To be published in *System* (Elsevier)

What is Second Language Pronunciation Proficiency? An Empirical Study

Yui Suzukida & Kazuya Saito[1]

Abstract

The current study set out to examine which segmental and suprasegmental factors discriminate different levels of global second language (L2) pronunciation proficiency. First, a total of 40 extemporaneous speech samples were elicited from Japanese learners of English with diverse experience/proficiency levels. Subsequently, experienced raters holistically assessed the global pronunciation qualities of the samples using the rubrics in IETLS Pronunciation Scale (Low to High). Finally, the dataset was submitted to a comprehensive set of segmental and suprasegmental measures. The results revealed that the raters attended to, in particular, the ratio of segmental errors with high communicative value (determined via the functional load principle) to distinguish between Low- and Mid-level L2 pronunciation proficiency. Other specific measures—segmental errors with low communicative value, the schwa vowel insertion in complex syllables, and the absence of word stress—played a significant role in the raters' decision to assign high ratings to identify High-level L2 pronunciation proficiency.

Key words: Second language pronunciation, pronunciation proficiency, segmentals, suprasegmentals

[1] Corresponding author: Kazuya Saito (k.saito@ucl.ac.uk)

One of the most extensively-researched topics in the field of second language (L2) acquisition has been how to define and assess the multifaceted characteristics of L2 oral proficiency according to various linguistic domains and different learner proficiency levels (De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012 for Communicative Adequacy; Iwashita, Brown, McNamara, & O'Hagan, 2008 for TOEFL Speaking Proficiency). For example, much attention has been given to examining what linguistic factors matter when raters evaluate the lexicogrammar aspects of overall L2 speech proficiency (e.g., Crossley, Salsbury & McNamara, 2015 for lexical proficiency; Révész, Ekiert, & Torgersen, 2016 for grammatical accuracy and complexity). However, few researchers have ever delved into the *pronunciation* aspects of L2 oral proficiency, i.e., global L2 pronunciation proficiency. In the current study, expert native speaking raters first assessed the overall pronunciation qualities of 40 Japanese learners of English with varied experience and proficiency levels with reference to the rubrics in IELTS Pronunciation Scale. Then, we explored which dimensions of pronunciation features—i.e., segmentals, syllables, word stress, and intonation—jointly interacted to affect the raters' subjective judgement of the low-, mid-, and high-level L2 pronunciation.

1. Background

1.1 Second Language Pronunciation Proficiency

Among the wide range of linguistic skills comprising L2 oral proficiency (e.g., lexicogrammar, fluency, discourse), attaining intelligible (but not necessarily nativelike) pronunciation is particularly crucial, as it directly impacts other native and non-native speakers' successful understanding of L2 speech in real life situations (Crowther, Trofimovich, Isaacs, & Saito, 2015a; Crowther, Trofimovich, Saito, Isaacs, 2015b; Derwing & Munro, 2015; Jenkins, 2002; Saito, 2021; Trofimovich & Isaacs, 2012). According to previous literature, many scholars have extensively examined which pronunciation errors are relatively relevant or irrelevant when novice listeners are asked to make *intuitive* judgements of global comprehensibility on a Likert scale (e.g., difficult to understand to easy to understand) or/and transcribe what they have heard (i.e., intelligibility). In general, L2 comprehensibility and intelligibility assessment of this kind is strongly associated with a range of suprasegmental features, such as word stress (Field, 2005), primary stress in sentences (Hahn, 2004), and choice of tone (Kang, 2008) and fluency features such as speech rate and number of filled pauses (Suzuki & Kormos, 2020). The phonological correlates of L2 comprehensibility/intelligibility can also be related to listeners' particular backgrounds, such as foreign-accent familiarity (Kennedy & Trofimovich, 2008), pedagogical training experience (Saito, Trofimovich, Isaacs, Webb, & Webb, 2017), and L1 vs. L2 listeners (Saito, Tran, Suzukida, Sun, Magne, & Ilkan, 2019).

When it comes to segmentals, certain individual sounds are likely to influence L2 comprehensibility more strongly than others. Such relative weights of segmental errors have been determined via the Functional Load principle (Brown, 1988). For example, Munro and Derwing's (2006) study adopted the FL principle as a coding scheme to identify the set of segmentals with relatively high communicative value. Munro and Derwing (2006) revealed that high FL errors (/l, ʃ, n, s, d/) negatively affect the judgments of comprehensibility to a greater extent. In contrast, the negative impact on comprehensibility caused by low FL errors (/ð, θ/) is much weaker (see also Suzukida & Saito, 2019). To date, however, it has thus far remained highly controversial and unclear which specific pronunciation features (e.g., high vs. low FL segmentals, syllables, word stress, intonation) differentially contribute to *expert* raters' overall pronunciation judgements, and what features distinguish between different pronunciation proficiency levels.

1.2 Assessing Second Language Pronunciation

Thus far, researchers have extensively examined which pronunciation features distinguish different proficiency levels in terms of listeners' *intuitive* judgements of pronunciation proficiency (for linguistics correlates of different accentedness and comprehensibility levels, see Isaacs & Trofimovich, 2012; Saito, Trofimovich, & Isaacs, 2016, 2017). However, it is notable that such intuitive judgements can inevitably be related to a range of linguistic errors (beyond pronunciation features), because raters are asked to evaluate L2 speech without any instruction or descriptors. Little is known about which segmental and suprasegmental features matter when *linguistically trained, expert* raters *intentionally* evaluate the *phonological* quality of L2 speech (for further discussion on the multifaceted nature of L2 pronunciation proficiency, see Saito & Plonsky, 2019). Such information of crucial features is especially valuable not only for language instructors but also for L2 learners who seek to achieve more than minimum communicative success (such as obtaining certain scores in high-stakes testing).

In the realm of L2 lexicogrammar, Crossley and his collogues have conducted a series of empirical studies to examine the linguistic influence on linguistically-trained raters' global L2 lexical proficiency judgements (Crossley, Salsbury & McNamara, 2015; Crossley, Salsbury, McNamara & Jarvis, 2011a, 2011b). Albeit limited in numbers, few studies have investigated linguistic correlates of global L2 pronunciation proficiency by using detailed and validated linguistic descriptors or pre-rated speech samples from testing organizations (e.g., Cambridge English in Galaczi, Post, Li, Baker & Schmidt, 2016; IELTS in Isaacs, Trofimovich, Yu, & Chereau, 2015; TOEFL in Iwashita et al., 2008). For example, with the speech samples of mixed L1speakers, Iwashita et al. (2008) identified target-like syllables, and fluency factors (speech rate, unfilled pauses, and total pause time) were particularly influential among other phonological measures to TOEFL iBT proficiency judgements. Similarly, Kang (2012) adopted a rubric of TOEFL iBT and demonstrated that suprasegmental features (rate, pause, stress, and pitch measures) produced by speakers of mixed L1 backgrounds (Chinese-Mandarin, Korean, Japanese, Saudi Arabian, Russian, Hindi, and Nepali) contributed to global L2 pronunciation proficiency judgements.

