ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 10, No. 1, pp. 141-149, January 2019
DOI: http://dx.doi.org/10.17507/jltr.1001.16

# Pronunciation Rating Scale in Second Language Pronunciation Assessment: A Review

Wenjun Zhong
School of English and International Studies, Beijing Foreign Studies University, Beijing, China

***Abstract*—By reviewing previous studies on pronunciation rating scale in second language pronunciation assessment, this article aims to summarize research gaps and weaknesses so as to contribute to the pronunciation rating scale research and development. Several research topics concerning construct, criterion, descriptor, scale length, scale format and scale users and suggestions with regard to participants, data collection methods and data analysis methods are provided for future research.**

***Index Terms*—Pronunciation rating scale, review, second language pronunciation, language assessment**

## I. Introduction

Pronunciation is one way to externalize language (Peng, 2014). It is important because it can facilitate communication (Gilner, 2008; Ketabi, 2015) and raise one's social status (Derwing, Rossiter & Munrol, 2002). In the field of second language assessment, although there is a resurgent interest in pronunciation assessment (Isaacs, 2013; Isaacs, Trofimovich, Yu, & Chereau, 2015; Yates, Zielinski, & Pryor, 2011), relevant research is still scare (Knoch, 2017; Yates, Zielinski, & Pryor, 2011), let alone studies concentrating on pronunciation rating scale exclusively.

A rating scale is essential for the successful execution of all kinds of language assessments. It is a manifestation of the underlying construct of language assessment (Isaacs & Thomson, 2013), the reference for raters to score test takers' performances (Harding, 2017; Isaacs & Thomson, 2013; Mcnamara, 2002; Underhill, 1987) and the guidelines for test takers or other score users to interpret the assessment results (Isaacs & Thomson, 2013). It is the same with pronunciation rating scale. Meanwhile, previous studies have identified some problems of pronunciation scales (Harding, 2013; Harding, 2016; Isaacs, 2013; Isaacs, Trofimovich, Yu, & Chereau, 2015), whether it is about scale design or about scale use. The scarcity of relevant studies, the importance of rating scale, and the need to improve pronunciation rating scale constitute the necessity for more research.

This article is aimed at reviewing previous studies that take the pronunciation rating scale in second language assessment as the research object. The author will compare their merits and drawbacks and tentatively put forward several future research directions.

In the following sections, several general key concepts relevant to pronunciation rating scale exploited in this article are elaborated. Next, relevant studies are categorized in accordance with research topics and are examined in details to discuss their strengths and weaknesses so as to put forward future research directions. The last section concludes this article by summarizing future research directions and suggestions about research methods.

## II. Key Concepts

The first key concept needs clarification is pronunciation. Across rating scales, this term is not used consistently (Isaacs & Trofimovich, 2012). Most studies regard it as the production of suprasegmental features along with the segmental ones (e.g., Chen & Li, 2017; Harding, 2017; Isaacs, Trofimovich, Yu, & Chereau, 2015). It is reasonable since suprasemental features are important for pronunciation, especially for achieving an intelligible and comprehensible pronunciation (Tian & Jin, 2015).

Currently, the intelligibility and comprehensibility goals are well recognized in the field of second language acquisition, instruction and assessment as a result of a globalized environment, the spread of English as the international communication tool and the resurgent desire for a more reasonable goal for second language learners (Ketabi, 2015). In addition, although some researchers maintain that pronunciation is an inalienable part of either intelligibility (e.g., Ketabi, 2015; Wen, Liu & Jin, 2005; Zeilinski, 2006) or comprehensibility (e.g., Isaacs, Trofimovich, & Foote, 2018), it does not imply that pronunciation should be defined in terms of intelligibility or comprehensibility. Instead, it may be better to treat it as one influential factor of intelligibility or comprehensibility. Based on these two reasons, this study defines pronunciation as the production of both suprasegmental and segmental features.

The second key concept is rating scale. It is an instrument that not only contains the operational definition of certain construct (Davies, Brown, Elder, Hill, Lumley, & McNamara, 1999; Mcnamara, 2002), but also provides "a series of constructed levels" (Davies et al., 1999, p.154) on which judgment about test takers' performance is based. Construct is "the trait or traits that a test is intended to measure" (Davies et al., 1999) and it is the third key concept in this article.

Any rating scale is supposed to operationalize certain construct and rating scales are usually the instantiations of the underlying construct (Harding, 2017).

Another element in a rating scale is the criterion and it is the fourth key concept. Generally it is defined as "a quality on which test performance is judged" (Davies et al., 1999, p. 38). Criterion is important because it reflects the underlying construct (Isaacs, Trofimovich, Yu, & Chereau, 2015). There are both linguistic and non-linguistic criteria and different selections of criteria and different weighting given to different criteria can lead to different rating results (Davies et al., 1999). Therefore, if a rating scale fails to present intended criteria or intended weighting of various criteria, it may be never able to perform its duty.

A criterion are usually displayed by descriptors of a rating scale (Yates, Zielinski, & Pryor, 2011), which are defined as statements that "describe the level of performance required of candidates at each point on a proficiency scale" (Davies et al., 1999, p.43). It is what is presented to scale users and represents both the construct and criteria.

