Pronunciation Assessment Criteria and Intelligibility

Okim Kang & Kevin Hirschi

Various aspects of second language (L2) speakers' pronunciation can be considered in the oral assessment of speaker proficiency. Over time, both segmentals and suprasegmentals have been examined for their roles in judgments of accented speech. Descriptors in the rating criteria often include speaker's intelligibility (i.e., the actual understanding of the utterance) or comprehensibility (i.e., easy of understanding) (Derwing & Munro, 2005). This paper discusses the current issues and rating criteria in L2 pronunciation assessment, and describes the prominent characteristics of L2 intelligibility. It also offers recommendations to inform assessment practices and curriculum development in L2 classrooms in the context of Global Englishes.

Issues in Pronunciation Assessment

L2 pronunciation has seen an emerging growth in the fields of Applied Linguistics and TESOL. It is now a revitalized field of inquiry with its own important implications and concerns. Part of this renaissance can be attributed to a shift in focus from perceptions of accentedness to broader aspects of performance, primarily intelligibility and comprehensibility (Kang & Ginther, 2017). When it comes to language assessment, assessing pronunciation is a critical issue because people tend to immediately judge native/non-native speaker status on the basis of pronunciation (Luoma, 2004). In fact, pronunciation is an essential aspect of the assessment of oral skills because it helps us understand the fundamentals in the process of the construction of spoken discourse in L2 performance.

For the last few decades, three speech constructs (i.e., intelligibility, comprehensibility, and accentedness) have dominated the L2 pronunciation literature. As seen from its wide practice in the field of L2 speech, intelligibility often refers to the extent to which the speaker's intended utterance is actually understood by a listener, whereas comprehensibility pertains to the degree of difficulty the listener experiences in attempting to understand an utterance. While Munro and Derwing (1995) found the first two constructs to be highly intercorrelated, accentedness (meaning the extent to which an L2 learner's speech is perceived to differ from a particular standard) was found to be only moderately or weakly correlated with comprehensibility or with intelligibility.

Notions that accentedness, comprehensibility, and intelligibility are related but partially independent constructs have paved the way for pronunciation teaching and research, which leads to the end goal of comprehensibility or intelligibility. The methods of assessing pronunciation also tend to reflect such perspectives (Kang & Ginther, 2017). As accent is indeed associated with the speaker's own identity, it adds further complexity and importance to focusing on intelligibility and comprehensibility. Relatedly, the interest in these inter-related constructs have guided researchers to seek and discover how they are best operationalized and measured, how they affect listeners' ratings, and how they should be addressed in the actual classroom. It further leads to validity or reliability questions about the measures of these constructs and their pedagogical and social consequences.

The discussion of validity is further related to the aspect of World Englishes in the global context, where the reliance on prestigious inner circle norms of native English (i.e., British English or American English) has been challenged. This issue of ecological validity examines the sociolinguistic realities of diverse language learners' actual use (Elder & Harding, 2008) and fairness and justice in language testing (Kunnan, 2014). Given that defining a standard norm is problematic in the era of globalization, where new Englishes are emerging (Taylor, 2006), the assessment of L2 pronunciation now seem to tackle more questions than ever before.

Kang et al.'s (2018) research demonstrated that as long as speakers were highly intelligible, test takers did not have any significant differences in their simulated-TOEFL test scores. A similar result was found in her recent study with a simulated-Duolingo English listening test (Kang et al., 2022) in which listening test scores from 160 test takers (Indian, Spanish, Korean, and Chinese) did not differ significantly even though there were some shared-L1 boosting effects. Indian listeners, for example, performed better when they listened to their own Indian accent. In fact, the background characteristics that raters bring to assessment tasks have been found to influence their evaluations of accentedness, comprehensibility, and intelligibility (Kang, 2012) and these factors explain about 18-21% of variance in their listening evaluations (Kang & Rubin, 2009). Therefore, in L2 speech assessment, listener characteristics and contextual influences should be carefully examined to promote pedagogical and theoretical accounts.

