Abstract

This study aims to develop a rating scale for evaluating speech prosody of learners of Chinese as a second language (CSL). The researchers first gathered 41 descriptors that were perceived as crucial indicators of prosody ability through interviewing ten CSL teachers, analyzing existing Chinese speaking proficiency scales from five universities in Mainland China. After rating the perception of the selected descriptors by ninety-four CSL teachers and consulting with four expert-teachers, 15 out of 41 descriptors remained to form a rating scale. Principal component analysis revealed that 15 descriptors with three different dimensions (prosodic strategic competence, fluency, prosodic naturalness) could meaningfully describe CSL prosody. Finally, using the 15 descriptors, 29 samples of CSL learners’ speech were evaluated by four raters. A combination of the structural equation modeling and the Many-Facets Rasch modeling confirmed that all the 15 descriptors fit well with the construct of prosody ability measured, demonstrating a good validity of this rating scale.

Index Terms: rating scale, second language prosody, Chinese as a second language (CSL)

1. Introduction

Measures of prosody are essential in distinguishing speaking proficiency levels [1]–[3]. Rogers (2018) finds that both experienced and novice raters appear to attend to specific prosodic features in a reliable way, which indicates that stress, rhythm, and intonation are essential parts on speaking proficiency scales [4]. Therefore, prosody ability is included in several existing pronunciation rating scales, which has been subsumed under “pronunciation” or “fluency” in most established assessment scales, such as ACTFL Proficiency Guidelines for Speaking, or exams, TOEFL iBT speaking test [5]. The Common European Framework of Reference for Languages (CEFR) provides detailed descriptors on prosodic features including intonation, rhythm and stress and speech rate/chunking [6]. In the CEFR phonological control scale, prosodic features are charity and thoroughly described from A1 to C2 [7]. For example, a B2 speaker “can employ prosodic features (e.g. stress, intonation, rhythm) to support the message he/she intends to convey, though with some influence from other languages he/she speaks” [7]. The CEFR phonological control scale highlights the importance of intelligibility and the role it plays in communication effectiveness. This orientation is also promoted in previous studies [8], [9], which push rating scales beyond the native speaker standard and bring light into what are often just intuitive impressions [10].

While prosodic features are incorporated into some existing rating scales of English assessment, they appear to be underrepresented in CSL acquisition frameworks and are not incorporated into the assessment scales that are used by major CSL tests, such as Hanyu Shuiping Kouyu Kaoshi (HSKK), an international standardized speaking test of Chinese language proficiency. It should be noted that Chinese differs from English in that it has different prosodic features in the aspects of rhythm, intonation and tonic accent [11]. It is well-researched that Chinese is a tone language, in which the meaning of a word is affected by its pitch. For example, dòng refers to “east” while dōng means “understand”. Also, a difference in rhythmic features in terms of isochrony can be observed between syllable-timed language such as Chinese, stress-timed languages such as English [12]. Since it is difficult to identify phonetic stress in Chinese words, it is claimed that Chinese has its stress in words with a neural tone. A neutral-toned syllable of disyllabic words may change the meaning of Chinese words/phrases. For example, dōngxi refers to “things, stuffs” while dǒngxi with a full tone in the second syllable means “east and west”. Besides the discussion of neural tone, studies on prosodic syntax consider that phonologically Chinese words has a foot structure though phonetically stress is difficult to detect [13], [14]. The way Chinese prosodic structure constrained syntax is by functioning as the morphology [15], which does not exist in the Chinese linguistic system. For example, “plant flower” is syntactically correct by “种植花草 (zhǒngzhī huācāo)”, “种花(zhǒnghuā)”, and “种花 (zhǒngzhī huā)” however, the last expression is unacceptable since it contradicts with prosody. Therefore, the role prosody plays in Chinese is crucial at the word and phrase level by functioning as syntax. In the case of intonation, English and Chinese both express affirmation with a falling pitch and question with a rising pitch. However, pitch variation is only used for intonation in English while it is used for both tone and intonation in Chinese. To maintain the tonal pattern, the boundary tone is perceived higher in questions and lower in affirmation when it is compared to the syllable read independently [16]. This interaction between tone and intonation is described as the “algebraic sum” by Zhao (1929). The variation of stress in Chinese is also based on the tonal pattern. For high pitch such as the level, rising and falling tones, the stress is expressed by the post focus compression; however, for low pitch falling-rising tone, the stress is expressed by lowering the pitch and increasing the following syllable [16].

