Running Head: L2 pronunciation assessment

Shifting sands in second language pronunciation assessment research and practice

Talia Isaacs, University College London

Article title: Shifting sands in second language pronunciation assessment research and practice

Corresponding author: Talia Isaacs
UCL Centre for Applied Linguistics
UCL Institute of Education, University College London
20 Bedford Way, London
United Kingdom WC1H 0AL
+44 (0) 207 612 6348
talia.isaacs@ucl.ac.uk

LAQ special issue: Conceptualizing and operationalizing second language speaking assessment: Updating the construct for a new century

Special issue editors: Gad Lim & Evelina Galaczi

Citation: Isaacs, T. (accepted). Shifting sands in second language pronunciation assessment research and practice. *Language Assessment Quarterly.*

Abstract

This article brings to the fore trends in second language (L2) pronunciation research, teaching, and assessment by highlighting the ways in which pronunciation instructional priorities and assessment targets have shifted over time, social dimensions that, although presented in a different guise, appear to have remained static, and principles in need of clearer conceptualization. The reorientation of the pedagogical goal in pronunciation teaching from the traditional focus on accent reduction to the more suitable goal of intelligibility will feed into a discussion of major constructs subsumed under the umbrella term of "pronunciation." We discuss theoretical gaps, definitional quagmires, and challenges in operationalizing major constructs in assessment instruments, with an emphasis on research findings on which pronunciation features are most consequential for intelligibility and implications for instructional priorities and assessment targets. Considerations related to social judgments of pronunciation, accent familiarity effects, the growth lingua franca communication, and technological advances, including machine scoring of pronunciation, pervade the discussion, bridging past and present. Recommendations for advancing an ambitious research agenda are proposed to disassociate pronunciation assessment from the neglect of the past, secure its presence as an integral part of the L2 speaking construct, and propel it to the forefront of developments in assessment.

Shifting sands in second language pronunciation assessment research and practice

From a historical perspective, it can be argued that pronunciation, more than any other component within the construct of second language (L2) speaking ability, has been subject to the whims of the time and the fashions of the day. That is, pronunciation, once dubbed "the Cinderella of language teaching" to depict its potentially glamorous yet marginalized existence (Kelly, 1969, p. 87), experienced a fall from grace after being a focal point of L2 instruction, teacher training, and testing during its heyday. This is a prime example of a pendulum swing in L2 teaching methodologies and pedagogical practices that has affected content coverage for learners in L2 classrooms, with likely detrimental effects when pronunciation, which encompasses both segmental (individual vowel/consonant sounds) and suprasegmental aspects of speech (e.g., rhythm, stress, intonation), poses a genuine impediment to oral communication (Derwing & Munro, 2015). Naturally, the aspects of L2 pronunciation that are accorded pedagogical value in the minds of teachers, researchers, and language testers have shifted over time (Levis, 2005). However, an aerial view of developments over the past century reveals the polarized nature of researchers' and educational practitioners' beliefs regarding the importance of pronunciation in L2 aural/oral instruction and assessment.

Pronunciation has experienced a resurgence of research interest and now has a higher profile within applied linguistics research than any other time over the past half century. There are also signs of its gradual reintegration into L2 classrooms despite limited teacher training (Baker & Burri, 2016) and of growing interest in language assessment circles after decades of being sidelined, including in relation to human and machine scoring of speech and moving beyond the NS standard (Isaacs, 2018). However, the role of pronunciation within the L2 speaking construct (or in notions of L2 proficiency more generally) is currently underconceptualized. To elaborate, there is no unitary construct of L2 speaking ability, as speaking ability is operationalized in different ways depending on the mode of assessment,

speech elicitation task, and scoring system (Fulcher, 2015).[1] However, pronunciation has received scant treatment in books on assessing speaking (e.g., Luoma, 2004). In addition, it was singled out as the only linguistic component relevant to the L2 speaking construct that the author of a research timeline on assessing speaking was "not able to cover" without any clear explanation or justification as to why (Fulcher, 2015, p. 201). Due to its role in mediating effective oral communication, particularly for L2 learners who struggle to make themselves understood, pronunciation can simply no longer be ignored in instruction and assessment (Harding, 2013). Its role within the construct of L2 speaking ability (however operationalized) and in relation to L2 proficiency and communicative language ability more generally would benefit from greater empirical exploration to move beyond its current undertheorized status (e.g., Galaczi, Post, Li, & Graham, 2012). This is essential to consolidating a place for pronunciation within mainstream L2 speaking assessment research and practice into the future.

The goal of this state-of-the-art article is to overview trends in L2 pronunciation research, teaching, and assessment within the broader context of developments in L2 speaking assessment as a springboard for advancing an ambitious research agenda that draws on different disciplinary domains (e.g., SLA, sociolinguistics, psycholinguistics, phonetics, speech processing) to drive the field forward. To this end, the article will first review the ways in which pronunciation instructional priorities and assessment targets have shifted over time to develop a historical consciousness and demonstrate aspects that have evolved, remained static, been rebranded, or are in need of clearer conceptualization. The social nature of judgments of accented speech and reorientation of the pedagogical aim in pronunciation teaching from the traditional goal of eradicating first language (L1) traces in target language productions (accent reduction) to the more suitable goal of intelligibility will feed into a discussion of major constructs subsumed under the umbrella term of "pronunciation" or that are often cited in L2 pronunciation research. Emphasis will be placed on theoretical gaps, definitional quagmires, and

challenges in adequately operationalizing the focal construct in assessment instruments for operational testing purposes and on implementation challenges in L2 classrooms. After discussing major trends in assessment-oriented L2 pronunciation research, the paper will set out a set of desirable future directions in light of technological advances, the rise in lingua franca communication due to transnational mobility, and the need to examine pronunciation performance on more interactional task types than have traditionally been researched in both psycholinguistically-oriented research, and phonetics experiments. The article will predominantly focus on English as a target language, in part because work on English dominates this research area. However, the core principles and recommendations apply to the learning, instruction, and assessment of other L2s.

The term "assessment" in this article is broadly interpreted to denote any information gathering activity that is used to make conclusions about an individual's language ability (Bachman, 2004) or that may be used to extrapolate other (nonlinguistic) characteristics of that person. This definition encompasses both instrumental measurements of speech using technology, and listeners' evaluative reactions to speech, whether in formal examination contexts, or informal interactional settings. Therefore, content coverage in this article includes not only the role of pronunciation in language tests, which are just one type of assessment, but also the phenomena of humans or machines arriving at (potentially spurious) conclusions about the linguistic or nonlinguistic characteristics of an L2 speaker based on their articulatory output (Lindemann, 2017; Solewicz & Koppel, 2006). Further, assessment is increasingly viewed as integral to teaching, learning, and achieving curricular goals (Turner & Purpura, 2016). In light of this broad view of assessment, insights from SLA, pronunciation teaching, speech sciences, sociolinguistics, and psycholinguistics are highly relevant to understanding the different facets of assessing pronunciation—an inherently interdisciplinary field (Isaacs & Trofimovich, 2017a). Generating conversations across these disciplines is essential for establishing the existing evidence base,

moving beyond silos, and truly advancing the field of pronunciation assessment. The next section will introduce the social angle of pronunciation assessment, underscoring the pervasiveness of making formal or informal evaluative judgments about the way someone sounds in different societies throughout history.

Accented speech, social judgments, and identity testing

Pronunciation assessment, whether formally or informally conducted, is arguably one of the most ubiquitous forms of human language assessment, stemming back to biblical times. As described in the Book of Judges, a single-item phoneme test involving oral production of the word "shibboleth" was used by the Gileadites to distinguish members of their own tribe from the warring Ephraimites. Pronunciation of an /s/ sound rather than an /ʃ/ sound word-initially was interpreted as signaling enemy status, resulting in 42,000 people being instantly killed in biblical accounts (Spolsky, 1995).

Although an extreme example, the biblical shibboleth test is underpinned by the notion that the sound properties of an individual's speech give clues about his/her community membership or geographic origin (Moyer, 2013). In fact, shibboleth (identity) tests are endemic in situations of inter-group conflict as a means of establishing insider and outsider status. A modern incarnation of the biblical shibboleth test is the Language Analysis for the Determination of the regional or social Origin of asylum seekers (LADO), in which decisions about the legitimacy of an asylum-seeker's claims are made based on analyses of his/her speech productions (McNamara, 2012). Linguistic analyses, often including an accent classification component, tend to be undertaken by government officials lacking linguistics qualifications or training, a consequence of which can be poor transcription quality of the speech (i.e., no adherence to phonetic transcription conventions), sometimes derived from poor quality recordings, underscoring the lack of scientific rigor in the analyses undertaken (Eades, 2005). This raises concerns about test validity and consequential decision-making based on flawed evidence, bringing to the fore

issues of social justice in such legal cases. In fact, listeners' perceptions of a speaker's identity could be influenced by their stereotyped expectations of the linguistic patterns that characterize a particular language variety, which could, in turn, be projected onto the speech sample regardless of the presence or absence of those features in actual productions (Kang & Rubin, 2009). Related to this, nonlinguistic social factors extraneous to the speech signal, such as attitudes toward the perceived ethnic identity of the speaker or politically-motivated considerations, could bias listeners' judgments or assumptions about the speaker (Moyer, 2013). In sum, making claims about an individual's social identity for legal reasons, particularly when conducted by nonlanguage experts, is highly problematic and could lead to unfair and discriminatory decision-making based on unsound evidence. As Fraser (2009) contends, even determinations of forensic linguists conducted in conjunction with evidence from computer-derived analyses of the speech are not error-proof, although more scientific and informed analyses should be invoked over a lay person's ad hoc reactions in legal cases.

