Article

LANGUAGE TESTING

Language Testing
2018, Vol. 35(2) 193–216

DOI: 10.1177/0265532217703433
journals.sagepub.com/home/ltj
SAGE

Developing a user-oriented second language comprehensibility scale for English-medium universities

Talia Isaacs
University College London, UK

Pavel Trofimovich
Concordia University, Canada

Jennifer Ann Foote
University of Alberta, Canada

Abstract

There is growing research on the linguistic features that most contribute to making second language (L2) speech easy or difficult to understand. Comprehensibility, which is usually captured through listener judgments, is increasingly viewed as integral to the L2 speaking construct. However, there are shortcomings in how this construct is operationalized in L2 speaking proficiency scales. Moreover, teachers and learners have little practical means of benefiting from research pinpointing the properties of learners' oral performance that optimize or hinder their ability to be understood. There is thus the need for a tool to guide teachers on what to focus on in instruction in order to target more effectively the linguistic factors that matter most for being understood and to raise learners' awareness about their abilities. To address this gap, this article reports on the development of an L2 English comprehensibility scale targeting the degree of perceived listener effort required for understanding L2 speech. The starting point was Isaacs and Trofimovich's (2012) preliminary 3-level empirically based L2 English comprehensibility scale, restricted for use with learners from one first language (L1) background on a single task. Through focus group consultations and piloting involving nine Canada- and UK-based English for Academic Purposes teachers (target end-users) rating international university students' speech samples drawn from Isaacs and Trofimovich's (2011) unpublished corpus, the instrument was expanded

Corresponding author:
Talia Isaacs, UCL Centre for Applied Linguistics, UCL Institute of Education, University College London, 20 Bedford Way, London WC1H 0AL, UK.
Email: talia.isaacs@ucl.ac.uk

to a 6-level scale through iterative revisions. The resulting formative assessment tool is intended for use with pre- and in-sessional university students from mixed L1 backgrounds on academic extemporaneous speaking tasks to support their oral language development.

Keywords
Assessing speaking, comprehensibility, English for Academic Purposes, pronunciation, rating scale development, second language learners

Challenges in operationalizing constructs in pronunciation research and assessment

Pronunciation – defined here as comprising both segmental (i.e., vowels and consonants) and prosodic (e.g., stress, rhythm, intonation) dimensions – is likely the linguistic component most amenable to diagnostic assessment. For example, discrete-point pronunciation items were prominently featured in Lado's seminal book *Language Testing* (1961) based on the assessment of differences between learners' first (L1) and second language (L2) inventories to promote mastery of L2 structures. Because sound- and rhythm-related pronunciation features can be easily itemized (e.g., aspiration, linking, syllable length), they have also been included in checklists, which represent a hallmark of diagnostic assessment (Alderson, Brunfaut, & Harding, 2015). More recent diagnostic approaches to pronunciation demonstrate how testing selected features, typically through transcriptions of learner speech, can be used to pinpoint learner perception and/or production errors to generate individualized pronunciation profiles (e.g., Gilbert, 2012). Although many pronunciation diagnostic tests suffer from task-specific limitations, such as unwanted spelling–sound correspondence effects or learners' use of avoidance strategies, pronunciation lends itself well to diagnostic assessment.

Nevertheless, there are several reasons why diagnosing pronunciation may not be practical or desirable when it is divorced from other skills making up the L2 speaking construct, particularly in communicatively oriented classrooms. Such reasons include the use of an integrated curriculum (i.e., it would be artificial to separate pronunciation from speaking and listening or from grammar and lexis); individual differences resulting in highly variable pronunciation performance; teachers with limited pronunciation training and thus difficulties in targeting the identified problem areas through tailor-made instruction; and constraints in instructional time (Foote, Holtby, & Derwing, 2011; Mora & Darcy, 2017). Moreover, there is a growing consensus within the applied linguistics community that the aim of pronunciation instruction should be to target the linguistic features that count the most for L2 speech being understandable to listeners as opposed to focusing on accent reduction (Harding, 2017). Thus, not all pronunciation elements diagnosed as non-native are worthy of being prioritized in instruction and assessment.

A practical alternative to using discrete-point diagnostic items to provide learners with feedback on their strengths and weaknesses in relation to the linguistic factors that matter most for being understood is to use empirically based rating scales. Educational stakeholders, particularly in Western post-secondary institutions (the context of this study) are familiar with scales, often used in high-stakes and classroom settings. Although

it is possible to construct user-oriented scales, developed through input from target users, such as teacher-raters (Turner, 2000), rating scales are not without limitations. For example, L2 speaking scale descriptors for extemporaneous tasks often fail to capture the complexities of oral performance, and even elaborated scale descriptors may underspecify the linguistic factors that feed into raters' scoring decisions (Lumley, 2005). Further, there is a need to achieve a balance between providing rich descriptors that can be used for diagnostic purposes on the one hand, and limiting the degree of detail within the descriptors on the other, so that the instrument is not too unwieldy for end-users.

Although the past decade has witnessed a surge in research on the linguistic features most relevant for promoting effective communication (Derwing & Munro, 2015), there is, as yet, little practical means for educational stakeholders to benefit from this insight specifically with regard to pinpointing the properties of learners' oral performance that optimize or hinder their ability to be understood by their interlocutors. This suggests the need for a pedagogically oriented tool to guide teachers on what to target in instruction to supplement existing pedagogical resources, which are often based on limited research evidence regarding the aspects of pronunciation that matter most for understanding (e.g., Rogerson-Revell, 2011). In the light of this gap, this qualitative study reports on developing an L2 English comprehensibility scale for use with learners from mixed L1 backgrounds in English for Academic Purposes (EAP) settings, drawing on EAP teachers' input to inform the evolution of the scale. This article's objective is to detail the instrument development process, which involved adapting and refining a preliminary data-driven L2 comprehensibility scale from a previous study (Isaacs & Trofimovich, 2012), and to describe and disseminate the resulting tool (research product).

