International Journal of TESOL Studies (2020)
Vol. 2 (4) 97-109 https://doi.org/10.46451/ijts.2020.12.09

Article

Development and Construct Validation of a Diagnostic Pronunciation Rating Scale by Many-Facet Rasch Analysis

Sen Liu*
East China Normal University, China

Zijie Niu
East China Normal University, China

Yuanyue Hao
University of Oxford, UK

Abstract

Assessment of pronunciation has long been established as an integral component of speaking assessment, usually combined with other dimensions such as fluency, lexico-grammatical resources and topic development to generate an overall score for the speaking section in major English tests. Few studies have focused on assessment of pronunciation *per se*, which plays a critical role in pedagogical context such as pre-service teacher training. This study attempts to develop a diagnostic rating scale of pronunciation for the purpose of pronunciation instruction in a pedagogical practice and provide supporting evidence for the construct validity of the scale by many-facet Rasch analysis. Results from statistical analysis suggest overall satisfactory construct validity of the scale. The statistical findings are further corroborated by focus group interviews with raters and examinees. This study has implications for second language pronunciation instruction, assessment and self-learning, as well as the development and validation of diagnostic rating scale of other aspects of organizational competence.

Keywords

Pronunciation assessment, diagnostic rating scale, construct validity, many-facet Rasch model, scale development and validation

1 Introduction

Pronunciation instruction plays a vital role in second and foreign language teaching. It helps language learners improve their speech fluency (Liu, 1999), develop their effective oral communicative abilities (Wang, 2005), and boost their confidence in communicating in English (Zhang, 2011). In this way, learners are expected to be proficient in conducting fluent and efficient language communication (Liu,

***Corresponding Author**.
Address: 500 Dongchuan Road, Minhang District, Shanghai, 200241, China
E-mail: sliu@english.ecnu.edu.cn

2013; Zhang, 2007), improve the accuracy and effectiveness of oral communication (Chen & Bi, 2015), and adapt to the complicated language environment (Wu, 2012). To achieve these goals, assessment as an important part in language teaching, can be used to evaluate students' learning achievement, monitor their progress, and provide diagnostic feedback for their future learning (Jones & Saville, 2016). Assessment can also bring positive washback effects to stimulate students' learning motivation, encourage students to have a better academic performance, and bring success to both teachers and students (Liu, 1991). In second language (L2) pronunciation research, however, there are a limited number of rating scales that are designated to assess pronunciation (Zhang, 2019). Among these few rating scales, only a handful of them have been subject to psychometric analysis of their construct validity by many-facet Rasch model (see Browne & Fulcher, 2017, for pronunciation and intelligibility scoring; Shintani, Saito, & Koizumi, 2019, for accentedness and comprehensibility scoring; and Isaacs & Thomson, 2013, for accentedness, comprehensibility, and fluency scoring). Another limitation with pronunciation rating scales in previous studies is that they are not accompanied by specific, detailed descriptors of speech performance, and thus are difficult to use in pronunciation instruction classroom by teachers to provide diagnostic feedback for L2 learners. For a rating scale to be used for diagnostic purposes, it should be comprised of a number of sub-scales that are described by precise, clear, and operationalizable terminology (Knoch, 2009). Therefore, to address these limitations, this research aims to develop and validate a diagnostic pronunciation rating scale to better facilitate L2 pronunciation teaching and learning.

2 Literature Review

2.1 Pronunciation teaching and assessment

Pronunciation teaching has been an integral component in second language learning and development. Since the end of the 19th century, English pronunciation teaching has been influenced by different approaches to foreign language teaching. As Le and Han (2016, p. 16) point out, pronunciation teaching has witnessed a shift in paradigm from focus on sound articulation to focus on supra-segmental features such as stress, rhythm, and intonation. This shift has been primarily driven by the popularity of communicative language teaching, as opposed to drilling and error correction.

In addition to the shift of focus in pronunciation teaching, the assessed quality of L2 English pronunciation has also seen a transition from nativelikeness to intelligibility and comprehensibility. This is partly due to the increasing use of English as a lingua franca (Crystal, 1997; Galloway & Rose, 2015; Jenkins, 2007), subsequent problematization of the concept of native speakerism (Davies, 2003), and great difficulty in acquiring nativelike accent when learning English as a foreign language (Levis, 2005). Results from empirical studies lend supporting evidence to this transition. Derwing and Munro (2009) found that it is intelligibility that plays an important role in successful communication, whereas accentedness is partially independent from intelligibility. As a result, the objective of pronunciation instruction has shifted from accent elimination to improvement in intelligibility and comprehensibility (Derwing & Munro, 2015; Levis, 2005).