Another important facet of a rating scale is its format (Knoch, 2017). While holistic scales contain general descriptions, analytic scales provide several criteria that are contributing to test takers' performance (Knoch, 2017).

A rating scale provides "a series of constructed levels" and the number of levels is known as the scale length. Research suggests that it is an important facet since it closely relates to the usability of the scale (Alderson, 1991; Flege & Fletcher, 1992; Fulcher, 1996).

The last concept is usability. Usability is defined as the easy of use (Harding, 2017). In the research of rating scale in language assessment, therefore, usability refers to whether a scale is easy to use for raters or for other scale users. Since a rating scale is designed to be used while a valid rating scale does not necessarily entail a useful rating scale (Harding, 2017), the usability of a rating scale can never be neglected.

## III. Review of Previous Studies

To ensure the studies reviewed are closely related to the pronunciation rating scale and the studies chosen are representative enough and are of high quality, the author conducted "key word" search on CNKI, Google Scholar, and EBSCOhost and finally selected thirteen most relevant studies. Two broad branches are identified: rating scale design and rating scale user. Studies that fall into the rating scale design branch are further divided into construct, criterion and descriptor, length and format. Studies which fall into the rating scale user branch are further divided into rater's characteristics and scale usability.

### *A. Rating Scale Design*

**construct**

Constructs that have been discussed in pronunciation assessment include comprehensibility (Derwing & Munro, 2009; Isaacs, Trofimovich, & Foote, 2018; Munro & Derwing, 1999; Trofimovich, & Isaacs, 2012), accentedness (Trofimovich, & Isaacs, 2012), and pronunciation and intonation (Chen & Li, 2017).

The first question under examination is how certain construct should be defined. Munro and Derwing (1999) distinguish comprehensibility, intelligibility, and accentedness conceptually. According to them, while both comprehensibility and intelligibility relate to how well a massage is understood, comprehensibility is more linked to listeners' subjective perception and intelligibility is more linked with the objective proportion of speech that is understood (Munro & Derwing, 1999). As for accentedness, it is more related to the comparison with native speakers' norm (Munro & Derwing, 1999). While this distinction is gaining momentum in recent years (Isaacs & Thomson, 2013), problems concerning definitions are still pervasive not only in pronunciation scale design, but also in pronunciation scale use.

Conflations between constructs in rating scales are found between accentedness and intelligibility (Anderson-Hsieh, Johnson & Koehler, 1992) and between accentedness and comprehensibility (Harding, 2013; Isaacs &Trofimovich, 2012). Intelligibility is also frequently used as a synonym of comprehensibility (Levis, 2006). It suggests that conceptual definitions are not enough for scale developers to understand what they really mean. More concrete definitions or at least descriptions of what certain construct is are therefore necessary for scale design and development. That is why studies that inquire into the most relevant linguistic correlates of certain construct are valuable.

Isaacs & Trofimovich (2012) and Trofimovich & Isaacs (2012) are representative ones. Both of them try to figure out what certain construct is by investigating what linguistic correlates comprise comprehensibility (Isaacs & Trofimovich, 2012) or by comparing respective linguistic features of comprehensibility and accentedness (Trofimovich & Isaacs, 2012). The first study combines the quantitative analysis of speech samples and qualitative analysis of introspective reports about linguistic correlates that raters rely on when judging comprehensibility. Lexical richness, fluency, grammar, discourse, and word stress errors emerge as the five correlates most frequently referred to. The popularity of lexical richness and grammar is also present in Trofimovich & Isaacs (2012). In that study, participants are asked to select measures that affect their ratings and to type in linguistic aspects they depend on for rating comprehensibility and accentedness respectively. Results illustrate that phonological features are more linked with accentedness. These two studies fill the research gap concerning the definition issue, which can be illustrating for rating scale design and development. However, what they actually probe into are constructs of the pronunciation assessment on a whole rather than certain underlying of a pronunciation rating scale. It is possible that constructs of a pronunciation assessment are

different from constructs of the rating scale employed. Therefore, what does certain construct embedded in a pronunciation rating scale refer to or what linguistic correlates consists certain construct embedded in a pronunciation rating scale requires further exploration.

While the two studies mentioned above inspect the most relevant linguistic correlates, Isaacs, Trofimovich, & Foote (2018) attempts to derive empirically the operational definition of comprehensibility construct in a rating scale. Focus group discussion is held after each rating session and comments made in the group discussion are brought together for scale revisions before the next rating session. Comprehensibility is eventually defined as the effort to understand the speech and it is deemed as the optimal operational definition. But there is an additional condition for this definition to operate: the raters should be professional English teachers with exposures to various second languages. Although this condition serves to minimize potential influences of some non-linguistic factors such as listener's attitudes (Kang & Rubin, 2009), real rating situations in most cases can seldom meet this high standard. It therefore raises the question as to whether the rating scale centering around comprehensibility can generate better rating effects or whether the operational definition is applicable.

