Applied Linguistics 2015: 1–25
doi:10.1093/applin/amv047

Using Listener Judgments to Investigate Linguistic Influences on L2 Comprehensibility and Accentedness: A Validation and Generalization Study

[1,*]KAZUYA SAITO, [2]PAVEL TROFIMOVICH, and [3]TALIA ISAACS

[1]Department of Applied Linguistics and Communication, Birkbeck, University of London, London, UK, [2]Department of Education, Concordia University, Canada and [3]University of Bristol, Graduate School of Education, UK
*E-mail: k.saito@bbk.ac.uk

The current study investigated linguistic influences on comprehensibility (ease of understanding) and accentedness (linguistic nativelikeness) in second language (L2) learners' extemporaneous speech. Target materials included picture narratives from 40 native French speakers of English from different proficiency levels. The narratives were subsequently rated by 20 native speakers with or without linguistic and pedagogical experience for comprehensibility, accentedness, and 11 linguistic variables spanning the domains of phonology, lexis, grammar, and discourse structure. Results showed that comprehensibility was associated with several linguistic variables (vowel/consonant errors, word stress, fluency, lexis, grammar), whereas accentedness was chiefly linked to pronunciation (vowel/consonant errors, word stress). Native-speaking listeners thus appear to pay particular attention to pronunciation, rather than lexis and grammar, to evaluate nativelikeness but tend to consider various sources of linguistic information in L2 speech in judging comprehensibility. The use of listener ratings (perceptual measures) in evaluating linguistic aspects of learner speech and their implications for language assessment and pedagogy are discussed.

Learning a second language (L2) has clearly become a necessity in the current global society. Although many teachers and their students around the world view nativelike linguistic abilities as the ultimate goal of L2 learning (e.g. Derwing 2003), previous research has convincingly shown that few adult learners can attain this goal, even if they start at an early age, and that a perceptible foreign accent is a common feature of L2 speech (e.g. Flege *et al.* 1995). Thus, it is important to set realistic instructional goals for learners, prioritizing comprehensibility, which refers to listeners' perception of how easy or difficult it is for them to understand L2 speech, over linguistic nativelikeness, typically measured through accentedness or listeners' perception of how closely speakers can approximate speech patterns of the target-language community (see Derwing and Munro 2009). Indeed, learners can communicate successfully in the vast majority of business and academic settings without

needing to sound nativelike (Derwing and Munro 2009). Underlying this view is an assumption that comprehensibility and accentedness are two interrelated yet separable constructs and that not all linguistic errors linked to accent equally hinder comprehensibility. While there is some evidence suggesting that comprehensibility is distinct from accentedness (e.g. Munro and Derwing 1995; Jenkins 2000; Kang *et al*. 2010), it is still relatively unclear which linguistic aspects primarily underlie comprehensibility and which are uniquely associated with accent. Furthermore, it has remained controversial whether and to what degree native speakers who are accustomed to listening to accented L2 speech (e.g. experienced ESL teachers, graduate students in applied linguistics) can perceive various linguistic dimensions of comprehensibility and accentedness, compared with those speakers who do not have much experience with foreign accents (Isaacs and Thomson 2013). Therefore, the goal of this study is twofold: (i) to identify 11 linguistic variables that can be used by linguistically experienced and inexperienced listeners to evaluate L2 speech; and (ii) to examine the contribution of these variables to L2 comprehensibility and accentedness.

WHY TARGET COMPREHENSIBILITY?

Before reviewing relevant background literature on the relationship between comprehensibility and accentedness, it is first important to clarify why the current study targeted comprehensibility rather than intelligibility as a measure of understanding. In a review of L2 intelligibility research, Levis (2006) outlined a distinction between broad and narrow views of intelligibility. In a narrow sense, intelligibility is conceptualized as a product of understanding and is operationally defined as accuracy with which listeners orthographically transcribe L2 speech (e.g. Munro and Derwing 1995) or answer comprehension questions related to its content (e.g. Hahn 2004). In a broad sense, however, intelligibility refers to listeners' *subjective* perception of how much or how easily they understand L2 speech. In this sense, according to Levis, intelligibility is 'not usually distinguished from closely related terms such as comprehensibility' (p. 252), in that both constructs are measured through listeners' scalar ratings, without reference to any *objective* measure of understanding (e.g. Varonis and Gass 1982). Indeed, outside research, most real-world applications of intelligibility, such as high-stakes assessment instruments (e.g. TOEFL, IELTS, CEFR), involve scalar ratings, which implies that the targeted construct is in fact comprehensibility. Thus, comprehensibility is subsumed within Levis's broad sense of intelligibility and represents a common and easy-to-use metric of understanding in both research and real-world contexts (e.g. Levis 2006). In keeping with this tradition of using a rated measure of understanding, the current study therefore targets intelligibility in its broad sense, focusing on comprehensibility, with the overall goal of distinguishing those linguistic dimensions of L2 speech that feed into comprehensibility from those that are linked to accentedness.

COMPREHENSIBILITY VERSUS ACCENTEDNESS

From a theoretical perspective, a focus on comprehensibility (rather than accentedness) is central to the idea that language learning is most efficient when learners are exposed to meaningful language, especially through interaction. The Interaction Hypothesis, for instance, posits that language learning primarily takes place in situations when communication is compromised during L2 conversational interaction (Long 1996). When interlocutors encounter communication breakdowns, they make intuitive or conscious effort to repair linguistic errors causing misunderstanding, engaging in negotiation for meaning. Assuming that certain linguistic features in learner speech are more likely than others to cause communication breakdowns and thus trigger negotiation for meaning (Mackey *et al.* 2000), the learning value of L2 conversational interaction will be greatest for those linguistic features that are tied to comprehensibility rather than those that only contribute to the perception of accent (Derwing and Munro 2009).

Although interaction itself is thought to make problematic features available to the learner, conversationally modified input and output appear to be facilitative of L2 learning only when learners are developmentally ready, that is, when they have some degree of metalinguistic awareness of the problematic features in question (Mackey and Philp 1998). Interaction also seems to play a facilitative role in L2 learning as a way of consolidating partially acquired knowledge rather than as a means of acquiring new knowledge (Shintani *et al.* 2013). Therefore, identifying and teaching the linguistic features that predominantly impact comprehensibility should equip learners with the kinds of knowledge that would be necessary for them to make the most of the L2 input and interaction. Comprehensibility, with its focus on the ease or difficulty of interlocutors' mutual understanding, thus becomes crucially important in enabling researchers and teachers to both isolate and target such linguistic dimensions.

Several studies have investigated which aspects of pronunciation, fluency, lexis, and grammar impact listener understanding through subjective judgments of comprehensibility or through more objective measures of intelligibility. With respect to pronunciation and fluency, listener understanding is associated with various aspects of L2 speech, including individual vowels and consonants with high functional load (Munro and Derwing 2006), sentence stress (Hahn 2004), word stress (Field 2005), speech rate (Derwing *et al.* 2004), as well as pitch range, stress, and pause or syllable length (Tajima *et al.* 1997; Kang *et al.* 2010; Winters and O'Brien 2013). With regard to lexis and grammar, which are a focus of a smaller volume of research, comprehensibility is associated with measures of grammar accuracy (Munro and Derwing 1995), and listener understanding is compromised when listeners are exposed to ungrammatical sentences (Varonis and Gass 1982) or poor word choice (Fayer and Krasinski 1987). In sum, listener understanding appears to be linked to a variety of linguistic variables, spanning the domains of pronunciation, fluency,

lexis, and grammar. Yet there exists a substantial overlap between the constructs of comprehensibility and accentedness in the domains of pronunciation and fluency, with such factors as segmental accuracy, temporal measures, syllable duration, stress, and pitch range also contributing to the perception of an L2 accent (Anderson-Hsieh *et al.* 1992; Winters and O'Brien 2013).