This shift can be substantiated by the fact that more detailed and specific descriptors on intelligibility have been provided in the Companion Volume to Common European Framework of Reference for Languages (CEFR) (2018). The new volume refrains itself from using the terms such as "native speakers", "non-standard accent", and "foreign accent" as in the original version published in 2001. In the new companion volume, assessment of pronunciation is primarily operationalized by the concept of intelligibility, which focuses on the effectiveness of meaning conveyance. Similarly, in China, the Ministry of Education published the China's Standards of English Language Ability (CSE), which provides a reference framework for English assessment, teaching and learning in the Chinese context. The phonological competence rating scale in the CSE stresses meaning-focused communication competence and no longer requires Chinese learners of English to acquire native-like accent (Ministry of Education, 2018).

The fundamental changes in those two documents were driven by a number of conceptual and empirical research on the relationship between accentedness, intelligibility, and comprehensibility (notably by Levis, 2005; Munro & Derwing, 1995; 2011; Saito & Saito, 2017). However, studies on pronunciation assessment receive relatively less attention from researchers (e.g. Harding, 2013; Isaacs & Trofimovich, 2017; Kang & Ginther, 2018). Pronunciation assessment plays an important role in judging whether the language communication is effective and meaningful. By reviewing previous studies on pronunciation rating scales in L2 assessment, Zhang (2019) summarizes the research gaps and weaknesses from the perspectives of the construct, criterion, and descriptors. She suggests that more research should be conducted to examine the reliability and validity of pronunciation rating scales.

In the current language assessment and teaching practice in China, little empirical research on pronunciation assessment has been carried out (Tian & Jin, 2015). The pronunciation rating scales used in some large-scale standardized language tests, such as the TOFEL, ILETS, CET-4, CET-6, are generic and sometimes ambiguous in their descriptors, thus impractical to be used in diagnostic language assessment that aims to identify language learners' strengths and weaknesses, and correspondingly provide individualized feedback and remedial learning (Alderson, 2005; Harding, Alderson, & Brunfaut, 2015; Lee, 2015). Scores obtained from the diagnostic language assessment are useful for making low-stakes decisions, charting profiles of language learners' abilities, providing feedback, and recommending future learning resources. It can be considered as an important part in learning-oriented assessment and learners' self-instruction (Lee, 2015).

Similar problems can be identified in classroom pronunciation teaching. Huang and Jia (2016) found that teachers usually provide corrective feedback to students' repeated mistakes, while students wish to have more detailed mistake-based feedback to improve their pronunciation. Such specific feedback is crucial for students' accurate self-assessment of their own pronunciation proficiency.

In this sense, it is of great significance to develop and validate a pronunciation rating scale accompanied by detailed performance descriptors that can be used to provide specific diagnostic feedback for L2 learners and facilitate their learner agency in pronunciation learning.

2.2 Rating scale construct validation by many-facet Rasch model

The validation of a rating scale follows an argument-based validity framework, which guides researchers to collect data and provide backing evidence for the claim that this scale is valid in score interpretation and use (Bachman & Palmer, 2010; Knoch & Chapelle, 2018). In this framework, the construct validity of a rating scale is of particular interest, as it is concerned with "the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores" (Bachman & Palmer, 1996, p. 21). Weigle (2002) defines construct validation as the process to ascertain whether a test is measuring what it intends to measure. Supporting evidence for the construct validity can come from different sources such as correlation analysis, factor analysis, and item response theory and Rasch model (Chapelle, Enright, & Jamieson, 2008; Knoch, 2009). Among these methods, many-facet Rasch model is commonly used in the construct validation of writing and speaking assessment to examine item fit, item difficulty, rater consistency, and rater severity (Fan & Ji, 2017).

Many-facet Rasch model (MFRM) specifies that the probability that an examinee is awarded a score on a rating criterion is jointly determined by a number of factors (facets), including examinee competence, rater severity, rating criterion difficulty, and relative difficulty of band levels (Bond, Yan, & Heene, 2020; Eckes, 2015). The model demonstrates all facets, as well as all elements in each facet on a single logit scale, which displays the distribution of estimated values of each facet and element vividly for the purpose of invariant measurement and direct comparison. More importantly, the MFRM can provide fit statistics of each rating criterion to investigate whether these criteria are measuring the same construct, thus providing empirical evidence for the construct validity claim.

3 Methods

3.1 Research questions

This study attempts to answer the following two research questions:

(1) To what extent does diagnostic pronunciation rating scale have construct validity?

(2) Does the rating scale have positive washback effect on pronunciation learning?

3.2 Data

The read-aloud task, which is widely used in pronunciation assessment (Thomson & Derwing, 2015), was used in this study to assess segmental and supra-segmental features of L2 speech by Chinese adult learners of English. Participants were first-year English-major undergraduate students in a university situated in eastern part of China. Recordings of read-aloud speech by 30 participants were obtained on their consent from their final examination of the course "English Phonetics and Pronunciation". Recordings of 15 male and 15 female students were selected. The passage used in the read-aloud task contains 89 words and its Flesch Reading Ease score is 61.6 (within standard range). The passage consists of one simple sentence, two compound sentences, and one complex sentence.