Another issue in pronunciation assessment is related to rating criteria and scale representation (validity). L2 pronunciation involves many features within the speech stream including vowels and consonants, pause, stress and rhythm, and Intonation. However, thus far, the rating features of pronunciation have not been clearly identified. For example, in iBT TOEFL rubrics, descriptors for Level 4 in the dimension of Delivery include "*Speech is clear. It may include min*or lapses, or *minor* *difficulties with pronunciation or intonation patterns, which do not affect overall intelligibility***".** But it is not clear what 'overall intelligibility' entails. In the case of IELTS Speaking, pronunciation is one of their four rating criteria (i.e., fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation). Their band score ranges from 1 to 9, and each band descriptor is somewhat difficult to follow. For instance, descriptors for Band 9 state *"uses a full range of pronunciation features with precision and subtlety; sustains flexible use of features throughout; is effortless to understand*". We can see comprehensibility (effortless to understand) being a part of the measurement construct, but it is rather ambiguous to identify what 'flexible use of features' actually means. In addition, descriptors for Bands 7, 5, 3 are not specified explicitly, as their descriptors overlap with adjacent band levels (i.e., Band 7: *shows all the positive features of Band 6 and some, but not all, of the positive features of Band 8*). Indeed, the relationship between pronunciation features and level-specific criteria is still very difficult to determine.

Finally, pronunciation is often assessed more holistically by mingling all constructs together, which may be methodologically inappropriate at times. Segmental and suprasegmental features contribute independently to different pronunciation constructs of L2 speech (Kang, 2010; Trofimovich & Isaacs, 2012), but in rubrics, they often appear as general pronunciation features. Detailed accounts are needed on how each of the pronunciation constructs should be evaluated, and on what construct can be used in the classroom context both for learners and teachers. Furthermore, with advances in speech science, computer-assisted instruments have aided in examining some elements of the pronunciation properties. The knowledge of these instrumentally analysed pronunciation features has potential to advance our understanding of speech production and inform rubric development and rater training in oral proficiency testing. Although it is exciting to see how the improvement of the Automatic Speech Recognition (ASR) approach can lead the field of pronunciation assessment in the future, we have to be mindful about the complexity of L2 speech characteristics and their relationships with speech evaluation.

Pronunciation Rating Criteria and Descriptors

Various speaking features have been shown to predict L2 speaking proficiency and/or cue accentedness. Earlier L2 research that focused on segmental features (i.e., consonant and vowel production) tended to measure the deviation from a native speaker norm (Flege & Port, 1981). More recent studies have highlighted the importance of suprasegmental features particularly in how much prosodic features could contribute to a listener's perception of a speaker's intelligibility (Kang et al., 2020) or comprehensibility (Kang, 2010). Still, identifying the linguistic components most conducive to non-native production of intelligible speech remains a challenge in the field of L2 speech.

Often, tests describe pronunciation errors as the speaker's deviation from a norm or use general terms that may group multiple pronunciation phenomena. On the Cambridge Advanced test, for instance, a speaker with a score of 3 is described as someone whose "individual sounds are articulated clearly" (Cambridge English Handbook for Teachers, 2015, p. 86). Other tests place pronunciation features within a more general descriptor. The TOEFL iBT speaking scale includes all of its pronunciation features in the "delivery" construct. This umbrella terms consists of the "flow" and clarity of the speech. The same can be said for the IELTS, which describes a proficient speaker as one who "uses a full range of pronunciation features with precision and subtlety" (IELTS Guide for Teachers, 2015, p. 18). Thus, teasing theses pronunciation features from others for evaluative purposes might prove to be difficult.

An empirical study (Kang & Moran, 2014) suggests that learners' segmental deviations, especially related to high functional load errors, affect their oral proficiency significantly. Kang and Moran analysed 120 test takers' spoken responses from the Cambridge English Language Assessment (CELA) and demonstrated that highly proficiency (C2) learners still make segmental errors, but their high functional errors dropped drastically as their proficiency increased. Functional loads (FL) rank segmental contrasts according to their importance in English pronunciation (Brown, 1991; Catford, 1987). The high FL errors had large effects on perceptual scales (e.g., p/b, l/r, or bit vs. bat, beet vs. bit), while the low FL errors had only a minimal impact (e.g., θ/ð, v/z or pooh/poor). This pattern also emerges in the assessment context.