Comprehensibility as a target construct for scale development

There are two main dimensions, both relevant to listeners' understanding of L2 speech, which could be targeted in scale development: intelligibility, which refers to listeners' actual understanding of L2 speech, often operationalized using listeners' orthographic transcriptions (i.e., the proportion of uttered words from an L2 speech sample that the listener correctly transcribes), and comprehensibility, which denotes listeners' *perceived* ease or difficulty in understanding L2 speech, operationalized through listeners' scalar ratings (Derwing & Munro, 2015). However, in operational assessments, this distinction is not clearly upheld (Levis, 2006). Many rating scales for standardized proficiency tests use the term "intelligibility" (e.g., TOEFL, IELTS) when what is, in fact, being measured is comprehensibility, or listeners' perceived understanding (Isaacs & Trofimovich, 2012). And for many language users, it is their subjective perceptions of processing ease or difficulty in dealing with linguistic input, more so than actual performance measures, that often predict a range of cognitive and linguistic behaviors toward the L2 speech (Oppenheimer, 2008). Therefore, as a construct common to many rating scales and likely reflective of language users' general experience with L2 speech, comprehensibility was chosen for scale development in this study. When speaking scale descriptors from standardized proficiency tests or benchmarking instruments are cited in this article in reference to the notion of ease of understanding, the original terminology used in the descriptors is retained.[1]

Cross-linguistic influence in pronunciation scale development

One of the challenges associated with modelling pronunciation in rating scales is for the scale to accommodate linguistic criteria that are applicable to speakers from different L1 backgrounds. This is, in part, because differences between speaker productions often occur as a result of transfer errors, with L1 effects on L2 production being more perceptually salient to listeners for pronunciation than for such skills as grammar and lexis (Major, 2012). Therefore, speakers from different L1s are likely to be highly variable in their segmental perception and production, which complicates teaching and testing heterogeneous cohorts, making it difficult to specify, in scale descriptors, the pronunciation features that apply to multiple groups (Derwing, 2008). To elaborate, problematic phonemic contrasts for one L1 group (e.g., /b/ vs. /p/ distinction for Arabic speakers of English) often do not generalize to other groups (Swan & Smith, 2001). Hence, such aspects of pronunciation are not universal enough to feature in a scale intended for speakers from multiple language backgrounds. Notably, segmental features tend to be more L1-specific than prosodic features, which have generally been viewed as more universal in benefiting speakers from different L1 backgrounds through instruction (Derwing & Munro, 2015). However, cross-linguistic influences involving forward transfer have also been detected for prosody, including intonation, rhythm, and stress (e.g., Li & Post, 2014). In sum, cross-linguistic differences, which are perceptually salient to listeners, pose difficulties for generating diagnostically oriented pronunciation descriptors that discriminate between different ability levels and which are also applicable to speakers from varied L1 backgrounds.

Furthermore, not all pronunciation errors "count" the same in pedagogical terms, with some being more detrimental to understanding than others. This has led to the view among applied linguists that the linguistic features most likely to interfere with speakers' ability to be understood should be emphasized in pronunciation instruction and, by extrapolation, assessment (Harding, 2017). Conversely, features that contribute to an L2 accent but are inconsequential for understanding should be left aside (Derwing & Munro, 2015). This argument acknowledges that sounding like a native speaker is an unrealistic goal for most adult learners (e.g., Moyer, 2013) and is unnecessary for integrating into a new society, excelling academically, or performing adequately in most jobs. In pedagogical terms, then, targeting the linguistic aspects of speech with the most bearing on speakers' ability to be understood should guide instructional and assessment priorities.

Why a universal pronunciation scale will be confined to generic descriptors

Data-driven rating scales that explicitly build teachers' perceptions into scale development have been subject to two opposing trends. On the one hand, there has been a move, within productive skills assessment, for scales to be task- and context-specific (e.g., Turner, 2000). On the other, efforts have been made to develop "a common, universally valid descriptive system," including scales that cut across all world languages and are not specific to any particular target language or L1 group (LTRC, 2014). Although the latter goal may seem appealing, the Common European Framework of Reference for Languages

(CEFR) demonstrates problems with modelling pronunciation on a common scale (Council of Europe, 2001). Pronunciation was excluded as a criterion in the global CEFR scale as a result of misfitting pronunciation indicators during scale development (North, 2000). This measurement-driven decision, brought on by teacher-raters' inability to consistently use the pronunciation descriptors, in part reflects the inadequacy of the descriptors themselves. The CEFR Phonological Control scale is also problematic (Harding, 2017). For example, if being understood, as opposed to whether or not someone has a discernible L2 accent, is what counts the most in real-world communication, then sounding native-like should not be a criterion for achieving the highest scale level. In the Phonological Control scale, because only the descriptors at levels B1 and below allow for L1 influence, the omission of L1 effects from the higher levels of the scale (B2 and C1/C2) suggests that speech needs to be accent-free at these levels. However, most L2 speakers at even advanced levels can have detectable accents (Moyer, 2013); thinking otherwise is unrealistic.

Instead of modelling pronunciation in a universal rating system applicable to all target languages, a more sensible starting point is to develop an empirically based scale that focuses solely on one L2. This has been the approach of the English Profile Programme, which describes the "criterial features" characterizing learner English at different CEFR levels in terms of grammatical, functional, and lexical features (e.g., Hawkins & Filipović, 2012). A similar approach was adopted here to develop a data-driven comprehensibility scale targeting one L2 (English) for use by learners from mixed L1 backgrounds.