3.3 Raters

Three raters were recruited in the research. One rater is an experienced teacher who is expert in pronunciation teaching and has taught English pronunciation for more than 30 years, while the other two raters are postgraduate students with research interest in L2 English pronunciation assessment. All raters are experienced in assessing L2 English pronunciation.

3.4 Rating scale

The rating scale in this study was designed with reference to the CEFR and CSE. The CEFR 2018 Companion Volume uses intelligibility, defined as "accessibility of meaning for listeners" (p. 134), to assess language learners' phonology control. The CSE, on the other hand, though not explicitly using the term "intelligibility", assesses learner's phonological competence in terms of their ability to use segmental and suprasegmental features to express meaning, emotions, and attitudes.

Informed by theoretical considerations in phonetics and phonology (Roach, 2000; Wells, 2000) and pronunciation teaching practice (Chen & Li, 2017; Liu, 2016; Ma & Zhao, 2017; Pei, 2014), the rating scale consists of seven rating criteria (i.e. vowels, consonants, word stress, consonant clusters, sentence stress, intonation, and pause and fluency), which can be broadly categorized into three factors (i.e. sound, word, and sentence and discourse). The rating scale takes the form of 4-point Likert scale. Raters are expected to assign scores from 1 to 4 on the seven rating criteria to speech recordings of learners' pronunciation performance according to the band descriptors.

3.5 Data analysis

Many-facet Rasch analysis was applied to investigate the construct validity of the rating scale. The MFRM used in this study can be specified as:

$\ln(P_{nrik}/P_{nri(k-1)})=B_n-T_r-D_i-F_k$, where

P_{nrik} = the probability of examinee *n* receiving a rating of *k* from the rater *r* on rating criterion *i*

B_n = examinees' pronunciation proficiency

T_r = rater severity

D_i = rating criterion difficulty

F_k = difficulty of receiving a score of *k* relative to *k-1*.

The rating data were analyzed by the software Minifac 3.81.2.

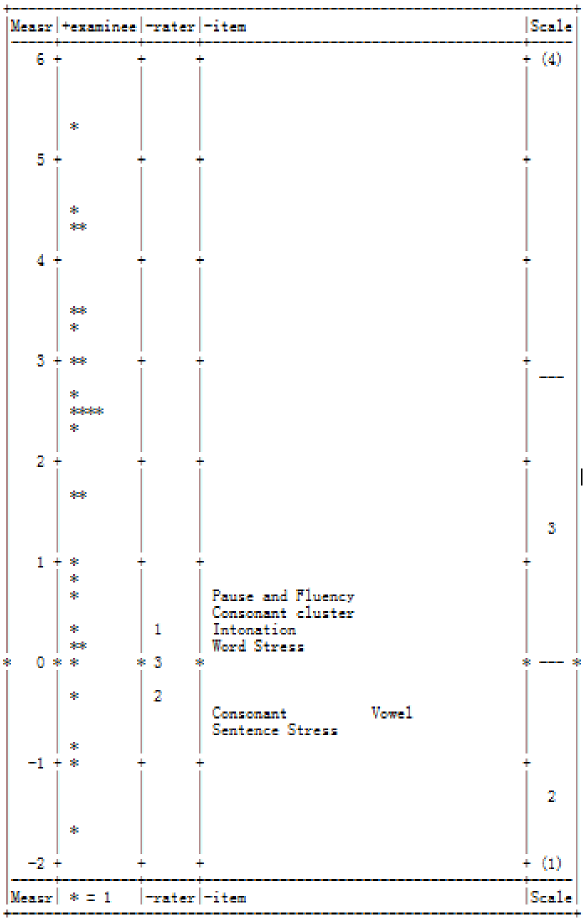
In addition, interviews with raters and examinees were conducted to explore the advantages and disadvantages of the application of this rating scale. Three raters and five examinees were invited to the interviews.

4 Research Results

4.1 Overall analysis

The correspondence among all elements in each facet can be displayed on a ruler-like variable map as in Figure 1. The figure can provide rich information regarding the distribution of elements of each facet and relative positions of all facets on the map.