1.3 Motivation for Current Study

In particular, the pronunciation qualities of the speech samples have been analyzed via the accurate use of segmentals, word stress, and intonation (e.g., Trofimovich & Isaacs, 2012). However, it is noteworthy that such segmental/suprasegmental factors could be further scrutinized via a wide range of other specific pronunciation measures from multiple angles. For instance, certain segmentals are suggested to be more detrimental to listeners' successful understanding of L2 speech (e.g., Munro & Derwing, 2006 for Functional Load; Jenkins, 2002 for Lingua Franca Core). There is also suggestive evidence that some suprasegmental errors may entail more communicative and acquisitional values. Native listeners' intelligibility judgements could be more negatively affected by "the deletion of segmentals" rather than "the insertion of schwa vowels" in complex syllables (Lin, 2001), and by "misplacement" rather than by "absence" of word stress (Field, 2005).

In the current investigation, we first recruited 40 Japanese learners of English with diverse proficiency levels (potentially covering a wide range of proficiency levels) and asked them to participate in a spontaneous speaking task (resulting in high-quality audio recordings suitable for fine-grained pronunciation analyses). Second, these samples were categorized by expert native speaking raters for Low, Mid, vs. High L2 pronunciation proficiency with reference to the publicly available descriptors in IETLS Pronunciation Scale. Finally, we scrutinized various dimensions of the 40 Japanese learners' speech in terms of segmentals (high FL, low FL), syllables (insertion, deletion), word stress (absence, misplacement) and intonation (absence, misplacement). The following research question was formulated:

- Which segmental and prosodic features influence expert raters' judgements of low-, mid-, and high-level L2 pronunciation proficiency (defined by IELTS Pronunciation Scale)?

As for predictions, we hypothesized that the frequency of high FL segmental errors would play a crucial role in distinguishing between low and mid-level L2 pronunciation proficiency; but the frequency of low FL segmental errors would be a key factor of distinguishing between mid- and high-level L2 pronunciation (Munro & Derwing, 2006). Given the general importance of suprasegmentals in L2 oral abilities (e.g., Trofimovich & Isaacs, 2012 for comprehensibility), we hypothesized that certain errors (e.g., misplacement rather than deletion) could be a determinant factor at every level of L2 pronunciation proficiency (low → mid → high).

The main objective of the current study did not relate to the validation of IELTS Pronunciation Scale. To define which of the 40 speech samples (produced by Japanese speakers of English) could be categorized as Low, Mid vs. High, we took an exploratory approach towards using the publicly available descriptors in IELTS Pronunciation Scale. Note that the in-house L2 pronunciation judgements in the current study (using the public version of IELTS Pronunciation Scale) could be essentially different from the way expert IELTS raters are usually trained to assess the pronunciation qualities of oral interviews using the non-public versions of IELTS Pronunciation Scale.

2. Method

2.1 Speech data

To cover the extensive range of pronunciation proficiency levels, a total of 40 Japanese learners of English with varied learning experiences and backgrounds were recruited in both Tokyo (Japan) and Calgary (Canada). Although all the participants had begun learning English in Japanese EFL classrooms from Grade 7, they widely varied in their age of testing (*M* = 25.6 years; *Range* = 18–53), as well as in the length of their experience abroad (*M* = 4.6 years; *Range* = 0–24) (summarized in Table 1). The sample size of the study (*N* = 40) was based on a range of precursor studies (e.g., Trofimovich & Isaacs, 2012 for *N* = 40 French speakers of English; Derwing & Munro, 1997 for *N* = 48 mixed ESL speakers).

Table 1. *Summary of 40 Japanese Participants' Experience and Age Profiles*

Length of the stay (months)	*n*	Age	*n*
0	16	18–19	5
0–10	16	19–20	16
10–20	5	20–30	11
20–30	3	30–53	8
Total	40	Total	40

The participants engaged in a monologue speaking task with the researcher. In order to elicit extemporaneous monologue, a decision was made to follow a style used in the standardised test, IELTS. A topic was tailored for the study (*Describe one of the toughest challenges in your life*) as well as suggestions of possible discussion points (*When? How old and where were you? Why did you encounter this challenge? Why was it so challenging? Did your friends/parents help you?*). The participants had one minute to prepare (while taking notes if necessary). Then, they gave their responses, speaking for 2 minutes. Afterwards, the researcher followed up each response with the following two round-off questions (*What did you learn from this experience? Would you like to go through the same experience again?*).

In keeping with L2 speech research standards (e.g., Derwing & Munro, 1997), and for the purpose of minimizing listeners' fatigue, the first 30 seconds of the approximately three-minute speeches (the participants' response to the prompt and round-off questions) were excised from each of the 40 audio recordings and saved as WAV files for the rating sessions.

2.2 Pronunciation Rating

In previous literature, global constructs of L2 oral proficiency have been measured via trained raters' scalar judgements. In Crossley et al. (2015), the in-house rating scale was elaborated based on the ACTFL guidelines. Similarly, our own rating scale for low, mid and high-level proficiency was created in conjunction with the publicly-available IELTS Pronunciation Scale (IELTS, 2017). A total of five raters with an extensive amount of ESL teaching experience were recruited to this end. They first received tailored training on the rating procedure by one of the researchers in order to familiarize themselves to the rating procedure, and then they evaluated the pronunciation quality of the 40 speech samples.

Note that our rationale of using the IETLS Pronunciation Scale here was to ensure that the raters clearly understood the nature of the task—i.e., making some form of global judgement while paying attention to pronunciation (rather than fluency and lexicogrammar) aspects of language without confusion. We were interested in eliciting the raters' agreed perception of what samples could represent low, mid and high-level L2 pronunciation proficiency—see the similar methodology and justifications in Crossley et al. (2015). As shown below, the raters demonstrated highly consistent judgements (α = .91), and we consider the results here to be satisfactory to the goal of our research. However, we acknowledge that future research is surely needed to examine precisely how much the IELTS Pronunciation Scale could be considered as a valid index and the extent to which the raters followed the descriptors—the topics beyond the scope of our current investigation (for the details of the validity and problems underling the IELTS Pronunciation Scale, see Isaacs, Trofimovich, Yu & Chereau, 2015).