Pronunciation and high-stakes English proficiency testing in higher education

One setting where stakeholders could benefit from greater guidance on the linguistic aspects of speech most crucial for understanding is the academic domain. This is due to rising numbers of university students studying abroad, the increasing use of English as a lingua franca on campuses, and the need for institutions to provide greater support to international students despite resource constraints (Jenkins, 2014). High-stakes productive skills assessments that rely on examiners' judgments (TOEFL, IELTS) or on automated scoring systems (PTE Academic) are often used to screen international students' English proficiency for university admissions (Ginther & Elder, 2014). However, passing this entrance screening by no means guarantees that students can cope with academic tasks in their L2 (Zhang & Goodson, 2011). One challenge students may face relates to expressing themselves in the language of instruction with communication breakdowns possibly jeopardizing their academic performance (Andrade, 2006).

Although universities and test-takers normally have access to overall scores and subscores from standardized tests, they provide end-users with little information about students' strengths and weaknesses, for example, to benchmark their ability level at the outset of EAP instruction, help students with awareness-raising about learning targets, or track their progress over time. The intention here is not to argue for this as a shortcoming of standardized tests, which are generally designed to inform admissions decisions and not to chart L2 learners' language development over time. Rather, our intent is to underscore a training gap for international students needing to function in their L2 in higher

education settings. Furthermore, "intelligibility" is often included among the assessment criteria in oral proficiency scales used for university admissions. However, there are shortcomings with how this construct is operationalized. For example, in the "Delivery" subscale of the 5-level TOEFL iBT Integrated speaking rubrics (ETS, 2014), the following descriptors are used: "problems with intelligibility" (level 1), "considerable listener effort" (level 2), or "overall intelligibility" characterized as "good" (level 3), or "high" (level 4). These characterizations of the speech are roughly associated with "pacing," "pronunciation" (seemingly referring only to segmental production), and "intonation" in the descriptors, with no published guidance on how to interpret these terms. In addition, the issue of which factors manifest more or less at different ability levels or have more or less bearing on intelligibility is not indicated. The nine-level IELTS Speaking band descriptors (IELTS, 2015) have similar limitations. At level 2 of the Pronunciation subscale, pronunciation is described as "often unintelligible," with no elaboration as to which features account for difficulties in understanding. Only levels 2, 4, 6, 8, and 9 have self-contained descriptors, posing problems for raters (Isaacs, Trofimovich, Yu, & Chereau, 2015). Levels 4, 6, 8, and 9 refer to the use of a "limited range," "range," "wide range," and "full range of pronunciation features," respectively, although the specific features, how they apply to different pronunciation levels, and how they relate to intelligibility are underspecified.

Finally, the PTE Academic is scored using an automated speech recognition algorithm trained on native listeners' ratings of a large corpus of speech samples (Pearson, 2012). Although the proportion of uttered words detected as unintelligible is specified at the three lowest levels of the 6-level scale for pronunciation and oral fluency, the descriptors of "nativelike" (level 5) and "non-English" speech (level 0) at the scalar extremes suggest that what is, in fact, being measured is deviations from what human raters consider to be native speaker norms (p. 21). It is also unclear which construct human assessors were rating as a basis for building the algorithm (e.g., degree of accent, extent of understanding, or both). Similarly, in PTE Academic machine scoring, there is no explicit mechanism for assigning greater weighting to the most important factors for intelligibility as distinct from nativeness (i.e., the production of linguistic features that may have no bearing on understanding). There is therefore a need to understand the linguistic factors that underlie understandable speech and to map these in a rating instrument that does not resort to the native speaker standard at the high end of the scale.

Research aims

The goal of this study is to present EAP teachers and their students with a formative assessment tool targeting comprehensibility, with scale descriptors applicable to learners from different L1 backgrounds on extemporaneous discourse-level oral production tasks. Such a tool, presented as a prototype in this article (see Appendix), could fulfill numerous pedagogical functions. First, it could be used to identify sources of students' strengths and weaknesses with respect to the aspects of speech most relevant to their comprehensibility, guiding EAP teachers in selecting the linguistic features to target in instruction and in providing feedback to students, including monitoring their progress pre- and post-instruction. The tool could also be used to enhance teachers' pronunciation literacy (i.e.,

familiarity with basic concepts in L2 pronunciation teaching and learning), since they may not have received teacher training in pronunciation and might feel unconfident about targeting it (Foote et al., 2011). Because comprehensibility is a broader construct than pronunciation, also incorporating the dimensions of fluency and lexicogrammar (e.g., Isaacs & Trofimovich, 2012; Saito, Trofimovich, & Isaacs, 2016), the instrument could help teachers integrate pronunciation with other skills when targeting academic speaking. Moreover, this instrument could foster students' self-awareness of their "comprehensibility profile," pinpointing areas on which to focus, and could assist with calibrating the ease with which listeners understand L2 speech with students' self-perceptions of their performance. This is important because high comprehensibility speakers tend to underpredict their own comprehensibility, whereas low comprehensibility speakers appear overconfident and may be oblivious to their interlocutors' difficulty in processing their message (Trofimovich, Isaacs, Kennedy, Saito, & Crowther, 2016).

These envisioned pedagogical uses of the scale guided the research team in instrument development, with the scale intended primarily for EAP teachers as end-users. Following from an earlier study (Isaacs & Trofimovich, 2012), which developed L2 comprehensibility scale guidelines that needed to be refined to accommodate a broader range of learner L1s, tasks, and academic settings, the exploratory research question for this study was as follows: How can the linguistic criteria featured in a L2 comprehensibility scale be refined and/or expanded with EAP teacher-oriented (non-technical) descriptors that can account for language performance by international students from different L1 backgrounds on extemporaneous academic oral production tasks?

