# Assessing Pronunciation Proficiency of Chinese Learners of English: Development and Validation of an Intelligibility Rating Scale

Zijie Niu*
East China Normal University, China
(Co-first-author)

Yuanyue Hao*
University of Oxford, UK
(Co-first-author)

Sen Liu*
East China Normal University, China
(Corresponding author. Email: sliu@english.ecnu.edu.cn)

Received: 24 August 2021; Accepted: 15 January 2022; Published: 5 March 2022
https://doi.org/10.46451/tc.20220103

**Abstract**
Language assessment is playing an increasingly important role in English as a Second Language teaching. During the past decades, researchers have made great efforts in assessing students' language proficiency in listening, speaking, reading and writing. However, few studies have focused on pronunciation proficiency assessment. The shift in the focus of pronunciation instruction from nativelikeness to intelligibility also calls for a new rating scale for pronunciation assessment. Therefore, the authors attempted to develop and validate a rating scale for pronunciation assessment with intelligibility as the construct. This paper reported the results from the piloting study which investigated the psychometric properties of the rating scale by adopting mixed methods design. The task of passage reading was used to elicit participants' ($N$ = 30) pronunciation performance, which was assessed by four raters using the proposed rating scale. The rating scale with seven dimensions demonstrated satisfactory reliability, and exploratory factor analysis revealed a high level of unidimensionality and internal consistency. The qualitative analysis suggested that the raters could effectively use the performance descriptors to guide their scoring decisions and well distinguish students with different levels of intelligibility. Despite small sample and preliminary results, the proposed rating scale could serve as a reliable and valid instrument to assess learners' pronunciation intelligibility. Since the development and validation of a rating scale is a dynamic procedure, future research should be conducted to glean more validity evidence from different perspectives using more assessment tasks and learners of different intelligibility levels.

**Keywords**
Pronunciation assessment, intelligibility, rating scale, validity, reliability, exploratory factor analysis

## Introduction

English, as a lingua franca, is playing an increasingly important role in daily life. Better pronunciation can lead to better communication and thus help people improve their perceived social status (Derwing, et al., 2002). To facilitate the acquisition of desirable pronunciation,

assessment is essential in that it can evaluate the effectiveness of pronunciation teaching, assess the learning progress of language learners, and provide individualized feedback for learners.

Different criteria have been proposed to assess the construct of pronunciation proficiency. In the 1960s and 1970s, native-like accent was suggested as the standard for good pronunciation and this criterion has been well-accepted by EFL teachers and used to assess students' pronunciation performance. By the end of the 20th century, Munro and Derwing (1995a) proposed accentedness, intelligibility, and comprehensibility to be the main assessing criteria for foreign-accented speech. In their study, accentedness focuses on the L2 accent, comprehensibility refers to the listeners' perceived easiness to understand an utterance, and intelligibility denotes the extent to which an utterance is actually understood. It should be noted that comprehensibility not only focuses on pronunciation but also on lexical and grammatical features, while intelligibility is more associated with the listeners' perception of the speakers' utterances (Pennington & Rogerson-Revell, 2019). Therefore, intelligibility has been operationalized as the criterion to assess pronunciation proficiency in language standards. For example, the China's Standards of English Language Ability (CSE), which sets a national standard for Chinese English learners, defines phonological competence at both segmental and suprasegmental levels as the ability to express intended meanings intelligibly (Ministry of Education of the People's Republic of China, 2018). According to the CSE, intelligibility instead of accentedness is emphasized as the criterion to assess phonological competence. In the past, pronunciation teachers usually counted the number of pronunciation mistakes in learners' speech, which was subtracted from the total score to generate the final score to assess students' pronunciation proficiency. The publication of the CSE has greatly promoted the changes in pronunciation teaching and assessment practice. However, the descriptors in the "Phonological competence" sub-scale in the CSE are relatively vague and thus difficult to be used by teachers and students to make assessment. In addition, the descriptors for some dimensions are defined in details (e.g., rising and falling tones), while others are cursory (e.g., stress, rhythm). As a result, it is possible and necessary to develop an easy-to-use rating scale to measure the intelligibility of pronunciation with more detailed descriptors for each of the rating dimensions.

Therefore, this study aims to develop and validate a rating scale for pronunciation proficiency based on intelligibility, and to report the development process and psychometric properties of the rating scale. Since this is a pilot study, we only used passage reading as the assessment task to elicit learners' pronunciation performance.

## Literature Review

### Pronunciation teaching

The late nineteenth-century witnessed the Reform Movement in English pronunciation teaching. "Reform must begin with the provision of an accurate description of speech based on the science of phonetics, and there must be a properly trained language teaching profession" (Howatt, 1999, p. 172). In 1888, the International Phonetics Association published the first version of the International Phonetics Alphabet, which later became popular across the globe.

With the popularity of the Audio-lingual Method and Oral Approach, teachers began to consider pronunciation as one of the key parts of English teaching. They used minimal pairs exercise to help students compare phonemes and imitate native-like accent.

In the mid-1970s, teachers began to embrace communicative approach. As a result, pronunciation teaching was emphasized because both parties need to make their speech less

accented to ensure communicative success. In this way, teachers began to use the model of Presentation, Practice and Production (PPP) as a way of pronunciation teaching (Lan, 2006). However, in real communication, linguists find it hard to only focus on the pronunciation of phonemes and words. Suprasegmental features such as liaison, stress, rhythm attach more importance to communication (Le & Han, 2006). The focus on suprasegmental features has gradually prompted the shift of pronunciation teaching from segmental feature to the combination of both segmental and suprasegmental features.

Coming into the 21st century, under the influence of globalization, English pronunciation teaching has shifted from the sheer pronunciation dimension to the comprehensive ability of pronunciation (Tian & Jin, 2015). Besides, to help teachers form an overall understanding of their students' learning process and help students know their pronunciation proficiency, teachers should focus on their students' attitudes, efforts, and learning improvements (Liu, 2016). Therefore, pronunciation assessment plays an important role in pronunciation teaching and learning.

**Pronunciation assessment**

Pronunciation receives increasing attention from language assessment researchers (Issacs, 2014). Pronunciation assessment used to focus on whether language learners could pronounce sounds and words in an accurate way with a native-like accent. However, as Munro and Derwing (1995b) and Harding (2017) point out, accentedness may not be the barrier in second language communication. It is acceptable for non-native speakers to communicate with foreign accents if their speeches can be recognized and understood by the listeners. Moreover, the Common European Framework of Reference for Languages (Council of Europe, 2018) urges teachers to focus more on intelligibility rather than accentedness in pronunciation teaching.

Unfortunately, little empirical research has been conducted on the intelligibility of the pronunciation assessment (Zhang, 2019). Most pronunciation assessment in China still focuses on the accuracy of pronunciation and raters tend to give the scores by counting the pronunciation mistakes, which is normally regarded as an effective way to evaluate students' pronunciation proficiency. Such assessment practice makes the assumption that second language learners should aim to speak like a native speaker. However, it is well accepted that pronunciation should serve for the purpose of communicative success (Liu, 2013). Lack of validated pronunciation rating scales that focus on intelligibility hinders effective teaching of pronunciation and oral communication. Thus, it is necessary to develop a rating scale for pronunciation proficiency under the criterion of intelligibility.