Studies that illustrate how scale users understand or interpret certain construct in a scale also shed some light on construct definition. Isaacs & Thomson (2013) focuses on comprehensibility construct while Chen & Li (2017) examines pronunciation and intonation constructs. Research methods of these two studies are both qualitative. Raters' interpretations or understandings are gathered through verbal reports (Isaacs & Thomson, 2013), stimulated recalls (Isaacs & Thomson, 2013), interviews (Chen & Li, 2017; Isaacs & Thomson, 2013) or questionnaires with open-ended questions (Chen & Li, 2017). Data are then analyzed through thematic analysis. Results reveal that the same construct is likely to receive different interpretations (Chen & Li, 2017; Isaacs & Thomson, 2013). These results are of great importance since diverse understandings of the construct mean different linguistic traits are assessed by scale users. If scale users do not agree upon this point, the results generated by such a scale can be highly unreliable. These two studies emphasize the urgent need for a clearer and more concrete definition of constructs. Nevertheless, they seem to treat constructs of their instrumental scales as fit for pronunciation assessment without questioning whether it is indeed the case. More research is therefore required to discover which construct is appropriate for assessing pronunciation.

Another issue deserves attention is construct validation. Construct validation is "an investigation of what a test actually measures and attempts to explain the construct." (Davies et al., 1999, p. 220). When it comes to pronunciation rating scale, Sawaki (2007) touches upon this problem by examining the construct validity of the speaking rating scale used in a Spanish language assessment. It is an analytic rating scale and pronunciation is counted as a sub-scale in this rating scale. Fifteen raters are asked to 214 speech samples and scores are given for further analysis. Confirmatory factor analysis is employed to find out the structural relationships among the five sub-scales while multivariate generalizability theory is accessed to find out the interrelationships among the sub-scales and between sub-scales and the overall rating. By doing these, the author is able to test the convergent validity, discriminant validity, and relationships among sub-scales and between scales and the overall rating. For the pronunciation scale, results show that it is less reliable and that contributes less information than grammar scale to the overall score. The author explains this by referring to the test purpose, according to which more importance is placed on grammar than on pronunciation. This study merits in its contribution to a highly important yet less explored field in language assessment. The clear research format of this study also allows future studies to learn from by examining other scales. However, in this study, pronunciation is regarded as a sub-scale of the speaking scale although it does not necessary to be so. Dimensions including segmental, rhythm, intonation, speech rate, etc. are all contributing to pronunciation so that it is probable to divide a pronunciation scale into sub-scales reflecting dimensions mentioned above. The recent trend from judging based on global features like comprehensibility or intelligibility to placing more emphases on more specific features (Yates, Zielinski, & Pryor, 2011) also suggests that future research should focus more on pronunciation specially. What, then, are the convergent validity, discriminant validity, and relationships among sub-scales and between sub-scales and the overall rating when only the pronunciation scale is focused on requires further illustration.

**Criterion**

The first question that is discussed by previous research is which criterion revealed in rating scales is appropriate for assessing pronunciation.

Intelligibility is treated as a criterion instead of a construct in Isaacs (2008). This study tests whether treating intelligibility as one criterion is appropriate for pronunciation assessment in academic context. Eight participants with various L1 backgrounds first record samples for rating. Next, the ratio of intelligible words of samples is calculated manually by researcher to form the quantitative data. Answers to open-ended questions from 18 native raters are analyzed to form the qualitative data. Results from that study proves a close correlation between the quantitative data and the qualitative data and lead to the conclusion that intelligibility is not a sufficient condition though it is an adequate criterion for assessing pronunciation proficiency of international teaching assistants. However, intelligibility depends on both sides of communication (Rajadurai, 2007) and is constructed by multiple factors, such as listeners' attitudes toward the speakers, communication context or even listeners' familiarity toward the topics discussed. In this study, these factors are neither examined nor controlled. How these non linguistic factors interact with intelligibility or influence the intelligibility as a criterion requires more research.

Isaacs (2008) examines the criterion of pronunciation assessment as a whole instead of restricting to the rating scale. By contrast, discovering appropriate criteria of a rating scale are the goals for both Isaacs, Trofimovich, Yu, & Chereau (2015) and Harding (2017). The first study tries to find the criteria that are discriminating enough for band 5 and band 7 in IELTS scale. Semantic differentiate scales are first developed based on focus group discussions. Raters are then asked to rate samples based on existing IELTS speaking scale for the first session and on semantic differentiate scales for the second session. Focus group discussions are held at the end of each session and during the third session, focus group discussion is held again to discuss linguistic features that influence raters' judgments most. Semantic differential scales are related back to the IELTS scale to discover linguistic features that best distinguish band 5 and band 7. Results suggest that grammar and lexical richness serve best as the discriminating criteria while segmental accuracy and comprehensibility, the worst. Test takers with higher proficiency are best distinguished by criteria including grammar accuracy, sentence structure, lexical richness and word stress. Qualitative data reveal that the existing IELTS pronunciation sub-scale fail to provide specific criteria at band 5 and band 7 and thus leading to the implementation difficulties. However, different results emerge from Harding (2017), a study which also utilizes focus group discussion as the research method. Raters in this study tend to exclude grammar criterion and accentedness criterion from pronunciation scale and combine pronunciation criterion with fluency criterion as one criterion and add intelligibility criterion into the scale. Differences may lie in that the instrumental scale in the former study is the IELTS rating scale (pronunciation sub-scale included) while the latter takes phonological control scale as the research subject so that some criteria such as grammar which are less relevant to pronunciation are not emphasized. Despite of their achievements, there are more to explore. Future studies can either select certain pronunciation scale instead of some speaking scales to investigate the most appropriate criteria, or applying similar research methods to examine other scales to challenge or substantiate results of previous research.