While influences of individual properties of speech on comprehensibility and accentedness are relatively well understood, it is still unclear how multiple linguistic dimensions *interact* and whether they affect comprehensibility differently from accentedness. One reason for this is that previous studies included only a handful of linguistic measures or a small number of listeners (e.g. Zielinski 2008), and few examined both comprehensibility and accentedness within the same data set (e.g. Munro and Derwing 1995). In an initial attempt to address this, our precursor research (Isaacs and Trofimovich 2012; Trofimovich and Isaacs 2012) examined 60 native English-speaking listeners' ratings of comprehensibility and accentedness for 40 adult native French speakers telling a picture story in English. Speech samples were additionally analyzed for 19 coded measures drawn from the domains of pronunciation (segmental, syllable structure, and word stress errors; pitch contour and range; vowel reduction ratio), lexis (lexical errors; token and type frequency), grammar (grammatical errors), and discourse structure (story cohesion; story breadth and depth). Whereas comprehensibility and accentedness were significantly related to several pronunciation measures (e.g. word stress, rhythm), comprehensibility was also associated with such variables as grammatical accuracy and lexical type frequency (number of distinct content words).

RATERS' JUDGMENTS AS MEASURES OF L2 SPEECH

The research reviewed above raises numerous further issues, including the pedagogical relevance of these findings for practitioners in L2 classrooms who need to make online and intuitive judgments of the quality of their students' speech. Our precursor research (Isaacs and Trofimovich 2012; Trofimovich and Isaacs 2012) analyzed learner speech through linguistic analysis conducted by trained coders via relevant software (e.g. pitch tracker) to derive various measures of pronunciation, fluency, lexis, grammar, and discourse structure. However, it is unclear what particular linguistic training and experience is required for reliable judgments of complex linguistic phenomena in L2 speech (e.g. vowel reduction or pitch movement) and whether naïve listeners can achieve this. Previous research focusing on rater experience (often defined in terms of linguistic training and/or teaching experience) has been inconclusive. Although experienced raters appear to be more consistent than inexperienced ones in their judgments (Calloway 1980), experienced raters have also been shown to assign higher accentedness ratings (Thompson 1991) and lower fluency ratings (Rossiter 2009) compared with inexperienced raters, albeit with some studies reporting no rater group differences (Bongaerts *et al.* 1997).

If only linguistic analysis, rather than intuitive listener judgments, can reveal component linguistic features of L2 global constructs such as comprehensibility and accentedness, such research findings would provide limited practical applications to most pedagogical settings. Researchers would still need to demonstrate sufficiently high intra- and inter-coding reliability by recoding a sizable proportion of the data, in some cases based on repeated listenings in a sound-attenuated environment—a highly time-consuming and labor-intensive endeavor. And if teachers are unable to obtain accurate judgments of their learners' speech, teachers may find it difficult to extrapolate research findings using listener-coded auditory or instrumental measures to their own teaching situations. In essence, this would threaten the entire premise of comprehensibility research, which is ultimately to enable practitioners to identify and integrate information about linguistic influences on comprehensibility into their teaching to enhance learners' success in communicative settings.

To date, empirical studies examining the validity of raters' intuitive judgments of *specific* areas of L2 speech have been confined to investigations of fluency (e.g. Derwing *et al*. 2004; Bosker *et al*. 2013), with only one study targeting both comprehensibility and accentedness (Derwing *et al*. 2004). This study focused on listeners without much teaching and linguistic experience (for definition, see Isaacs and Thomson 2013) and raised the possibility of using such inexperienced raters' global, impressionistic fluency judgments to measure the temporal aspects of L2 speech. In particular, Derwing *et al*. showed a significant relationship between the inexperienced raters' scalar ratings of fluency and several relevant instrumentally derived linguistic measures, such as mean length of run (MLR), articulation rate, and pausing. More importantly, listeners' fluency scores were correlated with comprehensibility more strongly than with accentedness, suggesting that inexperienced listeners can evaluate fluency as a component of comprehensibility. Derwing *et al*. concluded that 'rating data from even untrained listeners reflect properties inherent in the stimuli and are therefore useful in the evaluation of speech samples' (p. 672).

The current study

Motivated by this line of work, the current study focused on 11 rated variables drawn from the domains of phonology, fluency, lexis, grammar, and discourse structure created for use by listeners not accustomed to evaluating L2 speech, as opposed to trained coders. These variables were embedded in rater-training materials, in an effort to enable even inexperienced listeners to reliably apply them. The first objective was to determine the extent to which 20 native-speaking listeners with or without linguistic and teaching experience could use these 11 variables to evaluate L2 speech samples, and to examine how these listener-based ratings compare with the output of linguistic coding and analysis. To ensure comparability of research findings across studies,

these 11 listener-based categories were developed to closely match the 18 phonological, temporal, lexical, grammatical, and discourse measures used in our precursor research (Isaacs and Trofimovich 2012; Trofimovich and Isaacs 2012). Accordingly, the second objective was to examine the relationship between the 11 rated variables and listeners' judgments of comprehensibility and accentedness. To summarize, this study explored linguistic influences on comprehensibility and accentedness through listener judgments, as opposed to linguistic analysis. The overarching goal was to expand the practical relevance of L2 speech research to other research settings and pedagogical contexts.

METHOD

Participants

The participants (henceforth, raters) were 20 native English speakers (M_{age} = 28.0 years, *range* = 19–32), born and raised in English-speaking Canadian homes with at least one native English-speaking parent. They estimated using English 90 percent of the time (M = 86.5 percent for speaking and 91.0 percent for listening). At the time of the study, all raters were residents of Montreal, a bilingual French–English city, which was the same context where the target L2 speech samples had been recorded (see below). Because listener familiarity with L2 speech can impact their judgments (e.g. Isaacs and Trofimovich 2012; Trofimovich and Isaacs 2012) and because recruiting functionally monolingual speakers of English in Montreal is problematic, only the raters who reported 'high' familiarity with French-accented English were selected. This allowed for controlling the listener familiarity variable while also ensuring that both the raters and the speakers came from the same sociocultural context. Half of the raters, assigned to the inexperienced group (n = 10), had received no training in linguistics or phonetics and had no language teaching experience. Six were undergraduate students at an English-medium university while the others were unaffiliated with any postsecondary institution. The remaining 10 experienced raters, who, by definition, had linguistic and teaching backgrounds (Isaacs and Thomson 2013), were graduate students in applied linguistics at an English-medium university ($M_{L2\ teaching}$ = 3.7 years, *range* = 2–10). Six reported having taken a course in applied phonetics or pronunciation teaching, with the remaining four raters reporting no pronunciation-specific training experience. All reported having normal hearing.

Materials

Speech samples of 40 native French speakers of English (27 women, 13 men) from Montreal, Canada (M_{age} = 35.6 years, *range* = 28–61) from our precursor research (Isaacs and Trofimovich 2012; Trofimovich and Isaacs 2012) were used as stimuli, which allowed for comparison of findings across studies.

The speakers as a group reported a wide range of English speaking, listening, reading, and writing abilities (spanning the entire range of a 9-level proficiency scale, where 1 = *extremely poor*, 9 = *extremely proficient*) and using English to varying degrees daily (0–70 percent of the time, as measured on a percent scale). The speakers recorded a narrative based on an eight-frame picture story about two people colliding on a street corner and accidentally exchanging their suitcases, which were identical in appearance (Derwing *et al.* 2004). Because the original recordings ranged in length between 55 and 351 s, the beginning of each narrative (23–36 s) was excised from each audio file in line with previous L2 speech research (Derwing and Munro 2009). The resulting audio samples were normalized by matching peak amplitude across samples and removing initial dysfluencies (e.g. false starts, pauses). All samples were orthographically transcribed and verified for accuracy by another transcriber.

Speech rating

The raters were tested individually in a quiet room and performed three rating tasks in a fixed order, starting with a global rating of accentedness and comprehensibility, followed by ratings of the speech for five pronunciation variables and, finally, ratings of orthographic transcriptions of the speech for six lexical, grammatical, and discourse-level variables. To avoid rater fatigue, the rating was divided into two sessions scheduled on separate days, in line with previous research employing listener judgments (e.g. Derwing *et al.* 2004), with audio-based judgments completed in Session 1 and transcript-based judgments in Session 2. And to minimize unwanted order and speaker familiarity effects, the 40 audio files were presented to each rater in a unique randomized order, which ensured that at any given point in the rating, each rater's previous experience with audio files was different.