Method

Research-based rationale

The goal of the current scale development effort was not to recreate yet another general speaking proficiency scale; rather, it was to gain a better understanding of the linguistic dimensions of L2 oral productions that listeners use to evaluate *comprehensibility* so that these could be modelled in a scale. That is, the scale can be used to assess international students' comprehensibility (rather than overall proficiency) levels and to diagnose their strengths and weaknesses in relation to comprehensibility. The starting point was Isaacs and Trofimovich's (2012) preliminary empirically based L2 English comprehensibility scale. To generate this scale, 60 native English-speaking Canadian undergraduate students used a nine-point scale (easy/hard to understand) to rate the L2 comprehensibility of short speech samples by 40 adult L1 French speakers telling a picture story in English. Mean comprehensibility ratings for each L2 speaker were then correlated with 19 researcher-coded auditory and instrumental measures derived from the speech, spanning the domains of pronunciation, fluency, lexicogrammar, and discourse. By converging statistical analyses with three native English-speaking teacher-raters' introspective reports about the language variables contributing to their comprehensibility scoring decisions, it was possible to map a subset of linguistic variables that best discriminated between three comprehensibility levels in a preliminary data-driven scale. Lexical richness (types) and fluency (mean length of run) differentiated between low and intermediate

comprehensibility levels, whereas grammar (accuracy) and story breadth (number of distinct propositions) discriminated between intermediate and high levels. Word stress distinguished between all three levels. Although the scale descriptors were originally featured in a holistic scale, they are recast as an analytic scale in Table 1. This was the initial scale presented to focus group raters in this study.

Further research (not specific to rating scale development) has clarified the L2 comprehensibility construct, targeting the dimensions of L1 speaker background and task type. For example, Crowther, Trofimovich, Saito, and Isaacs (2015b) confirmed that both the dimensions of pronunciation (49%) and lexicogrammar (21%) account for the variance in L2 comprehensibility ratings. Whereas segmental errors were a barrier to English comprehensibility in the case of L1 Chinese speakers, for L1 Hindi-Urdu speakers, pronunciation fed into listeners' perceptions of how accented they sounded but did not impede their comprehensibility. Instead, it was lexicogrammar that contributed to Hindi-Urdu speakers' comprehensibility. With respect to task type, additional findings by Crowther, Trofimovich, Isaacs, and Saito (2015a) revealed an interplay between L1 background and task. L2 comprehensibility ratings in the more cognitively demanding TOEFL iBT integrated task were linked to a wider range of linguistic measures than in the less demanding IELTS long-turn task, where comprehensibility was solely associated with pronunciation and fluency for three of the four L1 speaker groups examined. Put differently, whereas pronunciation and fluency related to comprehensibility regardless of the task, lexicogrammar also played a role in a more complex task.

Scale development

The current scale development effort sought to adapt Isaacs and Trofimovich's (2012) preliminary scale by examining the generalizability of the linguistic criteria across L1 background and task type, with the primarily qualitative evidence complementing the quantitative findings from prior research. The scale was piloted in two contexts hosting large numbers of international students with different language norms (North American and British English).

L2 speech samples. Scale development proceeded in nine focus group sessions with EAP professionals evaluating and discussing multiple speech samples by university students (see below). The samples used in Session 1 included recordings of the 40 L1 French speakers performing the picture narrative from Isaacs and Trofimovich (2012). For all remaining sessions, samples were drawn from Isaacs and Trofimovich's (2011) unpublished corpus of 235 international university students' spoken English. Of these students, 150 were studying at a Quebec English-medium university in Canada (38 female, 20 L1 groups) and 85 at a UK university in southern England (65 female, 12 L1 groups). All students had resided in their host country for less than two years and thus had administratively valid English proficiency test scores. TOEFL iBT total scores for 83 of the students in the Canadian sample were low ($M = 88.75$, $SD = 9.66$) compared to the scores of only seven UK-based students who had taken the test ($M = 102.29$, $SD = 11.59$). Overall IELTS scores for the remaining students were similar for the Canada- ($M = 6.79$, $SD = .63$) and UK-based ($M = 6.66$, $SD = .63$) cohorts.

Table 1. Analytic L2 English comprehensibility scale used in Session 1.

Comprehensibility level	Fluency	Vocabulary & storytelling ability	Word stress	Grammar
High 3	Produces fluent stretches of speech. Generally only pauses or hesitates at the end of the clause.	Provides sufficient vocabulary to set the scene and propel the story plot forward. Lexical errors, if present, are not distracting.	Assigns word stress correctly in most instances.	Grammatical errors, which are infrequent, do not detract from the overall message.
Intermediate 2	Produces some fluent stretches of speech. Occasionally pauses or hesitates in the middle of the clause.	Experiences occasional lapses in vocabulary, although may roughly convey the setting or main plot of the story. Lexical errors are prevalent.	Is inconsistent in word stress placement.	Produces some grammatical errors that may detract from the overall message.
Low 1	Produces dysfluent stretches of speech. Frequently pauses or hesitates between lexical items.	Experiences frequent lapses in vocabulary that make the storyline unelaborated or indecipherable. High proportion of lexical errors, including L1 lexical influences.	Frequently misplaces word stress.	Produces frequent grammatical errors that are likely to detract from the overall message.

Four speaking task types from Isaacs and Trofimovich's (2011) corpus were used, one of which was the same picture narrative from Isaacs and Trofimovich's (2012) study. The remaining three were from publicly available standardized practice English proficiency tests. The first task was a speeded (45 second) graph description task from the retired Test of Spoken English (TSE), which was used to screen international teaching assistants' speaking ability prior to the introduction of the TOEFL iBT (ETS, 2011). The second task was the IELTS long-turn (presentation) task (1–2 minutes), in which the speaker interacted with a research assistant (who simulated an interviewer), responding to a scripted prompt about a familiar topic and, if time permitted, to a follow-up question (IELTS, 2009). The final task was the semi-direct TOEFL iBT integrated speaking task (ETS, 2009). After reading a short passage and hearing an academic lecture on a related topic, the speaker synthesized the information in a timed spoken response (1 minute). Both the TOEFL and IELTS are used as proficiency tests for university entrance purposes, making them appropriate for adapting an L2 comprehensibility scale for the academic domain.