The above discussion of shibboleth (identity) tests links to several discrete but related points. One is that accents are one of the most salient aspects of L2 speech. Listeners are highly sensitive to accented speech, to the extent that, in research settings, listeners with no prior linguistic training are able to distinguish native from nonnative speakers after listening to speech samples that are just 30 ms long (Flege, 1984), are played backwards (Munro, Derwing, & Burgess, 2010), or are in an unfamiliar language (Major, 2007). Foreign accents also tend to be persistent (fossilized) and perceptible to listeners, even in cases where native-like mastery is achieved in other linguistic domains, such as morphology, syntax, and lexis (Celce-Murcia, Brinton, Goodwin, with Griner, 2010). In a seminal article on age effects, Flege, Munro and MacKay (1995) detected a strong linear relationship between learner age of arrival in the country in which the target language was spoken, which was used as an index of age of L2 learning, and perceived foreign accent. Nevertheless, listeners were able to detect an

L2 accent in participants who had learned English well before what is traditionally considered to be the critical period (Scovel, 2000)—even as early as at 3.1 years of age in the case of one discerning listener.

Despite a hypersensitivity to accent, lay listeners tend to be relatively poor at correctly identifying the L1 background or ethnicity of speakers in recorded stimuli. For example, Lindemann (2003) found that American undergraduate students who heard read-aloud passages produced by native (Midwestern) English speakers and L1 Korean speakers correctly identified the L1 of the Korean speakers only 8% of the time, mistaking them for non-East Asians 41% of the time. More recently, Ballard and Winke (2017) provided native and nonnative undergraduate listeners at an American university with a fixed list of 13 L1 and L2 English accents in an accent identification task. Correct response rates were just 57% for the native listeners and 26% for the nonnative listeners, although their level of familiarity with various accents likely affected their accent identification performance (Huang, Alegre, & Eisenberg, 2016). In sum, listener sensitivity to the presence of a foreign accent does not appear to translate into accurate identification of that accent. The social dimension of this work is clearly revealed in a study by Hu and Lindemann (2009), who presented L1 Cantonese speakers of English in China with an audio recording produced by an American English speaker. Half were told that the speaker was American while the other half were told that the speaker was Cantonese. Respondents who were informed that they were listening to a Cantonese speaker superimposed linguistic properties stereotypically associated with Cantonese-accented English on the speech that were absent from the speech sample (e.g., unreleased or deleted word-final stops, such as the /g/ sound in /bɪg/). This suggests that listeners' associations of speech samples with particular stigmatized or idealized varieties and expectations of what they will hear can distort their perceptions, potentially threatening the validity of their informal or ad hoc observations made on the basis of the speech or formal evaluations in testing contexts. The next section of this paper moves beyond the discussion of shibboleth tests and ultimate

attainment to discuss the changing role of pronunciation in classroom instruction and high-stakes assessments in modern times.

The shifting role of pronunciation in L2 English teaching and assessment: Historical overview

Segmental primacy in traditional instruction and assessment and phonetic training

Pronunciation has had a fraught history in language teaching and standardized testing. At the turn of the 20th century, advocates of the Reform Movement rejected the presiding Grammar Translation Method (e.g., Sweet, 1899), with its sole focus on the written medium and emphasis on translation quality and grammatical correctness (Richards & Rodgers, 2014). Reform proponents heralded phonetics as foundational and central to teaching modern foreign languages, and phonetic transcriptions were emphasized as obviating the need for a native speaking teacher to model the accurate production of L2 sounds to learners. As Weir, Vidaković, and Galaczi (2013) document, an early instance of language teaching directly influencing tests in the Cambridge tradition was the incorporation of a mandatory written English Phonetics paper in the original Certificate of Proficiency in English (CPE) in 1913. The Phonetics paper required test-takers (language teachers) to phonetically transcribe written texts into both carefully enunciated speech, and the conversational speech of "educated persons" (p. 449). Additional items required test-takers to describe the place and manner of articulation of selected segments. Ultimately, the CPE Phonetics paper was short-lived. In an effort to make the test more attractive to prospective test-takers to increase registrations, the phonetics paper was dropped in the first round of test revisions in 1932 (Weir et al).

The importance placed by Reform Movement proponents on measuring aural/oral skills in modern foreign language teaching was echoed in numerous articles published by American authors in the Modern Language Journal in the 1920s to 1940s. Although the presiding view was that “the oral test will always be… the only real one for pronunciation” (Greenleaf, 1929, p. 534), in practice, this was

replete with practical challenges, many of which still resonate today. As Lundeberg (1929) stated in an article describing a phonetic perception test, “ear and tongue skills” are “less measurable because they are less tangible” (i.e., speech is ephemeral and nonvisible without the use of technology). In addition, scoring oral production is “subject to variation” (e.g., listeners may not agree whether sound has been accurately articulated) and “involve(s) the cumbersome and time-consuming expedient of the individual oral examination” (i.e., one-on-one testing and scoring is resource intensive; p. 195). Notwithstanding the “dearth of objective tests of pronunciation” in which responses can be mechanically scored as right or wrong (Tharp, 1930, p. 24), the articles discuss instruments or procedures for assessing pronunciation perception and production. An example of the latter is Bovée’s (1925) “score card” for rating L2 French students’ read utterances at syllable, word, and sentential levels for criteria such as “mouth position/purity” for vowels, “vibration or friction/explosion” for consonants, word stress, pausing (termed “breath group”), liaison, mute ‘e’ suppression, syllable length, and “facility” for sentence production (p. 16).

The argument for the need to speak and understand modern foreign languages was arguably accompanied by a greater sense of urgency in relation to the American war effort during the Second World War, specifically regarding the need for military trainees to demonstrate “oral/aural *readiness*” to communicate when abroad (Kaulfers,1944, p. 137, original emphasis). Kaulfer’s oral fluency test involved the test-taker orally translating language functions and the examiner recording ratings using two 4-level oral performance scales. The first, which assesses the information conveyed, reflects a functional view of language (Richards & Rodgers, 2014). The second, which assesses oral production quality, is perhaps the earliest instantiation of the construct of “intelligibility” in a rating scale. In this context, intelligibility is operationalized as how well “a literate native would understand” the speech, ranging from “unintelligible or no response” to “readily intelligible” (p. 144). The construct of

intelligibility has great relevance in current L2 teaching, research, and assessment, as discussed later in the article.

The emphasis on segmental features and articulatory phonetics continued during the Audiolingual era in the 1950s and early 1960s. This is reflected in Lado's seminal book, Language Testing (1961), with chapters on testing the perception and production of L2 segments, word stress, and intonation. Over half a century since its publication, Lado's work remains the most comprehensive practical guide to L2 pronunciation assessment, covering topics such as item writing, test delivery, and scoring, and, hence, is the existing authority on constructing L2 pronunciation tests, despite some concepts being outdated (Isaacs & Trofimovich, 2017a). Consistent with a contrastive analysis approach, Lado (1961) postulated that where differences exist between a learner's L1 and L2 systems, the L1 habits will be superimposed on the L2, leading to pronunciation problems (L1 transfer errors); however, these errors are predictable and need to be systematically tested.

L2 pronunciation perception and production inaccuracies are indeed often attributable to L1 effects (Derwing & Munro, 2015). However, a large volume of speech perception research has suggested a more nuanced relationship between the L1 and L2 than Lado (1961) maintained. For example, Flege's (1995) Speech Learning Model hypothesizes that learners' acquisition of L2 sounds is mediated by their ability to *perceive* the difference between L1 and L2 sounds. That is, more phonetically similar sounds are more likely to be inaccurately perceived (and, by implication, produced) than more phonetically dissimilar sounds, where learners are more likely to notice a difference. In the scenario where the learner perceives some phonetic difference between the L2 sound and the phonetically closest L1 sound, he/she will form a new L2 phonetic category (i.e., representation in the long-term memory) that is distinct from their existing L1 categories. Conversely, when the learner fails to discern any difference between the target L2 sound and their L1 sounds, he/she will simply substitute

a phonetically similar L1 sound for the target L2 sound, having deemed these sounds equivalent, instead of creating a new phonetic category for the L2 sound. Table 1 summarizes these tenents of Flege's model, which have received substantial empirical backing (Piske, 2013)[2]. Although this line of research has had little uptake in language assessment research, Jones (2015) demonstrates an application. His study aimed to extend the pervasive hVd stimuli (e.g., /hid/, /hɪd/, /hed/, /hɛd/, etc.) traditionally used in phonetics experiments to more authentic stimuli due to concerns about construct underrepresentation in terms of spectral variability (e.g., occurrence of L2 vowels in different phonetic environments than those tested in lab-based studies). Although only a set number of word and nonword pairs were tested, this unveils the possibility of using more naturalistic stimuli in diagnosing vowel perception and production in both experimental and assessment contexts.