In the figure below, the first column is the logit scale on which each facet element is estimated. The second column shows the distribution of the first facet, i.e. examinees' pronunciation proficiency. Each asterisk represents one element, i.e. one examinee.



```
+---------------------------------------------------------------------+
|Measr|+examinee|-rater|-item                                  |Scale|
|-----+---------+------+---------------------------------------+-----|
|   6 +         +      +                                       + (4) |
|     |         |      |                                       |     |
|     |         |      |                                       |     |
|     | *       |      |                                       |     |
|     |         |      |                                       |     |
|   5 +         +      +                                       +     |
|     |         |      |                                       |     |
|     |         |      |                                       |     |
|     | *       |      |                                       |     |
|     | **      |      |                                       |     |
|     |         |      |                                       |     |
|   4 +         +      +                                       +     |
|     |         |      |                                       |     |
|     |         |      |                                       |     |
|     | **      |      |                                       |     |
|     | *       |      |                                       |     |
|     |         |      |                                       |     |
|   3 + **      +      +                                       +     |
|     |         |      |                                       | --- |
|     | *       |      |                                       |     |
|     | ****    |      |                                       |     |
|     | *       |      |                                       |     |
|     |         |      |                                       |     |
|   2 +         +      +                                       +     |
|     |         |      |                                       |     |
|     | **      |      |                                       |     |
|     |         |      |                                       |     |
|     |         |      |                                       |  3  |
|     |         |      |                                       |     |
|   1 + *       +      +                                       +     |
|     | *       |      |                                       |     |
|     | *       |      | Pause and Fluency                     |     |
|     |         |      | Consonant cluster                     |     |
|     | *       |  1   | Intonation                            |     |
|     | **      |      | Word Stress                           |     |
*   0 * *       * 3    *                                       * --- *
|     |         |      |                                       |     |
|     | *       |  2   |                                       |     |
|     |         |      | Consonant            Vowel            |     |
|     |         |      | Sentence Stress                       |     |
|     |         |      |                                       |     |
|     | *       |      |                                       |     |
|  -1 + *       +      +                                       +     |
|     |         |      |                                       |     |
|     |         |      |                                       |  2  |
|     |         |      |                                       |     |
|     | *       |      |                                       |     |
|     |         |      |                                       |     |
|  -2 +         +      +                                       + (1) |
|-----+---------+------+---------------------------------------+-----|
|Measr| * = 1   |-rater|-item                                  |Scale|
+---------------------------------------------------------------------+
```

Figure 1. Variable map of each facet

The findings suggest that the examinees' pronunciation proficiency ranges from -2 logits to +6 logits. The third column depicts the distribution of estimated values of the second facet, rater severity. Rater 1 is found to be the most severe, while Rater 2 is the least severe. The severity measures spread within a much more limited range when compared with the second column, which indicates that the rater severity does not have great influence on the ratings (Myford & Wolfe, 2004). The fourth column shows the distribution of the third facet, rating criteria difficulty measures. Higher position on the logit scale suggests higher rating criterion difficulty. It can be observed that supra-segmental features, such as pauses and fluency, intonation, and word stress, are more difficult for the examinees than segmental features. The last column demonstrates the functioning of each rating category, i.e. band level in the rating scale, which will be discussed in detail in Section 4.3.

Summary statistics of the many-facet Rasch model are shown in Table 1. The separation index of the examinees is 3.78, which indicates that their pronunciation proficiency can be divided into four distinct levels, consistent with the number of rating categories in the scale. The reliability index is 0.93, suggesting that examinees' pronunciation proficiency can be reliably separated into four levels. As for the rater severity, its separation index is 1.78, indicating that it can be divided into two levels. Its reliability index is 0.75, which means that the separation of the rater severity is reliable. For further observation, it can be noticed that Rater 2 and 3 are close to each other in terms of their severity, while Rater 1 constitutes a distinct cluster. This is partly because Rater 2 and 3 are postgraduate students and novice teachers of pronunciation, and they don't have much experience in rating students' pronunciation proficiency compared with Rater 1, who is an experienced teacher.

Table 1.

Summary Statistics of the MFRM

	Examinee	Rater	Rating Criterion
Average Mean	1.86	.00	.00
S.E.	1.78	.23	.50
Chi-Square	429.3*	8.1*	42.0*
df	27	2	6
Separation	3.78	1.75	2.46
Reliability	.93	.75	.86

Note. * $p < .05$

4.2 Rating scale fit analysis

Fit statistics of the seven rating criteria are shown in Table 2. The second column "Measure" shows the estimated values of the difficulty of each rating criterion in the unit of logits, with higher value indicating higher difficulty. The difficulty measures of the seven rating criteria range from 0.7 to -0.69 logits. The most difficult criterion is "Pause and Fluency", while the easiest one is "Sentence Stress". In general, the supra-segmental features are more difficult than segmental features for Chinese learners of English. Results from Table 1 also indicate that there are statistically significant differences in terms of rating criterion difficulty ($\chi^2 = 42.0$, df = 6, $p < .05$), and that the rating criteria can be reliably divided into two levels (with separation index = 2.46 and reliability = 0.86). This can be explained by the fact that English pronunciation teaching in China has traditionally attached more importance to segmental features teaching, while neglecting supra-segmental features (Chen & Bi, 2015).

In addition, the results from Table 2 reveal that all Infit and Outfit mean-square statistics are between 0.5-1.5 logits, which can be interpreted as satisfactory fit of data to the mathematical model (Linacre, 2002). The results also suggest that all rating criteria are measuring the same construct (i.e. pronunciation proficiency) that the rating scale purports to measure, thus providing empirical backing evidence for the construct validity of this scale.