Rating sessions. EAP "domain experts" were recruited as target end-users of the scale to consult throughout scale development (nine sessions, totaling 23 hours). These were 10 experienced EAP teachers ($M_{experience}$ = 19.1 years, 4–44) from Canada (6) and the UK (4) with Master's degree qualifications in applied linguistics and affiliated with Education or Continuing Education departments. Seven reported having taken a course in pronunciation and four had taken an assessment course as part of their degree, with an additional teacher having served as an accredited IELTS examiner (UK2) and another having designed practice TOEFL iBT tests for a Korean publishing company (UK1). Each session, which was audio-recorded and transcribed, included two or three EAP teachers and two researchers, who were either running the session or taking notes. The corpus of transcribed focus group discussions included approximately 120,100 words across nine sessions, with EAP teachers contributing 3261 comments (136 comments per rater or 362 comments per session).

The approach to rater training was methodical and involved introducing only one new variable (L1s, tasks) per session, as shown in Table 2, with the researchers revising the scale iteratively and presenting a new version at each session. The first six sessions were conducted in Canada with two EAP teacher groups; the remaining three took place in the UK, also with two teacher groups. Table 2 summarizes the goals and content of the sessions and subsequent piloting, charting the a priori objectives guiding the evolution of the tool. Because the scale was in draft form through the course of the study, it was unnecessary for raters to score the entire audio corpus using the scale-in-progress. Rather, the speech samples needed to be purposefully selected for the goal of each session to be attained (e.g., evaluating scale criteria across L1s and tasks), while ensuring that the performances to be rated and benchmark samples used in rater training spanned the entire ability range of the target population. To select diverse speech samples in terms of L1 background and L2 comprehensibility level, the 150 Canadian and 85 UK students' performance samples were pre-rated by additional raters – experienced EAP teachers from either Canada (3) or the UK (3), matched for context and recruited from the same population as domain experts. The pre-ratings, obtained using the 9-point comprehensibility scale, were used to identify 30–60 samples as exemplars per session (six of which were used in rater training) based on rating means and samples with the highest rater agreement.

Table 2. Goal and content of the focus group and pilot sessions.

EAP raters	Session	L2 speakers	Task(s)	Goal and content
Focus group 1: Can1, Can2, Can3	1	Canadian L1 French speakers	Picture narrative	• To test how well descriptors from the preliminary scale work in relation to the speech samples used to generate them. • To discuss the number of distinguishable comprehensibility levels that are practical for implementing in the scale-in-development.
	2	Canadian international students, mixed L1s		• To generate L1-neutral descriptors for the scale to be used in mixed EAP classes.
	3		TSE graph, IELTS long-turn & TOEFL integrated[b]	• To make descriptors useable for any academic extemporaneous speech task (i.e., remove descriptors specific to the picture narrative).[a]
	4			• To identify benchmark samples for subsequent sessions by pinpointing speech samples that raters considered as typifying a particular band level.
Focus group 2: Can4, Can5, Can6	5			• To present the scale to new raters evaluating another set of academic task performances for cross-validation purposes (i.e., to remove idiosyncrasies associated with the particular sample of speakers and raters).
	6			• To finalize the scale for use with Canadian international students and identify benchmark samples in cases where consensus was achieved.
Focus group 3: UK1, UK2	7	UK international students, mixed L1s		• To adapt the scale for use with international students in the UK so that the final version could be used on both Canadian and British campuses. • Because Focus group 3 included IELTS (UK1) and TOEFL (UK2) experts, to check that the scale conformed to L2 assessment community norms.
	8			• To go over definitions of key terminology in a glossary for raters in relation to the operationalization of the comprehensibility construct.
Focus group 4: UK3, UK4	9			• To confirm that the scale would be intuitive to and manageable for EAP teachers relatively unfamiliar with rating systems from standardized tests.
Final piloting: UK5	10–12[b]	All international students		• To finalize the current version of the scale (Appendix). • To examine UK5's impressions after rating 250 samples, including the equivalence of Canadian and UK samples. No differences were noted in performance quality or scale applicability to either cohort.

Note: [a]Academic task performances were randomized for rating in the focus group sessions, whereas UK5 conducted ratings by task. [b]UK5 met the first author on three occasions for rater training or to share observations about the scale and rating process.

After being briefed about the purpose of the study and instrument in the focus group sessions, the teacher-raters were provided with the latest version of the scale, discussed its suitability for use in their teaching context and the quality of the descriptors, and received initial training using the benchmark samples. They then independently rated speech samples using individual laptops and headsets (self-paced task). Finally, they engaged in a post-rating debrief to discuss any issues and to compare the scores they had assigned. They also reflected on the alignment of the speech properties relevant to comprehensibility with the scale descriptors and, in some cases, variables extraneous to the construct (e.g., speakers' intelligence). This procedure was repeated for all nine sessions using two different focus groups per context and a unique set of speech samples for cross-validation purposes (Lane & Stone, 2006).