All ratings were collected using MATLAB, and the raters used a moving slider on a computer screen to assess the global and linguistic qualities in L2 speech samples for the various measures (for training materials, see Appendix A). If the slider was placed at the leftmost (negative) end, labeled with a frowning face, the rating was recorded as '0'. If it was placed at the rightmost (positive) end, labeled with a smiley face, the rating was recorded as '1,000'. Apart from the frowning and smiley faces and accompanying brief verbal descriptions (e.g. *difficult to understand*, *easy to understand*) to indicate the endpoints, the scale included no numerical labels or marked intervals. The relevant sliding scales for each set of judgments (accentedness and comprehensibility for global ratings; five pronunciation and fluency categories for audio ratings; six vocabulary and grammar categories for transcript ratings) were all visible simultaneously on a computer screen (for onscreen labels, see Appendix B). The raters were told that the speech samples represented variable English language proficiency and were encouraged to use the entire range of each scale. The slider on each scale initially appeared in the middle, and the raters were told that even a small movement of the slider may

represent a fairly large difference in the rating. Before proceeding to the next speech sample, the raters were allowed to adjust their judgments on all visible scales as many times as they wished until they felt satisfied with their decision.

At the beginning of Session 1, raters evaluated each of the 40 speech samples for accentedness and comprehensibility simultaneously. Following previous research, accentedness was defined as listeners' perceptions of the degree to which L2 speech is influenced by his/her native language and/or colored by other non-native features (Trofimovich and Isaacs 2012). Comprehensibility was defined as the degree of ease or difficulty in listeners' understanding of L2 speech (Derwing and Munro 2009). The raters first received detailed instructions about each construct on a printed paper and recapped orally (see Appendix A), then rated three practice speech samples (not included in the main data set) to familiarize them with the procedure, and then proceeded to evaluate the 40 randomly ordered speech samples, working at a pace that was consistent across all raters (approximately 30 min). Each sample was played once, following prior research (e.g. Trofimovich and Isaacs 2012), on the assumption that accentedness and comprehensibility tap into listeners' initial intuitions and impressions about L2 speech.

Following a short break, the raters then evaluated the same 40 speech samples for five pronunciation variables which represented segmental, suprasegmental, and fluency dimensions of speech. These variables included (i) vowel and consonant accuracy (substitution, omission, or insertion of individual sounds), (ii) word stress (misplaced or missing primary stress), (iii) intonation (appropriate, varied use of pitch moves), (iv) rhythm (alternation of stress between content and function words), and (v) speech rate (speed of utterance delivery). A total of 20 raters again received thorough instructions (see Appendix A) and then evaluated three practice samples. For each practice sample, they were asked why they made their decisions and then received feedback to ensure that the rated categories were understood and applied appropriately. The raters then proceeded to rate the 40 randomized speech samples at their own pace, which was comparable across all raters (approximately 60 min). They were allowed to replay each file if necessary, in view of the fact that they were rating several discrete measures (as opposed to global measures of accentedness and comprehensibility). Although all raters used this option during the practice sessions, few did so during the main rating.

In Session 2, the raters evaluated speech transcripts for six lexical, grammatical, and discourse-level variables (see Appendix A). Following Crossley *et al.* (2014), the raters evaluated written transcripts instead of listening to speech samples to ensure that they were not distracted by speakers' pronunciation accuracy (e.g. sound substitutions) or dysfluencies (e.g. filled and unfilled pauses). Therefore, using written transcripts (rather than actual speech samples) helped us tease apart and control the influence of pronunciation and fluency factors on rater judgment of lexical, grammatical, and discourse properties of L2 speech.

Each transcript was modified to remove obvious sound substitutions or omissions, especially those specific to French speakers of English (e.g. *then* spoken as *den* was transcribed as 'then', *have* spoken as *'ave* was transcribed as 'have'), and to eliminate orthographic marking of pausing (e.g. uh, um). Thus, cleaned-up versions of all transcripts were used in Session 2. The six variables included (i) lexical appropriateness (accuracy and precision of vocabulary), (ii) lexical richness (varied and sophisticated use of vocabulary), (iii) grammatical accuracy (errors in word order, grammar endings, agreement), (iv) grammatical complexity (use of sophisticated, non-basic grammar), (v) story richness (narrative sophistication and detail), and (vi) story cohesion (use of discourse markers). As in the speech rating session (Session 1), the raters scored three practice written samples following instructions. In each case, they were asked to explain their decisions and received feedback. Subsequently, the raters evaluated the 40 randomized written transcripts in a self-paced task, which was comparable in duration across all raters (approximately 30 min). After completing each session, all raters retrospectively evaluated the extent to which they understood the linguistic concepts in the 11 categories using a 9-point scale (1 = *I did not understand this concept at all*, 9 = *I understand this concept well*) in a post-task questionnaire.

Linguistic coding

To determine the extent to which the raters were able to assess the linguistic dimensions from the speech and transcripts using continuous sliding scales, their ratings were compared with the original coding of the same speech samples by linguistically trained coders from the precursor research (Isaacs and Trofimovich 2012; Trofimovich and Isaacs 2012). Because these coded measures are discussed in detail in the original study, only a brief description is provided here. Of the 18 coded measures selected for this comparison, the first nine targeted pronunciation and fluency and the remaining nine focused on various aspects of lexis, grammar, and discourse structure. A trained coder first conducted a linguistic analysis for each measure either through auditory measures (e.g. frequency counts) of the phenomenon being analyzed, or via relevant analysis software, such as *Praat* (Boersma and Weenink 2010) for speech measures and *Lexical Tutor* (Cobb 2010) for lexical measures. Then, another trained coder recoded 40 percent of the speech samples for each measure. Inter-coder agreement (Cronbach's α) exceeded .90 for all measures except lexical error ratio (.85), suggesting that coding was overall consistent.

1 *Segmental error ratio*. The total number of segmental (vowel, consonant) substitutions, divided by the total number of segments articulated.
2 *Syllable structure error ratio*. The total number of vowel and consonant epenthesis (insertion) and elision (deletion) errors, divided by the total number of syllables articulated.

3 *Word stress error ratio*. The total number of word stress errors (i.e. misplaced or missing primary stress) in polysyllabic words, divided by the total number of polysyllabic words produced.
4 *Intonation error ratio*. The number of correct pitch patterns produced at the end of phrases (i.e. syntactic boundaries) over the total number of phrases where pitch patterns are expected.
5 *Vowel reduction ratio*. The number of correctly reduced syllables over the total number of obligatory vowel reduction contexts in both polysyllabic words and function words (as a measure of English rhythmic timing).
6 *Articulation rate*. The total number of syllables produced excluding dysfluencies (e.g. filled pauses, repetitions, self-corrections, false starts), calculated over the total speech sample duration.
7 *Mean length of run*. The mean number of syllables produced between two adjacent filled or unfilled pauses (≥400 ms).
8 *Number of filled pauses*. The total number of non-lexical pauses (e.g. uh, um).
9 *Number of unfilled pauses*. The total number of silent pauses (≥400 ms).
10 *Lexical errors*. The number of incorrectly used lexical expressions, over the total number of words spoken.
11 *Token frequency*. The total number of words produced (i.e. all word occurrences in each narrative), normalized for sample duration.
12 *Type frequency*. The total number of unique words (i.e. unique word forms) produced, normalized for sample duration.
13 *Lexical sophistication*. The number of frequent words (i.e. first 2,000 word families based on the British National Corpus), over the total number of words spoken.
14 *Grammatical errors*. The number of words with at least one morphosyntactic error (errors in word order, agreement, morphological marking), divided by the total word count.
15 *Subordinate clause ratio*. The number of subordinate clauses, divided by the total number of clauses produced.
16 *Number of propositions*. The number of distinct propositions or storytelling elements (predicate, followed by another argument) in a speech sample, normalized for sample duration.
17 *Number of story categories*. The number of different proposition categories in a speech sample (e.g. setting, attempt, reaction), normalized for sample duration. This measure is based on the idea that a speech sample describing only the setting may be poorer in discourse structure than a sample that first focuses on setting and then describes the events and consequences.
18 *Number of cohesive devices*. The number of adverbials used as cohesive devices (e.g. *suddenly*, *but*, *hopefully*), normalized for sample duration. Cohesive devices help situate the listener in the story by establishing links between storytelling elements, propelling the storyline forward, or revealing the storyteller's attitude.