Raters. Five experienced native teachers were recruited in London, England, as raters. Their mean age was 36.4 (*Range* =29–56). All of the raters were born and/or raised in an English-speaking environment. Although one rater reported that both of his parents spoke Urdu as L1, he considered English to be his L1. They also confirmed that their major medium of the communication was English (*M* = 96%; *Range* = 90–100). With respect to their linguistic knowledge and teaching career, all the raters responded that they had obtained extensive linguistic/phonological knowledge through master degrees in Applied Linguistics, TESOL, or TEFL courses at universities in the UK, and had taught in EFL/ESL contexts (*M* = 10.7 years; *Range* = 3–25). This suggests that the raters could be considered relatively homogenous in their phonological and linguistic knowledge and their teaching experience. In light of Isaacs and Thomson's (2013) definition, our raters could be considered as "expert" rather than "novice." Based on a 6-point scale (1 = *not at all*, 6 = *very much*), all the raters showed relatively high familiarity with Japanese-accented English (*M* = 5.14; *Range* = 4–6), and had frequent contact with native speakers of Japanese (*M* = 5; *Range* = 4–6). In addition, four out of five raters reported experience visiting Japan (two weeks, one month, six months, and two years, respectively) while two of them had taken short Japanese language courses. Hence, the raters' familiarity with Japanese-accented English was considered as consistent — another significant factor that influences leniency in L2 speech judgements (Winke, Gass & Myford, 2013).

Procedure. All the rating sessions were individually conducted by one of the authors in a quiet room at a university in London. At the beginning of each rating session, the author carefully explained the rating procedure. In order to ensure that the raters adequately understood the procedure, they were first invited to join a practice session where they listened to three 30-sec speech samples (not included in the main dataset), rated them on a 9-point scale by referring to the publicly-available IELTS pronunciation scale (see APPENDIX), explained their rating decisions and received feedback from the author. The role of the author here was to make sure the raters' comfortable use of the rubric by eliciting the reasoning of the raters' evaluation. Thus, careful attention was paid to avoid any priming effect to the raters. After the raters felt comfortable with the procedure, and the author confirmed their adequate understanding of the rubric, they proceeded to the assessment of 40 simulated speech samples. During the session, the raters evaluated each speech samples via audio listening; however, they did not access any other resources such as waveforms and spectrograms. The samples were played in a randomized order via *Praat* (Boersma & Weenink, 2012). After hearing each sample once, the raters made their rating decisions without any time pressure.

2.3 Pronunciation Analyses

The authors conducted a detailed error analysis of the 40 speech samples by using a total of 12 specific pronunciation measures ranging from the segmental (individual vowels and consonants) to the suprasegmental level (i.e., syllable errors, misplacement of word stress, word stress absence, misplacement of intonation, intonation absence). Segmental errors were further subdivided according to the rank ordering list elaborated from the FL principle (Brown, 1988); accordingly, all of the segmental errors were sorted into either the high FL category or the low FL category.

Segmental Measures. Whereas many L2 pronunciation studies typically adopt native coders to conduct error analyses at a fine-grained level (for a methodological review, see Piske, MacKay & Flege, 2001), the decision was made to recruit non-native coders for the current study for the following reasons. When it comes to L2 pronunciation, it has remained extremely controversial and difficult for native coders to make dichotomous judgements regarding whether L2 learners have made mispronunciation and/or unclear pronunciation of target sounds (Jenkins, 2002). Given ample evidence that even early bilinguals can rarely attain "nativelike" pronunciation proficiency in an L2 (Abrahamsson & Hyltenstam, 2009), to our knowledge, there is no theoretically robust threshold that L2 pronunciation researchers can depend on to reliably decide what can be categorized as non-native speakers' pronunciation errors or not. To reflect this methodological concern, we used Riney et al.'s (2000) approach to analyze L2 pronunciation forms of the same-L1 talker group (Japanese learners of English). According to Riney et al. (2000), pronunciation errors were counted only when the talkers substituted their L1 Japanese counterparts for L2 English sounds (e.g., the Japanese tap sound instead of English /r/ nor /l/). Therefore, the current study restricted its focus on errors that attribute to L1 (but not developmental errors); if a coder perceived the talkers' effort to pronounce L2 English sounds (regardless of their "targetlikeness"), their pronunciation forms were not coded as "errors."

To this end, the primary author — a native speaker of Japanese with high-level proficiency in L2 English and extensive experience in L2 speech analysis of this kind — first conducted auditory evaluations of the 40 samples. The evaluation involved careful listening of each speech sample with transcription. In order to confirm the validity of the author's segmental analyses, two additional coders later conducted the same evaluations *separately*. Both coders had experience in speech coding, and more than five-years of English teaching experience, and demonstrated high-level proficiency in L2 English. The analysis of the author and the two coders demonstrated relatively high inter-rater reliability (*r* = .96). In the case of any disagreement, the author/coders discussed until they reached a consensus. The total number of segmental substitution errors was divided by the total number of segments produced for each sample.

To further examine the different gravity of the errors, we also divided the segmental errors into high and low functional load categories in accordance with Brown's (1988) FL theory. This theory states that certain segmental contrasts (e.g., English /r/ vs. English /l/) are assumed to entail more communicative value than others (e.g., English /s/ vs. /θ/), as these contrasts include more minimal pairs in frequently-used word contexts. In fact, high FL errors (e.g., /l, ʃ, n, s, d/) were found to make more negative impacts on native speakers' comprehensibility judgements than low FL errors (e.g., /ð, θ/) (Munro & Derwing, 2006). At the same time, we needed to adopt the FL theory in the current study. Since we coded segmental errors solely based on the presence of substitutions (English /r/ was pronounced as the Japanese tap), the original concept of the "contrast" (English /r/ was pronounced as English /l/) was not applicable in the context of the current study.