<INSERT TABLE 1>

The Structuralist linguistic approach influenced L2 English proficiency tests developed in the UK in the 1960s, including the English Proficiency Test Battery and English Language Battery, which included listening phonemic discrimination, intonation, and sentence stress items (Davies, 1984). In terms of pronunciation production, Lado (1961) echoed Lundeberg's (1929) concerns about the lack of objective scoring and acknowledged practical challenges as a potential deterrent to testing pronunciation. In instances when it was infeasible to administer a face-to-face oral pronunciation test (e.g., due to time, expense, resource), Lado suggested indirect testing through written fixed-response items (e.g., multiple choice) as a viable alternative. In a 1989 article entitled "Written tests of pronunciation: Do they work?" Buck tested Lado's (1961) proposal in Japan using an indirect English pronunciation test modelled on Lado's recommendations. Low correlations between written pronunciation test scores and ratings of test-takers' actual oral productions coupled with low internal consistency among items led Buck to respond to the question posed in the title of the article with an

emphatic “no.” Despite serious problems with the validity of indirect speaking test tasks, discrete-point written items modelled on Lado’s item prototypes that test segmental discrimination and stress placement are still in use today in the high-stakes Japanese National Center Test for University Admissions (Watanabe, 2013).

Deemphasis on pronunciation in communicative language teaching and testing, reconceptualization, and global constructs

Despite the pivotal role of pronunciation in Lado’s (1961) book, which is often taken to represent the birth of language testing as its own discipline (Spolsky, 1995), the focus on pronunciation in language testing was short-lived. During subsequent periods when teaching techniques closely associated with pronunciation (e.g., decontextualised drills symbolizing rote-learning) ran contrary to mainstream intellectual currents, pronunciation tended to be either shunned or ignored (Celce-Murcia et al., 2010). The Naturalistic Approach to teaching that emerged in the late 1960s at the onset of the Communicative era and continued into the 1980s deemphasised pronunciation in instruction, viewing it as ineffectual or even counterproductive to fostering L2 acquisition and helping learners achieve communicative competence (e.g., Krashen, 1981). The belief was that pronunciation, like other linguistic forms (e.g., grammar), could be learned by osmosis through exposure to comprehensible input alone, with no role for formal instruction, although metanalyses decades later showed positive effects for an explicit focus on pronunciation to counter these claims (e.g., Saito, 2012). Thus, pronunciation fell out of vogue for decades in applied linguistics in general and in language assessment in particular. The repercussions of this are evidenced in publications by pronunciation proponents from 1990 onwards citing the “neglect” of pronunciation in English teaching and learning (e.g., Rogerson & Gilbert, 1990). This discourse of neglect persists today (e.g., Baker & Burri, 2016) but has been absent in the area of pronunciation assessment in particular, where, until recently, few advocates have deplored its

marginalization as an assessment criterion in L2 speaking tests or from the collective research agenda. A research timeline on L2 pronunciation assessment (Isaacs & Harding, 2017) demonstrates the gap. Buck's (1989) article is the only timeline entry represented from the language testing literature from Lado (1961) until the emergence of a fully automated L2 speaking test (Phonepass) in 1999 (Bernstein, 1999).

From the mid-1990s until the early 21st century, pronunciation experienced a resurgence of interest among SLA-oriented applied linguists, with several pronunciation-focused articles appearing in prestigious SLA journals (e.g., Studies in Second Language Acquisition, Language Learning). The overarching focus of SLA research was on global constructs such as L2 intelligibility, comprehensibility, accentedness, and, later, fluency, L2 speaker background characteristics, and the linguistic properties of their productions (see Derwing & Munro, 2015, for a summary; see later in this section for definitions of key terms). There was little emphasis on rating scales and rater characteristics or behaviour and little, if any, concurrent pronunciation research in language testing during this period. Numerous indicators since around 2005 attest to the consolidation of pronunciation within mainstream applied linguistics research, including the emergence of pronunciation-specific journal special issues, invited plenaries and symposia, the establishment of a dedicated conference in 2009 (Pronunciation in Second Language Learning and Teaching), evidence syntheses on instructional effectiveness, and the launch of The Journal of Second Language Pronunciation in 2015.

To parallel this, there has been increased research activity in pronunciation assessment in the past decade compared to previous decades, building largely on earlier work primarily in SLA and sociolinguistics and spurred by the central role of pronunciation in the automated scoring of speech (Isaacs & Harding, 2017). For example, no published articles on pronunciation appeared in the journal, Language Testing, from 1984 (first volume) to 1988, compared to 0.54% of all published articles from

1999 to 2008 (Deng et al., 2009) and 4.45% from 1998 until 2009 (Levis, 2015). Similarly, Language Assessment Quarterly published no pronunciation-focused articles from its inception in 2004 until 2011, although at least five articles centering on major L2 pronunciation-related constructs (e.g., accentedness, intelligibility, and/or comprehensibility) have appeared in the years since (2012–17).

This revival of pronunciation research, also accompanied by increased emphasis on particularly suprasegmental aspects of pronunciation and a growing recognition of the need to bolster teachers' pronunciation literacy (Celce-Murcia et al., 2010), has been brought about, in part, by a reshift in focus and reconfiguration of thinking since the decontextualized, mechanical drills of the Audiolingual period (e.g., Lado, 1961). Levis's (2005) characterization of two "contradictory principles" in pronunciation teaching can be useful in elucidating this rebranding of pronunciation that has helped carry it forward into the 21st century (p. 370). The first principle, the nativeness principle, holds that the overall goal of L2 pronunciation teaching should be to help learners eliminate traces of their foreign accent to sound more native-like—a view that is compatible with treatment of the L1 as a bad habit in Audiolingual thinking. In fact, achieving accent-free pronunciation is an unrealistic goal for most L2 learners (Flege et al., 1995) and, furthermore, is unnecessary for integrating into society, achieving in academia, or succeeding on the job (barring, perhaps, serving as a spy or acting a role convincingly). Therefore, most applied linguists subscribe to Levis's (2005) second contrasting principle, the intelligibility principle, as the rightful goal of pronunciation teaching and, by implication, assessment (Harding, 2013). This principle holds that learners simply need to be able to produce L2 speech that is readily understandable to their interlocutors (as opposed to engaging in accent reduction), and that pronunciation pedagogy should target the most consequential features for getting the message across.

In L2 pronunciation research, the nativeness principle is most often operationalized by gauging listener perceptions of "accentedness" on a Likert-type scale (e.g., heavily accented/not accented at all at

the scalar endpoints), to measure the degree to which the L2 accent is perceived to deviate from the (presumed) standard language norm (Derwing & Munro, 2015). The treatment of the intelligibility principle is somewhat more complex due to the existence of numerous interpretations of terms such as intelligibility and comprehensibility and little consensus on how these constructs should be defined and operationalized (Isaacs & Trofimovich, 2012). Levis's (2006) distinction between broad and narrow senses of intelligibility provides a clear-cut characterization that accounts for at least some of the definitional confusion. "Intelligibility," in its broad sense, denotes the understandability of L2 speech in general and is used synonymously with "comprehensibility," often in relation to an instructional goal or assessment target. However, these terms are differentiated in their narrow sense in research contexts based on the way they are operationalized. In Derwing and Munro's (2015) pervasive interpretation (although see Smith & Nelson, 1985, for an alternative view), intelligibility, which is considered the more objective of the two terms, is most often measured by determining the accuracy of listeners' orthographic transcriptions after they hear an L2 utterance. Less frequently, intelligibility has also been operationalized by calculating the proportion of listeners' correct responses to true/false statements or comprehension questions (e.g., Hahn, 2004) or, more rarely, through reaction times measurement, with longer listener reaction times implying less intelligible speech (Hecker, Stevens, & Williams, 1966; Ludwig, 2012). Conversely, Derwing and Munro's (2015) notion of comprehensibility is operationalized by gauging listeners' perceived ease or difficulty of understanding an L2 utterance through the artifact of usually 9-point Likert-type scales.

Although this definitional distinction between intelligibility and comprehensibility is usefully applied in L2 pronunciation research, it is not adhered to in L2 speaking proficiency scales used in operational assessments (Isaacs, Trofimovich, & Foote, 2018). To elaborate, "intelligibility" is often referred to in rating scale descriptors when, in fact, what is being measured is Derwing and Munro's

(2015) “comprehensibility,” since scales and raters automatically imply comprehensibility (narrow sense). This is an example of an instrumental approach to construct definition, in which it is impossible to separate the instrument (scale) from the attribute itself, and is likely symptomatic of the lack of an underlying theory (Borsboom, 2005). However, the use of the term intelligibility in scales still conforms with Levis’ (2006) broad definition of ease of understanding. Therefore, in the remainder of this article, “intelligibility” will be used in its broad sense unless otherwise stated. “Comprehensibility” will be used in its narrow sense to refer to listeners’ scalar ratings of ease of understanding L2 speech, including in scale descriptors for high-stakes tests, unless a direct citation is provided, in which case the original terminology from the scale descriptor itself will be used.