**Criteria of pronunciation assessment**

Munro and Derwing (1995a) first proposed three criteria of pronunciation assessment, including accentedness, intelligibility and comprehensibility. Since then, pronunciation assessment has begun to shift from nativelikeness (accentedness) to understandable and intelligible speech.

Accentedness is defined as "how strong the talker's foreign accent is perceived to be" (Munro & Derwing, 1995b, p. 291), while comprehensibility focuses on how difficult the listeners feel in understanding a speech. Intelligibility mainly looks at the extent to which an utterance is actually understood (Munro & Derwing, 1995a).

To promote intelligibility in international communication, Jenkins (2002) proposed "Lingua Franca Core (LFC)", which is defined as a minimum core set of phonological features. There

are five main core features: consonant inventory, additional phonetic requirements, consonant clusters, vowel sound and production and placement of tonic stress. She (2012) argues that the focus of pronunciation teaching for L2 speakers should be intelligibility. Unlike the traditional criterion which aims for a native-like accent, the LFC emphasizes pronunciation intelligibility that helps the learners to keep their own pronunciation identity (Pei, 2014).

Intelligibility has been recognized as one of the major criteria to assess pronunciation proficiency. For instance, the new Companion to CEFR published in 2018 defines intelligibility as "accessibility of meaning for listeners, covering also the listeners' perceived difficulty in understanding" (Council of Europe, 2018, p. 134). The new companion to CEFR also points out that the focus on accent rather than intelligibility does harm to the development of pronunciation teaching.

Another related criterion is comprehensibility, which refers to the listener's perceived easiness to understand the utterance by the speaker (Munro & Derwing, 1995a). However, many empirical studies have revealed comprehensibility is not only associated with phonological features but also lexico-grammatical and discourse features. Crowther, Trofimovich, Saito and Issacs (2014) assessed comprehensibility of second language speech by 45 English learners from ten dimensions. They found that both phonological and lexico-grammatical variables are significant predictors of comprehensibility. Although speaking proficiency is argued to be a multi-faceted construct (de Jong et al., 2012), for the purpose of assessing pronunciation, intelligibility is considered to be a more valid criterion as it is mainly associated with phonological features (i.e., sound and prosody), without confounding influences by lexical and grammatical features.

**Validation of the rating scale**

Validation of a rating scale is a multi-stage process that provides supporting evidence for its validity through an argument-based framework (Knoch & Chapelle, 2018). During the validation process, different stakeholders such as raters, examinees, decision makers, and administrators should be involved to gather evidence for validity argument from the perspectives of task domain description, scale construction, scoring, interpretation and score use (Chapelle, Enright, & Jamieson, 2008). Of key importance is definition of the test construct and interpretation of test scores in relation to the construct (Bachman & Palmer, 2010).

Development of a rating scale should be informed by relevant theoretical framework (Bachman, 1990). For an analytic rating scale that consists of multiple dimensions, they should be representative of the latent trait and "cover the construct (i.e., no construct-irrelevance or under- representation)" (Knoch & Chapelle, 2018, p. 489). The dimensions included in the rating scales should be based on relevant models and theories. For instance, in writing assessment, Knoch (2011) argues that rating dimensions should be based on four models, which can be categorized as models of linguistic competence and models of rater behavior. The dimensions should be comprised of essential knowledge and skills according to models of linguistic competence and empirical research on how raters score using the dimensions. Therefore, to provide evidence for the validity argument of a rating scale, informed selection and inclusion of dimensions based on theoretical frameworks should be described and how raters use the rating scale to score examinees' performance should be reported. In addition, empirical research should be conducted to examine whether the dimensions are measuring the latent trait that developers claim to measure (McNamara, 1996). Different statistical analyses can be used to investigate the correspondence between the dimensions and latent trait, such as factor analysis and Rasch modelling (Knoch & Chapelle, 2018). Factor analysis is a statistical

technique that investigates the covariance between observed items and latent traits, and thus reveals the interpretable factor structure (Sims & Kunnan, 2016). A rating scale of satisfactory psychometric properties should demonstrate high factor loadings between the dimensions and intended latent trait(s). Therefore, factor analysis has been widely used in validation studies of language tests (Zhang & Luo, 2019).

This study attempts to develop and validate a rating scale to assess pronunciation intelligibility of adult learners of English, with the aim to refine the vague descriptors for phonological control in the CEFR and phonological competence in the CSE and to inform better pronunciation assessment. Specifically, this paper reports the preliminary results from a piloting study to answer the following research questions:

1. How is the rating scale developed based on relevant theories and research?
2. Does the rating scale demonstrate satisfactory psychometric properties?
3. Can raters effectively use the descriptors to assess learners' pronunciation performance?

## Methods

To answer the research questions, this study adopted a mixed methods design, which consists of content analysis of descriptors used in existing rating scales, quantitative analysis of psychometric properties of the rating scale, and qualitative analysis of rater behavior from think-aloud protocols and interviews.

### Participants

#### Examinees

Thirty first-year undergraduate students majoring in English Literature and Language at a university in Shanghai were invited to participate in the pronunciation test. There were 8 male (26.67%) and 22 female (73.34%) students. Their ages ranged from 18 to 19 with the average age of 18.28. Although the students started learning English as their second language at different ages, they all had been learning English for more than 8 years by the time of the test. They all passed the college entrance examination (CEE) to be enrolled in the university. Half of the students were required to take the oral exam during the CEE while the others were not. In order to eliminate the influence of any university courses, the test was taken in the first week of their college life. The students were required to read aloud a passage at the same time. The passage contained 114 words within 7 sentences. Its Flesch Reading Ease Score was 67.9, which is within the acceptable and standard range according to Flesch (1948).

#### Rating scale developers

There were four expert teachers involved in the development of the rating scale. Two of them are native speakers of English, while the other two are Chinese teachers. The two native speakers of English had teaching experience in English speaking for more than 5 years in China, and they are familiar with the pronunciation performance of Chinese learners of English. They both had the experience of developing rating scales to assess English speaking before. The other two expert teachers were professors from Chinese universities whose research interests were pronunciation teaching and assessment. Both of them had experience in teaching pronunciation and rating Chinese students' English pronunciation for more than 10 years and they have long been working on the rating criteria of pronunciation from accentedness to intelligibility to improve the current pronunciation teaching in China.

**Raters**

Four raters participated in scoring sessions after the development and finalization of the rating scale. Three of them were graduate students with research interests in pronunciation teaching and assessment, English phonetics, and teaching pedagogy. The other rater was professor in English phonetics and pronunciation, with teaching experience of more than 20 years. All raters had experience in teaching and assessing pronunciation of Chinese learners of English, and thus were familiar with Chinese-accented L2 English speech.