RESULTS

Rated judgments versus linguistically coded measures

The first objective of this study was to determine the extent to which native-speaking raters with and without linguistics and teaching experience could use 11 rated variables to evaluate L2 speech and to examine how these judgments compare with the output of linguistic coding by trained coders. To address this, the reliability of rating decisions was analyzed first, as a check of the consistency in an individual rater's behavior relative to that of the other raters. Table 1 lists interrater reliability indexes (Cronbach's α) for each rated measure, computed across all 20 raters and then separately for each rater group. The entire sample of 20 raters showed relatively strong agreement across all rated variables (α = .91–.97). However, when raters' linguistic and teaching backgrounds were considered, some group differences emerged, especially in ostensibly more subjective or conceptually more complex linguistic categories, such as suprasegmentals (i.e. word stress, intonation, rhythm), lexical appropriateness, and grammatical accuracy and complexity. In contrast, interrater agreement was high and comparable across the two groups ($\alpha > .90$) in more intuitive and conceptually simpler categories, such as global speech judgments (comprehensibility, accentedness), segmental accuracy, temporal fluency

Table 1: Interrater agreement (Cronbach's α) for global, audio- and transcript-based ratings

Rated variable	All raters ($n = 20$)	Inexperienced ($n = 10$)	Experienced ($n = 10$)
A. Global rating			
Comprehensibility	.97	.95	.94
Accentedness	.98	.95	.97
B. Audio rating			
Vowel/consonant errors	.96	.90	.93
Word stress	.95	.86	.93
Intonation	.93	.81	.91
Rhythm	.95	.88	.93
Speech rate	.97	.92	.94
C. Transcript rating			
Lexical appropriateness	.95	.88	.91
Lexical richness	.97	.91	.96
Grammatical accuracy	.96	.89	.94
Grammatical complexity	.94	.88	.91
Story richness	.97	.94	.95
Story cohesion	.91	.85	.81

(speech rate), and story richness. The only variable that elicited somewhat lower agreement was story cohesion, likely because the short audio excerpts featured relatively few cohesive devices ($M = 4.2$, *range* = 0–10), leaving raters with few items to evaluate. Thus, infrequent use of cohesive devices by speakers likely made it harder for raters to evaluate them in a highly consistent manner.

The next analysis targeted the relationship between the 11 rated variables and the 18 coded measures from the precursor research (Isaacs and Trofimovich 2012; Trofimovich and Isaacs 2012). This analysis was carried out using two sets of Pearson correlations, one for pronunciation and fluency (audio-based rating) and the other for lexis, grammar, and discourse structure (transcript-based rating), with alpha level for significance adjusted for the number of correlations computed in each analysis ($\alpha = .006$). Table 2 shows the correlations between the five rated pronunciation and fluency variables obtained in this study and the nine corresponding pronunciation and fluency coded measures from the precursor research (Isaacs and Trofimovich 2012; Trofimovich and Isaacs 2012). As this table illustrates, the raters' perception of vowel and consonant errors was significantly correlated with a coded measure of segmental accuracy ($r = .63$) but not with a measure of syllable structure errors ($r = .40$, $p > .006$). Their judgments for word stress, intonation, and rhythm were also significantly associated with word stress error ratio ($r = .70$), intonation error ratio ($r = .52$), and vowel reduction ratio ($r = .76$), respectively. As for fluency, raters' judgment of speech rate seemed to be particularly linked to MLR ($r = .78$) and to some extent to a number of unfilled pauses ($r = .47$). In terms of raters' linguistic and teaching backgrounds, the

Table 2: Correlations between audio-based ratings and coded linguistic variables from precursor research (Isaacs and Trofimovich 2012; Isaacs and Trofimovich 2012)

Rated variable	Coded dimension	All raters ($n = 20$)	Inexperienced ($n = 10$)	Experienced ($n = 10$)
Vowel/consonant errors	Segmental errors	.63*	.59*	.64*
	Syllable structure errors	.40	.38	.41
Word stress	Word stress errors	.70*	.64*	.72*
Intonation	Intonation errors	.52*	.47*	.54*
Rhythm	Vowel reduction	.76*	.75*	.74*
Speech rate	Articulate rate	.42	.39	.43*
	MLR	.78*	.75*	.79*
	Number of filled pause	.39	.36	.41
	Number of unfilled pause	.47*	.45*	.49*

Note: *$\alpha < .006$ (Bonferroni corrected).

ratings were more strongly associated with the coded measures for experienced than for inexperienced raters. Yet the difference between correlation coefficients for the two rater groups was not significant for any variable according to Fisher's *r*-to-*z* transformation ($p = .25–.40$).

Table 3 shows a parallel set of correlations, in this case between the six rated lexical, grammatical, and discourse-level variables and nine similar coded measures from the precursor research (Isaacs and Trofimovich 2012; Trofimovich and Isaacs 2012). As the table indicates, the ratings were overall associated with the relevant coded measures ($r = .50–.71$), except for grammatical complexity ($r = .37$) and discourse cohesion ($r = .15$), which involved weak and non-significant associations. The raters' judgment of lexical richness was mainly related to type ($r = .71$) and token ($r = .66$) frequencies, but was unrelated to the coded measure of lexical sophistication (infrequent word types). As with pronunciation and fluency judgments, experienced raters' judgments of lexis, grammar, and discourse structure yielded stronger correlations with the relevant coded measures than the ratings by the inexperienced raters. For instance, only the experienced raters' judgment of grammatical complexity was significantly correlated with the coded measure of subordinate clauses ($r = .44$). However, once again, a Fisher's *r*-to-*z* transformation resulted in no significant differences in the strength of correlation coefficients between experienced and inexperienced raters ($p > .05$).

Last, we analyzed the extent to which the raters felt they actually understood the five audio- and six transcript-based rated categories using the post-

Table 3: Correlations between transcript-based ratings and coded linguistic variables from precursor research (Isaacs and Trofimovich 2012; Isaacs and Trofimovich 2012)

Rated variable	Coded dimension	All raters ($n = 20$)	Inexperienced ($n = 10$)	Experienced ($n = 10$)
Lexical appropriateness	Lexical errors	.50*	.48*	.50*
Lexical richness	Token frequency	.66*	.61*	.69*
	Type frequency	.71*	.67*	.74*
	Lexical sophistication	.27	.31	.23
Grammatical accuracy	Grammatical errors	.67*	.64*	.67*
Grammatical complexity	Subordinate clause ratio	.37	.27	.44*
Discourse richness	No. of propositions	.64*	.63*	.64*
	No. of story categories	.27	.31	.29
Discourse cohesion	Cohesion devices	.15	.10	.19

Note: *$\alpha < .006$ (Bonferroni corrected).