Another global construct that has its roots in speech sciences research is “acceptability,” denoting how acceptable an utterance sounds (i.e., goodness of articulation), although there is limited evidence to show that it is a unitary construct distinct from accentedness (Flege, 1987). With a range of definitions, acceptability has also appeared under the guises of irritation, annoyance, and distraction, including to denote the extent to which the L2 speech deviates from language norms (e.g., Ludwig, 1982) or the extent to which those deviations affect intelligibility (e.g., Anderson-Hsieh, Johnson, & Koehler, 1992). Listener acceptability judgments have also been used in the context of synthesized (i.e., machine-generated) speech to capture their perceptions of how natural the synthetic speech sounds. For example, one method for distinguishing acceptability from intelligibility is to use reaction time measures of whether the response sounds like it was articulated by a human or a machine, although there are other operational measures (Nusbaum, Francis, and Henley, 1995). At the time of writing this article, acceptability in relation to synthetic speech has not yet been incorporated into L2 pronunciation assessment research but may gain currency in the future. For example, it could be expedient to examine in developing or validating a test consisting of dialogue systems with avatars (see Mitchell, Evanini and

Zechner, 2014, for an example of a spoken dialogue system for L2 learners). Notably, this construct should not be confused with "acceptability as a teacher," which has been recently used in Ballard and Winke's (2017) pronunciation assessment study in conjunction with other scalar measures (e.g., accentedness and comprehensibility) to denote listeners' "estimation of how acceptable the speaker is as an ESL teacher" (p. 128).

Linguistic features that should be prioritized in L2 instruction and assessment to promote intelligibility

Functional load and guarding against accent reduction resources that make unrealistic promises to consumers

A central challenge in current L2 pronunciation research for researchers who espouse the intelligibility principle is to empirically identify the linguistic components most conducive to learners' production of intelligible speech so that these can be targeted in instruction and assessment. Post-audiolingual communicatively-oriented pronunciation instruction has moved away from a sole focus on segmental aspects of pronunciation to emphasize the instruction of prosody—a term often used synonymously with "suprasegmentals" to refer to pronunciation features that are longer than the unit of a segment, such as word stress, rhythm, and intonation (Celce-Murcia et al., 2010). In one line of research, the approach has been to experimentally manipulate a pronunciation feature in isolation to examine its effects on intelligibility or comprehensibility (narrowly- defined), either by digitally manipulating the feature to create different spoken renditions using speech editing software (e.g., syllable duration; Field, 2005), or by having the speaker record different experimental conditions for the same passage (e.g., accurate versus inaccurate primary stress placement; Hahn, 2004). Overall, features related to stress and prominence have been shown to affect listener understanding, suggesting an

important role for prosody in achieving effective oral communication. However, only a limited number of prosodic features have, as yet, been examined.

In terms of segmental errors, some are more detrimental for intelligibility than others. For example, a substitution error involving pronouncing /i/ for /ɪ/ (e.g., “sheep” for “ship”) is more likely to result in a communication breakdown than pronouncing /f/ or /t/ for /θ/ (e.g., “fink” or “tink” for “think” (Derwing & Munro, 2015). A theory that can be used to guide the decision of which problematic contrasts, if any, to target in instruction and assessment is the functional load principle, which provides predictions about the communicative effect of mispronunciations of English sound contrasts. To ascertain error gravity of minimal pairs, functional load takes into account a series of factors, such as the frequency of the minimal pair in distinguishing between words, its position within a word, and the likelihood that the minimal pair contrast is upheld in different dialectical varieties of English, since listeners are more likely to be able to make perceptual adjustments for sound pairs that are subject to regional variation than for those that are not (Brown, 1988).

Kang and Moran (2014) demonstrate an application for assessment by classifying test-takers’ error types into high and low functional load on monologic speaking tasks from four Cambridge English exams targeting a range of levels from A2 (Cambridge English: Key) to C2 (Cambridge English: Proficiency) in the Common European Framework of Reference for Languages (CEFR). They found a significant drop in high functional load errors as proficiency level increased for all five levels. However, the result for low functional load errors was less robust, with the only significant difference detected between the highest and lowest levels. Further empirical investigation is necessary to be able to make more concrete recommendations about which contrasts to focus on and which not to other than those near the top and bottom of the rankings, which are obvious cases that have been subject to empirical backing (see Derwing & Munro, 2015). Functional load, which consists of two independently-created

ranking systems[3], would also benefit from some empirically-based consolidation to facilitate its incorporation into future L2 test, design, validation, or scoring procedures.

Once the target minimal pair contrasts have been identified through diagnostic assessment, computer-based applications such as Thomson's (2012) English Accent Coach can be used to draw learners' attention to the acoustic cues of the contrasting sounds to facilitate their formation of new L2 categories (Flege, 1995). This empirically-grounded resource is based on high variability phonetic training (i.e., recordings of different talkers producing the same sounds) to target accurate perception (although not directly production) of North American English segments. This is in striking contrast to accent reduction or elimination websites that mostly make unsubstantiated claims that their training will result in the end-user losing his/her accent in no time (Thomson, 2013). Such resources are often marketed to vulnerable consumers by so-called speech experts who know little about speech production and who may use pseudoscientific terminology or employ unhelpful or even counterproductive techniques in their teaching (e.g., practicing the /p/ sound using tongue twisters with marshmallows between the lips when /p/ is bilabial and can only be produced through lip closure). Further, if the intelligibility principle is espoused, it is incompatible to treat an L2 accent like a pathology that needs to be reduced or eliminated. However, learners may themselves wish to achieve L2 accent-free speech, especially since some L2 accents and regional varieties are stigmatized (Moyer, 2013). The next few paragraphs will leave accent reduction techniques behind and critically evaluate other approaches to identifying which linguistic features to prioritize in instruction and assessment that align with the intelligibility principle.

The Lingua Franca Core: Still not a viable alternative to supplanting the native speaker standard

One crucial topic that follows from the nativeness and intelligibility principles (Levis, 2005) is the issue of defining an appropriate standard for assessing L2 pronunciation proficiency. Jenkins (2002)

has presented the most elaborate set of pedagogical recommendations about pronunciation features that should be emphasized in instruction and, by implication, assessment, in a syllabus for an international variety of English called the Lingua Franca Core (LFC). In light of unprecedented transnational mobility and the pervasive use of English as a lingua franca across the globe, the argument for using an international variety of English that does not make reference to a NS variety and focuses instead on promoting mutual intelligibility is arguably timely and persuasive. Although some instructional materials and teacher training manuals draw heavily on Jenkins' recommendations (e.g., Rogerson-Revell, 2011; Walker, 2010), adopting the LFC uncritically is problematic in light of methodological shortcomings of this work. For example, the LFC was drawn from observational data of pronunciation error types that Jenkins (2002) interpreted as yielding communication breakdowns in learner dyadic interactions. However, the lack of systematicity in data collection and reporting (e.g., no description of the tasks, only some of which were audio recorded, nor an indication of the representativeness of the error types drawn from the dataset to derive the core features) is prohibitive for replication. In addition, the LFC was generated from a limited dataset of international students' interactions in England. Generalizing the resulting core features to all global contexts where English is used as the medium of communication likely overstates the case. More empirical evidence and validation work is needed before the LFC can be adopted as a standard for instruction and assessment that supplants the NS standard, which was integral to the conception of the LFC.

Overall, Jenkins' de-emphasis of the /θ/ and /ð/ sound in the LFC conforms with functional load research supporting the inconsequentiality of these sounds for intelligibility (Munro & Derwing, 2015). However, explicitly teaching L2 learners to substitute these sounds with /f/ and /v/ respectively, which the LFC recommends, has come under scrutiny from applied linguists and phoneticians (e.g., Dauer, 2005). More seriously, Jenkins' (2002) deemphasis of suprasegmental features such as word stress and

timing in the LFC contradicts a weightier body of research suggesting the importance of these features for intelligibility (e.g., Hahn, 2004). Thus, although the LFC is accessible and implementable as a guide for practitioners on which pronunciation features to target in the classroom, which could be extrapolated to assessment settings, it needs to be used with caution and in conjunction with additional research evidence on what counts the most for intelligibility.

"Unpacking" what makes an L2 speaker understandable by examining discrete linguistic features

Yet another approach to identifying which linguistic features to prioritize in instruction and assessment is to "disentangle" the aspects of speech that are most important for comprehensibility versus those that, while noticeable or irritating, do not actually impede listeners' understanding. This has been investigated by correlating listeners' mean comprehensibility and accentedness ratings either with researcher-coded auditory or instrumental measures (e.g., at segmental, suprasegmental, fluency, morphosyntactic, and/or discourse-levels; Isaacs & Trofimovich, 2012), or by eliciting listener ratings of discrete linguistic features using 0–1000 sliding scales (see Saito, Trofimovich, & Isaacs, 2017, for a validation study on examining the linguistic correlates of these ratings). Taken together, these studies have shown that comprehensibility cuts across a wider range of linguistic domains than previously expected, with a "pronunciation" dimension (segmental errors, word stress, intonation, speech rate) and a "lexicogrammar" dimension (lexical richness and appropriateness, grammatical accuracy and complexity, discourse measures), as identified in principal component analyses, both contributing to the variance in listeners' L2 comprehensibility ratings. By contrast, accentedness, which is chiefly related to segmental and prosodic (i.e., "pronunciation") features, appears to be narrower in its scope, at least on tasks which are not cognitively complex (Crowther, Trofimovich, Saito, & Isaacs, 2017).