The second question addressed by previous research is the definition of certain criterion. Previous studies reveal that the lack of precise and concrete definitions of criteria leads to different definitions construed by raters themselves (Orr, 2002; Wang, 2008; Gao, 2007; Brown, 2007; Yates, Zielinski, & Pryor, 2011). Most studies are qualitative since it is possible that raters define criteria differently yet arrive at similar quantitative scores (Wang, 2008). For instance, both Orr (2002) and Wang (2008) studies analyze data from verbal reports of raters and discover that during rating process using a rating scale, raters not only give different definitions to criteria, but also take other non-criterion information into consideration, thus undermining both the validity and reliability of assessment results. Similar results are also generated in Brown (2007) in which verbal protocol is exploited to collect raters' comments on previous holistic IELTS scale. These studies merit in their incorporation of raters' using experiences and reveal that without clear and concrete definitions of criteria, raters simply do not share similar understandings of criteria although the scores they give may achieve high correlation coefficient. The next step, then, is to discover what should be the appropriate definitions of individual criteria in a pronunciation scale so as to facilitate their understandings of criteria in rating scale and reduce individual differences in their interpretations.

The third topic that is important for criteria is the criterion-related validation. Only Gao (2007) attempts to analyze this validity of a pronunciation rating scale developed by herself. Students' samples of a reading-aloud task are first collected and annotated for mistakes. All mistakes are then categorized based on the type of competence (theory driven) they represent. The next step is to decide different weighing given to different types of mistakes and to decide how to calculate the scores. These steps form the ultimate rating scale that includes pronunciation, fluency, sentence structure and lexical richness as the four criteria. Criterion-related validation process of this rating scale is followed by conducting correlation analysis, factor analysis and variance analysis. Results favour this rating scale in term of criterion-related validity. However, this study seems to be too sloppy in the categorization of mistakes. For instance, inappropriate pause is regarded as sentence processing problem while failure to produce proper rhythm of the target language may also contribute to this problem (Meng & Wang, 2009; Zhou & Song, 2015; Yu, 2013; Zhu, 2011). Regarding faster speaking rate as higher proficiency is also problematic since it is simply not the case: a speaker can read very fast and produce unintelligible and non-comprehensible speech. Moreover, when assessing the criterion-related validity, students' CET 4 scores are utilized for coefficient analysis. Although reading aloud task can reflect the overall proficiency level (Gao, 2007), it is to the essence a speaking task which CET 4 does not assess. How reliable the analysis is then if the CET 4 scores are used as reference is questionable. Nevertheless, considering the scarcity of relevant studies, this study still stands out for probing into this issue. More research is therefore demanded considering both the problems of this research and the scarcity of relevant studies.

**Descriptor, Length and Format**

Although there is no research that exclusively examines descriptors, they are touched upon in many pronunciation rating scale studies. Relevant research generally gathers data concerning raters' perceptions and views through qualitative methods like verbal reports (Wang, 2008; Yates, Zielinski, & Pryor, 2011), stimulated recalls (Wang, 2008), open ended questionnaires (Yates, Zielinski, & Pryor, 2011) or focus group discussions (Isaacs, Trofimovich, Yu, & Chereau, 2015; Harding, 2017; Isaacs, Trofimovich, & Foote, 2018). Three facets are usually lamented about by raters: the lack of clear and exact wording of descriptors, the wording inconsistency of descriptors across different levels within a rating scale and the length of descriptors within a level.

Wang (2008) analyzes verbal reports after raters using the speaking rating scale (pronunciation as one criterion) for TEM 4 and indicates that some vague wording of descriptors in the rating scale will induce raters to form operational definitions based on their own understandings. Similar results are also obtained by Yates, Zielinski, & Pryor (2011) and Isaacs, Trofimovich, Yu, & Chereau (2015) which examine the IELTS pronunciation scale and by Harding (2017) which probes into CEFR Phonological control scale. The consistency problem is noted Isaacs, Trofimovich, & Foote (2018) when endeavouring to establish a comprehensibility scale in which a pronunciation sub-scale is included. The scale is developed by incorporating raters' comments after each rating session. The results imply that a scale with consistent wording among sub-scales and within each sub-scale is more friendly for raters. Similar remarks are given by Harding (2017). Although problems identified are illuminating, suggestions followed are not. For instance, to counter wording vagueness problem, researchers suggest that the wording of descriptors should be made more specific. Little is known about what kind of descriptors are considered to be specific enough and what scale designers and developers can do to reduce the vagueness of descriptors' wording. More research is therefore required in this respect. Lastly, while Isaacs, Trofimovich, Yu, & Chereau (2015) warns that the length of descriptors within each level should be neither too long / specific nor too short /generic, Harding (2017) specifies the length by saying that descriptors consisting of three to five clauses per level may be considered optimal. However, this conclusion needs further support not only due to the scarcity of relevant studies, but also due to the fact that how many clauses are required may depend on various factors such as test purpose, the stake of a test, rating time limit, etc.