As a remedy, we used the following exploratory approach. Although there are two studies of FL theory (i.e., Brown's and Catford's), a decision was made to use Brown's (1988)'s FL ranking due to its use of values instead of percentages (see Catford, 1987). Because we wanted to illustrate the importance of each segment instead of values of certain substitutions, we newly created a ranking of individual segmentals based on Brown's (1988) original rankings of segmental pairs (see Supporting Information for the original ranking). The procedure is illustrated in the following. In light of FL scores suggested in Brown's (1988) ranking, we first calculated a total score assigned to a segmental sound, then divided the score by the number of times the sound actually appeared in Brown's ranking in order to reflect the gravity of the segmental in the ranking (as summarized in Table 2). For example, English /s/ appeared four times in the following contrasts (score 8 for /s/-/z/; score 7 for /s/-/ʃ/; score 6 for /s/-/ʒ/; score 5 for /θ/-/s/). So the total score of /s/ was 26, and that was divided by four. Thus, English /s/ was assigned score 6.5. All the segmentals were ranked according to their averaged FL scores. They were divided into high and low functional load groups for consonants and vowels (for a similar approach of grouping FL segmentals into high and low, see Munro & Derwing, 2006).

Table 2. *Summary of High and Low Functional Load Grouping of Individual Segmentals Based on Brown (1988)*

Segmentals	Segmental Types	Total scores[a]	Times appeared	Average
A. Vowels				
eɪ	High Functional Load	9	1	9
ɑɪ		9	1	9
æ		34	4	8.5
ɪ		17	2	8.5
əʊ		25	3	8.33
ʌ		30	4	7.5
e		36	5	7.2
ɒ		35	5	7
ɜː		24	4	6
iː		11	2	5.5
ɔː	Low Functional Load	26	5	5.2
ɑː		25	5	5
ʊ		7	2	3.5
eə		6	2	3
aʊ		3	1	3
ɪə		5	2	2.5
uː		4	2	2
ʊə		1	1	1
ɔɪ		1	1	1
B. Consonants				
p	High Functional Load	20	2	10
m		10	1	10
l		20	2	10
r		10	1	10
h		9	1	9
k		9	1	9
g		9	1	9
b		17	2	8.5
n		25	3	8.3
w		8	1	8
d		14	2	7
v		28	4	7
f	Low Functional Load	27	4	6.75
z		20	3	6.67
t		13	2	6.5
s		26	4	6.5
ð		23	4	5.75

ŋ	5	1	5
θ	15	4	3.75
ʃ	11	3	3.67
ʒ	10	3	3.33
dʒ	9	3	3
tʃ	5	2	2.5
j	3	2	1.5

Note. [a] Total score was calculated by adding each ranking value (1-10) of a segmental illustrated in Brown (1988).

Suprasegmental Measures. Following the suprasegmental analysis method used in Isaacs and Trofimovich (2012), the same coders (the author and the two coders) also analyzed nine different suprasegmental aspects of L2 speech (see below). The analysis was conducted manually by hearing the speech samples with transcripts; no other information such as spectrograms were provided to purely focus on listening-based, objective error analysis. Syllable and word counts were first counted manually by the researchers as coders based on the transcripts, then the counts were double checked with a syllable and word counter *Wordcalc.com*. The transcript did not contain non-target like syllable forms but correct forms. For example, even if a speaker produced “it” as “ito” it was transcribed as “it.” The inter-rater reliability of the coders’ suprasegmental analyses was relatively high (r = .96). When their analyses demonstrated any disagreement, they discussed the analysis to reach a consensus (as in the segmental analyses).

- Syllable insertion error ratio was analyzed via dividing the overall number of syllable errors caused by schwa vowel insertion (e.g., *it* spoken as *ito*) by the overall number of syllables per sample.
- Syllable deletion error ratio was analyzed via dividing the overall number of syllable errors caused by deletion (e.g., *year* articulated without the initial /y/) by the overall number of syllables per sample.
- Overall syllable structure error ratio was analyzed via dividing the overall number of syllable errors (insertion, deletion) by the overall number of syllables per sample.
- Word stress misplacement error ratio was analyzed via dividing the overall number of word stress errors caused by the misplacement of primary stress (e.g., *CHA-llenge* spoken as *Cha-LLENGE*) by the overall number of polysyllabic words per sample.

- Word stress absence error ratio was analyzed via dividing the overall number of word stress errors caused by the absence of primary stress (e.g., *CHA-llenge* spoken as *cha-llenge or CHA-LLENGE*) by the overall number of polysyllabic words per sample. The absence of the primary stress includes both syllables pronounced with unreduced and reduced vowels.
- Overall word stress error ratio was analyzed via dividing the overall number of word stress errors (misplacement and absence) by the overall number of polysyllabic words per sample.
- Intonation misplacement error ratio was analyzed via dividing the overall number of misplacement errors — the inadequate choice of pitch (e.g., falling tone at the end of statement spoken in raising tone) — at the end the phrases by the overall number of obligatory contexts where raising, falling or level tone were expected. For a similar analysis, see Trofimovich and Isaacs (2012).
- Intonation absence error ratio was analyzed via dividing the overall number of absence errors — failing to produce any pitch movements (e.g., falling tone at the end of statement spoken with no tones) — at the end the phrases by the overall number of obligatory contexts per sample.
- Overall intonation error ratio was analyzed via dividing the overall number of intonation errors (misplacement, absence) at the end the phrases by the overall number of obligatory contexts per sample.

3. Results

We carried out a set of statistical analyses in order to examine which specific pronunciation measures (segmentals, syllables, word stress, intonation) were strongly associated with the IELTS pronunciation proficiency ratings.

3.1 Rater Consistency

First, the inter-rater reliability was checked among the five raters who assessed the overall pronunciation proficiency. Larson-Hall (2010) provided detailed guidelines specifically for L2 researchers to calculate inter-rater *consistency* for cases of judges rating persons (the main focus of the study). Different from orthodox analyses for inter-rater reliability (e.g., Cohen's Kappa, Fleiss Kappa), Larson-Hall recommended the use of Cronbach alpha as a way of *inter-class* correction because it takes into account not only the correlations between raters (reliability), but also the differences between raters' scores (consistency). Reliability indicates the similarity/difference in relative rating patterns, i.e., the extent to which rater behaviors follow consistent patterns. If raters simply agree on which samples should be considered as relatively low-, mid-, and high-proficiency in a certain order, their reliability could be high (although their actual rating scores may be different with certain raters being more lenient or strict than others). Further, consistency indexes the similarity/differences in absolute rating values, i.e., the extent to which raters' actual rating scores match with each other. Considering both reliability and consistency, the Cronbach alpha allows researchers to track different types of variances tied to participants, speech stimuli, and raters.