It is important to consider factors such as the L1 background of the speakers, raters, and task effects (among other variables) to inform what to target, how, and by whom in L2 pronunciation

assessment. For example, Crowther, Trofimovich, Saito, and Isaacs' (2015) study on learners' L1 background in relation to their L2 English speaking performance revealed that segmental errors are consequential for comprehensibility for L1 Chinese speakers, possibly due to the large crosslinguistic difference between English and Chinese. However, for Hindi/Urdu learners, segmental and prosodic errors, while associated with accent, were not significantly linked with comprehensibility. In particular, higher comprehensibility scores were associated with the use of nuanced and appropriate vocabulary, complex grammar, and sophisticated discourse structure for this group. Therefore, for L2 learners above a certain comprehensibility threshold, where pronunciation- and fluency-related variables (e.g., speech rate) do not interfere with comprehensibility, addressing lexicogrammar by helping learners use more precise vocabulary, accurate grammar, and so forth could add value to their degree of comprehensibility (Isaacs & Trofimovich, 2012). A key point here is that comprehensibility is not only about pronunciation. This argues for not construing comprehensibility as a pronunciation-only construct to the exclusion of other linguistic domains, nor confining it to the pronunciation subscale in analytic L2 speaking scales (e.g., IELTS, n.d.). For example, there is some evidence to suggest that grammar tends to be a factor particularly at higher L2 English comprehensibility levels and on more cognitively complex tasks, whereas fluency (e.g., speech rate) tends to be a factor at lower levels across all tasks (Crowther, Trofimovich, Isaacs, & Saito, 2015). Notably, considering L1 effects is relevant for rating scales that aim to cater to learners from mixed L1 backgrounds while, at the same time, avoiding using generic, relativistic descriptions—a key challenge that will be discussed further below.

One of the problems of research in this vein is that the findings are fragmented in different journal articles. The results need to be synthesized to develop coherent pedagogical and assessment priorities for teachers and testers to enhance their practical value. It should be noted that there are now metanalyses in SLA research on pronunciation instructional effectiveness (Lee, Jang, & Plonsky, 2015;

Saito, 2012). Due to the vacuum of practical recommendations for pronunciation assessment since Lado (1961), findings from such evidence syntheses could be a useful starting point for considering which pronunciation features to target in assessment to lead the way forward.

Beyond Lado: Advancing a research agenda for L2 pronunciation assessment into the future

The above paragraphs attest to the renewed activity on L2 pronunciation within the applied linguistics community. As suggested above, there are signs that the L2 assessment field is finally following suit, at least in small measure. The inclusion of pronunciation in the state-of-the-art on the speaking construct at the 2013 Cambridge Centenary Speaking Symposium and in this special issue is a case in point and implies that there is some acknowledgment that pronunciation is indeed an important part of L2 speaking construct. Further, in de Jong, Steinel, Florijn, Schoonen, and Hulstijn's (2012) influential psycholinguistic article, three pronunciation measures (segments, word stress, and intonation), obtained using discrete-point items and scored by judges as either correct or incorrect, were among the predictor variables included in a structural equation model examining different dimensions of the L2 speaking construct for or intermediate and advanced learners of Dutch as the target language. The major finding was that intonation, together with a measure of vocabulary knowledge, explained over 75% of the variance of speaking ability, suggesting a major role for pronunciation within the L2 speaking construct. This is an important precedent for further work on consolidating our understanding of what constitutes L2 speaking ability both within and across tests or tasks and the role that pronunciation may play.

Beyond the piecemeal contributions of individual researchers, a more sustained shift of attention back to pronunciation from the language assessment community is due, in part, to the introduction of fully automated standardized L2 speaking tests (e.g., Pearson's Versant, PTE Academic) and scoring applications (e.g., ETS's SpeechRater). These technologies place considerable weight on pronunciation

(Van Moere & Suzuki, 2018), feeding into field-wide debates on the implications of automated scoring for the L2 speaking construct, discussed in the next section.

Pronunciation, fully automated assessment, and implications of machine-driven scoring on the L2 speaking construct

Due to technological innovation, Lundeberg's (1929) and Lado's (1961) concerns about the lack of objective scoring for speaking and pronunciation can be fully addressed in modern times using automatic speech recognition and automated scoring technology. Essentially, a speech recording algorithm is optimized based on mean ratings carried out by a large cross-section of listeners, which averages out individual listener idiosyncrasies that would have factored into the assessment if only a small number of raters had scored the sample (e.g., 1-3), as would be likely in a nonautomated test scoring situation. This is a clear advantage of using an automated scoring system, which is trained on human raters, with correlations between automated speaking scores and human ratings (including through concurrent validity estimates) a key performance standard by which machine scoring systems are judged (e.g., Bernstein, Van Moere, & Cheng, 2010). Speech analysis software (e.g., Praat) can be used to obtain objective measures of speech. This could include spectral analyses of segments to examine and quantify acoustic properties, such as tongue raising and fronting (Deng & O'Shaughnessy, 2003), pitch range (Kang, Rubin & Pickering, 2010), and automated measures of fluency such as speech rate, obtained by detecting silent pause length and syllable nuclei (De Jong & Wempe, 2009). These machine-extracted measures are among those that could be considered for use by the speech recognizer to decode the speech in a fully automated speaking test and, furthermore, could feature in score reporting for test users in a description of auditory correlates of the automated measures (see Litman, Strik, & Lim, this issue, for background and an in-depth discussion of automatic speech recognition; see

Isaacs, 2018 and Van Moere & Suzuki, 2018, for dedicated chapters on the automated assessment of pronunciation).

In practice, spectral (frequency-based) and durational (time-based) measures lend themselves to automatic speech recognition more than do other pronunciation features. To elaborate, in the case of segmental production, deviations from the referent speech (i.e., training sample or corpus) are calculated by identifying the minimum number of segmental insertions, deletions, or substitutions needed to alter the utterance to find the best string match (Litman et al., this issue). Frequency cut-offs need to be set, for instance, taking into account the acoustic space required to disambiguate sounds (Deng & O'Shaughnessy, 2003). Because prosody is, by definition, longer in span than a segment and is subject to considerable sociolinguistic variation (e.g., by age, gender, social class, regional variety, etc.), it is comparatively difficult to identify an appropriate standard and compare test-takers' speech to that standard (e.g., acoustic correlates of intonation; van Santen, Prud'hommeaux, & Black, 2009).

One source of concern with fully automated scoring from within the L2 assessment community relates to the narrowing of the L2 speaking construct due primarily to restrictions in the task type, measures that can be automatically generated, and the preclusion of human interactions. Because pattern recognition is involved in automated scoring, the automated system can cope much more easily with highly predictable tasks such as read-alouds, utterance repetition, or sentence-building than with less controlled communicative tasks, where test-takers have more scope to respond creatively (Xi, 2010). In addition, establishing high correlations between machine scoring and human ratings in the current state of technology necessitates using highly constrained tasks and other trade-offs (Isaacs, 2018). A rigorous debate about the properties of test usefulness (Bachman & Palmer, 1996) in relation to the fully automated PhonePass test, a precursor to Pearson's Versant English test, was featured in Language Assessment Quarterly between the test reviewer and test development team between 2006 and 2008.

The test reviewer, who adopted a sociocultural perspective, decried the practical benefits of the test at the expense of authenticity, noting the discrepancy with direct speaking tests that capture a greater breadth of interactional patterns than the stimulus-response type questions in the Versant (Chun, 2008). Conversely, the testing team, who adopted a psycholinguistic perspective in their rebuttal, argued that having learners generate rapid responses in real-time is a facet of real-world communicative situations, and the ability to respond intelligibly and appropriately at a reasonable conversational pace is part of the test construct (Downey, Farhady, Present-Thomas, Suzuki, & Van Moere, 2006). Ultimately, the interactional versus cognitive approaches were irreconcilable in the context of the debate, as no agreement between the parties was reached on the merit of the test (although this example should not imply that these views need to be theoretically or practically polarized or dichotomized in other settings). Since the time of that exchange, the language assessment community has arguably arrived at a more pragmatic understanding that automated tests are here to stay (e.g., Xi, 2010), leading, in part, to more airtime for pronunciation at international language testing conferences and in scholarly journals, including in this special issue.

Although automated speaking tests may claim to assess intelligibility, emphasis tends to be on correspondence to NS norms or pronunciation accuracy, whereas some errors ‘count’ more than others in terms of their communicative effect (Isaacs, 2018). Therefore, automated speech scoring would also benefit from insight into the influence of different linguistic features on intelligibility. For example, an automated test that elicits and detects highly infrequent English consonant cluster strings (Pelton, 2013) is not likely to have much bearing on intelligibility. This suggests the importance of having applied linguists with a background in pronunciation and assessment team up with speech recognition programmers to ensure that the features the automated system is targeting are pedagogically sound (and

not simply selected because they are easy for the machine to detect), particularly if claims are being made about intelligibility.