Table 2.
Fit Statistics of Rating Criteria

Rating Criterion	Measure	S.E.	Fit Statistics			
			Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd
Pause and Fluency	.70	.19	1.16	1.0	1.10	.6
Consonant Cluster	.48	.20	.91	-.5	.86	-.8
Intonation	.30	.20	1.08	.5	1.25	1.4
Word Stress	.14	.20	.85	-.9	.87	-.7
Vowel	-.45	.21	1.24	1.4	1.28	1.3
Consonant	-.48	.21	.93	-.3	.90	-.4
Sentence Stress	-.69	.21	.69	-2.0	.64	-1.9

4.3 Effectiveness of the rating scale

To provide further evidence for the validity of the rating scale, the usage of the scores should also be taken into consideration. Linacre (2004) proposes five standards for the effectiveness of the rating scale: 1) the use frequency of each category on each criterion should be over 10; 2) average measure increases monotonically from the easiest category to the most difficult one; 3) the Outfit MnSq of each category in the scale is lower than 2; 4) the separation interval between each category can well discriminate the ability of the examinees; 5) each category has its independent top point in the Category Probability Plot.

Table 3.
Statistics of the Rating Categories in the Rating Scale

Scale	Frequency	Average Measures	Outfit MnSq	Rasch Thurstone Thresholds
1	16(3%)	-1.08	1.0	
2	121(22%)	-.15	.9	-2.72
3	220(41%)	1.57	1.0	.06
4	186(34%)	3.21	1.0	2.65

According to Table 3, raters use all of the four rating categories during the rating process and the frequency of each category is higher than 10. The average measures range from -1.08 to 3.21 logits, increasing monotonically from the easiest category to the most difficult one, which indicates that examinees with lower pronunciation proficiency receive lower scores when compared with their peers with higher proficiency. All categories' Outfit MnSq values are less than 2 and approximately equal to 1, which is a favorable fit result. The "Rasch Thurstone Thresholds" column indicates that the values range from -2.72 to 2.65, increasing monotonically from the easiest category to the most difficult one, which suggests effective discrimination of the pronunciation proficiency among the examinees. In addition, it can be observed from Figure 2 that each category has its own independent peak, which indicates the effectiveness of each category to assess an examinee's pronunciation proficiency.

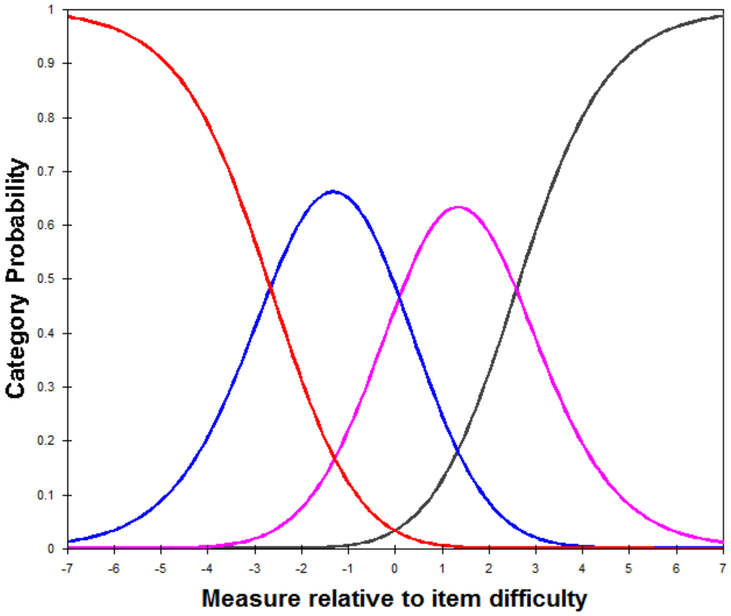


Figure 2. Category probability plot of each rating category

4.4 Interview analysis

To better understand the validity of the rating scale from the perspective of the raters, two raters were invited to interviews in order to explore their perceptions of and attitudes towards the rating scale. The interview was recorded, transcribed, and analyzed by two of the researchers, and the following two main themes can be identified from the interview:

4.4.1. A brand-new rating method

According to one of the raters, this rating scale operationalizes intelligibility as the construct of pronunciation. This method is more valid and efficient compared with the former rating method that counted the number of pronunciation mistakes in students' recorded speech. It can assess students' pronunciation proficiency more accurately. The other rater also points out that instead of using the ambiguous and complicated global rating scale, this scale presents a clearer and more concise rating method in pronunciation assessment. The descriptors in the rating scale can provide a detailed analysis of students' pronunciation proficiency, and can be used in individualized pronunciation tutoring and teaching. During the interviews, raters show a positive attitude towards this rating scale. They prefer to attach more importance to the effectiveness of communication facilitated by more intelligible L2 speech rather than traditional error-based pronunciation correction.