Table 4: Means and standard deviations for ratings of subjective understanding of all rated variables

Rated variables	All raters ($n = 20$)	Experienced ($n = 10$)	Inexperienced ($n = 10$)
Vowel and consonant errors	8.5 (0.8)	8.9 (0.3)	8.1 (1.0)
Word stress	8.3 (1.2)	8.7 (0.5)	7.8 (1.5)
Intonation	8.2 (1.0)	8.7 (0.5)	7.7 (1.2)
Rhythm	7.4 (1.7)	8.1 (1.3)	6.8 (1.9)
Speech rate	8.6 (0.6)	8.9 (0.3)	8.4 (0.7)
Lexical appropriateness	8.5 (0.8)	8.6 (0.7)	8.4 (0.7)
Lexical richness	8.6 (0.8)	8.9 (0.7)	8.3 (0.8)
Grammatical accuracy	7.9 (1.5)	8.8 (0.4)	7.1 (1.8)
Grammatical complexity	7.6 (1.4)	8.4 (0.7)	6.8 (1.6)
Discourse richness	8.4 (0.2)	8.8 (0.3)	9.0 (0.0)
Discourse cohesion	8.9 (1.0)	8.9 (0.4)	8.0 (1.2)

Note: 1 = I did not understand this concept at all; 9 = I understand this concept well.

task questionnaire (1 = *I did not understand this concept at all,* 9 = *I understand this concept well*). For the audio-based measures, the experienced and inexperienced raters' scores (summarized in Table 4) were submitted to a two-way analysis of variance (ANOVA) with group (experienced, inexperienced) as a between-subjects factor and category (vowel and consonant errors, word stress, intonation, rhythm, speech rate) as a within-subjects factor. The ANOVA yielded a significant main effect of group, $F(1, 17) = 7.40$, $p = .015$, and a significant main effect of category, $F(4, 68) = 5.19$, $p = .001$, but no significant two-way interaction, $F(4, 68) = 0.45$, $p = .67$. According to Bonferroni-corrected multiple comparisons, the experienced raters reported higher levels of understanding than the inexperienced raters ($M = 8.7$ vs. 7.8), and all found rhythm somewhat harder to understand ($M = 7.4$ for both groups). For the transcript-based measures, the ratings, submitted to a similar two-way ANOVA, yielded a significant main effect of group, $F(1, 17) = 6.81$, $p = .019$, a significant main effect of category, $F(5, 80) = 6.50$, $p < .001$, and significant two-way interaction, $F(5, 80) = 3.51$, $p = .006$. According to Bonferroni-corrected multiple comparisons, the inexperienced raters, compared with the experienced ones, showed more difficulty understanding grammatical accuracy ($M = 7.1$ vs. 8.8) and complexity ($M = 6.8$ vs. 8.4). Overall, the inexperienced raters judged their understanding of grammatical complexity ($M = 6.8$) significantly lower than their understanding of lexical appropriateness ($M = 8.4$) and richness ($M = 8.3$).

Linguistic influences on comprehensibility and accentedness

The second objective of this study was to examine the relationship between the 11 rated variables and listeners' judgments of comprehensibility and accentedness as a function of rater experience. For this analysis, experienced and inexperienced raters' mean comprehensibility and accentedness ratings were submitted to a two-way ANOVA, with group (experienced, inexperienced) as a between-subjects factor and rating (comprehensibility, accentedness) as a within-subjects factor. The ANOVA yielded a significant main effect of group, $F(1, 78) = 454.33$, $p < .0001$, and a significant main effect of rating, $F(1, 78) = 114.03$, $p < .0001$, but no significant two-way interaction, $F(1, 78) = 0.88$, $p = .35$. Regardless of rating type, the experienced raters overall gave higher (more positive) ratings than the inexperienced raters ($M = 587$ vs. 480 on a 1,000-point scale), and all raters, irrespective of experience, assigned higher ratings to comprehensibility than to accentedness ($M = 599$ vs. 468).

The next analyses directly targeted the overarching research aim, namely, to examine which of the 11 rated variables are associated with comprehensibility and which are linked to accentedness. Because interrater agreement for all rated linguistic dimensions was high (Cronbach's $\alpha = .91–.97$, Table 1) and because rater judgments (except story cohesion) were significantly correlated with the relevant measures by linguistically trained coders (see Tables 2 and 3), mean scores across all 20 raters were computed for each of the 11 rated variables in subsequent analyses. It was assumed that these composite scores would reflect 'global' judgments of native-speaking interlocutors varying in degree of linguistic and pedagogical experience. To examine the overall relationship between the 11 rated variables on the one hand and comprehensibility and accentedness on the other, a Principal Component Analysis with a Varimax rotation was first run to uncover any underlying factors among the 11 rated variables based on the clustering of these variables. This analysis generated a straightforward two-factor solution with eigenvalues greater than 1.0 (the Kaiser criterion for retaining eigenvalues; Stevens 2002), accounting for 89.2 percent of variance in raters' judgments. The first factor, which included the rated variables of vowel and consonant errors, word stress, intonation, rhythm, and speech rate, was labeled 'pronunciation'. The second

Table 5: Summary of a two-factor solution based on a principal component analysis of the 11 rated linguistic variables

Factor 1 (Pronunciation)	Vowel and consonant errors (.80), word stress (.83), intonation (.90), rhythm (.87), speech rate (.85)
Factor 2 (Lexicogrammar)	Lexical appropriateness (.83), lexical richness (.76), grammatical accuracy (.87), grammatical complexity (.79), story richness (.83), story cohesion (.85)

Note: All eigenvalues > 1.0.

factor, which included the rated variables of lexical appropriateness and richness, grammatical accuracy and complexity, story richness and cohesion, was labeled 'lexicogrammar' (see Table 5). Although this factor will henceforth be referred to as 'lexicogrammar', it encompassed not only measures of lexis and grammar but also discourse-level measures of story richness and cohesion.

The resulting factor scores for pronunciation (Factor 1) and lexicogrammar (Factor 2) for each of the 40 speech samples were then submitted to two separate stepwise multiple regressions, with comprehensibility and accentedness as the dependent variables. Although the two regression models accounted for roughly the same amount of variance (90.3 percent for comprehensibility, 88.1 percent for accentedness), the relative weights of the pronunciation and lexicogrammar factors differed (see Table 6). The pronunciation factor alone accounted for most variance in accentedness (60 percent explained). In contrast, the pronunciation (50 percent) and the lexicogrammar (40 percent) factors jointly contributed to comprehensibility.

In the final two analyses, two sets of partial correlations were performed to determine the relative contribution of the rated variables *within* the pronunciation and lexicogrammar factors to comprehensibility and accentedness. The first set examined which rated pronunciation variables (vowel and consonant errors, word stress errors, intonation, rhythm, speech rate) were related to comprehensibility and which to accentedness when the influence of the rated lexicogrammar variables was partialled out ($\alpha = .01$). As shown in Table 7, all pronunciation variables were significantly correlated with comprehensibility and accentedness. Fisher *r*-to-*z* transformations showed that the strength of associations between the five pronunciation variables and comprehensibility was comparable ($p > .05$). In contrast, vowel and consonant errors ($p = .001$) and word stress ($p = .01$) were related to accentedness more strongly than was speech rate. Thus, while all pronunciation variables were equally strongly linked with comprehensibility, vowel and consonant errors and word stress contributed to accentedness more than speech rate did.

A comparable set of partial correlations was performed to examine how the six rated lexicogrammar variables (lexical appropriateness and richness,

Table 6: Results of multiple regression analyses using the factors of pronunciation and lexicogrammar as predictors of comprehensibility and accentedness

Predicted variable	Predictor variables	*Adjusted* R^2	R^2 *change*	*F*	*p*
Comprehensibility	Pronunciation	.50	.50	40.62	.0001
	Lexicogrammar	.90	.40	147.49	.0001
Accentedness	Pronunciation	.60	.60	56.83	.0001
	Lexicogrammar	.88	.28	137.54	.0001

Note: The variables entered into the regression equation were the two factors obtained in the Principal Component Analysis reported in Table 4.

Table 7: Partial correlations between five rated pronunciation variables and comprehensibility and accentedness, with the influence of six lexicogrammar variables controlled

Pronunciation variable	Comprehensibility	Accentedness
Vowel/consonant errors	.75*	.88*
Word stress	.62*	.81*
Intonation	.54*	.57*
Rhythm	.79*	.70*
Speech rate	.66*	.47*

Note: *$\alpha < .01$ (Bonferroni corrected). The variables partialled out from each correlation include lexical appropriateness and richness, grammatical accuracy and complexity, and story richness and cohesion.