Due to the distinctive Chinese prosodic features, the phonological rating scale developed for English or other languages cannot be readily referred to in the case of assessing...
spoken Chinese. Therefore, a rating scale of speech prosody should be included in the pronunciation rating scales for CSL learners. The aim of this research, therefore, is to develop a prosody rating scale with detailed descriptions of key prosodic features. With a valid prosody rating scale, teaching and learning can be improved through anchoring the prosody performance with the rating scale descriptors. The following research questions (RQs) guided the present study:

RQ1 Which factors of speech prosody do CSL teachers consider important?

RQ2 To what extent is the proposed rating scale of speech prosody reliable, valid, and practical?

2. Method

This study is mainly divided into rating scale development and validation. Scale development drew on teachers’ intuitions to derive linguistic descriptions of prosody for developing a rating scale. Scale validation was conducted by using Many Facets Rasch Modeling (MFRM) analysis to investigate the quality of proposed rating scale.

2.1. Scale development

2.1.1. Creating a descriptor pool

A descriptor pool was generated from two sources. One source derived from the interview data of ten CSL teachers, whose subjective description about CSL learners’ prosody performance was collected. The average teaching length of these ten CSL teachers is 6.75 years (SD=2.39). They all have taught Chinese in both Mainland China and overseas. Five of them also have research background in CSL speaking and phonetics. Given the purpose of this study, the description of prosodic performance should be accessible to and comprehensible by non-expert users. Therefore, interviews were conducted to collect teachers’ intuitive and subjective interpretation of learners’ prosody ability. Ten major questions and the follow-up questions accordingly were included in the interview. For example, “What do you think foreign accent refers to?” “Is there any particularity in the foreign accent of learners from different countries?” “How would you describe the prosodic performance of novice-, intermediate- and high-level learners respectively?” “Which features of prosody would you consider the most difficult to acquire by learners?” “How would you describe a learner’s proper use of prosody?” etc. The other source of descriptors was an analysis of the available Chinese speaking proficiency scales from five universities in Mainland China. As such, a total of 96 descriptors (1120 Chinese characters), including dimensions on phonetic pronunciation, intonation, pauses, stress, the tone of speaking, rhyme and fluency, were generated from the two sources above.

2.1.2. Selecting and editing textual descriptors

This phase consists of three steps. Step one, consultation with four other CSL teachers to ensure that all the textual descriptors used in the pool are relevant and usable for the teachers. Four teachers, two of whom taught Chinese for more than ten years and another two of whom taught Chinese for four years with a research background in phonetics, read 96 textual descriptors together. They edited and moderated the textual descriptors that were found to be ambiguous, difficult to understand or negatively described. Descriptors were deleted and edited on the basis of the following principles. First, descriptors need to be relatively generic, describing performance in general, yet referring to each specific domain [17]. Second, descriptors should include the performance and criteria. Performance refers to the language operation itself and criteria refer to the intrinsic quality of the performance [18]. In Negishi’s (2012) research, conditions which refer to any extrinsic constraints or conditions defining the performance are included. However, in this study, as prosody with respect to sentence melody and rhythm was considered in assessing the pronunciation quality of CSL speakers, all descriptors are referred to as language actions a learner can perform. Therefore, prosodic performance descriptors would not be limited in specific contexts. For example, the descriptors were expressed as “can produce natural stress.” Third, the descriptors which are judged as duplicates were edited into one. For example, between “can pause accurately” and “there no inaccurate pause”, the former one was kept. Fourth, the wording of descriptors should be specific, where wordings like “mainly”, “may”, “might” and “usually” were avoided [19]. Some descriptors which describe similar features such as “can pause accurately” and “can pause appropriately”, “can express with a sense of sentence rhythm” and “can express with natural rhythm” will be both kept into the next step. Altogether 47 descriptors were then selected. Step two was conducted to obtain teachers’ intuitive impression of the selected descriptors on empirical questionnaire data. Ninety-four CSL teachers, among whom 43 teachers have more than ten years teaching experience, were asked to rate their perceptions of the descriptors in five-point scale questionnaires. All teachers were from the background in Chinese teaching, and ten of them were also researchers in phonetics or phonology fields. The rationale for this procedure was to test the clarity and appropriateness of the descriptors. Each descriptor was rated by the 94 teachers with respect to its acceptance and supportiveness on a five-point scale ranging from 1 for “completely disagree” to 5 “completely agree”. Rating was conducted via an online survey and the participants were given a small amount of honorarium as a token of appreciation. The average score of 94 teachers’ rating was calculated to obtain intuitive supportive scores. Given descriptors’ average scores of 3.5 or above, which means they were perceived as crucial indicators of prosody ability, the questionnaire data found that 41 out of 47 could survive. Step three, four expert teachers in step one were sitting together again to re-select the remaining descriptors on the principle of independency. The descriptors remained should be independent of each other. Therefore, the descriptors which are describing the same features and performance will only keep one based on the teachers’ intuitive supportive scores and the experts’ consistent conclusion through discussion. For example, “can chunk in sentence appropriately” was kept among these seven descriptors through the intuitive scores and experts’ consistent decision.