Immediately after each session, the research team conferred about strategies for revising the scale using the researchers' written observation notes from the session. Due to the logistics of needing to schedule focus groups in close succession, in some cases with only a day between sessions, it was not possible to transcribe and code the audio data between sessions as a basis for making revisions to the scale. Instead, the first author iteratively revised the scale to address the raters' comments within a day of each session, summarizing the rationale for scale modifications based on the observation notes and incorporating feedback from the research team on draft revisions. Of the entire corpus of focus group discussions, 217 comments were identified as directly relevant to scale development (80 from Sessions 1–6, 137 from Sessions 7–9). These comments included teacher-raters' clarifications or justifications for rating decisions relative to current scale descriptors, observations related to teacher literacy, questions about terminology, and suggestions for scale descriptors or scale improvement. At the beginning of the next session, teacher-raters were presented with the revised scale and summary of all changes and discussed whether the points reflected their intended changes. Additional feedback on the quality of the scale was iteratively incorporated to evolve the instrument further.

Finally, one British EAP teacher (UK5) with 20 years of teaching experience, who was also a qualified IELTS trainer leading standardization exercises for accredited IELTS examiners worldwide, was recruited to provide further feedback on the scale in a similar procedure but with a larger sample set to rate. After reviewing the scale and confirming that it was fit-for-purpose for EAP teachers following a final round of minor modifications, UK5 used the instrument to rate 250 speech samples (18 hours). This consisted of 60 Canadian students performing the IELTS long-turn and TOEFL integrated tasks, which were the same materials used in Crowther et al. (2015a, 2015b) to examine task and L1 effects. A parallel sample of 60 UK students performing the same tasks was also used, plus a subset of five speakers from each cohort completing the TSE graph task to examine the suitability of the scale for this retired academic task. UK5 recorded written memos that arose while rating to complement focus group raters' comments.

Results and discussion

Table 3 documents the major changes to the scale or to the operationalization of comprehensibility following each session, along with a selection of teacher-raters' verbatim

Table 3. Summary of major structural or conceptual changes to the L2 English comprehensibility scale.

Session (group)	Summary of changes	Raters' quotes (verbatim comments) to justify changes
1 (1)	• Expanded the 3-level analytic scale to a 5-level scale, splitting the highest and lowest levels in two.	• Can3 [referring to classroom settings with mixed proficiency levels]: In the bunch of intermediates, some are going to be better than others. More gradations. • Can2 to Can3: I agree with you, it does need finer distinctions. • Can1: It might be interesting to think about two levels for high, two for intermediate and two for low.
2 (1)	• The "Word stress" subscale was rebranded as the "Pronunciation" subscale to make it applicable for learners from different L1 backgrounds.	• Can1: We don't have a category for difficulty in pronunciation that causes confusion, and I thought there was confusion here. • Can2: I think that what we seem to be saying is that pronunciation is hitting us harder than word stress.
3 (1)	• Clarified that speech does *not* need to sound native-like to receive the highest comprehensibility score at the top level of the Pronunciation subscale to remove ambiguity about this.	• Can1: We need to qualify this … Because even that Chinese girl that we loved, her intonation was not all, it was not native-like. I mean it was very coherent, very cute, and this is a 5, but this [descriptors stating "natural or authentic"] is a bit too strict… They're not perfect, so… • Can2 to Can1: So let's not say native-like but generally sounds authentic? • Can3 to Can2: Authentic is okay.
4 (1)	• Reordered the subscales from Fluency, Vocabulary, Pronunciation, and Grammar to Pronunciation, Fluency, Vocabulary, and Grammar due to raters' view that pronunciation was most essential for being able to communicate coherently at the lowest scale level. Grouping pronunciation and fluency together followed by vocabulary and grammar was based on raters' proposal for logically ordering these criteria and conformed to statistical clustering (e.g., Isaacs & Trofimovich, 2012; Saito et al., 2016).	• Can1: I didn't hear very many grammatical problems, is that because I couldn't understand what he was saying? But… there were problems with pronunciation. And in a case like this we're not sure if there are vocabulary problems, but for sure I would say there's dysfluent speech. • Can2 [about producing comprehensible language]: It would be the thing that opened the door for the rest of them. It should be that if this criterion [producing enough comprehensible language] is … if the answer is no, the speaker does not, then we do not perceive anything else in the language because… this is the door that opens the rest of the grid.

(Continued)

Table 3. (Continued)

Session (group)	Summary of changes	Raters' quotes (verbatim comments) to justify changes
5 (2)	• Added "0" as the lowest level of the scale to describe unassessable speech (labelled UR for Unable to Rate so as not to demotivate students), resulting in a 6-level scale.	• Can6: Zero should be an option, right? I'm just saying, if somebody cannot speak if they're just monosyllabic or just not able to complete the task? • Can5 to Can6: This is a diagnostic not an evaluation scale, so wouldn't it be that if people come in and can't speak, you just say they're not rateable, I mean, instead of a zero?
6 (2)	• Used shading to reflect the raters' contention that Vocabulary and Grammar subscales are relatively less important for comprehensibility than the Pronunciation and Fluency subscales, in line with prior findings (e.g., Isaacs & Trofimovich, 2012; Saito et al., 2016). • Added a space for raters to assign an overall numerical comprehensibility score capturing the "overall effort required" to understand the speech.	• Can5: In terms of comprehensibility, I found I was using the first two categories the most… 'cause these are two different things… The grammar errors didn't really affect comprehensibility but I could tell you that this person needed help in their grammar. But I could understand their message and that could be because we're language teachers that we just… don't hear the errors. But for me it was usually the rate of speech, the rhythm of speech and the hesitations that gave me problems 'cause I had to listen more closely, like, you had to struggle to hear. • Can6 to Can5: Vocabulary was almost the least important to me, not the least but…
7 (3)	• Expanded the notion of assigning an overall numerical comprehensibility score by adding a global "overall description of comprehensibility" scale that summarizes the effortfulness entailed in understanding L2 speech in terms of the frequency and impact of errors on listener processing. • Included and highlighted the statement about the degree of effortfulness into each cell of the analytic scale to signal its importance as the essential criterion for placing a speaker at a particular level. Clarified descriptors to provide extra detail about the language expected at each level to support or complement the main description about effortfulness. • Added a glossary of keywords used in the scale (e.g., appropriate junctures, clauses, complex sentences).	• UK2: Actually in IELTS… they have an overarching statement. Just like you said, "understanding the message is effortless although minor pronunciation errors may be present." And then typically the indent goes in further, there is a greater indent for the elements underneath that. • UK2: It would be an idea to have a sheet of some kind that explains what it means and then where you're giving examples and that would, that should make it clear… if you were to give training.