Table 8: Partial correlations between six rated lexicogrammar variables and comprehensibility and accentedness, with the influence of five pronunciation variables controlled

Lexicogrammar variable	Comprehensibility	Accentedness
Lexical appropriateness	.68*	.23
Lexical richness	.55*	.31
Grammatical accuracy	.52*	.23
Grammatical complexity	.51*	.37
Discourse richness	.62*	.30
Discourse cohesion	.40	.23

Note: *$\alpha < .008$ (Bonferroni corrected). The variables partialled out from each correlation include vowel and consonant errors, word stress, intonation, rhythm, and speech rate.

grammatical accuracy and complexity, story richness and cohesion) related to comprehensibility and accentedness when the influence of the five rated pronunciation variables was partialled out ($\alpha = .008$). As shown in Table 8, most lexicogrammar and discourse variables (except story cohesion) played an important role in comprehensibility, but none of these variables were significantly associated with accentedness. Fisher *r*-to-*z* transformation identified no significant difference in correlation strength either for comprehensibility or for accentedness.

DISCUSSION

The first objective of this study was to examine how experienced and inexperienced listeners use 11 rated categories to assess phonological, lexical,

grammatical, and discourse aspects of L2 speech. Results indicated that the 20 raters' judgments for all but one variable (story cohesion) were internally consistent and were also significantly associated with the relevant coded measures based on linguistic analyses of L2 speech. In line with Derwing *et al.*'s (2004) suggestion that rater judgments can be used as a dependable measure of L2 temporal fluency, the findings of this study showed that raters can also reliably evaluate multiple linguistic aspects of L2 speech, including vowel/consonant accuracy, word stress, intonation, rhythm, speech rate, lexical appropriateness, lexical richness, grammatical accuracy, and story richness. However, the raters reported more difficulty understanding the linguistic concepts in certain categories (e.g. grammatical complexity, rhythm) than others (e.g. lexical appropriateness and richness, speech rate). In addition, the raters with linguistic and pedagogical experience, compared with inexperienced raters, overall (i) provided higher (more lenient/positive) judgments of accentedness and comprehensibility, (ii) better understood the linguistic concepts of the rated measures, and (iii) were more consistent in evaluating complex and less intuitive linguistic variables, such as grammatical complexity. These results extend previous research, showing that rater experience impacts L2 speech judgments (Calloway 1980; Thompson 1991) and that characteristics of rater training (e.g. access to appropriate terminology) and background (e.g. familiarity with accents) may bias rater behavior in unwanted ways, for instance, by compromising the reliability of aural testing (Isaacs and Thomson 2013; Winke *et al.* 2013).

The second objective was to determine how raters' judgments of 11 linguistic variables related to the constructs of comprehensibility and accentedness. Although comprehensibility and accentedness were once again shown to be overlapping dimensions, much of the variance in the raters' accentedness judgments (60 percent out of 88 percent of variance explained) was accounted for by the pronunciation rather than the lexicogrammar factor. And within the pronunciation factor, vowel and consonant errors and word stress most strongly related to accentedness, compared, for instance, with speech rate. As for comprehensibility, the raters seemed to rely on both the pronunciation (50 percent of variance) and the lexicogrammar (40 percent of variance) factors (see Table 5).

First and foremost, these findings are consistent with the claims that comprehensibility and accentedness are generally overlapping but essentially distinct constructs (Derwing and Munro 2009). In particular, they show that, for listeners, the extent of accentedness or perceived nativelikeness is likely determined mostly by the accuracy of segmental and word stress production, rather than by how fluent the speech sounds (see also Derwing *et al.* 2004). These results, which are based on rater judgments, are also in agreement with findings from previous analyses of the same speech samples by linguistically trained coders (Isaacs and Trofimovich 2012; Trofimovich and Isaacs 2012). Taken together, this research points to the general conclusion that listener perception of accentedness is strongly linked to phonological (pronunciation)

aspects of speakers' oral production, whereas perceived comprehensibility is related to several linguistic variables spanning the dimensions of pronunciation, lexis, grammar, and discourse structure.

These differences in rater behavior with respect to comprehensibility versus accentedness likely reflect the demands of each rating task. In terms of understanding, listeners likely attend to every piece of linguistic information available in L2 speech to extract as much meaning as possible from it, and the listening effort associated with this is reflected in listeners' comprehensibility judgment. For accentedness, listeners may pay particular attention to pronunciation, at the expense of other linguistic dimensions, to indicate how L2 speech differs from their own speech or that of their linguistic community. That is, in line with the operational definitions used here, there appears to be an empirical basis for claiming that comprehensibility is associated with all aspects of speech that contribute to listener effort in extracting the overall meaning of an utterance, whereas accentedness has more to do with the *manner* of speakers' productions (as opposed to its associative content). For the listener, accentedness judgments are likely 'automatic' in that they are fast and effortless and require little linguistic content to be performed. For example, Munro *et al*. (2010) showed that native-speaking listeners could reliably detect accented speakers in content-masked speech (utterance played backward) from a single word available for judgment. From an assessment standpoint, then, comprehensibility can be used to capture the extent to which L2 speakers have reached a certain threshold of phonological, lexical, grammatical, and discourse structure requirements for their conversational partners to comprehend their speech more easily. In contrast, accentedness can be viewed as an index of the extent to which L2 speakers have mastered the phonological aspects of L2 speech, notably, segments and prosody (e.g. word stress). In fact, the finding that all raters in this study overall rated the speakers higher in comprehensibility than in accentedness is consistent with this view. A speaker may reach a threshold of comprehensibility while still being fairly accented (Munro and Derwing 1995; Jenkins 2000; Derwing and Munro 2009).

PEDAGOGICAL IMPLICATIONS

These findings have several implications for teaching L2 oral skills. In this study, a brief rater training session featuring a few speech samples for practice enabled novice raters to reliably rate various domains of L2 speech. Therefore, teachers should be made aware that they can evaluate their learners' speech for a variety of linguistic dimensions, including individual sounds, suprasegmentals, fluency, vocabulary, grammar, and discourse, using simple scalar judgments on Likert or Likert-type scales. And because comprehensibility is linked to numerous linguistic dimensions, teachers should also be encouraged to address multiple linguistic aspects of learner speech through instruction, as a way of helping learners improve comprehensibility and to promote their

communicative success in real-time aural/oral interactions. Although many teachers have started to pay attention to comprehensibility in the teaching of L2 speaking (Foote *et al*. 2010), teachers are often unaware of how they can reliably assess linguistic domains relevant to comprehensibility (rather than accent). The findings of the study suggest that teachers can feel confident in their own ability to assess comprehensibility based on relatively short audio excerpts (e.g. 30s in length), and even short amounts of training can enable teachers and even untrained listeners to assess pronunciation, fluency, vocabulary, grammar, and discourse aspects of L2 speech. However, more specific pedagogical implications of the current findings will require further research. For example, it would be interesting to determine if teachers can conduct live and ongoing assessment during in-class activities throughout instructional sequences (Saito 2015) or to examine the role of L2 learners' self- and peer-assessment/feedback in their development of L2 comprehensibility (Trofimovich *et al*. 2015).

Next, as suggested by this study, in order to promote learners' communicative success through instruction, teachers should not restrict their teaching only to pronunciation targets, such as individual sounds (e.g. Munro and Derwing 2006), syllable structure (e.g. Couper 2006), or word stress (e.g. Field 2005). Teachers should also focus on fluency, illustrating how it affects comprehensibility (Derwing *et al*. 2004), and on grammar, especially because listeners find grammar errors in L2 speech both serious and distracting (Derwing *et al*. 2002). In addition, teachers may consider targeting lexical aspects of learner speech. For instance, previous research has shown that L2 learners may need to increase their vocabulary size beyond the first 2,000 word families to be able to understand everyday spoken discourse (e.g. Van Zeeland and Schmitt 2013) and other speech genres (e.g. Webb and Rodgers 2009). However, in this study focusing on L2 vocabulary in speech output, it was not how many infrequent words speakers produced (i.e. beyond the first 2,000 word families) but rather how many different words they included in oral narratives (i.e. type frequency) that was linked to lexical richness and, ultimately, to comprehensibility. In essence, those speakers who chose conceptually and contextually appropriate words and delivered them at an optimal rate, speaking without much hesitation and repetition, were judged to be more comprehensible. Thus, our findings suggest that L2 learners should not only expand their vocabulary beyond frequent and familiar words but also broaden their accurate and fluent use of different types of frequent and familiar words (see Saito *et al*. 2015c).