In fact, Cronbach alpha has been used in previous literature (e.g., Derwing & Munro, 1997) and considered as the standard analysis for inter-rater consistency in L2 speech research (see Saito, 2021 for a research synthesis). The alpha of the five raters was .91. In conjunction with Larson-Hall's (2010) benchmark in the field of applied linguistics, the inter-rater consistency reported here met the acceptable level (above α = .70), and therefore could be considered satisfactory, suggesting that the five raters appeared to demonstrate a sufficient amount of agreement while evaluating the pronunciation qualities of the 40 samples.

3.2 Classifying Groups

In order to classify the 40 audio samples into different proficiency levels, the researchers first averaged the five raters' scores (based on 9-point scale) assigned to each sample and rounded up (for a similar method, Derwing & Munro, 1997). For example, if one sample received 7, 8, 7, 8 and 8 from the five raters, respectively (the average score for the sample was 7.6), it was considered to belong to the Band 8 group. Instead of looking at each rater's behavior, we decided to highlight their average scores, because the goal of the current study was to elicit any generalizable patterns in the five raters' pronunciation judgements.

As summarized in Table 3, the samples widely ranged from Bands 5 to 9. Due to the small sample size for Band 9, we created a composite group — i.e., Bands ≥ 8. In a similar fashion, we found seven samples for Bands 5 and 6, respectively. Therefore, to conduct robust statistical comparisons of different groups with approximately similar sample size, the decision was made to create another composite group including samples from both Bands 5 and 6 — i.e., Bands ≤ 6. In conjunction with the main objective of the current study (i.e., measuring the segmental and suprasegmental correlates of different levels of L2 pronunciation proficiency), and following the British Council's reference of the CEFR benchmarks (British Council, 2018), the Bands ≥ 8 group was roughly labeled as Upper-Proficient users of L2 pronunciation (representative of C2), the Band 7 group as Lower-Proficient users of L2 pronunciation (representative of C1), and the Bands ≤ 6 group as Basic-Independent users of L2 pronunciation (representative of B1, B2). For the sake of the analyses presented in this study, we labeled Upper-Proficient as "High," Lower-Proficient as "Mid" and Basic-Independent as "Low."

Table 3. *Results of Group Classifications According to Different Pronunciation Proficiency Levels*

Global scores	Number of Speech samples	Classification	Number of speech samples in the proficiency groups
5	7	Low	14
6	7		
7	15	Mid	15
8	9	High	11
9	2		

3.3 Inter-relations Between Pronunciation Measures

In the current study, a comprehensive set of 12 pronunciation-specific measures was adopted. Table 4 illustrates comprehensive summaries of all of the error analyses in value and ratio. With the averaged rating scores and the result of error analysis in ratio, a set of correlations was computed. According to the results of the Spearman's correlation analyses (summarized in Table 5), strong correlations were found for overall and sub-constructs of segmentals (overall vs. high/low FL), syllables (overall vs. insertion), word stress (overall vs. misplacement, and overall vs. absence) and intonation (overall vs. misplacement). In contrast, the overall segmental, syllable, word stress and intonation measures were not significantly related to each other except for segmental and syllable errors. Regarding the sub-measures, strong relations were found between overall segmental errors and syllable insertion errors, low FL segmental error and overall syllable errors, and overall syllable errors and word stress absence errors. As conceptualized earlier, the results here generally indicated that the 12 pronunciation measures in the current study were assumed to mirror four different specific dimensions of L2 pronunciation proficiency — segmentals, syllables, word stress and intonation.

Table 4. *Summaries of Participants' Errors of 12 Pronunciation Measures*

Pronunciation measures	Absolute value				Ratio value (%)			
	M	*SD*	95%CI		*M*	*SD*	95%CI	
			Lower	Upper			Lower	Upper
Segmentals (overall)	2.54	2.38	1.80	3.28	1.64	0.63	1.14	2.15
Segmentals (high FL)	1.23	1.58	0.74	1.71	0.81	1.12	0.46	1.16
Segmentals (low FL)	1.25	1.37	0.82	1.68	0.83	0.90	0.55	1.11
Syllables (overall)	2.05	2.55	1.26	2.84	3.63	4.70	2.17	5.08
Syllables (insertion)	1.70	2.50	0.92	2.48	3.08	4.68	1.63	4.53
Syllables (deletion)	0.35	0.74	0.12	0.58	0.73	1.47	0.28	1.19
Word stress (overall)	3.13	2.54	2.34	3.91	27.81	23.08	20.66	34.96
Word stress	1.05	1.51	0.59	1.52	8.37	12.24	4.58	12.17
Word stress (absence)	2.08	2.26	1.38	2.78	19.44	21.31	12.84	26.04
Intonation (overall)	0.50	0.91	0.22	0.78	8.08	13.85	3.79	12.37
Intonation	0.32	0.67	0.12	0.53	7.86	13.88	3.56	12.16
Intonation (absence)	0.03	0.16	-0.02	0.07	0.19	1.22	-0.18	0.57

Table 5. *Inter-Relationships Between 12 Specific Pronunciation Measures*

	1		2		3		4		5		6		7		8		9		10		11	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
1. Segmentals (overall)																						
2. Segmentals (high FL)	.78*	< .001																				
3. Segmentals (low FL)	.78*	< .001	.29	.067																		
4. Syllables (overall)	.44*	.004	.14	.371	.48*	.002																
5. Syllables (insertion)	.46*	.002	.23	.140	.39	.013	.84*	< .001														
6. Syllables (deletion)	.02	.857	-.10	.521	.15	.344	.30	.057	-.11	.480												
7. Word stress (overall)	.33	.033	.24	.125	.25	.115	.26	.104	.26	.104	-.01	.932										
8. Word stress (misplacement)	.03	.829	.19	.219	-.15	.339	-.16	.300	-.11	.484	-.25	.116	.44*	.004								
9. Word stress (absence)	.33	.034	.12	.432	.40	.01	.41*	.009	.35	.035	.14	.363	.83*	< .001	.01	.979						
10. Intonation (overall)	-.16	.297	-.11	.469	-.20	.200	-.07	.650	-.07	.665	.27	.085	.11	.479	-.16	.312	.22	.158				
11. Intonation (misplacement)	-.15	.354	-.07	.633	-.21	.182	-.05	.746	-.02	.875	.22	.159	.13	.417	-.12	.428	.21	.178	.97*	< .001		
12. Intonation (absence)	-.07	.638	-.16	.324	.02	.861	-.08	.604	-.18	.255	.21	.194	-.06	.700	-.14	.380	.05	.761	.14	.370	-.09	.551

Note. *indicates $p < .01$

3.4 Between-band Comparisons for the Pronunciation Scales

In order to examine which of the pronunciation measures (segmentals, syllables, word stress, intonation) distinguished between the different pronunciation proficiency levels measured via holistic rating, we performed a set of one-way ANOVAs with the results of the 12 pronunciation measures as dependent variables and the group categories (Low, Mid, and High) as independent variables. The results of Levene's tests pointed out the violation of homogeneity of variance in all the pronunciation measures; thus, multiple comparisons —Low vs. Mid vs. High— were carried out using Tamhane's post-hoc tests.