**Procedure**

**Stage 1: Development of the rating scale**

The development of the rating scale mainly focuses on the selection and determination of dimensions and descriptors, which were based on previous studies and exemplars of examinees' performance. The former was a top-down process through which the developers analyzed and selected descriptors from existing rating scales, scoring rubrics, language standards, course syllabus and relevant studies, while analysis of exemplars was a bottom-up process that invited expert teachers and scholars to write and modify the descriptors (Zhang & Deng, 2019).

The developers analyzed the dimensions and descriptors in language standards, teaching syllabus and the rating scales in English tests. Three different kinds of standards were selected: Chinese local standard (CSE), the European standard (CEFR) and the American standard (ACTFL). The teaching syllabi were the College English Major Syllabus published in 2000 and the latest English Major Syllabus (2020). The rating scale was the scoring rubric in the oral test for Test for English Majors: Band 4 (TEM 4).

In terms of exemplar analysis, seminars were held among the four expert teachers who discussed key concepts and criteria to clarify the definition of "intelligibility". Then, four expert teachers were asked to listen to ten recording samples. Informed by theoretical considerations in phonetics and phonology (Roach, 2000; Wells, 2000), they were required to list the important factors that might influence the intelligibility of the pronunciation from three different aspects: words, phrases and discourse. This process was done independently without any distraction. Then the teachers discussed their opinions on the rating dimensions and drafted the initial version of the rating scale.

**Stage 2: Validation of the rating scale**

Both quantitative and qualitative methods were applied to investigate the validity of the rating scale. Exploratory factor analysis (EFA) was applied to examine whether all proposed rating dimensions were highly associated with the latent construct, i.e., intelligibility. Think-aloud protocols (TAPs) and interviews were adopted to probe into the rating process.

**Results**

**Development of the rating scale**

**Descriptor analysis**

The first step of the development was to collect and analyze the descriptors from different sources. Microsoft Excel was used to collect and organize the descriptors, which were subsequently processed in three different ways: 1) unmodified: the descriptors met the levels of students' current pronunciation proficiency and were thus retained as they were; 2) rewritten: the descriptors were rewritten if there were some inappropriate descriptions; 3) translated: the descriptors were translated into English if they were originally written in Chinese. Some examples of descriptor analysis and processing are shown in Table 1.

It was found that most descriptors were measuring pronunciation performance at the advanced level (e.g., score of four) from the perspectives of sound, word and discourse, while there were few descriptors at beginning and intermediate levels. literature describes the top level of students' pronunciation proficiency. Though the descriptors did not explicitly mention "intelligibility", similar terms such as "clarity" and "effortless to understand" were indicative of the transition of rating criteria from nativelikeness to intelligibility. Moreover, the rating scale should describe work instead of judgment of examinees' performance (Brookhart, 1999). In this sense, modification was made to rewrite "good pronunciation" into "few segmental mistakes" to describe examinees' performance.

Table 1
*Examples of Descriptor Analysis and Processing*

| Descriptors | Source | Original descriptors | Ways of processing | Classification | Score |
|---|---|---|---|---|---|
| Accurate production of sounds | CEFR | Can articulate virtually all the sounds of the target language with clarity and precision. | Rewriting | Sound | 4 |
| Natural use of suprasegmental features that makes for fluency and comprehensibility of the utterance. | CSE | Can appropriately use stress, intonation, pitch, and volume to express meaning and attitude. | Rewriting | Discourse | 4 |
| communicate with accuracy and fluency in order to participate fully and effectively in conversations | ACTFL | Few segmental and suprasegemental errors and speech is effortless to understand. | Rewriting | Discourse | 4 |
| Correctly grasp the regular stress patterns of multisyllabic words, compounds and sentences. | Syllabus (2000) | Accurate production of the word stress | Translation | Word | 4 |
| Appropriate use of tones and intonations | Syllabus (2020) | Appropriate use of pauses, rhythm and the variety of tones | Rewriting | Discourse | 4 |
| A few pronunciation mistakes | TEM 4 | Minor mistakes in pronunciation | Translation | Sound | 2 |

**Finalized rating scale**

After the analysis and processing of existing descriptors, the four expert teachers conducted a seminar to finalize the descriptors in the proposed rating scale. The experts raised several suggestions for the improvement of the rating scale. Expert L made the following comment on the rating criterion:

> Since both "word" and "discourse" sections mention the use of suprasegmental features, what is the difference between these two sections? Moreover, if students add or delete sounds when pronouncing the words, how should we rate them?

Expert R made comments on the format of the rating scale:

> In my previous teaching experience, we designed a rating scale, using different colors to show different levels and underline the keywords in the assessment so that the raters can clearly and quickly understand the rating scale.

The experts also made a heated discussion on the number of bands in the rating scale. Expert M said that:

> We should classify students' pronunciation proficiency into five levels, which is a common way of scoring.

However, considering the central tendency bias of the five-point scale (Nadler et al., 2015), Expert C suggested that:

> I think we should still apply the four-point rating scale with scores from one to four, with score one as the lowest pronunciation proficiency and four as the highest. In this way, we can eliminate the effect of central tendency bias.

The four experts commented on different aspects of the improvement of the rating scale. Based on the review of existing descriptors and the experts' comments, the initial version of the rating scale was established. The rating scale contains seven dimensions with four bands. The seven dimensions are vowel, consonant, word stress, consonant cluster, sentence stress, intonation and pause and fluency.

**Psychometric properties of the rating scale**

The quantitative analysis focuses on the psychometric properties of the rating scale, including rater consistency and (uni-)dimensionality. Rater consistency examines whether different raters are making similar judgements on the same examinee on a given rating criterion (Bachman, 2004). Dimensionality explores the structure of the rating scale, i.e., how each of the rating dimensions is associated with the latent variable(s), and thereby investigates to what extent the dimensions are measuring the construct(s) that test developers intend to measure (Knoch & Chapelle, 2018). For this study, the seven dimensions claim to measure the intelligibility of pronunciation. Thus, for this rating scale to be of high validity, all of the seven dimensions should be significantly loaded into one latent variable, i.e., intelligibility.

**Rater consistency**

The first psychometric property examined is inter-rater reliability, which assesses to what extent the scores endorsed by one rater on one dimension are consistent with other raters on the same dimension. To investigate inter-rater reliability, Cronbach's $\alpha$ and intra-class correlation coefficient (ICC) were computed to measure the degree of agreement among four raters in rating the seven dimensions for 30 students. As Table 2 shows, all seven dimensions demonstrated a satisfactory level of rater consistency, with Cronbach's $\alpha$ exceeding the suggested threshold value of .70 (Larson-Hall, 2010). In addition, ICC was computed as a more

conservative measure of inter-rater reliability (Multon, 2012). Since students were a random sample from the population and raters were fixed, the two-way mixed effects model was used to calculate the absolute agreement among the four raters. Six dimensions demonstrated a good level of inter-rater reliability (i.e., between .75 and .90), while only "Sentence stress" showed a moderate level of reliability that fell between .50 and .75 (Koo & Li, 2016). Therefore, raters can be considered to be consistent in rating students' performance on all seven dimensions. A composite score was thus calculated for each of the dimensions by taking the average over the four raters.