Updating theoretical models and improving the operationalization of pronunciation-relevant constructs in rating scales

Although the field of language testing has moved beyond Lado's (1961) skills-and-elements model as the dominant theoretical view (Bachman, 2000), in some ways, pronunciation is still stuck in a time warp. For instance, "phonology/graphology" in Bachman's (1990) influential model of communicative language ability was excluded from the multi-trait multi-method analysis that informed its development (Bachman & Palmer, 1982). In publications of their eventual model, the authors provide no rationale for reintegrating phonology/graphology back into the model after it had been omitted from their analysis, nor for pairing "phonology" with "graphology" (i.e., legibility of handwriting)—an apparent remnant of Structuralist models of the past (e.g., Carroll, 1961; Lado, 1961). Arguably, in the modern digital age, goodness of penmanship would seem to be obsolete, whereas the need to pronounce words in a way that an interlocutor can understand is indispensable for achieving effective communication. The role of phonology should be more carefully conceptualized in future models of communicative language ability.

Developing an evidential basis for operationalizing pronunciation features in rating scales is also essential for generating more valid ratings and advancing the field. In the case of pronunciation, an intuitive or experiential approach to rating scale development has led to considerable shortcomings in the quality of the descriptors in L2 speaking proficiency scales (Isaacs et al., 2018). For example, intuitively-developed pronunciation descriptors compiled from numerous scales were excluded from the global CEFR scale in a measurement-driven decision (erratic statistical modelling; North, 2000), which partially reflects the shortcomings of the descriptors themselves. Pronunciation is relegated to the status

of a stealth factor or "a source of unsystematic variation" in cases when it does affect listeners' rating but is excluded from the descriptors (Levis, 2006).

Current L2 speaking proficiency scales that *do* include pronunciation are also problematic. Some haphazardly reference behavioral indicators across scale levels (e.g., ACTFL, 2012). Others are so vague or general that the specific linguistic features that constitute level distinctions are often unclear (e.g., IELTS public version, IELTS, n.d.; TOEIC, ETS, 2010). The TOEFL iBT speaking rubrics arguably provide more concrete level distinctions than longer scales (e.g., the scales cited earlier in this paragraph consist of 8–10 levels) by roughly associating "pronunciation," "intonation," "pacing," and "articulation" with varying degrees of intelligibility across four bands (ETS, 2014). However, there is no published guidance on how these terms are defined. Still other scales either implicitly or explicitly equate increasing intelligibility with a more native-like accent or present foreign accent-free speech at the high end of the scale (e.g., CEFR Phonological control scale, Council of Europe, 2001; the now retired Cambridge ESOL common scale for speaking, Taylor, 2011).[4] This contradicts robust research evidence over the past two decades showing that, in fact, perfectly intelligible speech does not preclude the presence of a detectable L2 accent, whereas a heavy accent is a hallmark of unintelligible speech (Derwing & Munro, 2015). Thus, the construct of intelligibility (broadly speaking) needs to be unpacked and not conflated with accent in scale descriptors, and accent needs to be left aside.

The Likert-type scale and sliding scales used in L2 pronunciation research are also limiting in that they only provide relativistic descriptors at the scalar anchors (e.g., very accented/not accented; very easy/difficult to understand). Although such scales can be used reliably with listeners who have no prior linguistic training or rating experience, raters' variable interpretations of the constructs raise questions about construct validity, even if they are only used for low-stakes research purposes. For example, in the absence of a clear operational definition, comprehensibility could be differentially interpreted by

teacher-raters as referring to their understanding of every single word of an utterance or, alternatively, to their understanding of the overall message. And rather than basing their scoring decisions on how much of the information they think they have understood, their focus may be on the degree of effort they feel they have expended in deciphering what was said (i.e., perceived cognitive load). Comprehensibility judgments also could be made from their perspective as a teacher who has had some exposure to the speaker's accent and/or the speaking task, or from the perspective of a naïve listener who has no familiarity with speaker's accent and/or the context of the L2 speaking prompt (Isaacs & Thomson, 2013). Thus, in research and assessment settings, it is important to clarify for raters whether comprehensibility refers to word- or sentence-level understanding or, rather, to the gist of the message and whether listeners should rate from their own perspective, or should attempt to compensate for their experience by pretending that they are a different target listener[5].

Isaacs et al. (2018) demonstrate an approach to developing an L2 English comprehensibility scale intended for formative assessment purposes to guide teachers on the linguistic features to target at different L2 English comprehensibility levels. The starting point for this work was Isaacs and Trofimovich's (2012) data-driven, three-level L2 English "Comprehensibility scale guidelines," restricted for use with learners from one L1 background on a picture narrative task. Through extensive piloting of different iterations of the scale with focus groups of teachers (target end-users), who informed its development, the tool was expanded to a 6-level holistic and analytic scale for use with international university students from mixed L1 backgrounds performing monologic academic speaking tasks. One caveat is that relativistic descriptors of how "effortful" the L2 speech is to understand are accompanied by examples of the same error types that may impede understanding at the lowest three levels ("misplaced word stress, sound substitutions, not stressing important words in a sentence"), meaning that these bands cannot be distinguished based on error type alone. A notable design decision

was to specify that "sounding nativelike is *not* expected" at the top level of the scale. This was included to explicitly clarify for raters that it is possible to have a detectable L1 accent and still reach the highest level of the scale, since international university students need not sound like native speakers (NSs) to excel in academia. This instrument needs to undergo rigorous validation but sets the stage for further work on data-driven scale development that aligns with the intelligibly principle. It is also an improvement on scales that either state or imply that learners at intermediate or advanced levels do not have L1 traces in their speech, which is unrealistic (e.g., CEFR Phonological control scale ≥B2).

Rater effects in judgments of L2 speech and the possibility of integrating human and machine scoring

Another fruitful area for further research is on construct-irrelevant sources of variance (e.g., individual differences in rater cognitive or affective variables) that have the potential to influence rater judgments of L2 comprehensibility and accentedness (e.g., Mora & Darcy, 2017). For example, possible bias in undergraduate students' judgments of International Teaching Assistants has been a source of some L2 pronunciation research (e.g., Kang, 2012). Further research using rating scales from high-stakes tests scored by accredited examiners (as opposed to scales designed for research purposes and used by lay listeners) could extend existing work and enhance its practical relevance for assessment practitioners. Within the past decade, there has also been growing research on the potential biasing effects of raters' accent familiarity on their assessments of overall L2 speaking proficiency, degree of foreign accent, or intelligibility. The findings have been mixed. Some studies have found that raters with greater rater familiarity or exposure to a given L2 accent are significantly more lenient in their ratings than raters with less familiarity or exposure (e.g., Carey, Mannell & Dunn, 2011; Winke, Gass, & Myford, 2013). This parallels findings from listening assessment research that L2 test-takers who are familiar with the accent of the speaker in a recorded listening test prompt may be at a performance

advantage compared to those without less familiarity or exposure (Harding, 2012; Ockey & French, 2016).

Other studies on assessing L2 speaking have failed to detect significant differences as a function of rater familiarity. In one such study by Huang et al. (2016), possible reasons for this null result include statistical underpower and too much overlap between the groups (intact classes) in terms of their exposure to members from the relevant L1 community. The fact that raters in that study *perceived* that they were more lenient in their scoring as a result of greater accent familiarity suggests the need for further investigations that overcome these methodological limitations. In Xi and Mollaun's (2011) study on the TOEFL iBT speaking, the rater training provision was much more extensive than in the other familiarity studies cited above, which could partially account for the contradictory finding. Rigorous rater selection, training, and certification procedures, coupled with supplementary benchmarking for the potentially problematic L1 group to score appears to have improved rater confidence and mitigated familiarity bias. Taken together, these results suggest that high-stakes speaking tests need to take rater familiarity into account in examiner screening or training. Similarly, research studies should attempt to control for raters' accent familiarity (Winke et al., 2013), although this may be difficult to achieve in practice when this variable is not the main focus of the study.

One way of eliminating rater effects is by opting for automated scoring of speech, although there are trade-offs discussed above (Xi, 2010). In addition, automatic speech recognition systems are not foolproof and are subject to error in the form of false positives (i.e., system scores the production of a correctly pronounced L2 sound as an error) and false negatives (i.e., system fails to detect the production of an incorrect L2 sound; see Litman et al., this issue). To mitigate the limitations of machine-driven measurement, future automatic and human scoring systems could conceivably complement each other in a single integrated system (Isaacs, 2018). For example, one approach could be for the machine to

measure the elements that it scores most effectively (e.g., spectral and durational measures), allowing raters to focus their attention on other elements that the machine is less adept at measuring (e.g., task fulfillment, cohesion, appropriateness). This could mitigate construct underrepresentation entailed in purely automated scoring, although the issue of what raters should focus on and how it could complement machine scoring while promoting efficiency gains (e.g., reducing raters' cognitive load when scoring multidimensional constructs, cost considerations, etc.) would need to be carefully considered, as would implications for the nature of the L2 speaking construct being measured. Notably, a hybrid machine-human scoring approach has successfully been operationalized in the writing section of the TOEFL (Ramineni, Trapani, Williamson, Davey, Bridgeman, 2012). It is likely only a matter of time before such an approach is implemented in the context of large-scale speaking tests as well.