According to the ANOVAs (summarized in Table 6), the raters' pronunciation proficiency ratings were significantly related to segmental errors (overall, high FL, low FL), syllable errors and word stress errors but intonation errors at a *p*. < 01 level. In fact, a significant group distinction could be observed in overall segmental errors, *F*(2, 37) = 17.937, *p* = .001; high FL segmental errors, *F*(2, 37) = 12.399, *p* = .001; low FL segmental errors, F(2, 37) = 6.134, *p* = .005, overall syllable errors, *F*(2, 37) = 7.244, *p* = .002; syllable insertion errors, *F*(2, 37) = 5.804, *p* = .006; overall word stress errors, *F*(2, 37) = 9.162, *p* = .001; and word stress absence errors, F(2, 37) = 6.811, *p* = .003. Post-hoc multiple comparison analyses further revealed that these pronunciation measures uniquely differentiated between three proficiency levels (Low, Mid, and High). According to the result shown in Table 8, the low FL segmental errors distinguished between Low and High (*p* = .012), whereas the high FL segmental errors differentiated between Low and Mid (*p* = .004) but not between Mid and High. Same distinctions were observed with overall segmental errors: Low and Mid were distinguished (*p* = < .001) but not Mid and High. Regarding suprasegmental measures, while syllable insertion errors differentiated between Low and High (*p* = .03), overall syllable and word stress errors distinguished Mid and High: overall syllable errors distinguished Low and High (*p* = .015) and marginally differentiated Mid and High (*p* = .086), overall word stress errors marginally distinguished Low and High (*p* =.054) and Mid and High (*p* = .088), and word stress absence errors distinguished Mid and High (*p* = .057) but not Low and Mid or Low High.

Table 6. *Summary of Group Differences for Low, Mid and High Levels of Pronunciation Proficiency*

Pronunciation measures	ANOVA results				Significant group differences
	F(2, 37)	*p*	η_p^2	Power	
Segmentals (overall)	17.937	.001*	.48	.999	Low < Mid = High
Segmentals (high FL)	12.399	.001*	.40	.993	Low < Mid = High
Segmentals (low FL)	6.134	.005*	.24	.841	Low < High
Syllables (overall)	7.244	.002*	.28	.891	Low = Mid < High
Syllables (insertion)	5.804	.006*	.23	.822	Low < High
Syllables (deletion)	.382	.685	.02	.836	*n.a.*
Word stress (overall)	9.162	.001*	.33	.667	Low = Mid < High
Word stress (misplacement)	.673	.516	.03	.415	*n.a.*
Word stress (absence)	6.811	.003*	.26	.367	Mid < High
Intonation (overall)	.886	.421	.04	.252	*n.a.*
Intonation (misplacement)	1.010	.371	.05	.281	*n.a.*
Intonation (absence)	1.341	.274	.06	.271	*n.a.*

Note. *indicates *p*. < 05

4. Discussion and Conclusion

By incorporating a comprehensive set of 12 pronunciation-specific measures (spanning various dimensions of segmentals, syllables, word stress and intonation), the main objective of the current study was to thoroughly scrutinize the segmental and suprasegmental correlates of different levels of pronunciation proficiency (low to high). To this end, a total of 40 speech samples produced by Japanese learners were divided into Low-, Mid-, and High-level global L2 pronunciation proficiency, and submitted to a set of acoustic analysis. Using the functional load principle, we also determined the relative weights of segmental errors in successful communication. Segmental errors with higher functional load are supposed to make more negative impact on listeners' understanding of foreign-accented speech.

First and foremost, the results showed that the segmental factor differentially impacted the raters' pronunciation proficiency ratings according to two types of errors (high vs. low communicative value). Whereas the number of errors with high communicative value played a significant role in the distinction between Low- and Mid-level learners, the errors with low communicative value differentiated between Mid- and High-level. Second, the syllable (*insertion* rather than *deletion*) and word stress (*absence* rather than *misplacement*) factors distinguished between Mid- and High-level learners. Finally, the intonation factor was not significantly associated with the raters' judgements at least in the current study's range of pronunciation level (Low, Mid, and High).

With speech of Japanese learners of English, the raters' evaluation (who have no experience seemed to be affected by a segmental factor with more communicative value (i.e., errors with high communicative value) to identify speech performance that meets Mid-level of L2 pronunciation proficiency. However, that appeared not to be the case in terms of segmental features with less communicative value (i.e., errors with low communicative value) and prosodic features. The reason why the raters showed the specific rating pattern in regard to the role of segmental accuracy while distinguishing different levels of L2 pronunciation proficiency may be due to the relatively high saliency of segmental errors in the listeners' perception when evaluating the speech. According to the studies of speech perception, the foreign accentedness of L2 learners' speech has extreme saliency to native speakers of the target language regardless of its comprehensibility (e.g., Derwing & Munro, 1997). In addition, segmental accuracy is found to be one of the sources of accentedness (see also Saito et al., 2016, 2017). Therefore, it is reasonable to interpret that segmental errors produced by the speakers in the current study may involve in both stages (Low- vs. Mid-level, and Mid- vs. High-level) of the listeners' evaluation.

In the current study, it is noteworthy that segmental and prosodic errors differentially influenced raters' judgments. When it comes to assigning the speech performances to High-level, our raters' judgements appeared to have been affected by segmentals with less communicative value (low communicative/FL errors) and suprasegmentals (syllable insertion errors, word stress absence errors). This indicates that while frequency and quality of high FL segmental errors are consequential to distinguishing Mid-level L2 speakers from Low-level L2 speakers, word stress, syllable, and low FL segmental accuracy may be a prerequisite for reaching the High-level of L2 pronunciation in the current study.