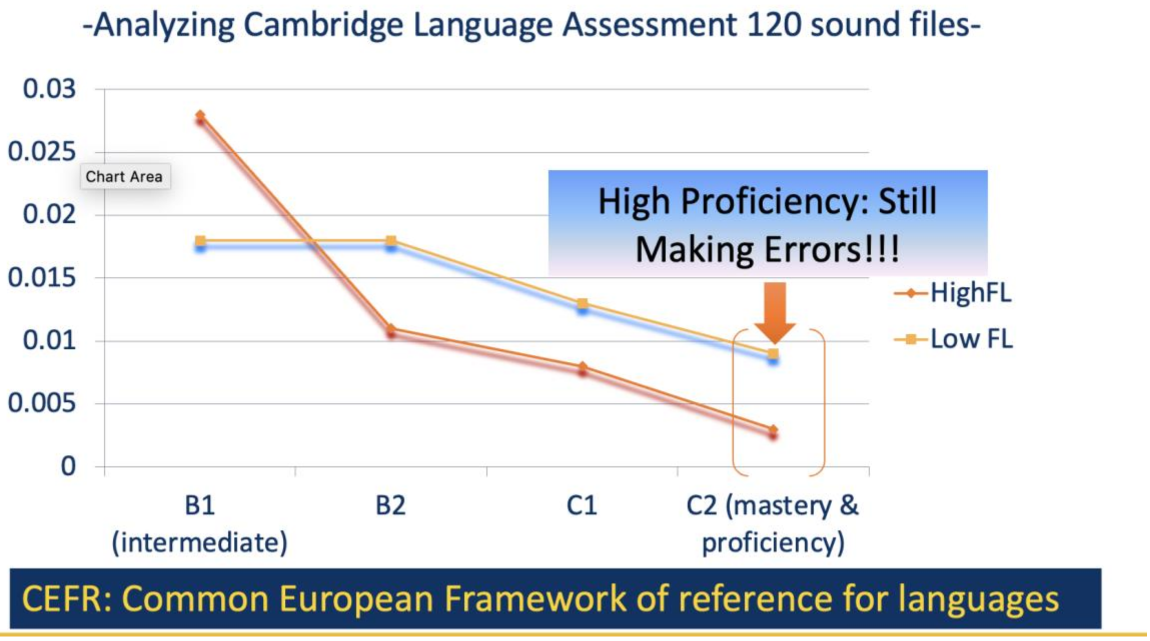

Figure 1. Functional Load Segmental Errors (a revised version, cited in Kang & Hirschi, 2022; Originally Kang & Moran, 2014)

Kang (2013) also proposes a hierarchical priority in pronunciation features particularly in L2 oral assessment after analysing the 120 CELA speech files for a comprehensive list of pronunciation features. In Figure 2, stress/pitch features and fluency are strong contributors to the common European framework of reference (CEFR) proficiency (over 58%) whereas segmental errors and intonation somewhat weakly contributed to proficiency judgments. Indeed, there appears to be a clear hierarchical structure; i.e., stress and pitch were first ranked, followed by fluency measures, segmental errors, and tone choices at the end. Accordingly, English language teachers may need to prioritize pronunciation features in classroom instruction for the promotion of intelligibility or for the preparation of high-stakes speaking test. This knowledge can be further applied to develop scoring criteria for L2 oral assessment.