Another length issue is concerned with the number of levels within a rating scale. Isaacs & Thomson (2013) directly addresses this topic. Raters are invited to use two IELTS pronunciation scales (one with five levels, one with nine levels) to evaluate 38 speech samples. Scores that they give are then processed by quantitative methods. Although the mean scores among raters generated from a five-level scale and a nine-level scale are not significantly different, more levels lead to more difficulties in differentiating adjacent levels. This is different from Yates, Zielinski, & Pryor (2011) in which raters are quite positive about the 9-point pronunciation scale. But it can be explained by the fact that what Yates, Zielinski, & Pryor (2011) reports is raters' perception before using the scale and no comment is made in terms of the length after their use. But what is agreed upon is a scale should neither be too long nor too short since short one can be overly crude (Brown, 2006; Cumming et al., 2002) while long one can be clumsy for raters to discriminate between adjacent levels (Alderson, 1991; Flege & Fletcher, 1992; Fulcher, 1996; Van Moere, 2013). Six levels are advocated both by Harding (2017) and Isaacs, Trofimovich, & Foote (2018) as the optimal choice. However, Harding (2017) derives the conclusion from raters' comments after using the scale only once. In other words, raters only believe that six levels are optimal for a pronunciation rating scale. Whether it is really the case needs further studies by applying the revised scale again. Isaacs, Trofimovich, & Foote (2018) has advantages over Harding (2017) in this respect. Several turns of rating sessions, discussions and revisions are held before the authors claim their findings, making the conclusion more convincing. Besides, the number "six" falls within the magical spectrum, namely, seven plus or minus two. This spectrum is known as the limits on our brains for processing information (Miller, 1956). However, due to the scarcity of relevant research and for the convenience of developing scale, still more research is required to test whether six levels is the optimal choice for a pronunciation rating scale.

For format, usually two types are examined: holistic and analytic. While holistic scales contain general descriptions, analytic scales provide several criteria that are contributing to test takers' performance (Knoch, 2017). Despite of its importance, none pronunciation rating scale study has ever made an effort to compare these two formats. Instead, these two terms are more often used to describe two rating styles. For instance, both Wang (2008) and Chen & Li (2017) reveal that raters prefer holistic rating style over analytic one. The lack of precise and specific descriptors and criteria is employed to account for this phenomenon. For most studies, scales under examination are analytic ones. It is understandable since an analytic scale is believed to be better. An analytic scale format is congruent with the current trend of viewing language ability as being composed by various components (Bachman, Lynch & Mason, 1995) and it allows raters to attend to more specific criteria which they may ignore when using a holistic scale (Brown & Bailey, 1984). However, whether an analytic scale is better is more dependent on various factors such as constructs, criteria, test purposes, the stake of a test or rating conditions. Besides, how pronunciation scale format influences rating process also deserves more attention.

### *B. Rating Scale User*

#### Rater's Characteristics

Raters are important since they are users of rating scales. It is impossible to expect that two raters are completely the same. Individual differences always exist (Douglas, 1994; Lumley, 2005). Raters vary in their previous rating experiences and teaching experiences (Isaacs & Thomson, 2013; Schairer, 1992), degrees of severity in rating (Schairer, 1992), their views on language (Gao, 2007), their attitudes toward test takers' accents (Kang & Rubin, 2009), etc. These differences will affect how they apply the rating scale (Douglas, 1994; Isaacs & Ron I. Thomson, 2013; Orr, 2002; Wang, 2008; Brown, 2007), thus greatly undermining the validity and reliability of rating results. In spite of their importance, when it comes to pronunciation rating scale, only Isaacs & Thomson (2013) specially focuses specially on this issue. Data are collected from verbal protocols and interviews and are analyzed by thematic analysis. Results reveal that experienced raters and novice raters indeed differ in rating strategies, including how they utilize rating scales. However, although it reports that rating experiences can lead to different rating behaviors, how individual

characteristics including rating experiences interact with the use of rating scale is still not clear (Knoch, 2017). For instance, which individual characteristics are most influential to rating scale use, which criteria are more subject to certain individual characteristic and what are the best ways to control individual differences in using certain rating scale? For the last question, rater training and automated rating can be the answers (Isaacs, 2013). However, it is not clear if rater training is really efficient as expected and if the automated scoring can achieve scoring results similar to those given by human raters. More research is therefore necessary.

