | | 100% | 0% | 0% | 100% |

**Table 9** Holistic CEFR levels awarded by two experienced raters and six teachers of English

| No | Extract code | Learner's L1 | Rater 1 | Rater 2 | Teacher 1 | Teacher 2 | Teacher 3 | Teacher 4 | Teacher 5 | Teacher 6 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 007 | Chinese | **B1+** | B1+ | B1- | B1 | B1- | B1 | B1 | B1 |
| 2 | 008 | Chinese | **B2-** | B2- | B2 | B1+ | A2+ | B1 | B1 | B1 |
| 3 | 003 | Chinese | **B2** | B2+ | B1 | B1+ | B1- | B2+ | B2 | B2 |
| 4 | 004 | Chinese | **B2** | B2 | B1+ | B2- | B1 | C1- | B2 | B1 |
| 5 | 009 | Czech | **B2** | B2 | B2 | B2 | B2- | C1 | C1 | B2 |
| 6 | 006 | Chinese | **C1-** | C1- | C1- | B2+ | C1 | C2 | C1 | C1 |
| 7 | 005 | Czech | **C1** | C1 | B2+ | C1- | C2 | C2 | C1 | B2 |
| 8 | 010 | Czech | **C1** | C1 | C1 | C1- | C2 | C2+ | C1 | C2 |
| 9 | 001 | Czech | **C1+** | C1+ | B2+ | C2- | C1 | C1 | C1 | C1 |
| 10 | 002 | Czech | **C1+** | C1+ | C1 | C1 | C1- | C2 | C1 | C2 |

(1993), T1, a native speaker of English, does not appear to be more tolerant with learners than the non-native raters are.

In terms of the benchmarks set by R1 and R2, T1 and T2's rating performance is better than that of T3, T4, and T6. The rater standardization training could have helped enhance T1 and T2's rating performance; they said in the anonymous questionnaire that the training session had satisfied them completely. T3, T4, and T6's rating performance, on the one hand, raises doubts about the teachers' ability to judge students' oral proficiency on the CEFR scales and on the other, suggests that the descriptors of CEFR are so vague that raters' interpretations can vary greatly. If teachers of English are required to rate learners' speaking using CEFR, it is strongly recommended that rater standardization training be offered.

### *Feedback from raters*

All the raters were invited to give feedback during the training and after rating. R1 raised the issue of the length of extracts. He pointed out that a 5-min extract was much shorter than the speaking test in IELTS, in which raters awarded a band score for speaking proficiency after a 12-min interview. The trainer, however, commented that a 5-min recording was long enough for one's memory span and enough to justify a CEFR score.

After the rating task, R2 mentioned that among the five competencies, Taiwanese learners' grammar was less accurate than that of their Czech counterparts. This aspect is worth investigating by measuring learners' error rates.

Some teacher raters reported their rating process. T3 commented that marking five competencies one by one had facilitated the awarding of global proficiency levels. It seemed that analytic scores could be used to guide raters towards a global rating. T4 rated speaking based on the CEFR descriptors and set up her own standards by constantly referring to her exposure to native speech. T6 admitted to be affected by personal preference (e.g., dislike for hesitation markers). She also listened several times to the audio files, since she had difficulty in making scoring decisions, even though the guidelines required raters to award levels directly the recording was played.

T5 also found that making judgements immediately was the most challenging. Some training should have been provided for the rating task in CEFR. It was difficult,

particularly for coherence, to distinguish between B2 and C1 and between C1 and C2. Similarly, T3 suggested that scoring aids, such as exemplars of each level, should be offered. This method might have compensated for the vague description in the CEFR table.

## Conclusions

The high inter-rater agreement between the two experienced raters suggests that previous rating experience, relevant qualifications, and attendance at a session of rater standardization training are the main conditions for successful assessment. The speaking proficiency of the 100 learners in LINDSEI-CZ and LINDSEI-TW are pinned down on the scales of CEFR. This study is a pioneer among those using the LINDSEI sub-corpora. To our knowledge, these two sub-corpora and the French one are probably the only 3 out of 20 that have been assessed in CEFR. This study also, in support of Alderson's (2007) call, recommends that more learner corpora of European languages should be constructed to investigate how language proficiency develops, in order to provide empirical evidence for the CEFR scales.