Examining the role of pronunciation on dialogic tasks and in lingua franca communication

One final major research priority in need of exploration in relation to L2 pronunciation performance is to examine the nature of communication breakdowns and strategies on more authentic tasks and interlocutor effects. To elaborate, although intelligibility has traditionally been considered an attribute of L2 speakers, some researchers have construed it as "hearer-based" (Fayer & Krasinski, 1987), emphasizing listeners' role in assuming communicative responsibility. Still others have depicted intelligibility as "a two-way process" (Field, 2005), underscoring the "interactional intelligibility" element of communication (Smith & Nelson, 1985). In practice, the notion of intelligibility as a bidirectional process is not reflected in most current L2 pronunciation assessment research, which tends to elicit performance on monologic, nonreciprocal tasks. Thus, the nature of intelligibility breakdowns, strategies for circumventing such breakdowns, and self-, peer-, and external observers' perceptions of communicative success during interactional exchanges have been vastly underexplored. For example, there is a dearth of L2 assessment research on phonological accommodation (i.e.,

convergence/divergence of speech patterns to an interlocutor to establish solidarity or adjust social distance; Moyer, 2013). Jaiyote's (2016) work on peer dyadic interaction among test-takers from shared- versus different-L1 backgrounds, although not specifically focusing on pronunciation, is the kind of research that is needed to catalyze developments in examining key global pronunciation constructs (e.g., intelligibility, accentedness) in relation to overall L2 speaking proficiency. Future research should move beyond lab-based contrived speaking tasks that elicit relatively controlled output to those that, while introducing interlocutor effects, allow for a more communicative orientation with different interactional patterns. Among the key considerations emerging from this work could be the issue of joint scoring for interactional intelligibility and fair test-taker pairing practices (May, 2011).

One related promising area for future investigation relates to redefining an appropriate standard for L2 pronunciation proficiency in lingua franca contexts, in which the interlocutors' shared language is the medium of communication. Although research in this vein broadly rejects the nativeness principle in favor of an international or interactional variety of intelligibility for assessment (e.g., Sewell, 2013), what exactly this entails needs to be more clearly conceptualized (e.g., is Jenkins' 2002 LFC relevant?) Finally, further research on the most important factors for (interactional/international) intelligibility for target languages other than English is essential for understanding which linguistic (particularly pronunciation) features are specific to English and which extend to the many other world languages (Kennedy, Blanchette & Guénette, 2017).

Concluding remarks

The revival of pronunciation research in L2 assessment circles after a long period of hibernation is now indisputable. The recent and forthcoming publication of two edited volumes dedicated to pronunciation assessment (Isaacs & Trofimovich, 2017b; Kang & Ginther, 2018) will ideally promote a shared understanding of central issues and research priorities in assessing pronunciation to different

research communities (e.g., speech sciences, psycholinguistics, sociolinguistics, SLA, pronunciation pedagogy, lingua franca communication, signal processing) to open up the conversation and promote interdisciplinary approaches. This includes better understanding the role of human nature in oral communication, including identifying potential sources of bias in social judgments regarding the way someone sounds, finding effective ways of attenuating those effects to promote fairer assessments (e.g., through rater screening or training), and promoting research that more closely resembles authentic communicative situations. There is also a pressing need to consolidate existing evidence about L2 pronunciation in a practical reference manual for teaching and assessment practitioners, including on the linguistic (including segmental and prosodic) features to prioritize in instruction and assessment using an intelligibility-based approach to guide test development and validation. The publication of such a resource is among the most pressing priorities for advancing the field and, ultimately, firmly establishing the place of pronunciation as an essential part of the L2 speaking construct.

Endnotes

1. For example, the speaking construct being measured in direct speaking tests (e.g., IELTS) tends to be markedly different than in both semi-direct speaking tests that are human scored (e.g., TOEFL iBT), and in fully automated speaking tests (e.g., Versant; see Lim & Galaczi, this issue). This is most obviously reflected in the different way that the speaking ability is scored across tests, which often draw on qualitatively different assessment criteria. Pronunciation and pronunciation-relevant constructs, are, in turn, differentially operationalized in relation to each given speaking ability measure.

2. Flege's (1995) model posits these hypotheses at an abstract level without specifying which particular substitutions will take place for each sound, although, based on the general principles of phonetics, this could be presumed to relate to the place and manner of articulation (consonants) or to tongue height, frontness, and lip rounding (vowels; Reetz & Jongman, 2009).

3. Two independently established functional load systems prioritizing minimal pairs in terms of error gravity, as proposed by Brown (1988) and Catford (1987), are normed on Received Pronunciation and General American English, respectively (i.e., use standard NS varieties as their point of reference). The former presents rank orders of 10 contrasts with which learners often have difficulty using a 10-point scale, having first determined the probability of occurrence of phonemes and their likelihood of being conflated. The latter represents functional load on a percent scale and describes a different process for selecting and ordering these contrasts. Notably, these hypotheses about which contrasts are most and least problematic are English-language specific and cannot generalize to other target languages.

4. Notably, the new incarnation of the Cambridge ESOL Common Scale for Speaking, the Overall speaking scales, lists intelligibility as a criterion without referring to accent or nativeness (Cambridge English, 2016). However, the "phonological features" that lead to varying degrees of intelligibility is vaguely defined in that scale, which introduces a different set of challenges in using the scale.

5. It is, as yet, unclear how accurately and consistently raters are able to channel the views of imagined or idealized listeners while rating. In the absence of further evidence, the general recommendation has been to involve raters from the target audience(s) in conducting ratings, including screening raters for the desired individual difference characteristics where possible (Isaacs & Thomson, 2013; Winke et al., 2013).

References

ACTFL. (2012). *ACTFL proficiency guidelines*. Alexandra, VA: ACTFL.

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*(4), 529–555.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*, 1–42.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*, 449–465.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Baker, A., & Burri, M. (2016). Feedback on second language pronunciation: A case study of EAP teachers' beliefs and practices. *Australian Journal of Teacher Education, 41*, 1–19.

Ballard, L., & Winke, P. (2017). The interplay of accent familiarity, comprehensibility, intelligibility, perceived native-speaker status, and acceptability as a teacher In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 121–140). Bristol, UK: Multilingual Matters.

Bernstein, J. (1999). *PhonePass testing: Structure and construct*. Menlo Park, CA: Ordinate Corporation.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*, 355–377.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Boveé, A. G. (1925). A suggested score for attainment in pronunciation. *The Modern Language Journal*, *10*, 15–19.

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

Buck, G. (1989). Written tests of pronunciation: Do they work? *ELT Journal*, *43*, 50–56.

Cambridge English. (2016). *Cambridge English First for schools: Handbook for teachers for exams from 2016*. Cambridge: Cambridge English Language Assessment.

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.

Carroll, J. B. (1961). Fundamental considerations in testing English proficiency of foreign students. In Center for Applied Linguistics (Ed.), *Testing the English proficiency of foreign students* (pp. 30–40). Washington, DC.

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* (pp. 87–100). Washington, DC: TESOL.

Celce-Murcia, M., Brinton, D., Goodwin, J., with Griner, B. (2010). *Teaching pronunciation: A course book and reference guide* (2nd ed.). Cambridge: Cambridge University Press.

Chun, C. W. (2008). Comments on "evaluation of the usefulness of the *Versant for English* test: A response": The author responds. *Language Assessment Quarterly*, *5*, 168–172.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

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

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2015). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly, 49*(4), 814–837.

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2017). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition*, 1–15. Advance online publication. doi:10.1017/S027226311700016X

Dauer, R. (2005). The Lingua Franca Core: A new model for pronunciation instruction? *TESOL Quarterly, 39*, 543–550.

Davies, A. (1984). Validating three tests of English language proficiency. *Language Testing, 1*, 50-69.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*, 385–390.

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*(1), 5–34.

Deng, J., Holtby, A., Howden-Weaver, L., Nessim, L., Nicholas, B., Nickle, K., . . . Sun, S. (2009). English pronunciation research: The neglected orphan of second language acquisition studies? *Working Paper WP05-09*. Edmonton, AB: Prairie Metropolis Centre.

Deng, L., & O'Shaughnessy, D. (2003). *Speech processing: A dynamic and optimization-oriented approach*. New York: Marcel Dekker.

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

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., & Van Moere, A. (2008). Evaluation of the usefulness of the Versant for English test: A response. *Language Assessment Quarterly, 5,* 160–167.

Eades, D. (2005). Applied linguistics and language analysis in asylum seeker cases. *Applied Linguistics, 26,* 503–526.

ETS. (2010). *TOEIC® user guide: Speaking & writing*. Princeton, NJ: Educational Testing Service.

ETS. (2014). *TOEFL iBT® Test: Integrated speaking rubrics*. New York: McGraw-Hill.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37,* 313–326.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly, 39,* 399–423.

Flege, J. E. (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America, 76,* 692–707.

Flege, J. E. (1987). The production and perception of foreign language speech sounds. In H. Winitz (Ed.), *Human communication and its disorders* (Vol. II, pp. 224-401). Norwood, NJ.: Ablex.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233-277). Timonium, MD: York Press.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America, 97,* 3125–3134.

Fraser, H. (2009). The role of "educated native speakers" in providing language analysis for the determination of the origin of asylum seekers. *International Journal of Speech Language and the Law, 16,* 113–138.

Fulcher, G. (2015). Research timeline: Assessing second language speaking. *Language Teaching*, *48*, 198–216.

Galaczi, E., Post, B., Li, A., & Graham, C. (2012). Measuring L2 English phonological proficiency: Implications for language assessment. In J. Angouri, D. M & J. Treffers-Daller (Eds.), *The impact of applied linguistics: Proceedings of the 44th annual meeting of the British Association for Applied Linguistics* (pp. 67–72). London: BAAL.