### Table 1: descriptors about “pause”.

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can pause normally</td>
<td>3.89</td>
<td>1.047</td>
</tr>
<tr>
<td>Can pause accurately</td>
<td>3.94</td>
<td>.987</td>
</tr>
<tr>
<td>Can express with necessary pauses</td>
<td>3.81</td>
<td>.936</td>
</tr>
<tr>
<td>Can pause appropriately</td>
<td>4.05</td>
<td>.864</td>
</tr>
<tr>
<td>Can pause properly</td>
<td>3.97</td>
<td>.853</td>
</tr>
<tr>
<td>Can express with natural pauses</td>
<td>3.89</td>
<td>1.026</td>
</tr>
<tr>
<td>Can chunk in sentences appropriately</td>
<td>4.19</td>
<td>.756</td>
</tr>
</tbody>
</table>

After step three, 15 out of 41 descriptors remained. The internal consistency coefficient of reliability (Cronbach α) reached .929.
which means the remaining 15 descriptors obtained a satisfactory internal consistency. They together measure a single latent aspect of individuals.

2.1.3. Dimensionality and verification of the dimensionality

The collected questionnaire data were then processed so that what 94 teachers perceived as the prosodic dimensions could be accordingly extracted. Principal component analysis (PCA) was conducted to process the responses to the questionnaires. To further verify the rationality of the dimensions and the embedded descriptors, structural equation modelling) confirmatory was conducted to verify the dimensionality.

2.2. Validation of the rating scale

Aiming to answer RQ2, four raters assessed 29 speech sample representatives of different CSL learners’ prosodic abilities by using the prosody scale developed from previous steps.

The validation of the assessing procedure was mainly examined in five aspects. First, whether the raters used all the score categories (1-4) in the scale. Second, whether the average measures value of each category increases unidirectionally from low value to high. Third, whether the design of four-score scale is acceptable for rating. Fourth, whether each score category in the scale has a separation distance, which is sufficient to distinguish students’ proficiency. Fifth, whether the probability curve of each score category has a separate peak (Linacre, 2005).

2.2.1. Speech samples

Speech samples were obtained from speaking test performances of 29 CSL learners with varied degrees of Chinese proficiency. Learners are from 12 typologically different L1 backgrounds and three HSKK levels (11 of them were certified as novice, 12 were intermediate, 6 were advanced). The reason to select these samples is to stimulate the real rating situations, where a range of L1 backgrounds and of different proficiency levels were normally included.

For the speaking test, learners have a free talk task by choosing a topic out of six. Five minutes were given to prepare for the task. Topics such as “what do you think true friendship looks like?” and “what’s the most important thing in a relationship?” were provided to elicit speakers’ spoken output. Learners were allowed to produce as long as they want. Yet, each speech sample was approximately one minute long that was intercepted from the test recordings. All the recording sessions took place in a university-based phonetic laboratory using the same hardware and recording software.

2.2.2. Raters

All four raters had research background on CSL learners’ speaking development and had more than three years’ experience of CSL teaching. They did not participate in any steps of the previous phases.

Each learners’ recording was rated by four raters independently. The raters assigned individual scores (1-4) according to each descriptor. The score of 1 represents that the learner cannot achieve the described performance at all; 2 represents that the learner can marginally achieve the described performance; 3 represents the learner can fairly achieve the described performance; 4 represents that the learner can definitely achieve the described performance with ease.

2.2.3. Statistical analysis

This study utilized MFRM analysis to investigate the quality of rating using the scale. The MFRM analysis is one of the most commonly used quantitative methods in validating rating scales, which provides a leeway to compare the ability of test takers, severity of raters and difficulty of tasks with the same logit scale [20]. FACETS (version 3.81.2) [21] was employed to carry out the MFRM analysis.

3. Results

3.1. Results of dimensionality analysis

PCA was conducted to process the questionnaire response, with the anticipation that certain descriptors might cluster together accounting for a single, unitary yet latent assessment domain. SPSS 24.0 was used to process the data. To determine whether the dataset is suitable for factor analysis, the researchers checked KMO and Bartlett’s Test of Sphericity, both indices of which test the data fitness for PCA. In this case, the KMO value is .892, indicating the suitability of the dataset for factor analysis. As for Bartlett’s Test of Sphericity, the approximate chi-square distribution of Bartlett spherical test is 850.939 (df=105, p<.001), reaching the statistical significance level. Without limiting the number of factors, PCA and varimax rotation with Kaiser normalization extracted three factors, whose eigenvalues are greater than 1.