Table 2
*Inter-rater Reliability Indices for Seven Dimensions*

| Dimensions | Cronbach's $\alpha$ | Intraclass Correlation | |
|---|---|---|---|
| | | Coefficient | 95% Confidence Interval |
| Vowel | .891 | .881 | [.779, .940] |
| Consonant | .827 | .808 | [.644, .902] |
| Word stress | .842 | .835 | [.700, .916] |
| Consonant cluster | .866 | .854 | [.731, .926] |
| Sentence stress | .718 | .702 | [.524, .817] |
| Intonation | .852 | .801 | [.575, .906] |
| Pause & fluency | .813 | .779 | [.580, .890] |

**Unidimensionality**

Another psychometric property that a valid scale should possess is unidimensionality, which refers to the quality of items measuring a single latent construct or trait (Knoch, 2009). Unidimensionality can provide evidence for the construct validity of the rating scale (McNamara, 1996). It reveals to what extent "expected scores are attributed to the defined construct" (Knoch & Chapelle, 2018, p. 482). Two commonly used statistical techniques to examine unidimensionality are factor analysis and Rasch modelling. Due to the small sample size in this study, exploratory factor analysis (EFA) using principal axis factor analysis was used to investigate the factor structure of the rating scale that consists of seven dimensions. The correlation matrix was presented in Table 3 for the purpose of verifiability and reproducibility of the EFA results.

The suitability of EFA was assessed prior to analysis. There were moderate to strong correlations between the dimensions. None of the correlation coefficients exceeded .80, suggesting that each dimension is likely to function independently and no serious issue of multicollinearity should be noted. The overall Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) was .87, well above the suggested cutoff value of .60 to assess the factorability of the sample data (Kaiser, 1974). Individual KMO values for three dimensions were marvelous according to Hutcheson and Sofroniou (1999): .90 for "Vowel", .91 for "Consonant", and .90 for "Consonant cluster", and the other four dimensions were meritorious: .87 for "Word stress", .80 for "Sentence stress", .89 for "Intonation", and .83 for "Pause & fluency". The correlation matrix was significantly different from an identity matrix, as assessed by Bartlett's test of sphericity, $\chi^2(21) = 148.49$, $p < .001$), indicating that the data were likely factorizable. Parallel analysis suggested a one-factor solution, which was visually corroborated by the scree plot where there was a sharp decline at the point of inflexion at the second factor (Field, 2013). Factor analysis using the principal axis factoring method revealed that the extracted factor with eigenvalue of 4.56 explained 65.14% of the total variance. The one-factor solution revealed satisfactory model fit, as assessed by RMSR = .07, RMSEA = .12, the Tucker-Lewis Index (TLI) = .92. Among these fit indices, RMSR and TLI were used to assess

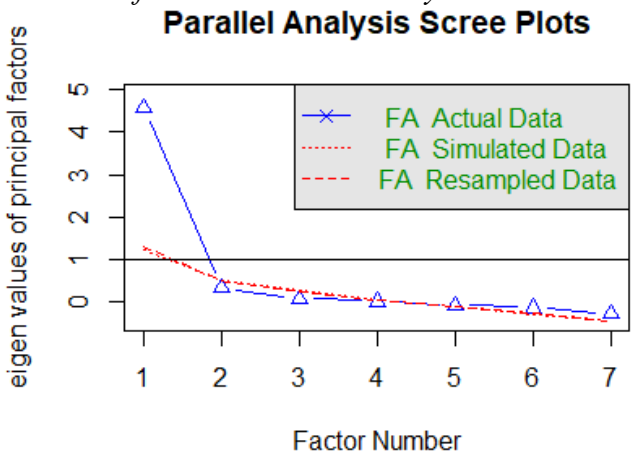
model fit, since the RMSEA tends to be large with small samples (Hu & Bentler, 1999). They suggested that the RMSR needs to be close to 0 and the TLI close to .95 for maximum likelihood factor analysis. Therefore, a relatively good model fit was obtained for the model with one latent factor.

Table 3
*Correlation Coefficients between the Dimensions*

| Dimensions | Vowel | Consonant | Word stress | Consonant cluster | Sentence Stress | Intonation | Pause & fluency |
|---|---|---|---|---|---|---|---|
| Vowel | 1 | | | | | | |
| Consonant | .558** | 1 | | | | | |
| Word stress | .627** | .654** | 1 | | | | |
| Consonant cluster | .721** | .728** | .760** | 1 | | | |
| Sentence stress | .618** | .425* | .462* | .580** | 1 | | |
| Intonation | .628** | .658** | .769** | .778** | .663** | 1 | |
| Pause & fluency | .577** | .603** | .545** | .681** | .788** | .755** | 1 |

*Note.* **: statistically significant at the level of .01 (two-tailed); *: statistically significant at the level of .05 (two-tailed)

Figure 1
*Scree Plot from the Parallel Analysis*




The factor loadings and communalities for the seven dimensions were presented in Table 4. No rotation method was applied, since there was only one extracted factor. For the unrotated EFA, the standardized factor loading reported in Table 4 is essentially both the correlation coefficient and partial standardized regression coefficient between the dimension and the latent factor. The factor loading reveals the unique contribution of one dimension to the factor. All of the factor loadings were well above the threshold value of .30, indicating that the seven dimensions loaded significantly onto the latent factor (Field, 2013). The communality refers to the amount of variance in the dimension that can be explained by the factor. For each of the dimensions in the rating scale, more than half of the variance in the observed scores can be explained by the latent construct, i.e., intelligibility. Finally, the rating scale demonstrated a satisfactory level of internal consistency, as evident by the squared multiple correlations of scores with factors (SMC). The SMC value was .942, indicating that the factor (i.e., intelligibility) was stable and well defined by the seven dimensions (Tabachnick & Fidell, 2013).

Table 4
*Factor Pattern Matrix for EFA with One-Factor Solution*

| Dimensions | Factor loadings | Communalities |
|---|---|---|
| Vowel | .764 | .584 |
| Consonant | .746 | .557 |
| Word stress | .792 | .627 |
| Consonant cluster | .895 | .801 |
| Sentence stress | .721 | .520 |
| Intonation | .897 | .805 |
| Pause & fluency | .816 | .666 |

**Analysis of rater behavior**

To answer the third research question, think-aloud protocols (TAPs) and interviews with raters were conducted to examine rater behavior in-depth, with the aim to provide additional evidence for the validity argument of the rating scale. TAPs were used to investigate the functioning of the rating scale during the rating process, and interviews were conducted to unearth raters' problems and feedback.

**Rating procedure**

TAPs were primarily concerned with attentional focus of raters during the rating procedure (Connor & Carrell, 1993). We analyzed TAPs reports from four raters, revealing their rating procedures and use of the rating scale (see Table 5, which shows the procedures and average rating time).