4.4.2. The washback effect of the rating scale on pronunciation learning

Both raters believe that this rating scale will bring more positive washback to pronunciation teaching and learning. As one of the raters points out, the descriptors in the rating scale can be used to provide diagnostic information for students. Their pronunciation performance is assessed on a range of rating criteria that are accompanied by detailed descriptors that are written with reference to the CEFR and CSE. The scores, together with corresponding performance descriptors, are communicated to students in a meaningful way, in which students can be aware of their weaknesses and strengths of their pronunciation. They can refer to the rating scale descriptors for the explicit descriptions of their current pronunciation proficiency in seven aspects, and for learning guidance if they want to improve their pronunciation and progress into a higher level. The rating scale can be also applied by students in self and peer assessment, which plays an important role in facilitating their learner agency.

In addition to the raters, five students were also invited to the interviews. They made positive comments on the effects of the rating scale on their pronunciation learning. They all talked about the "accurate placement" of their pronunciation proficiency on the scale. One student said, "In the past, I wondered whether my pronunciation was good or not and how to improve my pronunciation. A pronunciation rating scale would be a good tool for me to quantify my pronunciation proficiency." The scale can also be used to track students' longitudinal development of their pronunciation proficiency, as is pointed out by another student that "it can keep track of my progress, which will build up my confidence and strongly motivate me to improve and practice my pronunciation." In short, the rating scale not only helps the examinees to locate their own pronunciation proficiency, but also records their growth and progress in the process of learning, and promotes their self-confidence and motivation.

5 Discussions and Conclusion

5.1 The validity of the rating scale

This research aims to provide an alternative assessment method for the current and future pronunciation teaching in China, by developing and validating a diagnostic pronunciation rating scale with detailed descriptors. It echoes the shift in the 2018 CEFR Companion Volume from the traditional criterion of nativelikeness to intelligibility that is a key to successful oral communication. By means of MFRM and interviews, the researchers are able to gather positive supporting evidence for the validity argument of the diagnostic rating scale.

However, some limitations with this rating scale were also mentioned by raters during the interviews, such as the clarity of the descriptors, and the addition and reduction of the rating dimensions. Some words in the descriptors are subject to individualistic and impressionistic interpretation such as "accurate", "natural", and "moderate". Also, it should be noted that the sample size is relatively small and the examinees are all English major undergraduate students from the same university. Although there are some individual differences among the examinees, they still have limited representativeness, so more examinees from diverse educational backgrounds should be included in the analysis to provide more convincing and reliable research results.

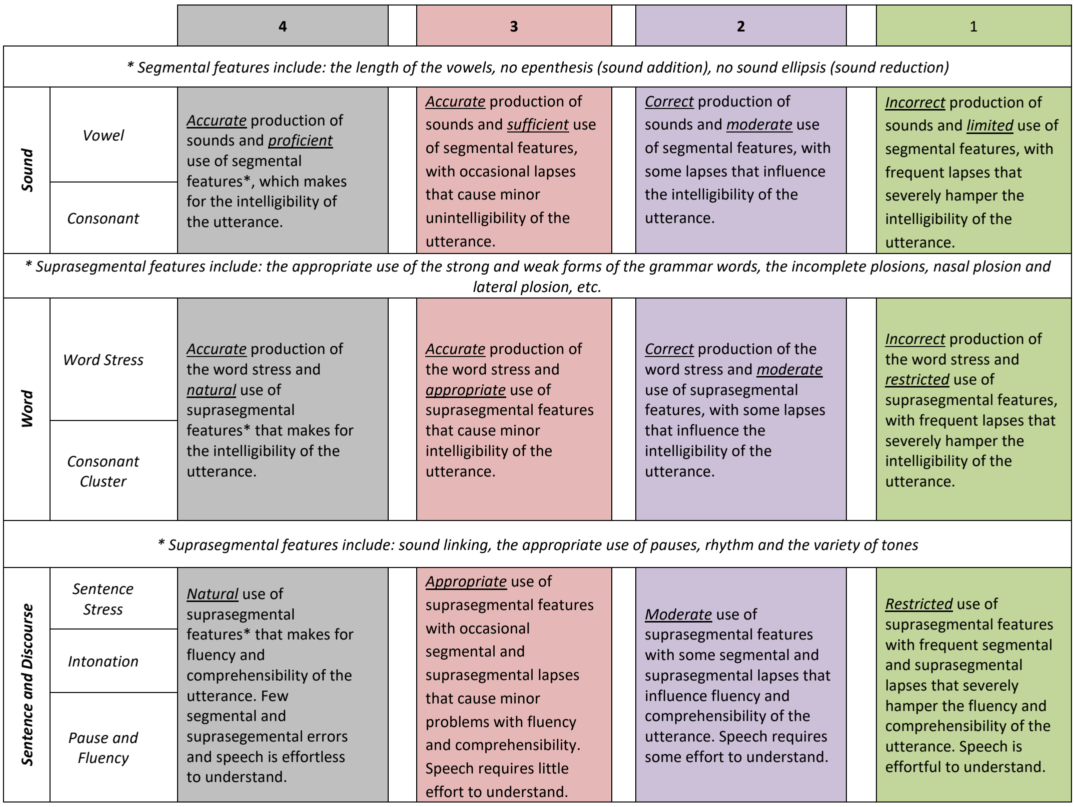
5.2 Application of the rating scale

The feedback from both raters and examinees shows that the application of this rating scale can greatly facilitate pronunciation teaching. Based on our practice, we suggest that the rating scale should be applied in self and peer assessment in a multi-dimensional pronunciation assessment system (Liu & Niu, 2018) to motivate students' learning and build up their confidence and promote their learning progress.