Lastly, the relevance of both pronunciation and lexicogrammar factors to comprehensibility leads us to call for a more integrative teaching approach which highlights a communicative focus on pronunciation, vocabulary, and grammar form, especially in the context of meaning-oriented and content-based classrooms. This goal can be achieved through communicative activities designed to create obligatory contexts for eliciting learners' use of a specific phonological, lexical, and grammatical feature (Ellis 2003), or through such

corrective feedback techniques as recasts (reformulation of an erroneous utterance without the error) or prompts (elicitation of the correct utterance) targeting phonological, lexical, and grammatical errors in learners' speech (Lyster and Saito 2010). These and other similar pedagogical techniques have already been featured in many teaching materials focusing on pronunciation (Celce-Murcia *et al*. 2010), vocabulary (Nation 2008), and grammar (Nassaji and Fotos 2011). However, to our knowledge, very few teacher-friendly textbooks have to date taken an integrative approach to improving L2 learners' speaking ability with a focus on phonological, temporal, lexical, grammatical, and discourse-related aspects of language (e.g. Grant 2010; Gorsuch *et al*. 2012). Building on such existing resources, material developers are expected to further extend and elaborate the integrative approach toward teaching L2 speaking.

CONCLUSION

Two broad conclusions can be drawn from the findings. The first conclusion is that even naïve listeners can reliably evaluate phonological, lexical, grammatical, and discourse structure aspects of L2 speech using scalar ratings, and those with teaching and linguistic experience, in particular, can be more consistent in evaluating complex and less intuitive linguistic categories. The second conclusion is that listeners chiefly rely on L2 pronunciation (vowel/consonant errors, word stress) when judging accentedness. However, they consider a much broader range of linguistic dimensions (vowel/consonant errors, word stress, fluency, lexis, grammar) when judging comprehensibility. These findings suggest that ease of understanding (comprehensibility) and linguistic nativelikeness (accentedness) constitute two overlapping but distinct goals of L2 oral skill development, and that learners may need to attain a certain threshold of oral ability—not only in terms of pronunciation but also lexicogrammar—in order to achieve L2 communicative success.

These results suggest several promising lines of future research. First, the current data set included only speech samples by French learners of English in a picture description task. In addition, these speech samples were rated by listeners from Montreal who reported relatively high familiarity with French-accented English. Therefore, future research needs to test the generalizability of listener ratings to other contexts by targeting L2 learners from different first-language backgrounds (e.g. Indo-European vs. Asian languages: Crowther *et al*. 2014) and proficiency levels (e.g. beginner vs. intermediate vs. advanced: Saito *et al*. 2015a) under multiple task conditions or interactional patterns (e.g. monologue vs. interview: Crowther *et al*. 2015) and with various kinds of raters (e.g. native vs. non-native, familiar vs. unfamiliar with L2 speech: Saito and Shintani 2015; Saito *et al*. 2015b). In particular, future research on the rater facet will reveal a number of implications for L2 speech assessment and teaching in a globalized society where non-native

speakers interact not only with native speakers but also with other non-native speakers (Jenkins 2000).

Furthermore, our findings (especially those pertaining to measures of lexicogrammar and discourse) need to be interpreted with caution due to relatively short speech samples used in this study (approximately 30 s). Although short audio recordings are sufficient for pronunciation and fluency analyses of speech (Derwing and Munro 2009), robust lexical analyses may require longer speech samples (e.g. 3–5 min in Crossley *et al.* 2014; cf. Saito *et al.* 2015c, 2015d). Finally, because the majority of the 11 linguistic variables were found to be accessible to even inexperienced listeners with the provision of user-friendly definitions and a brief training session, it would be interesting to examine the degree to which various linguistic dimensions of L2 speech are associated with comprehensibility for non-native listeners. It would also be worth investigating whether providing L2 learners with focused instruction targeting several linguistic aspects of comprehensibility can help learners to notice, practice, and ultimately automatize various linguistic dimensions of speech (i.e. accurate segmental pronunciation with varied and adequate prosody, optimal speech rate, proper lexicogrammar usage). This and other research has the potential to identify the kinds of instructional materials and methods that will help learners improve their comprehensibility as a way of attaining greater L2 communicative success in real-time communication.

SUPPLEMENTARY DATA

Supplementary material is available at *Applied Linguistics* online.

REFERENCES

Anderson-Hsieh, J., R. Johnson, and **K. Koehler**. 1992. 'The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure,' *Language Learning* 42: 529–55.

Boersma, P., and **D. Weenink**. 2010. *'Praat: Doing Phonetics by Computer [computer program]* (version 5.1.29),' available at www.praat.org.

Bongaerts, T., C. van Summeren, B. Planken, and **E. Schils**. 1997. 'Age and ultimate attainment in the pronunciation of a foreign language,' *Studies in Second Language Acquisition* 19: 447–65.

Bosker, H. R., A. F. Pinget, H. Quené, T. Sanders, and **N. H. de Jong**. 2013. 'What makes speech sound fluent? The contributions of pauses, speed and repairs,' *Language Testing* 30: 159–75.

Calloway, D. R. 1980. 'Accent and the evaluation of ESL oral proficiency' in J. W. Oller Jr and K. Perkins (eds): *Research in Language Testing*. Newbury House, pp. 102–15.

Celce-Murcia, M., D. Brinton, J. Goodwin, and **B. Griner**. 2010. *Teaching Pronunciation: A Course Book and Reference Guide*. Cambridge University Press.

Cobb, T. 2010. 'The compleat lexical tutor [website],' available at www.lextutor.ca.

Couper, G. 2006. 'The short and long-term effects of pronunciation instruction,' *Prospect* 21: 46–66.

Crossley, S. A., T. Salsbury, and **D. S. McNamara**. 2014. 'Assessing lexical proficiency using analytic ratings: a case for collocation accuracy,' *Applied Linguistics* 35. Advance online publication. doi: 10.1093/applin/amt056

Crowther, D., P. Trofimovich, T. Isaacs, and **K. Saito**. 2015. 'Does speaking task affect

second language comprehensibility?,' *Modern Language Journal* 99: 80–95.

Crowther, D., P. Trofimovich, K. Saito, and **T. Isaacs**. 2014. 'Second language comprehensibility revisited: investigating the effects of learner background,' *TESOL Quarterly*. Advance online publication. doi: 10.1002/tesq.203.

Derwing, T. M. 2003. 'What do ESL students say about their accents?,' *Canadian Modern Language Review* 59: 545–64.

Derwing, T. M. and **M. J. Munro**. 2009. 'Putting accent in its place: Rethinking obstacles to communication,' *Language Teaching* 42: 476–90.

Derwing, T. M., M. J. Rossiter, and **M. Ehrensberger-Dow**. 2002. 'They speaked and wrote real good: Judgements of nonnative and native grammar,' *Language Awareness* 11: 84–99.

Derwing, T. M., M. J. Rossiter, M. J. Munro, and **R. I. Thomson**. 2004. 'L2 fluency: Judgments on different tasks,' *Language Learning* 54: 655–79.

Ellis, R. 2003. *Task-Based Language Learning and Teaching*. Oxford University Press.

Fayer, J. M. and **E. Krasinski**. 1987. 'Native and nonnative judgments of intelligibility and irritation,' *Language Learning* 37: 313–26.

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

Flege, J., M. Munro, and **I. R. A. MacKay**. 1995. 'Factors affecting degree of perceived foreign accent in a second language,' *Journal of the Acoustical Society of America* 97: 3125–34.