As shown in Table 5, the average time spent on rating the passage reading varied across the four raters. Rater 2, 3, and 4 spent approximately the same time on rating the passage reading, while Rater 1 spent more than four minutes on each recording. As for the rating procedure, all raters were found to choose to identify pronunciation mistakes when they listened to the recordings. After listening, Rater 2 and 3 made decisions on the scores based on different dimensions and then formed a general impression of the examinees' pronunciation proficiency.

Table 5
*Summary of the Raters' Rating Procedure*

| Rater | Rating procedure | Average time |
|---|---|---|
| 1 | listen while find the mistakes → relisten → give general impression → get scores based on different dimensions | 4.01 mins |
| 2 | listen while find the mistakes → get scores based on different dimensions→ give general impression | 2.50 mins |
| 3 | listen while find the mistakes → get scores based on different dimensions→ give general impression | 2.39 mins |
| 4 | listen while find the mistakes → give general impression → get scores based on different dimensions | 2.37 mins |

For example, Rater 3 reported that:

> Rater 3 - Student 24:
>
> Oh, the pronunciation of "th" was vague. (listen to the record). The consonant ellipsis here. (listen to the record) ...Why did he delete the sound /s/ in "has to do?" "th" again ... "th" was not good, so the consonant should be rated as 3 ... Generally speaking, the student's reading was good. I can understand it.The sense group and pause was good, but he read too fast. Only 3 points.

However, Rater 1 underwent a different process. Like Rater 3, Rater 1 also listened to the recordings and identified the mistakes at the same time, but later, Rater 1 decided to relisten the recordings and then formed a general impression of the students' pronunciation proficiency, which might influence his scoring decisions. The following is an episode of the rating process of Rater 1:

Rater 1 - Student 24:

> Psycho'logical? Psy'chological? Ah, the word stress was not correct .... (listen to the record) ... Generally speaking, the student performed well in passage reading. But why did he read so fast? Too fast. I cannot get accustomed to its fast speed from the very beginning ... As for consonant, 4 points, although he cannot pronounce "th" clearly, it doesn't matter.

As for Rater 4, who has expertise in language testing, he also applied the same rating procedure: first listening to the recordings, identifying pronunciation mistakes, and then forming a general impression of the students' pronunciation proficiency before assigning the scores to each dimension. As he said:

Rater 4 - Student 24:

> Umm, the pronunciation of "pond" was wrong (listening to the recording). The "th" was not very clear. In general, the student read too fast and there were some difficulties in figuring out the sense group which influenced the intelligibility to some extent, so the sense group and pause should be rated as 2. Although the "th" was vague, it had nothing to do with the intelligibility, so the consonant should be rated as 4.

In terms of the rating procedure, despite different average time spent on rating students' passage reading, all four raters underwent some similar steps, i.e., listening to the recording, identifying mistakes, forming a general impression, and giving scores to each dimension, except with different orders.

**Problems that raters met in the rating process**

During the rating process, the raters encountered some problems of which they were unsure in assigning the scores, which has negative impact on both the rating efficiency and rating quality. The TAPs reports showed their problems during the rating process. The first problem that the raters encountered was the difficulty in giving a certain score. As Rater 1 pointed out that:

Rater 1 - Student 29:

> As for the vowel dimension, umm, to be honest, there were so many mistakes, like "error", "relative", "tell" and so on. Give 1 or 2? umm, maybe 2? Time is limited, and no more hesitation.

From the TAPs reports, it can be found that some raters had some difficulties in deciding a certain score, and in particular, they were confused about the adjacent bands. In the future study, the researchers will make clearer the descriptors and differences between two adjacent bands to reduce raters' confusion.

Another problem was relatively infrequent use of the descriptors when raters made their scoring decisions. Despite clear descriptors listed in the rating scale, the raters did not always refer to the rating scale during the rating process except when they encountered some confusion.

**Analysis of raters' opinions**

This section of the qualitative analysis reports the results from interviews with raters, with focus on raters' comments on their general impression and on the rating dimensions and the rating scores.

**General impression**

Raters mainly commented on the practicality and usability of the rating scale. As for the practicality, all raters thought that the rating scale was easy to apply to rate the passage reading. Rater 2 commented that:

> Generally speaking, the rating criteria are easy to understand and the rating scale is practical because each score on each dimension has a clear description.

However, raters also pointed out some issues which need further improvements. Some raters found it difficult to tell the difference between adjacent rating scores. According to Rater 1:

> Here is a confusing problem: I find it hard to distinguish between Score 2 and 3.

Moreover, raters also met some problems on some rating dimensions. As Rater 3 pointed out:

> When there are some problems in pronunciation, I cannot judge whether the problems are caused by mispronunciation or by students' unfamiliarity with the words, especially in some vowel problems. Besides, students make a lot of mistakes in weak form, sense group and pause, so I can only give scores based on my general impression.

Despite some problems to be improved in future research, the rating scale was believed to be practical and usable by raters, who had positive and favorable impression of the rating scale.

**Rating dimensions**

Most raters thought that the rating dimensions are representative of the construct of intelligibility. In the meantime, they also suggested some improvement in terms of the rating dimensions, including the combination of consonant and consonant clusters, as Rater 1 and 2 pointed out:

> I think we should combine consonant and consonant clusters together; otherwise, there may be a higher proportion. (Rater 1)
>
> I think there is no sharp difference in consonant cluster, so in my opinion, it can be deleted. The mistakes in consonant clusters can be attributed to the mistakes in consonants. (Rater 2)

On the other hand, raters also thought that there should be some addition to the rating scale. For example, Rater 1 advised that "rhythm or stress timing" could be added to the rating scale, and Rater 3 thought that "the addition of familiarity with vocabulary is important, because it will have effects on the vowel and consonant mistakes."

Based on the raters' scoring experience, the dimensions in the rating scale still need improvement. However, we should make a cautious adjustment since the current seven dimensions are the important and difficult teaching points in the current pronunciation syllabus.

**Rating scores**

The interviews also examined discrimination and central tendency of the scores. All raters were in favor of the four rating scores, thinking that the rating scale "basically describes students' pronunciation proficiency and can discriminate their pronunciation level clearly" (Rater 3).

As for the difficulty of rating scores, the raters all agreed that Score 1 and 4 were easy to score, as Rater 2 said that:

> Score 1 and 4 are easy to score because their criteria are easy to grasp and their features are also very distinct.

But, as for Score 2 and 3, all raters faced some challenges during the rating procedure, such as unclear differentiation between the scores. According to Rater 3:

> The descriptors in Score 2 and 3 are too subjective. It reads "sufficient, moderate, appropriate, etc." How can I define these descriptors clearly?

In addition to score discrimination, the interviews also aimed to investigate whether there was central tendency during the rating process. The raters said that they tended to give Score 3 to students' pronunciation proficiency. Rater 2 commented that:

> I tend to give Score 3 because, to be honest, most students have no severe problems in the passage reading. I may focus more on the details. Once there are two or three mistakes, the score will go down to 3. Moreover, I think Score 3 is a medium score, giving this score a sense of relative reasonableness or fairness.