**Scale Usability**

Another area concerning users' effects is the scale usability. Research questions of this branch mainly include how raters perceive or view certain pronunciation rating scale from the perspective of usability, what problems are identified based on scale using experiences and what suggestions can one derive correspondingly. Raters' perceptions about the usability of IELTS pronunciation rating scale are examined by Yates, Zielinski, & Pryor (2011) and Galaczi, Lim and Khabbazbashi (2012). Raters' feelings about the usability of IELTS pronunciation rating scale are extracted from qualitative comments in open-ended items and from verbal protocol. Both of them report that raters are less confident and less comfortable in using the pronunciation rating scale. While the former also reveals raters' using experiences before using the scale, the latter differs in its comparison between the pronunciation scale and other component scales within the IELTS speaking scale. Their results echo conclusions of Brown and Taylor (2006).

Raters report difficulties in applying the rating scale due to the vague descriptors (Yates, Zielinski, & Pryor, 2011), vague concepts (Galaczi, Lim & Khabbazbashi, 2012) and overlap with other scales (Yates, Zielinski, & Pryor, 2011) or in terminology (Galaczi, Lim & Khabbazbashi, 2012). The recommendations given in Harding (2017) also reflect these problems and are further divided into technical ones and construct-related ones. Another problem is observed in Isaacs, Trofimovich, & Foote (2018) and Harding (2017), namely, the consistency of descriptors. Raters in these two studies maintain that the inconsistency of descriptors across levels renders the scales less user-friendly. In these two studies, different scales are explored and focus group discussion is used instead of verbal protocol. While the former aims to develop a comprehensibility scale (pronunciation sub-scale included) by incorporating scale using experiences, the latter aims to improve existing scale by identifying problems mentioned by scale users. These studies merit in their focus on usability. In many cases, feelings, experiences, opinions, perceptions and views of using scales are seldom used as data in pronunciation assessment research (Harding, 2017). Meanwhile, raters should never be ignored wince they are the ultimate users of a rating scale. By examining raters' using experiences, these studies fill this gap to some extent. However, relevant studies are still scarce.

Further studies can also focus on score users that have seemingly been neglected in pronunciation rating scale research. This group of users relies on scales to interpret what scores mean. For individual test takers, making inference about their pronunciation proficiency by resorting to a pronunciation rating scale can help them locate problems to be solved for self-improvement. For some institutions, making inference about some test takers' pronunciation proficiency by resorting to a pronunciation rating scale can help them grasp both strengths and weaknesses of test takers and make decisions accordingly. If rating scales are not well-developed and not user-friendly, difficulties and inconvenience can arise for this user group. Exploring how score users perceive certain pronunciation rating scale from the perspective of usability might therefore generate more interesting results for both scale design and scale development, at least for the design and development of the public version of the scale which is usually accessible for most score users.

### *C. Suggestions Concerning Research Methods*

Two sections above reveal several future research directions. In this section, the author will illustrate what can future research learn from methodological problems of previous studies.

Firstly, for the selection of participants, the biggest problem centres on rater's prior rating experiences. While there are some studies that specially attend to raters' experiences (e.g., Isaacs & Thomson, 2013), be it rating experiences or more general experiences (native language, age, education background, teaching experiences, etc.), most studies allow great variance of experiences of raters while focusing on other issues. For instance, in Harding (2017), some raters have more than ten-year assessment experience while others have got none (at least not mentioned in this article). It is doubtful since raters' experiences have been reported to affect the rating results (Isaacs & Thomson, 2013). If the rating experience is not the research aim, its influences are supposed to be reduced to the minimum. Harding (2016) defends himself by suggesting that the variance ensures the ecological validity which requires that instruments and methods employed in a research should resemble the real-world under-examination to the maximum (Brewer, 2000). However, if the variance is overly huge, whether the results can represent the real assessment situation is questionable since real assessment situation always tries to minimize differences among raters, including raters' experiences. The same problem also resides in Isaacs & Trofimovich (2012), Trofimovich & Isaacs (2012), and Isaacs, Trofimovich, Yu, & Chereau (2015). To enhance the reliability of research results, future research should take great care of raters' experiences when designing a study.

Secondly, rating process is an inalienable part for scale relevant studies. However, rating process in studies reviewed above seems to be not authentic enough. Some rate samples by watching recorded videos of authentic test (e.g., Orr, 2002) and others by listening to the samples (e.g., Isaacs, 2008; Harding, 2017). It is highly possible that results obtained will be altered in real assessment situations where raters may attend to different criteria and cast different

interpretations of criteria. Future research can attempt to incorporate data from more authentic rating process to shed more light on the mystery of pronunciation rating scale.

Thirdly, although most studies reviewed above utilize verbal protocol to peep into raters' mind during rating process, it is actually not valid enough as a data collection method since it somehow intrudes the rating process and therefore cannot truthfully represent what happens during rating process (Brown, 2007). Other methods are desired along with the verbal protocol. For example, future studies can, if possible, record videos of authentic rating sessions and ask the very raters in the videos to participate in the research immediately after the assessment. These participants can be invited to watch recorded videos and recall their rating processes. In this way, researchers can not only observe the posture, facial expressions, body movements and any linguistic behaviours of raters in the video, but also ask the very raters to provide verbal report when watching the video with regard to what they are thinking during the rating process. This multi-modal approach may generate more information about how raters use the rating scales, how they understand the construct, how they define and use criteria and how they perceive the descriptors, length and format of the scale during rating process.