![Scree plot for PCA](image)

Figure 1: Scree plot for PCA

Therefore, three components were extracted after PCA. The components accumulatively explained 66.176% variance.

Table 2: Principal component analysis-total variance explained

<table>
<thead>
<tr>
<th>Rotation Sums of Squared Loadings</th>
<th>Component</th>
<th>Total % of variance</th>
<th>Cumulative %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.474</td>
<td>23.160</td>
<td>23.160</td>
</tr>
<tr>
<td>2</td>
<td>3.376</td>
<td>22.505</td>
<td>45.666</td>
</tr>
<tr>
<td>3</td>
<td>3.143</td>
<td>20.955</td>
<td>66.620</td>
</tr>
</tbody>
</table>

The three components were named according to the shared features of prosodic performance. The remaining 15 descriptors formed the finalized version of prosody ability rating scale descriptors.

Table 3 Rotated Component Matrix

<table>
<thead>
<tr>
<th>Component</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can use stress to emphasize purpose (x1)</td>
<td>.835 .308 .144</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Can use intonation to express mood (x2)</td>
<td>.792 .119 .320</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Can use stress to express emotions and attitudes (x3)</td>
<td>.788 .272 .320</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Can produce accurate tone (x4)</td>
<td>.540 .456 .116</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Can produce natural intonation (x5)</td>
<td>.229 .845 .119</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Can produce natural stress (x6)</td>
<td>.288 .793 .120</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Can express with natural rhythm (x7) .022 .768 .444
Can produce natural mood (x8) .436 .563 .110
Can produce natural word connection (x9) .412 .518 .445
Can produce accurate intonation (x10) .365 .417 .303
Can chunk in sentences appropriately (x11) .435 .114 .770
Can express with appropriate speech rate (x12) .308 .020 .756
Can express with a sense of sentence rhythm (x13) -.151 .367 .751
Can express without frequent breaks (x14) .421 .262 .630
Can express without long pauses (x15) .351 .276 .502

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization
Rotation converged in 8 iterations.

Therefore, the three components were named and tentatively defined as follows: 
1. Prosodic strategic competence (x1-x4), which refers to whether the speaker can use appropriate prosodic measures to express emotions, attitudes, and/or realize other communicative purposes; 
2. Prosodic naturalness (x5-x10), which refers to the natural rhythm, stress, mood and word connection; 
3. Fluency (x11-x15), which demonstrates an appropriate speech rate, appropriate sentence chunking, a sense of sentence rhythm, no frequent breaks and/or no long pauses.

To further verify the rationality of the dimensions and their descriptors, lavaan (version 0.6-3) package [22] in R 3.6.0 was adopted to carry out SEM-based confirmatory factor analysis. 

SEM statistics revealed that the fit value of the model is 197.619, df = 87, p < .001. Figure 2 presents the confirmatory model of the three dimensions with their descriptors. It can be seen that the loading of each descriptor is within the expected range (between .5 and .95), which indicates satisfactory fitness of the model. All the descriptors have high loadings on specific dimensions, among which the highest is .91 (x3 Can use stress to express emotion and attitude on dimension prosodic strategic competence) and the lowest is .50 (x10 Can produce accurate intonation). There is no unsatisfactory estimate as the absolute value of scaling regression coefficient is greater than 1.

3.2. Results of MFRM analysis

The overall analysis based on MFRM displays 29 examinees, 4 raters and 15 descriptors on the same Rasch logit scale. As is shown in Figure 3. The estimates of students’ prosody proficiency displayed that student 4 demonstrated the best proficiency (6.76 logits) while student 24 was the least proficient (-2.91 logits). The third column (rater) displays the rating severity or leniency of each rater with more severe rater appearing higher and more lenient one lower in the column. Four raters are evenly located around .00 logit, which indicates satisfactory inter-rater consistency. The distribution range of raters (1.53 logits) is less than ¼ of the distribution range of students (9.67 logits), indicating that differences in raters’ severity do not have decisive effects on the scores obtained by students [23]. The fourth column (descriptors) makes comparisons in terms of difficulties among the 15 descriptors in the rating scale. The higher the difficulty measure of a particular descriptor, the harder it is for students. Thus, x4 (Can produce accurate tone) rank the most difficult descriptor (.89 logits), x11 (Can chunk in sentences appropriately) rank the easiest descriptor (-0.77 logits).