**Discussion**

This study aims to develop a rating scale to assess intelligibility of passage reading by Chinese learners of English and reports its construction and preliminary validation processes. The construction of the rating scale was based on theories of phonetics and phonology and informed by descriptors in existing rating scales, language standards and teaching syllabi. The preliminary validation revealed that the rating scale demonstrates satisfactory inter-rater reliability and unidimensionality, despite some problems to be improved in the future studies as suggested by raters in the interviews.

**Construction of the rating scale**

Rating scales play an important role in performance-based language assessment, as they work as "the de facto test construct" (Knoch, 2011, p. 81). Therefore, the construction of rating scales is a meticulous and iterative process. This study follows the principle of theoretically-based and empirically-developed rating scale development and validation as suggested by Knoch (2009). The proposed rating scale echoes the shift in pronunciation teaching and assessment from accentedness to intelligibility (Isaacs, 2014; Munro & Derwing, 1995a) and attempts to provide more detailed descriptors of both segmental and suprasegmental features to assess intelligibility. The construction of the rating scale is based on phonetic and phonological theories and informed by existing descriptors in language tests, teaching syllabi and language standards. Detailed ways of analyzing and processing descriptors were reported to demonstrate the content validity of the rating scale (Bachman & Palmer, 2010).

This rating scale reflects the latest trends in pronunciation teaching and assessment in that it 1) focuses on intelligibility rather than accentedness and 2) assesses both segmental and suprasegmental features of L2 English speech.

Intelligibility has been proposed as a more realistic learning goal for second language learners of English than nativelikeness (Levis, 2005). This change is partly due to the increasing use of English as a Lingua Franca across the world in many areas such as international business, communication and academics (Seidlhofer, 2011), where speaking intelligibly rather than like a native speaker of English is the key. For example, Isaacs (2008) proposed intelligibility as an

adequate criterion to assess non-native English speech to screen international teaching assistants in a Canadian university.

Since the transition from nativelikeness to intelligibility (Munro & Derwing, 1995a), many researchers have attempted to identify key features that are crucial for intelligibility. One of the most cited works is the Lingua Franca Core (LFC) proposed by Jenkins (2000). Despite its popularity, LFC has been criticized for its over-emphasis on segmental features. In fact, suprasegmental features have been found to be significant predictors of intelligibility (Kang, Thomson, & Moran, 2020) and speaking proficiency (Kang & Johnson, 2018). Suprasegmental features such as stress, intonation and rhythm are important factors in facilitating speaking proficiency (Iwashita, Brown, McNamara, & O'Hagan, 2008; Kang, Rubin, & Pickering, 2010) and oral communication (Saito, Trofimovich, & Isaacs, 2016). However, little research has investigated suprasegmental features of L2 English speech by Chinese learners, especially in terms of how these features influence speech intelligibility. Zhou, Deterding and Nolan (2019) analyzed linguistic features in English speech produced by L2 learners in central China and found that the main phonological features that impacted intelligibility were mistakes in syllables and consonant clusters. The only suprasegmental feature investigated in their study was misplacement of word stress. This study contributes to research on L2 speech of Chinese learners of English by attempting to assess intelligibility at the levels of both segmental and suprasegmental features, which also echoes with recent research on L2 pronunciation.

**Validation of the rating scale**

After construction, the validity of the rating scale was examined, with particular focus on its psychometric properties such as reliability and construct validity. The results suggested that the scale has high inter-rater reliability (Cronbach's $\alpha$ and ICC between .70 and .90) and satisfactory correspondence between the dimensions and construct (factor loadings between .70 and .90). Think-aloud protocols and interviews with raters revealed that the scale was perceived to be practical and valid in assessing intelligibility with representative and comprehensive descriptors, despite some issues to be improved in future studies.

High inter-rater reliability on all of the seven dimensions indicates that the raters had consistent understanding and interpretation of the descriptors in the scale. Their agreement in assigning scores is partly due to the detailed descriptors with sufficient description of examinees' performance and clear distinction between adjacent band levels. Despite satisfactory inter-rater reliability, raters also disclosed some confusion and problems during their rating. These issues merit further improvement in future research. In fact, establishing the number of band levels, or rating scale length, is worthy of individual empirical research (Isaacs & Thomson, 2013).

Investigation of dimensionality by exploratory factor analysis (EFA) reveals that all rating dimensions are significantly loaded onto one latent construct. The one-solution factor structure with satisfactory model fit indices suggests that the seven dimensions are all associated with one single construct, i.e., intelligibility. This means that all of the seven dimensions are significantly correlated with intelligibility only and nothing else. Such unidimensionality provides supporting evidence for the construct validity of the rating scale. The EFA results also indicate that the proposed seven dimensions are valid and adequate criteria to assess intelligibility. The results are consistent with previous studies on intelligibility which found both segmental and suprasegmental features are significant predictors of L2 speech intelligibility (Kang et al., 2020; Saito & Plonsky, 2019). This study also corroborates the criticism against Jenkins' Lingua Franca Core (2000) for its neglect of the role of

suprasegmental features in facilitating L2 speech intelligibility. The findings from this study suggest that a more balanced view should be adopted when assessing intelligibility.

Moreover, interviews with raters reveal that they think highly of the proposed rating scale because it is not only practical to use but also a valid way to assess students' pronunciation proficiency. Since the focus of pronunciation assessment has shifted from accentedness to intelligibility, this rating scale is handy to use for L2 English teachers who find it hard to evaluate students' pronunciation proficiency.

However, the raters also suggested some improvements to the rating scale. As they pointed out, they found it difficult to tell the differences between Score 2 and Score 3. Despite pre-sessional rater training, the raters still used their previous rating experiences to understand and use the rating scale. More attention should be paid to the rater training procedure. Besides, the raters also wanted further clarity of the descriptors in the rating scale. Future improvement of the rating scale could involve quantification of the mistakes so that the raters can find it more practical during the rating process. Some raters also suggested addition of some dimensions. However, too many rating dimensions may distract raters' attention and the results cannot fully reflect the students' actual proficiency.

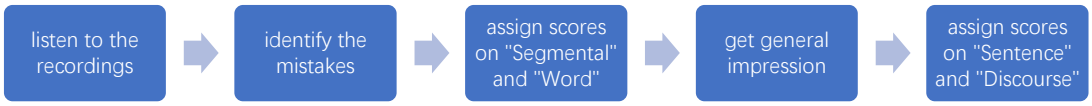
**Rating process**

We analyzed the TAPs report to explore the raters' rating procedures and the problems that they encountered during the rating process.

In terms of the rating procedure, all four raters underwent the following rating procedure: 1) listening to the recording and identify mistakes; 2) assigning scores on each dimension; 3) giving a general comment. In addition, Rater 1 would also relisten to the recording before giving the scores. Combining the rating procedure of four raters, we put forward the following general rating procedure, as in Figure 2.