The rating scale application should be further improved. As the raters point out in the interviews, the scale should be used more specifically in different learning phases, different learning tasks, and different teaching objectives. Before being put into use, the rating scale should be validated to ensure that it can measure students' pronunciation proficiency both accurately and reliably. Further improvement will be made to satisfy the needs for different teaching phases to conform to the teaching objectives and requirements. Moreover, the rating scale should be supplemented by further learning tips, instruction, and learning resources. Based on their scores on each rating criterion in the diagnostic rating scale, students should be able to avail themselves of corresponding learning materials that have been carefully analyzed, selected, and labelled. In this way, when students are diagnosed with certain weaknesses in some dimensions of pronunciation, they can refer to the targeted learning materials to remedy their weaknesses via self-instruction. Finally, future research should be conducted to develop and validate a pronunciation diagnostic assessment, which includes a variety of test tasks that are designed specifically for the diagnostic purpose. Test tasks should be effective and valid in identifying strengths and weaknesses of pronunciation of Chinese L2 learners of English.

To conclude, pronunciation itself is a multi-componential concept and deserves a finer treatment. Future research should endeavor to propose more useful, proper, and concrete definitions of specific pronunciation features and global constructs such as intelligibility and comprehensibility, discuss which constructs and criteria are appropriate to assess pronunciation, provide more insights into construct validation and criterion-related validation of pronunciation rating scale, explore the relative weighting of sub-scales within a pronunciation rating scale, and investigate how non-linguistic dimensions such as individual differences may interact with pronunciation ratings.

Appendix



		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.	*Accurate* production of sounds and *sufficient* use of segmental features, with occasional lapses that cause minor unintelligibility of the utterance.	*Correct* production of sounds and *moderate* use of segmental features, with some lapses that influence the intelligibility of the utterance.	*Incorrect* production of sounds and *limited* use of segmental features, with frequent lapses that severely hamper the intelligibility of the utterance.
	Consonant				
** Suprasegmental features include: the appropriate use of the strong and weak forms of the grammar words, the incomplete plosions, nasal plosion and lateral plosion, etc.*					
Word	*Word Stress*	*Accurate* production of the word stress and *natural* use of suprasegmental features* that makes for the intelligibility of the utterance.	*Accurate* production of the word stress and *appropriate* use of suprasegmental features that cause minor intelligibility of the utterance.	*Correct* production of the word stress and *moderate* use of suprasegmental features, with some lapses that influence the intelligibility of the utterance.	*Incorrect* production of the word stress and *restricted* use of suprasegmental features, with frequent lapses that severely hamper the intelligibility of the utterance.
	Consonant Cluster				
** Suprasegmental features include: sound linking, the appropriate use of pauses, rhythm and the variety of tones*					
Sentence and Discourse	*Sentence Stress*	*Natural* use of suprasegmental features* that makes for fluency and comprehensibility of the utterance. Few segmental and suprasegemental errors and speech is effortless to understand.	*Appropriate* use of suprasegmental features with occasional segmental and suprasegmental lapses that cause minor problems with fluency and comprehensibility. Speech requires little effort to understand.	*Moderate* use of suprasegmental features with some segmental and suprasegmental lapses that influence fluency and comprehensibility of the utterance. Speech requires some effort to understand.	*Restricted* use of suprasegmental features with frequent segmental and suprasegmental lapses that severely hamper the fluency and comprehensibility of the utterance. Speech is effortful to understand.
	Intonation				
	Pause and Fluency				

References

Alderson, C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). New York: Routledge.

Browne, K., & Fulcher, G. (2017). Pronunciation and intelligibility in assessing spoken fluency. In

T. Isaacs & P. Trofimovich (Ed.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 37–53). Bristol; Blue Ridge Summit: Multilingual Matters / Channel View Publications.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Ed.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

Chen, H., & Bi, R. (2015). *The research on the phonological structure of intonation for Chinese English learners*. Beijing: Foreign Language Teaching and Research Press.

Chen, H., & Bi, R. (2008). English majors' pronunciation and intonation in read speech: A longitudinal study. *Journal of PLA University of Foreign Languages*, 4, 43–49.