Table 3. (Continued)

Session (group)	Summary of changes	Raters' quotes (verbatim comments) to justify changes
8 (3)	• Added a dotted line separating the Pronunciation and Fluency subscales from the Vocabulary and Grammar subscales following the suggestion that the latter two criteria appear on an overleaf for optional (non-mandatory) use by EAP teachers at their discretion. Focusing only on pronunciation and fluency could mitigate raters' cognitive overload when rating, but they would still be given the option of providing learners with feedback on vocabulary and grammar.	• UK2: The first thing that comes to mind for me was the vocabulary and grammar. It was more difficult to … attend to them. I was focusing more on pronunciation and fluency and … vocabulary and grammar almost seemed … superfluous. Even though there may have been instances where I couldn't comprehend the candidates, I had a gut feeling that their grammar was probably not bad. • UK1: You sometimes find that if you're giving a 2 for pronunciation and a 2 for fluency and you're saying well how about the grammar and vocabulary. So while they were both 3s so there was no issue there. But [the overall comprehensibility score] it's not a 3 score, it's not a borderline score, you feel it's definitely a 2. So … I mean, if you could weight vocabulary and grammar … • UK1 [on which subscales to include]: It should be pronunciation and fluency for sure. I don't know if you could use grammar and vocabulary as … a citing factor if there's an issue. If you're borderline, you can say well actually his grammar and vocabulary is quite good so that's a reason for having a higher or lower score.
9 (4)	• No major changes were made (only minor rewording of existing content).	• –
10 (UK5 pilot)	• Added "depending on the task" to the descriptors at the top level of the Vocabulary and Grammar subscales to acknowledge that more complex tasks might elicit more sophisticated oral expression than easier tasks (Crowther et al., 2015a). This wording should allow for speakers to attain the top level regardless of the task. • Alphabetized the glossary, added new terms, and deleted terms no longer included in the scale.	• These suggestions arose in unrecorded pilot feedback.

comments that led to the changes. As Table 3 shows, the original analytic scale evolved into a 5- and subsequently 6-level scale after Sessions 1 and 5, respectively. The addition of a global scale to allow for summarizing speakers' overall comprehensibility level, intended to be rated prior to and independently of the analytic scale ratings, arose as a result of input from Session 7. These changes were accompanied by minor modifications to the descriptors, incorporating raters' direct wording suggestions to result in the final product (see Appendix).

The comprehensibility construct in the scale

Comprehensibility was defined for teacher-raters as follows:

> Comprehensibility is broadly defined in the research literature as how easily a listener can understand L2 speech. In this study, we will define comprehensibility in terms of listener effort in processing the speech. Comprehensibility is *not* being defined in terms of your ability to understand every word that is said…Be sure to evaluate the speech from your own perspective as a teacher who is familiar with the speaking task. Do *not* pretend that you are a naïve listener… who is unfamiliar with the context when assigning scores.

In the glossary of key terms (developed after Session 7), comprehensibility was more succinctly defined as "how effortful processing the L2 speech is from your perspective as an ESL/EFL professional, who has had presumed exposure to various L2s; that is, you are rating the degree of effort required for *your* understanding of the speech." This definition does not ask raters to compensate for their familiarity with L1 accented English, even though greater familiarity could reduce processing load and positively affect scoring (Carey, Mannell, & Dunn, 2011), which is a limitation of the present study. The decision to instruct teacher-raters to score from their own perspective rather than asking them to score as though they were lay listeners was because of teachers' familiarity with academic oral communication demands, their status as major stakeholders in student training, and their role as the intended users of the scale. The goal was to make the scale as user-friendly as possible for EAP teachers to incorporate in their classrooms, encouraging them to rely on their expertise when assessing students. In the final operationalization of comprehensibility, the level descriptor in the summary scale, rearticulated in each band of the analytic scale, ranges from "speech is effortless to understand" (level 5) to "speech is painstakingly effortful to understand" (level 1). This is described in terms of the frequency and (more crucially) the impact of errors on teachers' processing load.

One point emphasized during rater training was that the rating tool is an oral production scale. Assessing speakers' written or aural comprehension of the speaking prompt (receptive skills) based on the appropriateness of their response is beyond the remit of the scale and extraneous to the comprehensibility construct being measured. As UK5 noted, the scale cannot account for the truth-value of the output (e.g., factual accuracy of graphical data that the student describes). It is also beyond its scope to assess topic development or the strength of evidence underpinning an argument beyond the linguistic properties in the descriptors.

Design features of the scale

One strong thread emerging from the sessions and subsequent piloting, as reflected in Table 3, was that the pronunciation and fluency subscales are more important for comprehensibility than are the vocabulary and grammar subscales (e.g., see Table 3 comments by Can5, Sessions 6, and UK2, Session 8). Raters' views about the importance of pronunciation and fluency for comprehensibility are consistent with empirical findings suggesting that a greater proportion of the variance in L2 comprehensibility ratings is explained by these factors than by a lexicogrammar dimension. For instance, whereas temporal fluency distinguishes between low and intermediate comprehensibility speakers, grammatical accuracy sets higher performers apart (Isaacs & Trofimovich, 2012; Saito et al., 2016). This seems consistent with Can1's suggestion that L2 speakers need to have control of certain linguistic features (e.g., fluency) to "open the door" for raters' attention to other aspects of speech (see Table 3).