It is reasonable for the CEFR descriptors to be general in order to accommodate the variety of languages in Europe. In practice, detailed descriptions and exemplars of each level are needed to extend their usefulness; however, it will certainly take some time for research of this kind to inform CEFR. For the time being, teachers of English are advised to be aware of varying interpretations of the CEFR descriptors.

Although teachers' rating performance does not diverge greatly from the benchmark, the proportion of correctly assigned levels is unsatisfactory. A lack of experience in rating speaking tests seems to be a disadvantage, for which lay assessors' many years of teaching experience have probably not compensated. Without the provision of rater standardization training to teacher raters, it appears that the descriptors on the CEFR scales are applied inconsistently by different assessors. It is therefore concluded that rating with the CEFR descriptors incurs a great deal of subjective judgment from assessors unless they are trained. This study establishes that a rating training framework using spoken learner corpus data can be used as a model and offered to (prospective) teachers of English in tertiary education.

## Endnotes

[1]The updated version (Council of Europe 2018) was not yet available when this study was conducted. In the event, the descriptors for range, accuracy, fluency, and coherence were not changed. The section on phonological control in the 2001 version was expanded in the 2018 version.

[2]The authors gratefully acknowledge the contributions of audio-taped samples from LINDSEI partners: Dr. Gaëtanelle Gilquin (French sub-corpus), Dr. Sandra Götz (German sub-corpus), and Dr. Lea Meriläinen (Finnish sub-corpus).

[3]One of the training materials, Rating Procedure, was obtained from the Centre for English Corpus Linguistics at the Catholic University of Louvain, Belgium. This was applied in the rating of five random samples from each of the first 11 sub-corpora of LINDSEI (Gilquin et al. 2010).

## Appendix

**Table 10** CEFR descriptor scales for linguistic competence (Council of Europe 2001, pp. 28, 29, 117)

| Linguistic competence | A2 (KET) | B1 (PET) | B2 (FCE) | C1 (CAE) | C2 (CPE) |
|---|---|---|---|---|---|
| Range | Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations. | Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events. | Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so. | Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say. | Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms. |
| Accuracy | Uses some simple structures correctly, but still systematically makes basic mistakes. | Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations. | Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes. | Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur. | Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions). |
| Fluency | Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident. | Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production. | Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses. | Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language. | Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it. |
| Phonological control[a] | Pronunciation is generally clear enough to be understood despite a noticeable foreign accent, but conversational partners will need to ask for repetition from time to time. | Pronunciation is clearly intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur. | Has a clear, natural, pronunciation and intonation. | Can vary intonation and place sentence stress correctly in order to express finer shades of meaning. | Can vary intonation and place sentence stress correctly in order to express finer shades of meaning. |
| Coherence | Can link groups of words with simple connectors like | Can link a series of shorter, discrete simple elements into a connected, | Can use a limited number of cohesive devices to link his/her | Can produce clear, smoothly flowing, well-structured | Can create coherent and cohesive discourse |

**Table 10** CEFR descriptor scales for linguistic competence (Council of Europe 2001, pp. 28, 29, 117) *(Continued)*

| Linguistic competence | A2 (KET) | B1 (PET) | B2 (FCE) | C1 (CAE) | C2 (CPE) |
|---|---|---|---|---|---|
| | "and", "but" and "because". | linear sequence of points. | utterances into clear, coherent discourse, though there may be some "jumpiness" in a long contribution. | speech, showing controlled use of organisational patterns, connectors and cohesive devices. | making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices. |
| Global assessment | Relates basic information on, e.g. work, family, free time etc. Can communicate in a simple and direct exchange of information on familiar matters. Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident. Can describe in simple terms family, living conditions, educational background, present or most recent job. Uses some simple structures correctly, but may systematically make basic mistakes. | Relates comprehensibly the main points he/she wants to make. Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair may be very evident. Can link discrete, simple elements into a connected, sequence to give straightforward descriptions on a variety of familiar subjects within his/her field of interest. Reasonably accurate use of main repertoire associated with more predictable situations. | Expresses points of view without noticeable strain. Can interact on a wide range of topics and produce stretches of language with a fairly even tempo. Can give clear, detailed descriptions on a wide range of subjects related to his/her field of interest. Does not make errors which cause misunderstanding. | Shows fluent, spontaneous expression in clear, well-structured speech. Can express him/herself fluently and spontaneously, almost effortlessly, with a smooth flow of language. Can give clear, detailed descriptions of complex subjects. High degree of accuracy; errors are rare. | Conveys finer shades of meaning precisely and naturally. Can express him/herself spontaneously and very fluently, interacting with ease and skill, and differentiating finer shades of meaning precisely. Can produce clear, smoothly-flowing, well-structured descriptions. |