One of the reasons of the differences in its influence on raters' judgements between segmental (high FL) and prosodic errors could be due to the raters' high familiarity with

Japanese accented English. According to studies of listeners' perceptions (e.g., Carey, Mannel & Dunn, 2011; Kraut & Wulff, 2013; Saito & Shintani, 2016; Winke et al., 2013), raters with high familiarity in speakers' L1 and/or foreign accents in general tend to make lenient judgements. In the case of the current study, due to the raters' high familiarity with Japanese accented English, they may have well tolerated with learners' Katakana sounds (i.e., English loanwords that trigger word stress errors and syllable insertions, see Ohata, 2004; for a comprehensive summary of problematic L2 English pronunciation features for Japanese speakers, see Saito, 2014) that otherwise hamper their comprehensibility judgement.

Taken together, the current study has provided learning/teaching priorities for improving L2 pronunciation proficiency. First, segmental accuracy (especially segmentals with high communicative value) deserves to be the first requirement to reach Mid-level from Low-level proficiency. Then this could be followed by suprasegmental accuracy (especially word stress and syllable production). Suprasegmental-based training may help Mid-level L2 speakers reach High-level in L2 pronunciation.

Interestingly, our tentative suggestions here are in line with the previous literature. Suprasegmental errors tend to make a negative impact on native speakers' overall proficiency judgements of L2 learners from mixed L1 groups (e.g., Derwing & Munro, 1997) as well as Japanese learners of English (Ohata, 2004 for unnecessary vowel insertions after consonants; Saito, 2014 for conflations of English loanwords with correct English words). It has also been shown the intelligible pronunciation of certain segmental features (e.g., English /r/ and /l/) seems to be strongly tied to, in particular, Japanese learners' overall oral proficiency. For instance, Saito et al. (2016) demonstrated that Japanese learners' segmental errors significantly affected native rater's judgements of both comprehensibility and accentedness compared to other linguistic features. Furthermore, Riney et al. (2000) found that Japanese learners' substitution of English /ɹ/ and /l/ caused negative accentedness rating.

However, it is probably more crucial to point out that our study provided clear evidence that the raters' perceptions were influenced by specific segmental (high vs. low FL errors) and suprasegmental (schwa vowel insertions, word stress absence) information when making decisions to provide different pronunciation ratings (Low → Mid → High). The results here suggest that it is vital for L2 learners and teachers to set clear learning priorities (segmental → suprasegmental) to achieve specific proficiency levels L2 learners aim.

Finally, we would like to turn our discussion to the clear discrepancy between previous studies and the current study regarding the phonological correlates of L2 proficiency. That is, whereas intonation accuracy influenced L2 speech judgments in the previous studies (e.g., Pickering, 2001), the intonation factor in the current study did not significantly relate to any part of the raters' pronunciation proficiency judgements. One potential reason for this could be related to the methodological choice in the current study. In the current study, we made dichotomous judgements in obligatory contexts (correct or incorrect patterns) according to two different types of errors (misplacement, absence) while other studies such as Pickering (2001) combined instrumental and auditory analysis (cf. Crowther et al., 2015a, 2015b for trained coders' subjective judgements on a 1000-point scale [*1 = unnatural*, *1000 = natural*]). Interestingly, it has remained open to discussion which types of intonation errors could be relatively detrimental (or irrelevant) to overall intonation proficiency (Trofimovich & Isaacs, 2012), and which outcome measures could be appropriate for such perceived intonation accuracy (Trofimovich & Baker, 2006).

Another reason for the lack of any significant associations between intonation errors and overall pronunciation ratings in the current study could be due to the raters' perceptual salience of segmental, syllable, and word stress errors. Previous studies of Japanese learners of English have indicated that they are prone to make perceptible errors in terms of segmental (Riney et al., 2001) and word stress (Ohata, 2004). In fact, based on our casual observations during the tailored, in-house training sessions for the raters, all of the raters seemed to pay particular attention to segmental errors without any difficulty. Considering the research evidence that speakers' L1 could affect listeners' judgements (Darcy et al., 2012), it can be reasonable to assume that the raters' attention towards the absence or incorrect pitch changes might have been sacrificed due to the relative salience of errors of individual sounds and stress patterns. In this regard, more research is called for regarding the interaction between listeners and speakers' L1 backgrounds.

To close, it is crucial to acknowledge a range of methodological limitations and provide promising future directions:

- In the current study, we elaborated on our in-house rating scale of L2 oral proficiency based on the IELTS Pronunciation Scale, and our own training procedure. This was done so, because we followed previous L2 oral proficiency literature (e.g., Crossley et al., 2015 for ACTFL Guidelines). Thus, it would be intriguing if future studies further investigate the phonological correlates of L2 pronunciation proficiency using a range of other rubrics used in the standardized tests (e.g., TOEFL).

- The discussion presented throughout this paper was limited to the particular instance of L2 speech learning, i.e., Japanese speakers of English. One obvious research direction concerns the generalizability of the findings to other L1-L2 pairings. In fact, there is emerging empirical evidence that the relative weights of segmental and suprasegmental information in overall L2 pronunciation proficiency may vary according to different L1 backgrounds (e.g., Crowther et al., 2015b; but see Saito & Akiyama, 2018 for the phonological correlates of L2 Japanese comprehensibility).
- Although the sample size was based on the previous literature (e.g., Trofimovich & Issacs, 2012), the post-hoc power analysis (presented in the Results section; Table 5) revealed the varied degree of statistical power (.252-.999). This indicates that the generalizability of the findings needs to be re-examined with larger sample size (e.g., Saito et al., 2016 for 120 L2 speakers).

References

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language learning, 59*. 249–306.

Boersma, P. & Weenink, D. (2012). Praat: doing phonetics by computer [Computer software]. Retrieved from http://www.fon.hum.uva.nl/praat/ (accessed 20 July 2016).

British Council. (2018). Our levels and the CEFR. Retrieved from https://www.britishcouncil.pt/e n/our-levels-and-cefr (accessed 21 June 2018).

Brown, A. (1988). Functional load and the teaching of pronunciation. *TESOL Quarterly, 22*. 593–606.

Brown, A. (2006). An examination of the rating process in the revised IELTS Speaking Test. *IELTS Research Reports, 6*. 1–30.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*. 201–219.