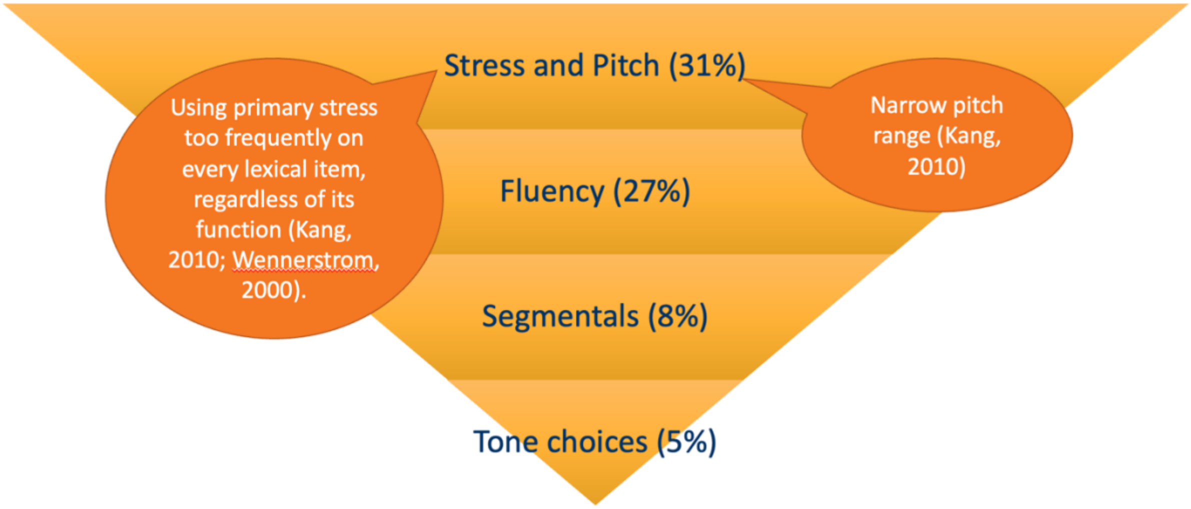
Stress and Pitch (31%)
Using primary stress too frequently on every lexical item, regardless of its function (Kang, 2010; Wennerstrom, 2000).
Narrow pitch range (Kang, 2010)
Fluency (27%)
Segmentals (8%)
Tone choices (5%)

Figure 2. Hierarchical Priority in Pronunciation Features (a revised version, cited in Kang & Hirschi, 2022; originally in Kang, 2013)

Overall, certain errors may make the speech accented, but not make the speech unintelligible. Of course, both segmentals and suprasegmentals are important, but fluency and stress seem to be particularly more critical in some assessment cases. Also, research-informed descriptors of specific features can be used for enhancing scoring. However, note that it might be very difficult for a human rater to account for all the variables and features when evaluating a short oral production. In order to limit the cognitive load and make an informed decision about constructs, the constructs assessed should be task- and context- specific. Detailed description of selected task-related features might need to be considered rather than supply a large selection with general and vague definitions. It is also essential to think about consistency in the use and definition of pronunciation constructs among different speaking descriptors (e.g., IELTS 9 bands). Finally, although ASRs have proven their ability in including a significant number of features, their less-than-perfect accuracy has still not eliminated the need for human raters. Therefore, when choosing the relevant constructs for any pronunciation assessment, it is still imperative to choose the right features that can fit well with rater ability and task variability.

Characteristics of L2 Intelligible Speech

What constitutes intelligibility remains under-defined, but existing literature provides a useful starting point for considering which features contribute to intelligible L2 speech. In the realm of segmentals (i.e., vowels and consonants), scholars (Brown, 1991; Catford, 1987; Kang & Moran, 2014; Munro & Derwing, 2006) argue for the importance of functional loads. Likewise, Jenkins (2003) based her description of what constitutes a Lingua Franca Core for English as an International Language on empirical research, also having found that some phonemes are relatively less important to successful communication than are others.

At the same time, the eminence of non-native English speakers' prosody (suprasegmentals) has been underscored in L2 pronunciation assessment (e.g., Kang et al., 2010). For example, incorrect nuclear stress (i.e., emphasizing the "wrong" words) (Field, 2005) or missed placed and missing prominence (Hahn, 2004) can affect listeners' judgements of L2 speech. Similarly, the use of stress to emphasize every word, regardless of its function or semantic importance, causes difficulty for listeners (Wennerstrom, 2000). Poor intonational structure (e.g., narrow pitch range in Kang, 2010) and a disturbance in prosodic composition can also considerably affect NS listeners' ability to understand the intended message (Pickering, 2001). In particular, how a speaker applies rising, falling, or level pitch on the focused word of a tone unit can affect both perceived information structure and social cues in L2 discourse. Insufficient differentiation in syllable duration between stressed and unstressed syllables, thereby creating an unnational speech rhythm (Setter, 2006), can affect intelligibility as well.