Greenleaf, J. J. (1929). French pronunciation tests. *The Modern Language Journal*, *13*, 534–537.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*, 201–233.

Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing*, *29*, 163–180.

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

Hecker, M. H. L., Stevens, K. N., & Williams, C. E. (1966). Measurements of reaction time in intelligibility tests. *The Journal of the Acoustical Society of America*, *39*, 1188–1189.

Hu, G., & Lindemann, S. (2009). Stereotypes of Cantonese English, apparent native/non-native status, and their effect on non-native English speakers' perception. *Multilingual and Multicultural Development*, *30*, 253–269.

Huang, B., Alegre, A., & Eisenberg, A. (2016). A cross-linguistic investigation of the effect of raters' accent familiarity on speaking assessment. *Language Assessment Quarterly*, *13*(1), 25–41.

IELTS. (n.d.) IELTS Speaking band descriptors (public version). Retrieved January 30, 2017, from https://takeielts.britishcouncil.org/sites/default/files/IELTS_Speaking_band_descriptors.pdf

Isaacs, T. (2018). Fully automated speaking assessment: Changes to proficiency testing and the role of pronunciation. In O. Kang, R. I. Thomson, & J. Murphy (Eds.), *The Routledge handbook of English pronunciation* (pp. 570–584). New York: Routledge.

Isaacs, T., & Harding, L. (2017). Research timeline: Pronunciation assessment. *Language Teaching, 50*, 347–366.

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. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition, 34*, 475–505.

Isaacs, T., & Trofimovich, P. (2017a). Key themes, constructs, and interdisciplinary perspectives in second language pronunciation assessment. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 3–11). Bristol, UK: Multilingual Matters.

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

Isaacs, T., Trofimovich, P., & Foote, J. A. (2018). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing, 35*, 193–216.

Jaiyote, S. (2016). *The relationship between test-takers' L1, their listening proficiency and their performance in pairs*. Unpublished PhD thesis, University of Bedfordshire, UK.

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

Jones, J. (2015). *Exploring open consonantal environments for testing vowel perception*. Unpublished Master's thesis, University of Melbourne, Australia.

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.

Kang, O., & Ginther, A. (Eds.). (2018). *Assessment in second language pronunciation*. Abingdon: Routledge.

Kang, O., & Moran, M. (2014). Functional loads of pronunciation features in nonnative speakers' oral assessment. *TESOL Quarterly*, *48*, 176–187.

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*(4), 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.

Kaulfers, W. V. (1944). Wartime development in modern-language achievement testing. *The Modern Language Journal, 28*, 136–150.

Kelly, L. G. (1969). *25 centuries of language teaching: An inquiry into the science, art, and development of language teaching methodology, 500 B.C.-1969*. Rowley, MA: Newbury House.

Kennedy, S., Blanchet, J., & Guénette, D. (2017). Teacher-raters' assessments of French lingua franca pronunciation. In T. Isaacs & P. Trofimovich (Eds.), *Interfaces in second language pronunciation assessment: Interdisciplinary perspectives* (pp. 212–236). Bristol, UK: Multilingual Matters.

Krashen, S. (1981). *Second language acquisition and second language learning*. Oxford: Pergamon Press.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*, 345–366.

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

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.

Levis, J. (2015). Pronunciation trends across journals and the Journal of Second Language Pronunciation. *Journal of Second Language Pronunciation*, *2*, 129–134.

Lim, G. S., & Galaczi, E. D. (this issue).

Lindemann, S. (2003). Koreans, Chinese or Indians? Attitudes and ideologies about non-native English speakers in the United States. *Journal of Sociolinguistics*, *7*, 348–364.

Lindemann, S. (2017). Variation or 'error'? Perception of pronunciation variation and its implications for assessment. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 193–209). Bristol, UK: Multilingual Matters.

Litman, D., Strik, H., & Lim, G. S. (this issue). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly*.

Ludwig, A. (2012). *Interlanguage speech intelligibility benefit for non-native listeners of English*. Unpublished Master's thesis. Universitat de Barcelona, Spain.

Ludwig, J. (1982). Native-speaker judgments of second-language learners' efforts at communication: A review. *Modern Language Journal*, *66*, 274–283.

Lundeberg, O. K. (1929). Recent developments in audition-speech tests. *The Modern Language Journal, 14*, 193–202.

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

Major, R. C. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition, 29*, 539–556.

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*, 127–145.

McNamara, T. (2012). Language assessments as shibboleths: A poststructuralist perspective. *Applied Linguistics, 33*, 564-581.

Mitchell, C. M., Evanini, K., & Zechner, K. (2014). A trialogue-based spoken dialogue system for assessment of English language learners. *Proceedings of the 5th International Workshop on Spoken Dialog Systems*, Napa, CA.

Mora, J. C., & Darcy, I. (2017). The relationship between cognitive control and pronunciation in a second language. In T. Isaacs & P. Trofimovich (Eds.), *Interfaces in second language pronunciation assessment: Interdisciplinary perspectives* (pp. 95-120). Bristol, UK: Multilingual Matters.

Moyer, A. (2013). *Foreign accent: The phenomenon of non-native speech*. Cambridge: Cambridge University Press.

Munro, M. J., Derwing, T. M., & Burgess, C. S. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication, 52*, 626–637.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang.

Nusbaum, H. C., Francis, A. L., & Henly, A. S. (1995). Measuring the naturalness of synthetic speech. *International Journal of Speech Technology*, *2*, 7–19.

Ockey, G., & French, R. (2016). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*, *37*, 693–715.

Pelton, G. (2013). *Intelligent tutoring of pronunciation consonant cluster problems.* Cambridge English centenary symposium on speaking assessment (pp. 25–26). Cambridge: Cambridge English Language Assessment.

Piske, T. (2013). Flege, James. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Hoboken, NJ: Wiley-Blackwell.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). *Evaluation of the e-rater scoring engine for the TOEFL independent and integrated prompts*. Research Report 12-06. Princeton, NJ: ETS.

Reetz, H., & Jongman, A. (2009). *Phonetics: Transcription, production, acoustics, and perception*. Malden, MA: Wiley-Blackwell.

Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching* (3rd ed.). Cambridge: Cambridge University Press.

Rogerson, P., & Gilbert, J. B. (1990). *Speaking clearly*. Cambridge: Cambridge University Press.

Rogerson-Revell, P. (2011). *English phonology and pronunciation teaching*. London: Continuum.

Saito, K. (2012). Effects of instruction on L2 pronunciation development: A synthesis of 15 quasi-experimental intervention studies. *TESOL Quarterly*, *46*, 842–854.

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

Scovel, T. (2000). A critical review of the critical period research. *Annual Review of Applied Linguistics, 20,* 213–223.

Sewell, A. (2013). Language testing and international intelligibility: A Hong Kong case study. *Language Assessment Quarterly*, *10*, 423–443.

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

Solewicz, Y. A., & Koppel, M. (2006). *Automatically correcting bias in speaker recognition systems.* Proceedings of the 2006 IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing (pp. 186–191). Maynooth, Ireland: IEEE.

Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford: Oxford University Press.

Sweet, H. (1899). *The practical study of languages: A guide for teachers and learners*. London: Dent.

Taylor, L. (Ed.). (2011). *Examining speaking: Research and practice in assessing second language speaking*. Studies in Language Testing, 30. Cambridge: UCLES/Cambridge University Press.

Tharp, J. B. (1930). The effect of oral-aural ability on scholastic achievement in modern foreign languages. *The Modern Language Journal, 15,* 10–26.

Thomson, R. I. (2012). Improving L2 listeners' perception of English vowels: A computer-mediated approach. *Language Learning, 62,* 1231−1258.

Thomson, R. I. (2013). Accent reduction. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Hoboken, NJ: Wiley-Blackwell.

Turner, C. E., & Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 255−274). Boston: Walter de Gruyter.

Van Moere, A., & Suzuki, M. (2018). Using speech processing technology in assessing pronunciation. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 137–152). Abingdon, UK: Routledge.

van Santen, J. P. H., Prud'hommeaux, E. T., & Black, L. M. (2009). Automated assessment of prosody production. *Speech Communication*, *51*, 1082−1097.

Walker, R. (2010). *Teaching the pronunciation of English as a Lingua Franca*. Oxford: Oxford University Press.

Watanabe, Y. (2013). The National Center Test for University Admissions. *Language Testing, 30*, 565–573.

Weir, C. J., Vidaković, I., & Galaczi, E. (2013). *Measured constructs: A history of Cambridge English language examinations 1913-2012*. Cambridge: Cambridge University Press.

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.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing, 27*, 291–300.

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning, 61*, 1222–1255.

Table 1

Summary of hypotheses from Flege's (1995) Speech Learning Model on crosslinguistic perception and L2 segmental acquisition

Does the learner perceive a difference between the L1 and L2 sounds?	Predicted action	Predicted perception accuracy for the L2 sound
Yes	Learner creates a new L2 sound category distinct from his/her existing L1 category	High – a new L2-specific category has been created
No	Learner substitutes his/her existing L1 sound for the L2 sound without creating a new L2 category	Low – equivalence classification blocks new L2 category formation