Chen, H., & Li, J. (2017). Reflections on the current situation of phonological assessment: A survey among raters of standardized oral tests in China. *Foreign Languages and Their Teaching*, 5, 81–87.

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

Crystal, D. (1997). *English as a global language*. Cambridge: Cambridge University Press.

Davies, A. (2003). *The native speaker: Myth and reality* (2nd ed.). Clevedon: Multilingual Matters.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*(4), 476–490.

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Philadelphia, PA: John Benjamins.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Frankfurt am Main: Peter Lang.

Fan, J., & Ji, P. (2017). Teacher-, self- and peer-assessment in translation teaching: A many-facet Rasch modeling approach. *Foreign Language World*, 4, 61–70.

Galloway, N., & Rose, H. (2015). *Introducing global Englishes*. London: Routledge.

Hall, C., & Hastings, C. (2017). *Phonetics, phonology & pronunciation for the language classroom*. London: Palgrave Macmillan.

Harding, L. (2013). Pronunciation assessment. In C.A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Hoboken, NJ: Wiley-Blackwell.

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, *32*(3), 317–336.

Heaton, J. B. (1975). *Writing English language tests*. London: Longman.

Huang, X., & Jia, X. (2016). Corrective feedback on pronunciation: Students' and teachers' perceptions. *International Journal of English Linguistics*, *6*(6), 245–254.

Isaacs, T. (2013). Assessing pronunciation. In A. J. Kunnan (Ed.), *The companion to language assessment. Hoboken*, New Jersey: John Wiley & Sons.

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.

Isaacs, T., & Trofimovich, P. (Ed.). (2017). *Second language pronunciation assessment: Interdisciplinary perspectives*. Bristol: Multilingual Matters.

Jenkins, J. (2007). *English as a lingua franca: Attitude and identity*. Oxford: Oxford University Press.

Jones, N., & Saville, N. (2016). *Learning oriented assessment: A systemic approach*. Cambridge: Cambridge University Press.

Kang, O., & Ginther, A. (Ed.). (2018). *Assessment in second language pronunciation*. London; New York: Routledge.

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

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

Lei, J., & Han, T. (2006). Review on teaching pronunciation in foreign languages: Comments on jazz chants as a method of English pronunciation teaching. *Foreign Language World*, 1, 16–21.

Lee, Y.-W. (2015). Diagnosing diagnostic language assessment. *Language Testing*, *32*(3), 299–316.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*(3), 369–377.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith & R. M. Smith (Ed.), *Introduction to Rasch measurement* (pp. 258–278). Maple Grove, MN: JAM Press..

Liu, R. (1991). *Language testing and its methods*. Beijing: Foreign Language Teaching and Research Press.

Liu, S. (1999). Teaching design of the pronunciation course for English majors at normal universities. *Foreign Language Teaching Aboard*, 2, 3–5.

Liu, S. (2013). *Better pronunciation for better communication*. Shanghai: 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.

Liu, S., & Niu, Z. (2018). An empirical research on the improved multidimensional pronunciation teaching assessment model. *Foreign Language Learning Theory and Practice*, 4, 62–68.

Ma, Q., & Zhao, Y. (2017). Phonology, phonetics and pronunciation teaching. *Journal of Beijing International Studies University*, 4, 40–55.

Ministry of Education of the People's Republic of China. (2018, November 30). *China's standards of English language ability*. http://www.neea.edu.cn/html1/report/18113/2797-1.htm

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97.

Munro, M. J., & Derwing, T. M. (2011). The foundations of accent and intelligibility in pronunciation research. *Language Teaching*, *44*(3), 316–327.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189–227.

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

Roach, P. (2009). *English phonetics and phonology: A practical course* (4th ed.). Cambridge: Cambridge University Press.

Saito, Y., & Saito, K. (2017). Differential effects of instruction on the development of second language comprehensibility,word stress, rhythm, and intonation: The case of inexperienced Japanese EFL learners. *Language Teaching Research*, *21*, 589–608.

Shintani, N., & Saito, K., & Koizumi, R. (2019). The relationship between multilingual raters' language background and their perceptions of accentedness and comprehensibility of second language speech. *International Journal of Bilingual Education and Bilingualism*, *22*(7), 849–869.

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, *36*(3), 326–344.

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.

Wang, G. (2005). Revisiting the goal of EFL pronunciation teaching in the Chinese context. *Foreign Languages in China*, *2*(6), 18–23.

Wells, J. C. (2000). *English intonation: An Introduction*. Cambridge: Cambridge University Press.

Zhang, L. (2007). *English pronunciation course*. Wuhan: Central China Normal University Press.

Zhang, N. (2011). *English pronunciation: Mystery solved*. Shanghai: Fudan University Press.

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

Sen Liu is an associate professor and the Director of Oral English Teaching and Research Centre of East China Normal University. 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.

Zijie Niu is a post graduate student 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. His research interest includes language testing and assessment, L2 pronunciation teaching and assessment, and data science in applied linguistics.