Figure 3 MFRM overall analysis of rating

In descriptors measurement report, the Separation Index is 2.37 and Reliability is .85, indicating that 15 descriptors are significantly different in difficulty, which can be further confirmed by the Fixed chi-square (90.2, df = 14, p < .001). Also, the Infit MnSq value of all descriptors are all within the acceptable range (0.85-1.34), indicating that 15 descriptors are of satisfactory consistency when they were utilized in rating. The Infit MnSq value indicates the extent to which the data (the difficulty of the descriptors) fit the Rasch model (the logit range of 0.5-1.5 adopted as the criteria for acceptable fitness [24]. The results of the calculation of Separation Index (2.37), indicating that the difficulty of descriptors can be divided into approximately 2-3 different levels. This result also confirmed that all the descriptors fit well with the construct of prosody ability measured.

To investigate the practicality of the developed rating scale, category of the scores was further explored. The frequency of each score category (1-4) the raters used were 343, 553, 630 and 214 times respectively, which satisfied the threshold requirement. Also, this study’s four categories’ values ranging from low to high indicate that low score corresponds to less capable students while high score corresponds to more capable students. Third, the Outfit MnSq value of each score category in the scale is less than 2 (the ideal value being 1). The
corresponding values are with the range of 0.9-1.1, indicating that the design of four-score scale is acceptable for rating. Fourth, the ideal separation distance should be between 1.4-5 logits. The distance between adjacent score categories is between 1.4-5 logits. This indicates four categories clearly distinguished students’ prosody ability. Lastly, the rating scale functionality can be visually examined in the category probability graph from FACETS. The category probability graph shows that each category has an obvious peak and presents a unidirectional increasing trend (see Figure 4).

4. Discussion

The above analysis indicates that the scales designed in this study are generally ideal and can effectively distinguish students’ prosody ability. In response to RQ1, the scale includes the following three dimensions: prosodic strategic competence, fluency and prosodic naturalness. According to the measurements report, the ranking of descriptor difficulty is different (Figure 3). The higher of the ranking of the descriptor, the more difficult it is for the students. Therefore, prosodic strategic competence and prosodic naturalness are the dimensions that is comparatively difficult for students, whereas fluency is easier. The results are in accordance with the research on the learning difficulties of specific prosody aspect. Trofimovitch and Baker (2006) conclude that L2 speech fluency measures such as infrequent and short pauses might be learned early in L2 acquisition process. Other aspects such as word stress might require several years of L2 exposure whereas the tonal peaks might not be learned even with a decade of L2 experience. As shown in the ranking (Figure 3), some fluency descriptors rank the least difficulty. In comparison, intonation, stress and rhythm pose challenges to students. This confirms that prosody learning varies according to the prosodic dimensions studied, thus again justifying the necessity of developing a separate rating scale for demystifying prosody ability.

To answer RQ2, this study conducted an MFRM analysis to examine the reliability, validity and practical utilization of the scale. Based on the MFRM results, it can be claimed that the scale has satisfactory discriminating power. The results reveal that the developed rating scale shows good internal discriminating power because its 15 descriptors are significantly different in difficulty but with good fitness indices. The data fit to the MFRM model is usually regarded as an important criterion to judge the validity of rating scales[25]. The comparatively satisfactory inter- and intra-rater consistency based on Separation Index and InFit MnSq values also indicate that raters’ performance in rating the 15 descriptors are of satisfactory consistency. Based on the calculation of Separation Index (6.93), students’ proficiency of prosody ability can be divided into approximately 7 different levels. This finding, though based on a small subset of measures for prosody ability, provides evidence in scaling prosody ability across seven levels. The validation of the scoring criteria also indicates that the design of four score categories for each descriptor are acceptable for rating.

5. Conclusions

Constructing a rating scale is an ongoing process. This present study, as part of this process, developed and validated the prosody rating scale for CSL learners. The application of this scale should be further studied. Whether the guidelines are presented in a way that is straightforward and immediately accessible to language teachers and whether it fits different CSL contexts of teaching L2 prosody remain to be explored.

6. Acknowledgements

This study was supported jointly by Shanghai municipal foundation for philosophy and social science (2020EYY002), Innovative research team of Shanghai International Studies University (2020114050). The corresponding author is Mingwei Pan.

7. References


冯胜利,《汉语的韵律、语法与句法（修订版）》, 北京大学出版社, 2009.

林茂灿,《汉语语调实验研究》, 中国社会科学出版社, 2012.