Lastly, for the data analysis method, the reliability of analysis process needs to be addressed. For instance, both Isaacs, Trofimovich, Yu, & Chereau (2015) and Harding (2017) employ thematic analysis without checking the reliability of the coding procedure. Future research therefore needs to make an improvement in this respect.

## IV. Conclusion

Previous research suggests that pronunciation rating scales seldom receive attention it deserves. For studies that are relevant, pronunciation rating scales are more often examined as a sub-scale of other larger scales, such as a speaking scale or a comprehensibility scale. However, pronunciation itself is a multi-componential concept and deserves a finer treatment. Moreover, the recent trend from judging based on global features like comprehensibility or intelligibility to placing more emphases on more specific features (Yates, Zielinski, & Pryor, 2011) also suggests future research to focus more on the pronunciation rating scale exclusively.

For construct and criterion, future research can endeavour to find out more useful, properer, and more concrete definitions, discuss which constructs and criteria are appropriate for pronunciation rating scale, provide more insights on construct validation and criterion-related validation of pronunciation rating scale, and explore the relative weighting of scales within a pronunciation rating scale. Moreover, how non-linguistic factors interact with criteria also requires further exploration.

For descriptors and scale length, future research can strive to provide more concrete suggestions concerning the wording and the length of descriptors and the length of a pronunciation rating scale.

For scale format, how various factors such as test purpose, construct or criteria affect the choice of format and how format influence raters rating decisions can be further explored.

For users, the interactions between users' characteristics and criteria weighting and interpretation require more research. How score users perceive certain pronunciation rating scale from the perspective of usability can also be illustrative. It is also not clear if individual differences reduction methods such as rater training on using scale and automated scoring system developed based on rating scales are effective or not.

For research methods, future studies are suggested to take great care of individual differences of participants, to incorporate data from more authentic scale using settings, to employ other methods along with the verbal protocol to tap into raters' mind and to conduct reliability check for the data analysis process.

Lastly, to date, IELST pronunciation rating scale has received most attention and English pronunciation scale has been more often explored than others. Future research therefore can either substantiate or challenge previous conclusions by examining other pronunciation rating scales. For instance, in the context of China, future research can take the pronunciation scale in China's Standards of English Language Ability as the instrumental scale to explore its underlying construct, criteria, descriptors, length, format, and users' effects, etc.

To conclude, pronunciation assessment is still a relatively less fertile land while cultivating this land from the angle of pronunciation rating scale is promised to yield fruitful results.

## References

[1] Alderson, J. C. (1991). Bands and scores. in J. C. Alderson & B. North (eds.) *Language testing in the 1990s: The communicative legacy,* pp. 71-86. London, UK: Macmillan.

[2] Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning, 42,* 529-555.

[3] Bachman, L., Lynch, B., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12,* 238-57.

[4] Brewer, M. B. (2000). Research design and issues of validity. in Reis, H. T., & Judd, C. M. (eds.) *Handbook of research methods in social and personality psychology*, pp. 3-16. New York: Cambridge University Press.

[5] Brown, A. (2006). An examination of the rating process in the revised IELTS Speaking Test. in *International English Language Testing System (IELTS) Research Reports* (Vol. 6), pp. 41-65. Canberra: IELTS Australia, Canberra and British Council.

[6] Brown, A. (2007). An investigation of the rating process in the IELTS oral interview. in Taylor, L, & Flavey, P. (eds.) *IELTS collected papers: research in speaking and writing assessment,* pp 98-139. Cambridge: Cambridge University Press.