Catford, J. C. (1987). Phonetics and the teaching of pronunciation: A systemic description of the teaching of English phonology. In Morley Joan (ed.), *Current perspectives on pronunciation: Practices anchored in theory,* 83–100. Washington DC: TESOL.

Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*, 570-590.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011a). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly*, *45*, 182-193.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011b). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing, 28*, 561-580.

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015a). Does a speaking task affect second language comprehensibility?. *The Modern Language Journal*, *99*(1), 80-95.

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2015b). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL quarterly*, *49*, 814-837.

Darcy, I., Dekydtspotter, L., Sprouse, R. A., Glover, J., Kaden, C., McGuire, M., & Scott, J. H. (2012) Direct mapping of acoustics to phonology: On the lexical encoding of front rounded vowels in L1 English–L2 French acquisition. *Second Language Research, 28*. 5–40.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*, 5-34.

Derwing, T. M. & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility. *Studies in Second Language Acquisition, 19*. 1–16.

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

DeVelle, S. (2008). The 2007 revised IELTS pronunciation scale. *Cambridge ESOL Research Notes, 34*. 36–39.

Galaczi, E., Post, B., Li, A., Barker, F., & Schmidt, E. (2016). Assessing second language pronunciation: Distinguishing features of rhythm in learner speech at different proficiency levels. In T. Isaacs, & P, Trofimovich (eds.), *Second language pronunciation assessment: Interdisciplinary perspectives*, 157–182. Bristol, UK: Multilingual Matters.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*. 201–223.

IELTS. (2017). *Speaking: Band descriptors*. Retrieved from https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx?la=en (accessed 16 December 2017).

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

Isaacs, T., Trofimovich, P., Yu, G. &Chereau, BM. (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS Pronunciation scale. *IELTS Research Reports Online Series 48*. 1–48.

Iwashita, N., Brown, A., McNamara, T. & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct?. *Applied Linguistics, 29*. 24–49

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics, 23*. 83–103.

Kang, O. (2008). Ratings of L2 oral performance in English: Relative impact of rater characteristics and acoustic measures of accentedness. *SPAAN FELLOW, 6*. 181–205.

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly, 9*. 249–269.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review*, *64*(3), 459-489.

Kraut, R. & Wulff, S. (2013). Foreign-accented speech perception ratings: a multifactorial case study. *Journal of Multilingual and Multicultural Development, 34*. 249–263.

Kyle, K. & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*. 757–786.

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

Levis, J. M. (2006). Pronunciation and the assessment of spoken language. In R. Hughes (ed.), *Spoken English, TESOL and applied linguistics*, 245–270. New York: Palgrave Macmillan.

Lin, Y. H. (2001). Syllable simplification strategies: A stylistic perspective. *Language Learning, 51*. 681–718.

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34*. 520–531.

Ohata, K. (2004). Phonological Differences between Japanese and English: Several Potentially Problematic. *Asian EFL Journal, 6*. http://www.asian–efl–journal.com/december_2004_KO.php (accessed 13 August 2018).

Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly, 35*. 233–255.

Piske, T., MacKay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29*. 191–215.

Riney, T. J., Takada, M. & Ota, M. (2000). Segmentals and global foreign accent: The Japanese flap in EFL. *TESOL Quarterly, 34*. 711–737.

Saito, K. (2014). Experienced teachers' perspectives on priorities for improved intelligible pronunciation: The case of Japanese learners of English. *International Journal of Applied Linguistics*, *24*(2), 250-277.

Saito, K. (2021). What characterizes comprehensible and native-like pronunciation among English-as-a-Second-Language speakers? Meta-analyses of phonological, rater, and instructional factors. *TESOL Quarterly, 55, 866-900.*

Saito, K., & Akiyama, Y. (2017). Linguistic correlates of comprehensibility in second language Japanese speech. *Journal of Second Language Pronunciation*, *3*(2), 199-217.

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

Saito, K., & Shintani, N. (2016). Do native speakers of North American and Singapore English differentially perceive comprehensibility in second language speech?. *TESOL Quarterly*, *50*(2), 421-446.

Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (2019). How do second language listeners perceive the comprehensibility of foreign-accented speech?: Roles of first language profiles, second language proficiency, age, experience, familiarity, and metacognition. *Studies in Second Language Acquisition*, *41*(5), 1133-1149.

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, *37*, 217-240.

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*(4), 439-462.

Saito, K., Trofimovich, P., Isaacs, T., & Webb, S. (2017). Re-examining phonological and lexical correlates of second language comprehensibility: The role of rater experience. *Second language pronunciation assessment: Interdisciplinary perspectives*, 141-156.

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*(1), 143-167.

Suzukida, Y., & Saito, K. (2019). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the functional load principle. *Language Teaching Research*, 1362168819858246.

Trofimovich, P & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28*. 1–30.

Trofimovich, P. & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition, 15*. 905–916.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*. 231–252.

Supporting Information: Functional load ranking in Brown (1988)

	vowels		consonants
10	/e, æ/	10	/p, b/
	/ æ, ʌ/		/p, f/
	/ æ, ɒ/		/m, n/
	/ ʌ, ɒ/		/n, 1/
	/ ɔː, əʊ/		/1, r/
9	/e, ɪ/	9	/f, h/
	/e, eɪ/		/t, d/
	/ɑː, aɪ/		/k, g/
	/ɜː, əʊ/		
		8	/w, v/
8	/iː, ɪ/		/s, z/
7	—	7	/b, v/
			/f, v/
6	/ ɔː, ɜː/		/ð, z/
	/ɒ, əʊ/		/s, ʃ/
5	/ɑː, ʌ/	6	/v, ð/
	/ɔː, ɒ/		/s, ʒ/
	/ɜː, ʌ/		
		5	/θ, ð/
4	/e, eə/		/θ, s/
	/æ, ɑː/		/ð, d/
	/ɑː, ɒ/		/z, dʒ/
	/ɔː, ʊ/		/n, ŋ/
	/ɜː, e/		
		4	/θ, t/
3	/iː, ɪə/		
	/aː, aʊ/	3	/tʃ, d3/
	/uː, ʊ/		
		2	/tʃ, ʃ/
2	/ɪə, eə/		/ʃ, ʒ/
			/j, ʒ/
1	/ɔː, ɔɪ/		
	/uː, ʊə/	1	/f, θ/
			/dʒ, j/