[a]Phonological control is adopted to replace interaction because interaction with the interviewer is not required in the recordings for rating.

### Abbreviations

2BF: Two Band Fit (or Match); ACTFL: American Council on the Teaching of Foreign Languages; CEFR: Common European Framework of Reference for Languages; CELTA: Certificate in Teaching English to Speakers of Other Languages; CLC: Cambridge Learner Corpus; DELTA: Diploma in Teaching English to Speakers of Other Languages; L2: Second language; LINDSEI: Louvain International Database of Spoken English Interlanguage; LINDSEI-CZ: The Czech sub-corpus of the Louvain International Database of Spoken English Interlanguage; LINDSEI-TW: The Taiwanese sub-corpus of the Louvain International Database of Spoken English Interlanguage; R1–3: Raters 1–3; T1–6: Teachers 1–6; TESOL: Teaching English to Speakers of Other Languages; TKT: Teaching Knowledge Test

### Acknowledgements

We would like to thank the efforts of our raters, Mr. Chris Tkach, Mr. Steve Wright, Mr. Steven Lung-hsun Hsu, Mr. Michael Rossiter, Ms. I-ping Lin, Ms. Hui-ching Huang, and two other anonymous participants. Our gratitude also goes to the LINDSEI team at the Centre for English Corpus Linguistics of the Université Catholique de Louvain, Belgium, in particular, Dr. Gaëtanelle Gilquin and Dr. Amandine Dumont, for their experience in researching learner corpus data.

**Funding**
This paper was based on the project "Contrastive Interlanguage Analysis: Fluency in the spoken English of learners and native speakers," which was financially supported by the Ministry of Science and Technology, Taiwan, under grant number MOST105-2628-H-158-001.

**Availability of data and materials**
The data that support the findings of this study will be published in the second version of Louvain International Database of Spoken English Interlanguage (LINDSEI) by Presses Universitaires de Louvain, Belgium.

**Authors' contributions**
L Huang served as the principal investigator, conducting the main study and performing the data analysis. She was a major contributor in writing the manuscript. SK devised the rater standardization training and drafted the rater standardization training in the section of Methods. NK reviewed and wrote the relevant literature. L Hsu was one of the teacher raters. He also participated in the data analysis and created Tables 1, 2, and 7. All the four authors read and approved the final manuscript.

**Competing interests**
The authors declare that they have no competing interests.

## Publisher's Note



**Author details**
[1]Shih Chien University, 200 University Road, Nei-men, Kaohsiung 845, Taiwan. [2]Pu Tai Senior High School, Nantou, Taiwan. [3]University of Vaasa, Vaasa, Finland.

Received: 13 June 2018 Accepted: 31 July 2018
Published online: 17 September 2018