In addition, the relationship between fluency and L2 pronunciation assessment has been well documented as well. Tavakoli and Skehan (2005) suggest that fluency can be characterized along three separate dimensions (1) speed and density per time unit (speaking rate); (2) breakdown fluency (number and length of pauses); and (3) repair fluency. Research indicates that speech rate (Kormos & Denes, 2004), breakdown fluency (Derwing et al., 2004), and repair fluency (Iwashita et al., 2008) are all associated with assessments of speakers' oral proficiency. Thomson (2015) reports a strong correlation between oral fluency and comprehensibility ratings, although he highlights the fact that in rare cases, a speaker can be fluent in a language, yet not very easy to understand. Kang et al. (2020) also point out that when it comes to break-down fluency (pauses), it is not always about the number of pauses (quantity), but the location of pauses (quality) matters more sometimes. While much research reveals the importance of such features in L2 intelligibility, it is still uncertain what really characterizes intelligible L2 speech.

Intelligibility in Global Contexts

Although the field has stressed intelligibility or comprehensibility over accentedness, pronunciation assessment still tends to juggle these two contradictory principles: the nativeness principle vs. the intelligibility principle (Levis, 2005). In nativeness principle, L2 speech is judged against a 'native-like' pronunciation standard.

However, English is a lingua franca in this era of globalization, and we all have accents. Currently, the number of English users who speak varieties historically considered to be standard (e.g., British, American, etc.) are now in the minority (Crystal, 2003). In line of this movement, scholars argue that proficiency tests should incorporate other varieties of English in addition to the traditionally dominant British and American models (Kang et al., 2019). Global Englishes are becoming new norms and may gradually be used as assessment and pedagogical models (Taylor, 2006).

Admittedly, the intelligibility principle is multifaceted and complex. First, being intelligible is not synonymous with being accent free. Second, intelligibility does not reside solely in the speaker or the listener, but rather in the interaction between the two (Smith & Nelson, 1985), or amongst different audiences (Brown, 1991). Correspondingly, mutual indelibility is key to the successful communication in global contexts. Nevertheless, many learners still seem to prefer to model native speakers from inner-circle countries such as the UK or from North America (Kang, 2015). The majority of learners consider speaking with perfect native pronunciation to be a desirable goal; hence, they are dissatisfied with their current curriculum of learning pronunciation due to misunderstanding of various models and accents made available to them (Kang, 2014). Not surprisingly, a very recent study has confirmed that teachers are not necessarily comfortable with the idea of incorporating accent varieties as pedagogical models in their classroom (Dalman et al., 2022).

Given this disparity between research findings and classroom practices, intentional and real-world efforts from three parties (learners, teachers, and researchers) would be desirable. Such efforts can include explicit training about the practicality of global intelligibility for both teachers and learners. First, learners need to be educated about what features can make their speech intelligible and why intelligible speech is more achievable than native-like accent. Next, teachers should be aware of learners' realistic goals and try to assess their pronunciation development from intelligibility-based perspectives. Finally, researchers should make evidence-based research findings public and accessible to both teachers and learners so that their assessment practice and curriculum development can be informed. Such ongoing triangulation works among these three parties can essentially ensure leaners' successful communication in global contexts.

References

Brown, A., (1991). Functional load and the teaching of pronunciation. In A. Brown (Ed.), *Teaching English Pronunciation: A Book of Readings*. Routledge, London, pp. 221–224.

Catford, J.C., (1987). Phonetics and the teaching of pronunciation: A systemic description of English phonology. In J. Morley (Ed.), *Current Perspectives on Pronunciation: Practices Anchored in Theory*. TESOL, Washington, DC, pp. 87–100.

Cambridge English Handbook for Teachers, 2015, Cambridge English Language Assessment.

Crystal, D. (2003). English as a global language (2nd ed.). Cambridge: Cambridge University.

Dalman, M., Yaw, K., & Kang, O. (2022). Global perspectives on English teachers' Attitudes and perceptions of World Englishes in TESOL classrooms. *TESOL Quarterly.*

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, *39*(3), 379-397.

Derwing, T. M., Rossiter, M. J., Munro, M. J., Thompson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54*, 655-679.
Derwing, T. M., Thomson, R. I., & Munro, M. J. (2006). English pronunciation and fluency development in Mandarin and Slavic speakers. *System*, *34*(2), 183-193.