Foote, J., A. Holtby, and **T. Derwing**. (2011). 'Survey of the teaching of pronunciation in adult ESL programs in Canada, 2010,' *TESL Canada Journal* 29: 1–22.

Gorsuch, G., C. Meyers, L. Pickering, and **D. Griffee**. 2012. *English Communication for International Teaching Assistants*. Waveland Press.

Grant, L. 2010. *Well Said: Pronunciation for Clear Communication*. Thomson/Heinle & Heinle.

Hahn, L. 2004. 'Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals,' *TESOL Quarterly* 38: 201–23.

Isaacs, T. and **R. I. Thomson**. 2013. 'Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions,' *Language Assessment Quarterly* 10: 135–59.

Isaacs, T. and **P. Trofimovich**. 2012. 'Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings,' *Studies in Second Language Acquisition* 34: 475–505.

Jenkins, J. 2000. *The Phonology of English as an International Language*. Oxford University Press.

Kang, O., D. Rubin, and **L. Pickering**. 2010. 'Supresegmental measures of accentedness and judgments of language learner proficiency in oral English,' *Modern Language Journal* 94: 554–66.

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*. Palgrave Macmillan, pp. 245–70.

Long, M. H. 1996. 'The role of the linguistic environment in second language acquisition' in W. C. Ritchie and T. K. Bhatia (eds): *Handbook of Language Acquisition: Second Language Acquisition*. Academic Press, pp. 413–68.

Lyster, R. and **K. Saito**. 2010. 'Corrective feedback in classroom SLA: A meta-analysis,' *Studies in Second Language Acquisition* 32: 265–302.

Mackey, A. and **J. Philps**. 1998. 'Conversational interaction and second language development: Recasts, responses and red herrings?,' *Modern Language Journal* 82: 338–56.

Mackey, A., S. Gass, and **K. McDonough**. 2000. 'How do learners perceive interactional feedback?,' *Studies in Second Language Acquisition* 22: 471–97.

Munro, M. and **T. Derwing**. 1995. 'Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech,' *Language and Speech* 38: 289–306.

Munro, M. and **T. Derwing**. 2006. 'The functional load principle in ESL pronunciation instruction: An exploratory study,' *System* 34: 520–31.

Munro, M. J., T. M. Derwing, and **C. Burgess**. 2010. 'Detection of nonnative speaker status from content-masked speech,' *Speech Communication* 52: 626–37.

Nassaji, H. and **S. Fotos**. 2011. *Teaching Grammar in Second Language Classrooms: Integrating Form-focused Instruction in Communicative Context*. Routledge.

Nation, I. S. P. 2008. *Teaching Vocabulary: Strategies and Techniques*. Thomson Heinle.

Rossiter, M. J. 2009. 'Perceptions of L2 fluency by native and non-native speakers of English,' *Canadian Modern Language Review* 65: 395–412.

Saito, K. 2015. 'Communicative focus on L2 phonetic form: Teaching Japanese learners to perceive and produce English /r/ without explicit instruction,' *Applied Psycholinguistics* 36: 377–409.

Saito, K. and **N. Shintani**. 2015. 'Do native speakers of North American and Singapore English differencially perceive second language comprehensibility?,' *TESOL Quarterly*. Advance online publication. doi: 10.1002/tesq.234.

Saito, K., P. Trofimovich, and **T. Isaacs**. 2015a. 'Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels,' *Applied Psycholinguistics.* Advance online publication. doi: 10.1017/S0142716414000502.

Saito, K., P. Trofimovich, T. Isaacs, and **S. Webb**. 2015b. 'Re-examining phonological and lexical correlates of second Language comprehensibility: The role of rater experience' in T. Isaacs and P. Trofimovich (eds): *Interfaces in Second Language Pronunciation Assessment: Interdisciplinary Perspectives*. Multilingual Matters.

Saito, K., S. Webb, P. Trofimovich, and **T. Isaacs**. 2015c. 'Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness and sense relations,' *Studies in Second Language Acquisition*. Advance online publication. doi: 10.1017/S0272263115000297.

Saito, K., S. Webb, P. Trofimovich, and **T. Isaacs**. 2015d. 'Lexical correlates of comprehensibility versus accentedness in second language speech,' *Bilingualism: Language and Cognition*. Advance online publication. doi: 10.1017/S1366728915000255.

Shintani, N., S. Li, and **R. Ellis**. 2013. 'Comprehension-based versus production-based grammar instruction: A meta-analysis of comparative studies,' *Language Learning* 63: 296–329.

Stevens, J. P. 2002. *Applied Multivariate Statistics for the Social Sciences*, 4th edn. Lawrence Erlbaum.

Tajima, K., R. Port, and **J. Dalby**. 1997. 'Effects of temporal correction on intelligibility of foreign-accented English,' *Journal of Phonetics* 25: 1–24.

Thompson, I. 1991. 'Foreign accents revisited: The English pronunciation of Russian immigrants,' *Language Learning* 41: 177–204.

Trofimovich, P. and **T. Isaacs**. 2012. 'Disentangling accent from comprehensibility,' *Bilingualism: Language and Cognition* 15: 905–16.

Trofimovich, P., T. Isaacs, S. Kennedy, K. Saito, and **D. Crowther**. 2015. 'Flawed self-assessment: Investigating self- and other-perception of second language speech,' *Bilingualism: Language and Cognition.* Advance online publication. doi: 10.1017/S1366728914000832.

Van Zeeland, H. and **N. Schmitt**. 2013. 'Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension?,' *Applied Linguistics* 34: 457–79.

Varonis, E. M. and **S. Gass**. 1982. 'The comprehensibility of nonnative speech,' *Studies in Second Language Acquisition* 4: 114–36.

Webb, S. and **M. P. H. Rodgers**. 2009. 'The vocabulary demands of television programs,' *Language Learning* 59: 335–66.

Winke, P., S. Gass, and **C. Myford**. 2013. '''Raters'' L2 background as a potential source of bias in rating oral performance,' *Language Testing* 30: 231–52.

Winters, S. and **M. G. O'Brien**. 2013. 'Perceived accentedness and intelligibility: The relative contributions of F0 and duration,' *Speech Communication* 55: 486–507.

Zielinski, B. W. 2008. 'The listener: No longer the silent partner in reduced intelligibility,' *System* 36: 69–84.

NOTES ON CONTRIBUTORS

Kazuya Saito is a lecturer in second language acquisition in the Department of Applied Linguistics and Communication at Birkbeck, University of London, London, UK. His research highlights the complex mechanism underlying adult second language (L2) speech in terms of assessment (e.g., which linguistic errors are relatively detrimental to native speakers' L2 speech judgements?), development (e.g., how do adult L2 learners differentially enhance their oral ability in various naturalistic/classroom contexts?), and teaching (e.g., how can we optimize adult L2 learning processes via various types of focus-on-form instructional options). *Address for correspondence*: Kazuya Saito, Department of Applied Linguistics and Communication, Birkbeck, University of London, London, UK. <*k.saito@bbk.ac.uk*>

Pavel Trofimovich is a professor of applied linguistics in the Department of Education at Concordia University, Canada. His research focuses on cognitive aspects of second language (L2) processing, phonology, sociolinguistic aspects of second language acquisition, and the teaching of L2 pronunciation. He currently serves as the editor of *Language Learning*. *Address for correspondence*: Pavel Trofimovich, Department of Education, Montreal, QC, Canada.

Talia Isaacs is a Senior Lecturer in Education and Director of the Second Language Speech Lab at the University of Bristol. With work at the interface between SLA and assessment, her research focuses on rater behaviour, scale validation, and oral performance in high-stakes testing and classroom settings. Talia is currently an expert member of the European Association for Language Testing and Assessment and serves on the Editorial Boards of Language Assessment Quarterly, Language Testing, and Journal of Second Language Pronunciation. Her work has appeared in *Applied Psycholinguistics, Health Communication, Language Assessment Quarterly, Studies in Second Language Acquisition,* and *TESOL Quarterly*. *Address for correspondence*: Talia Isaacs, Graduate School of Education, University of Bristol, Bristol, UK.