Figure 2

*Rating Procedure for Assessing the Task of Passage Reading*




As shown in Figure 2, the raters would first listen to the recordings and try to find out the pronunciation mistakes at the level of segmental and word. Since the dimensions at the levels of sentence and discourse cannot be decided by the pronunciation mistakes, the raters would first form a general impression on the examinees' passage reading performance and then give the scores on the "Sentence" and "Discourse" dimensions. If encountered with any confusion, the raters would listen to the recording again before making the final decision.

For improvement of the rating scale, more attention could be paid to refining the descriptors so that the raters can distinguish the adjacent scores more easily. As for the reported underuse of the rating scale during the rating procedure according to the TAPs results, it does not necessarily mean that the raters did not base their decisions on the descriptors in the rating scale. As Barkaoui (2011) argues, the TAP may not reveal the whole rating procedure, which means that the raters might fail to mention referring to the rating scale but actually they did. In the future research, the rating can be conducted online to investigate the actual rating process via techniques such as eye-tracking and mouse-click tracking.

## Conclusion

This study aims to develop and validate a rating scale to assess intelligibility of Chinese learners' English speech. The development of the rating scale is based on theories in phonetics and phonology and previous research on L2 pronunciation to inform the inclusion and wording of descriptors in the rating scale. The analysis of psychometric properties indicates satisfactory inter-rater reliability and factor structure that is intended by the researchers. The seven dimensions that include both segmental and suprasegmental features in the scale are valid to assess intelligibility. Think-aloud protocols and interviews with raters reveal that they considered the rating scale as practical, adequate and valid to assess intelligibility of passage reading. They also suggested some points to be improved in future research. It should be noted that this study reports the results from the piloting phase of development and validation of the rating scale with a limited number of raters and students. Therefore, the results should be interpreted with caution and future research involving more participants and tasks should be conducted to provide more generalizable results.

## Appendix 1

*Rating Scale for English Pronunciation Proficiency*

| Dimensions \ Grades | | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|
| * Segmental features include: the length of the vowels, no epenthesis (sound addition), no sound ellipsis (sound reduction) | | | | | |
| Sound | Vowel | Accurate production of sounds and proficient use of segmental features*, which makes for the intelligibility of the utterance. There are less than two pronunciation mistakes. | Accurate production of sounds and sufficient use of segmental features, with occasional lapses that cause minor unintelligibility of the utterance. There are three to six pronunciation mistakes. | Correct production of sounds and moderate use of segmental features, with some lapses that influence the intelligibility of the utterance. There are seven to nine pronunciation mistakes. | Incorrect production of sounds and limited use of segmental features, with frequent lapses that severely hamper the intelligibility of the utterance. There are more than ten pronunciation mistakes. |
| | Consonant | | | | |
| * Suprasegmental features include: the appropriate use of the strong and weak forms of the grammar words, the incomplete plosions, nasal plosion, lateral plosion, etc. | | | | | |
| Word and Sentence | Word Stress | Accurate production of the word stress and natural use of suprasegmental features* that makes for the intelligibility of the utterance. There are less than two pronunciation mistakes. | Accurate production of the word stress and appropriate use of suprasegmental features that cause minor intelligibility of the utterance. There are three pronunciation mistakes. | Correct production of the word stress and moderate use of suprasegmental features, with some lapses that influence the intelligibility of the utterance. There are four pronunciation mistakes. | Incorrect production of the word stress and restricted use of suprasegmental features, with frequent lapses that severely hamper the intelligibility of the utterance. There are more than five pronunciation mistakes. |
| | Sentence Stress (word stress in sentence, i.e., notional stress and logical stress) | Accurate production of the sentence stress and natural use of suprasegmental features* that makes for the intelligibility of the utterance. There are less than two pronunciation mistakes. | Accurate production of the word stress and appropriate use sentence suprasegmental features that cause minor intelligibility of the utterance. There are three to five pronunciation mistakes. | Correct production of the sentence stress and moderate use of suprasegmental features, with some lapses that influence the intelligibility of the utterance. There are more than five pronunciation mistakes. | Incorrect production of the word stress and restricted use of suprasegmental features, with frequent lapses that severely hamper the intelligibility of the utterance. There is no sentence stress. |
| * Suprasegmental features include: use of intonations, pause between sense groups, self-repair, and hesitations | | | | | |
| Discourse | Sense Group and Pause | Natural use of suprasegmental features* that makes for comprehensibility of the utterance. Appropriate pause between sense groups, appropriate and varying use of intonations according to the intended meaning. The speech is fluent with few self-repairs and hesitations. There are less than two mismatches between the use of suprasegmental features and the intended meaning. | Appropriate use of suprasegmental features with occasional lapses that cause minor problems with comprehensibility. Some inappropriate pauses within sense groups, some inappropriate use of intonations according to the intended meaning. The speech is relatively fluent with some self-repairs and hesitations. There are less than five mismatches between the use of suprasegmental features and the intended meaning. | Moderate use of suprasegmental features with some lapses that influence comprehensibility of the utterance. Frequent inappropriate pauses within sense groups, some inappropriate use of intonations according to the intended meaning, and limited range of intonations. The speech is disfluent with many self-repairs and hesitations. There are less than seven between the use of suprasegmental features and the intended meaning. | Restricted use of suprasegmental features with frequent lapses that severely hamper the comprehensibility of the utterance. Little awareness of pauses between sense groups, many inappropriate uses of intonations according to the intended meaning, and overuse of one or two intonations. The speech is fragmental with frequent self-repairs and hesitations. There are more than seven mismatches between the use of suprasegmental features and the intended meaning. |
| | Intonation | | | | |
| | Fluency and Rhythm | | | | |

**Appendix 2**

*Extract (Bands 5-9) of the Sub-scale of Phonological/Graphological Competence in the China's Standards of English Language Ability*

| | |
|---|---|
| CSE 9 | • Can appreciate and comprehend prosodic features of English drama and poetry.<br>• Can appropriately use pronunciation, intonation, and variation of rhythm to reinforce expressive effect on academic or professional occasions. |
| CSE 8 | • Can appropriately use stress, intonation, pitch, and volume to express meaning and attitude.<br>• Can appreciate humour, exaggeration, satire, and other effects expressed through pronunciation, intonation, and rhythm. |
| CSE 7 | • Can understand major varieties of English pronunciation with little to no effort (e.g. American English and Indian English).<br>• Can use correct pronunciation and intonation to communicate in academic and professional settings.<br>• Can effectively use stress, intonation, rhythm, and liaison to express meaning and to emphasise opinions. |
| CSE 6 | • Can understand meaning expressed through varying intonation.<br>• Can use stress and rhythm to express views and attitudes. |
| CSE 5 | • Can appropriately use stress, liaison, reduced voice, and loss of plosion in daily communication.<br>• Can use tone and pitch to express emotions (e.g. surprise, anger, amazement).<br>• Can attract and maintain the listeners' attention by using stress, tone, pitch, volume, etc.<br>• Can use rising tone or falling tone to indicate beginning or ending of speech. |

**References**


Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

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

Brookhart, S. M. (1999). *The art and science of classroom assessment: The missing part of pedagogy*. ASHE ERIC Higher Education Report, *27*(1). Washington, DC: The George Washington University Graduate School of Education and Human Development.