Elder, C. & Harding, L. (2008). Language testing and English as an international language: Constraints and contributions. *Australian Review of Applied Linguistics 31* (3), 34.1–34.11.
Field, J. (2005). Intelligiblity and the listener: the role of lexical stress. *TESOL Quarterly*, *39*, 399-423.

Flege, J. E., & Port, R. (1981). Cross-language phonetic interference: Arabic to English. *Language and speech*, *24*(2), 125-146.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*(2), 201–223.

IELTS Teacher Guide, 2015, Cambridge ESOL.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct?. *Applied linguistics*, *29*(1), 24-49.

Jenkins, J. (2003) Intelligibility in Lingua Franca discourse. In J. Burton and C. Clennell (eds) Interaction and Language Learning (pp. 83-97). Alexandria, VA: TESOL.

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, *38*(2), 301-315.

Kang, O. (2012). Impact of rater characteristics on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly,* 9, 249-269.

Kang, O. (2013). Relative impact of pronunciation features on non-native speakers' oral proficiency. In J. Levis & K. LeVelle (Eds.), *Proceedings of the Pronunciation in Second Language Learning and Teaching*. Iowa State University.

Kang, O. (2015). Students' perceptions of pronunciation instruction in the three circles of World Englishes. *TESOL Journal, 6* (1), 59–80. DOI: 10.1002/tesj.146

Kang, O., & Ginther, A. (Eds.). (2017). *Assessment in second language pronunciation*. New York: Routledge.

Kang, O., & Hirschi, K. (2022). Pronunciation assessment criteria and intelligibility. Workshop. IATEFL, PronSIG IATEFL, Online.

Kang, O., & Moran, M. (2014). Pronunciation features in non-native speakers' oral performances. *TESOL Quarterly, 48*,173-184.

Kang, O., & Rubin, D. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology, 28,* 441-456.

Kang, O., Rubin, D., Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal, 94*, 554-566.

Kang, O., Thomson, R., & Moran, M. (2019). The effects of international accents and shared L1 on listening comprehension tests. *TESOL Quarterly.* 53, 56-81.

Kang, O., Thomson, R., & Moran, M. (2020). Which Features of Accent affect Understanding? Exploring the Intelligibility Threshold of Diverse Accent Varieties. *Applied Linguistics,* 41 (4), 453-480.

Kang,O., Yan, X., Kostromitina, M., Thomson, R., & Isaacs, T. (2022, under review). Fairness of using different English accents: The effect of shared L1s in listening tasks of the Duolingo English Test.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*, 145–164.

Kunnan, A. (2014). Fairness and justice in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1–17). Wiley.

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

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press: UK, Cambridge.

Munro, M. & Derwing, T. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38*, 289-306.

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

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

Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *English Language Teaching Journal, 60*(1), 51–60.

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

Setter, J. (2006). Speech rhythm in world Englishes: The case of Hong Kong. *TESOL Quarterly, 40,* 763-787.

Smith, L., & Nelson, C. (1985). International intelligibility of English: Directions and resources. *World Englishes, 4,* 333-342.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and Task Performance in a Second Language* (pp. 239-276). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Thomson, R. I. (2015). Fluency. In Reed, M. & Levis, J. (Eds.). *The Handbook of Pronunciation.* (pp. 209-226). Hoboken, New Jersey: Wiley.

Wennerstrom, A. (2000). The role of intonation in second language fluency. In H. Riggenbach, (Ed.) *Perspectives on fluency*. (pp. 102-127.) Ann Arbor, MI: University of Michigan.

Okim Kang is a Professor in the Applied Linguistics Program and Director of Applied Linguistics Speech Lab at Northern Arizona University, Flagstaff, AZ, USA. Her research interests are L2 pronunciation and intelligibility, speech production and perception, L2 oral assessment and testing, automated scoring and speech recognition, World Englishes, and language attitude. Email: Okim.Kang@nau.edu.

Kevin Hirschi is a doctoral candidate in Applied Linguistics at Northern Arizona University in Flagstaff, Arizona. His research is centered on L2 pronunciation perception and development and the uses of technology for pronunciation training and language analysis. Email: kevinhirschi@nau.edu

(Editor's note. For full details of the APA referencing system go to http://library.nmu.edu/guides/userguides/style_apa.htm)