[7] Brown, J. D. & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34*, 21-42.
[8] Brown, A., & Taylor, L. (2006). A world survey of examiners' views and experience of the revised IELTS Speaking Test. *Cambridge ESOL: Research Notes, 26*, 14–18.
[9] Chen, H., & Li, J. N. (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.
[10] Cumming, A., Kantor, R. & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(1), 67-96.
[11] Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). Dictionary of language testing (Vol. 7). Cambridge: Cambridge University Press.
[12] 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.
[13] Derwing, TM & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication, *Language Teaching, 42*, 1-15.
[14] Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*, 125-144.
[15] Flege, J. E., & Fletcher, K. L. (1992). Talker and listener effects on degree of perceived foreign accent. *Journal of the Acoustical Society of America, 91*, 370-389.
[16] Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*, 208-238.
[17] Galaczi, E. D., Lim, G., & Khabbazbashi, N. (2012, November). Descriptor salience and clarity in rating scale development and evaluation. Paper presented at Language Testing Forum, Bristol, UK.
[18] Gao, X. (2007). An analysis and design of a reading-aloud assessment marking scheme. *Foreign Language Teaching Abroad*, (4), 15-20.
[19] Gilner, L. (2008). Assisting fluency development through task-based activities. *Journal of School of Foreign Languages, 34*, 155-177.
[20] Harding, L. (2013). Pronunciation assessment, in Chapelle, C. A. (ed.) *The Encyclopedia of Applied Linguistics*, pp. 1-6. Blackwell Publishing Ltd. DOI: 10.1002/9781405198431.wbeal0966.
[21] Harding, L. (2017). What do raters need in a pronunciation scale? The user's view. in Isaacs, T, & Trofimovich, P. (eds) *Second language pronunciation assessment*, pp. 12-34. Bristol: Multilingual Matters.
[22] 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.
[23] Isaacs, T. (2013). Assessing pronunciation. in Kunnan, A. J. (ed) *The companion to language assessment* (Vol. 1), pp. 140-155. Malden, MA : Wiley-Blackwell.
[24] 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.
[25] Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition, 34*(3), 475-505.
[26] Isaacs, T., Trofimovich, P., & Foote, J. A. (2018). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing, 35*(2), 193-216.
[27] Isaacs, T., Trofimovich, P., Yu, G., & Chereau, B. M. (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. in *IELTS research reports online series* (Vol. 4), pp. 1-48. British Council, Cambridge English Language Assessment and IDP: IELTS Australia.
[28] Kang, O.,& Rubin, D. L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology, 28*, 441-456.
[29] Ketabi, S., & Saeb, F. (2015). Pronunciation Teaching: Past and Present. *International Journal of Applied Linguistics and English Literature, 4*(5), 182-189.
[30] Knoch, U. (2017). What Can Pronunciation Researchers Learn From Research into Second Language Writing? in Isaacs, T, & Trofimovich, P. (eds) *Second language pronunciation assessment*, pp. 54-71. Bristol: Multilingual Matters.
[31] Levis, J. M. (2006). Pronunciation and the assessment of spoken language. in R. Hughes (ed.) *Spoken English, TESOL and Applied Linguistics: Challenges for Theory and Practice*. pp. 245-270. New York: Palgrave Macmillan.
[32] Lumley, T. (2005). Assessing second language writing: The rater's perspective. Frankfurt: Peter Lang.
[33] McNamara, T. (2002). Discourse and assessment. *Annual Review of Applied Linguistics, 22*, 221-242.
[34] Meng, X. J., & Wang, H. M. (2009). Boundary tone patterns in Chinese English learners' read speech. *Foreign Language Teaching and Research, 41*(6), 447-451.
[35] Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.
[36] Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 49*, 285-310.
[37] Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System, 30*(2), 143-154.
[38] Peng, N. H. (2014). A study on the differences in the use of pronunciation learning strategies between English and non-English majors. *Foreign Language Research*, (2), 105-110.
[39] Rajadurai, J. (2007). Intelligibility studies: A consideration of empirical and ideological issues. *World Englishes, 26*, 87-98.
[40] Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing, 24*(3), 355-390.
[41] Schairer, K. E. (1992). Native speaker reaction to non-nativespeech. *Modern Language Journal, 76*, 309-319.
[42] Tian, Z. X., & 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.

[43] Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition, 15*(4), 905-916.
[44] Underhill, N. (1987). Testing spoken language. Cambridge: Cambridge University Press.
[45] Van Moere, A. (2013). Raters and ratings. in Kunnan, A. J. (ed) *The companion to language assessment* (Vol. 3), pp. 1358-1374. Malden, MA: Wiley-Blackwell.
[46] Wang, H. Z. (2008). Raters' understanding and application of the TEM-4 oral rating scale. *Foreign Language Learning Theory and Practice*, (2), 33-39.
[47] Wen, Q.F., Liu, X. D., & Jin, L. M. (2005). Native and nonnative judgements of Chinese learners' English public speaking ability. *Foreign Language Teaching and Research, 37*(5), 337-342.
[48] Yates, L., Zielinski, B., & Pryor, E. (2011). The assessment of pronunciation and the new IELTS pronunciation scale. in Osborne, J. (ed.) *IELTS Research Reports* (Vol. 12), pp. 1-46. Manchester and Manchester: IDP IELTS Australia and British Council.
[49] Yu, J. (2013). Rhythm patterns in the speech of Chinese EFL learners---A phonetic study based on Hangzhou-accented learners. Doctor's dissertation, Zhejiang University.
[50] Zhou, W. J. & Song, H. P. (2015). A study on Chinese college students' productive phonetic competence in English. *Foreign Languages and Their Teaching*, (3), 1-7.
[51] Zhu, P. P. (2011). A Study on the Characteristics of the Rhythm Patterns in Chinese English EFL Learners' Reading-aloud Production. Master's thesis, Xuzhou Normal University.
[52] Zielinski, B. (2006). The intelligibility cocktail: An interaction between speaker and listener ingredients. *Prospect, 21*, 22-45.

**Wenjun Zhong** was born in Mianzhu, China in 1994. She received her bachelor's degree in translation and interpretation from Tianjin Foreign Studies University, China in 2016.

She is currently a student pursuing the master's degree in English Linguistics and Applied Linguistics in the School of English and International Studies, Beijing Foreign Studies University, Beijing, China. Her research interests include phonetics, phonology, language assessment and second language acquisition.