Carr, N. T. (2006). The factor structure of test task characteristics and examinee performance. *Language Testing*, *23*(3), 269–289. https://doi.org/10.1191/0265532206lt328oa

Chapelle, C., Enright, M., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Connor, U., & Carrell, P. (1993). The interpretation of tasks by writers and readers in holistically rated direct assessment of writing. In J. G. Carson & I. Leki (Eds.), *Reading in the composition classroom: Second language perspectives* (pp. 141–160). Heinle and Heinle.

Council of Europe. (2018). *Companion volume to the common European framework of reference for languages: Learning, teaching, assessment*. Retrieved January 25, 2022, from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2015). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly*, *49*(4), 814–837. https://doi.org/10.1002/tesq.203

de Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(1), 5–34. https://doi.org/10.1017/S0272263111000489

Derwing, T. M., Rossiter, M. J., & Munro, M. J. (2002). Teaching native speakers to listen to foreign-accented speech. *Journal of Multilingual and Multicultural Development*, *23*(4), 245–259. https://doi.org/10.1080/01434630208666468

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics* (4th ed.). SAGE Publications.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Harding, L. (2017). What do raters need in a pronunciation scale? The users' view. In T. Issacs & P. Trofimovich (Eds.), *Second language pronunciation assessment* (pp. 12–34). Multilingual Matters.

Howatt, A. P. R. (1999). *A history of English language teaching*. Shanghai Foreign Language Education Press.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hutcheson, G., & Sofroniou, N. (1999) *The multivariate social scientist: Introductory statistics using generalized linear models*. Sage Publication. https://doi.org/10.4135/9780857028075

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. (2014). Assessing pronunciation. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 140–155). Wiley-Blackwell.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

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. https://doi.org/10.1093/applin/amm017

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford University Press.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics*, *23*(1), 83–103. https://doi.org/10.1093/applin/23.1.83

Jenkins, J. (2012). English as a Lingua Franca from the classroom to the classroom. *English Language Teaching Journal*, *66*(4), 486–494. https://doi.org/10.1093/elt/ccs040

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*, 31–36. https://doi.org/10.1007/BF02291575

Kang, O., & Johnson, D. (2018). The roles of suprasegmental features in predicting English oral proficiency with an automated system. *Language Assessment Quarterly*, *15*(2), 150–168. https://doi.org/10.1080/15434303.2018.1451531

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, *94*(4), 554–566. https://doi.org/10.1111/j.1540-4781.2010.01091.x

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

Knoch, U. (2009). Diagnostic writing assessment: The development and validation of a rating scale. Peter Lang.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, *16*(2), 81–96. https://doi.org/10.1016/j.asw.2011.02.003

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

Koo, T., & Li, M. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Lan, R. (2006). *Towards pronunciation teaching in Chinese universities* [Unpublished master's thesis]. Shanghai International Studies University.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. Routledge.

Le, J., & Han, T. (2006). Revision and current situation of foreign language pronunciation teaching: Principles of technique of teaching English pronunciation of jazz chants. *Foreign Language World*, 1, 16–21.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*(3), 369–377. https://doi.org/10.2307/3588485

Liu, S. (2013). B*etter pronunciation for better communication.* Shanghai Foreign Language Education Press.

Liu, S. (2016). A practical study on the English pronunciation and intonation contest in college English pronunciation teaching. *Foreign Language Learning Theory and Practice*, 3, 49–54.

McNamara, T. (1996). *Measuring second language performance*. Pearson Education.

Ministry of Education of the People's Republic of China. (2018, February 12). China's Standards of English Language Ability. Retrieved January 25, 2022, from https://www.neea.edu.cn/res/Home/1908/0c96023675649ac8775ff3422f91a91d.pdf

Multon, K. D. (2012). Interrater reliability. In N. J. Salkind (Ed.), *The encyclopedia of research design* (pp. 627-628). SAGE.

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 & Speech*, *38*(3), 289–306. https://doi.org/10.1177/002383099503800305

Nadler, J. T., Weston, R., & Voyles, E. C. (2015). Stuck in the middle: The use and interpretation of mid-points in items on questionnaires. *The Journal of General Psychology*, *142*(2), 71–89. https://doi.org/10.1080/00221309.2014.994590

Pennington, M. C., & Rogerson-Revell, P. (2019). *English pronunciation teaching and research: Contemporary perspectives*. Palgrave Macmillan.

Pei, Z. (2014). English phonetics teaching models: Theories, selection and reflections. *Foreign Language World*, 3, 88–96.

Roach, P. (2000). *English phonetics and phonology*. Cambridge University Press.

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708. https://doi.org/10.1111/lang.12345

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at

different ability levels. *Applied Psycholinguistics*, *37*(2), 217–240. https://doi.org/10.1017/S0142716414000502

Seidlhofer, B. (2011). *Understanding English as a Lingua Franca*. Oxford University Press.

Sims, J. M., & Kunnan, A. J. (2016). Developing evidence for a validity argument for an English placement exam from multi-year test performance data. *Language Testing in Asia*, 6(1), 1–14. https://doi.org/10.1186/s40468-016-0024-x

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson Education.

Tian, Z., & Jin, T. (2015). Recent development of English pronunciation assessment and testing studies: World trends and their messages for the teaching in China. *Foreign Languages in China*, *12*(3), 80–86.

Wells, J. C. (2000). *Longman pronunciation dictionary* (2nd ed.). Pearson Education.

Zhang, W. & Deng, H. (2019). Development of China's Standards of English Language Ability (CSE) taking the construction of descriptor pool for the writing ability scale as an example. *Foreign Language Testing and Teaching*, 4, 1-10+39.

Zhang, L., & Luo, W. (2019). Application of exploratory factor analysis in language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment volume I: Fundamental techniques* (pp. 243–261). Routledge.

Zhong, W. (2019). Pronunciation rating scale in second language pronunciation assessment: A Review. *Journal of Language Teaching and Research*, *10*(1), 141–149. https://doi.org/10.17507/jltr.1001.16

Zhou, W., Deterding, D., & Nolan, F. (2019). Intelligibility in Chinese English spoken in central China. *Chinese Journal of Applied Linguistics*, *42*(4), 449–465. https://doi.org/10.1515/CJAL-2019-0027

***Zijie Niu*** is a post graduate student of ECNU and member of the Oral English Teaching and Research Centre of East China Normal University. His research interest lies in pronunciation teaching and assessment.

***Yuanyue Hao*** is a doctoral student at Department of Education, University of Oxford and member of the Oral English Teaching and Research Centre of East China Normal University. His research interest includes language testing and assessment, L2 pronunciation teaching and assessment, and data science in applied linguistics.

***Sen Liu*** is professor and the Director of Oral English Teaching and Research Centre of East China Normal University (ECNU). She also acts as the Vice Chairman of Pronunciation Teaching and Phonetics Research Committee, China Association for Comparative Studies of English and Chinese (PTPRC). Her major interests are English Phonetics, Spoken English and Public Speaking teaching and research.