**References**

2013 members of the Audacity development team. (2013). Audacity. (2.0.3 ed.).
Alderson, CJ. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*(4), 659–663.
Alderson, CJ, Clapham, C, Wall, D (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
Alderson, CJ, Clapham, C, Wall, D (2001). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
American Council on the Teaching of Foreign Languages (2012). *The ACTFL proficiency guidelines 2012*. Yonkders: ACTFL.
Bachman, LF (2004). *Statistical analysis for language assessment*. Cambridge: Cambridge University Press.
British Council. (2015). Guide for teachers IELTS. Manchester, Melbourne, Cambridge, Los Angeles: British Council, IDP: IELTS Australia, Cambridge English Language Assessment.
Callies, M, Díez-Bedmar, MB, Zaytseva, E (2014). Using learner corpora for testing and assessing L2 proficiency. In P Leclercq, A Edmonds, H Hilton (Eds.), *Measuring L2 proficiency*, (pp. 71–90). Bristol: Multilingual Matters.
Carlsen, C (Ed.) (2013). *Norsk Profil. Det europeiske rammeverket spesifiert for norsk. Et forste steg*. Oslo: Novus forlag.
Cohen, J (1988). *Statistical power analysis for the behavioral sciences*, (2nd ed., ). Hillsdale: Lawrence Earlbaum Associates.
Council of Europe (2001). *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.
Council of Europe (2018). *Common European framework of reference for languages: learning, teaching, assessment companion volume with new descriptors*. Strasbourg: Council of Europe.
Davis, L. (2015). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117–135.
Gilquin, G (2018). LINDSEI Partners. https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei-partners.html. Accessed 21 July 2018.
Gilquin, G., Bestgen, Y., & Granger, S. (2016). Assessing the CEFR assessment grid for spoken language use: a learner corpus-based approach. Paper presented at The 37th International Computer Archive of Modern and Medieval English Conference (ICAME 37), The Chinese University of Hong Kong, 25-29 May 2016
Gilquin, G., De Cock, S., & Granger, S. (Eds.). (2010). LINDSEI Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.
Götz, S (2013). *Fluency in native and non-native English speech*. Amsterdam: John Benjamins.
Gráf, T (2015). *Accuracy and fluency in the speech of the advanced learner of English*. Prague: Charles University.
Granger, S, Gilquin, G, Meunier, F (Eds.) (2015). *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press.
Harrison, J, & Barker, F (Eds.) (2015). *English profile in practice*. Cambridge: Cambridge University Press.
Hawkins, JA, & Filipovic, L (2012). *Criterial features in L2 English: specifying the reference levels of the Common European Framework*. Cambridge: Cambridge University Press.
Huang, LF. (2014). Constructing the Taiwanese component of the Louvain International Database of Spoken English Interlanguage (LINDSEI). *Taiwan Journal of TESOL*, *11*(1), 31–74.
Hulstijn, JH. (2007). The shaky ground beneath the CEFR: quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, *91*(4), 663–667.
Hyland, K, & Anan, E. (2006). Teachers' perceptions of error: the effects of first language and experience. *System*, *34*(4), 509–519.
Iwashita, N, Brown, A, McNamara, T, O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49.

Koster, CJ, & Koet, T. (1993). The evaluation of accent in the English of Dutchman. *Language Learning*, *43*(1), 69–92.
Luoma, S (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
Magnan, S. (1988). Grammar and the ACTFL oral proficiency interview: discussion and data. *The Modern Language Journal*, *72*(3), 266–276.
Rose, RL (2017). Differences in second language speech fluency ratings: native versus nonnative listeners. In *Proceedings of the International Conference "Fluency & Disfluency Across Languages and Language Varieties"*, (pp. 101–103). Louvain-la-Neuve: Catholic University of Louvain.
Shechtman, Z. (1992). Interrater reliability of a single group assessment procedure administered in several educational settings. *Journal of Personnel Evaluation in Education*, *6*(1), 31–39.
Shohamy, E, Gordon, CM, Kraemer, R. (1992). The effects of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, *76*(1), 27–33.
Weigle, SC. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197–223.
Wester, M, & Mayo, C (2014). Accent rating by native and non-native listeners. In *2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (pp. 7749–7753).
Winke, P, Gass, S, Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231–252.
Wisniewski, K. (2017). Empirical learner language and the levels of the common European framework of reference. *Language Learning*, *67*, 232–253.
Xi, X. (2017). What does corpus linguistics have to offer to language assessment? *Language Testing*, *34*(4), 565–577.
Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: a mixed-methods approach. *Language Testing*, *31*(4), 501–527.
Zhang, B, Xiao, Y, Luo, J. (2015). Rater reliability and score discrepancy under holistic and analytic scoring of second language writing. *Language Testing in Asia*, *5*(5), 1–9.