In terms of task effects, UK5 noted that "the IELTS is a nonacademic task … [which] might impact on the need to produce a range of vocab/grammatical structures." Unlike IELTS Reading and Writing, IELTS Speaking does not have an academic module and is used for immigration and university gatekeeping purposes (IELTS, 2015). UK5's view aligns with findings from Crowther et al. (2015a) that for most L1 speaker groups, "comprehensibility was associated solely with pronunciation and fluency categories" in the IELTS task (p. 80), whereas the more cognitively demanding TOEFL integrated task drew on a broader range of features linked to comprehensibility (e.g., lexicogrammar). The raters, who were unaware of these results, thus provided convergent evidence for the quantitative results from the published studies.

Although EAP teachers assigned the vocabulary and grammar subscales only a secondary role, they were neither in favor of deleting them nor of collapsing them into a single lexicogrammar subscale, which they felt would conflate too many elements. The consensus was that teachers should be given the option to use these subscales. This is visually represented by the dotted line separating the pronunciation and fluency subscales from the vocabulary and grammar subscales in the comprehensibility scale (see Appendix), which could either be folded over and disregarded or used at the teacher's discretion, depending on students' needs, comprehensibility level, and the speaking task characteristics (e.g., length, complexity).

An additional point that emerged across the different focus groups was the teacher-raters' desire for streamlining descriptors to make the scale more user-friendly, as demonstrated through the following quotes:

- It's good to use the same expressions over the scale because it makes it easier to use in class (Can2, Session 4).
- I know you made a big effort overall to use parallel structures, but the odd time that you didn't I get a little bump in my head. I would like it to be as much the same thing, and then, oh that's the one that's different (Can4, Session 6).
- You've decided to kind of like, explain pronunciation. I'm just wondering whether those wordings should really not been there … I'm just thinking about the consistency. Under fluency, you haven't put anything in brackets indicating what you mean by it. Same with vocab and grammar (UK2, Session 7).

Although the issue that UK2 raised was resolved by introducing a glossary, one challenge in elaborating the descriptors was to make the wording consistent between and within subscales while creating clear-cut distinctions between adjacent levels. For example, the researchers' summary of changes to the scale after Session 4 clarified:

> We now add the word "frequent" to "produces *frequent* pronunciation errors" in levels 1 and 2 of the pronunciation subscale to parallel descriptors in the fluency subscale. However, we add "and/or" to signal that pronunciation inaccuracies for meaning-laden words might alternatively account for the possibility that a single serious content word error could jeopardize the rater's understanding of the utterance (Isaacs & Trofimovich, 2012).

Similarly, the co-occurrence of multiple pronunciation errors (e.g., stress placement and substitution errors) could make the speech more difficult or effortful to parse (Zielinski, 2008). Following Can4's comment in Session 5 that error frequency and its impact on the listener should be reflected across scale levels, the wording further evolved to contrast "frequently confusing" (level 1) with "occasionally confusing" (level 2; subject to further changes in later sessions).

Ultimately, reference to error frequency in the descriptors, as shown in the Appendix, was modified following an anonymous *Language Testing* reviewer's comments during the peer-review process of this manuscript, with which we fully concur. The issue was that the pronunciation criteria in the previous version of the assessment tool confounded the number of errors with their severity due to the use of "and/or" in the descriptive statements. For example, with the previous wording, a score of 2 for pronunciation could have been awarded if any of the following applied: (a) errors are frequent; (b) errors are detrimental to the message; and (c) errors are both frequent and detrimental to the message. However, in instances where pronunciation errors are frequent but not detrimental to the message, the speaker should not be penalized for reduced comprehensibility. Therefore, it is the effect of the error on the listeners' understanding and not its frequency that needs to be highlighted. This change is consistent with the notion that the presence of even a large number of benign errors in a speaker's output is much less important than a single serious error. To parallel this change to the pronunciation subscale and until there is more evidence to the contrary, we also removed any reference to error frequency in the vocabulary and grammar subscales. In sum, the scale development process was organic and was informed by teacher-raters' comments (including to enhance user-friendliness), the researchers' knowledge of relevant research evidence, and anonymous external experts' feedback on the tool as a final check. This enabled us to develop a principled basis for incorporating recommendations that emerged from the focus groups and academic experts.

Future research and conclusion

In Session 8, UK1 and UK2 felt that pronunciation and fluency subscores should be weighed more heavily than vocabulary and grammar subscores if they are to be summed as an overall comprehensibility score (see Table 3). However, it is premature to take up this suggestion without having completed quantitative piloting. In the quantitative instrument validation that will follow this scale development project, a larger group of EAP teachers will use the scale to rate speech samples from Isaacs and Trofimovich's (2011) corpus to

enable various analyses, including the distribution of scores across levels, the over- or underuse of scale bands, interrater reliability and fit statistics, the relationship between the scores and known linguistic properties of the speech from previous research, and the extent to which scalar ratings predict students' TOEFL or IELTS university entrance scores for speaking. The assumption behind the latter analysis is that the speaking sections of these tests, although not targeting comprehensibility, should nonetheless capture this element of speaking ability, providing a benchmark for validating the prototype tool.

The results of this study and the follow-up validation study will then inform the development of user manuals and materials preceding the roll-out of the scale to stakeholders. These will include a user guide (description of the construct, intended uses and potential misuses of the scale, glossary), audio-recordings of benchmark samples for teachers to calibrate their ratings, and a simplified score reporting and feedback sheet to use with students for diagnostic assessment and to foster students' self-awareness of their comprehensibility profile. Within a regular teaching cycle, the scale could be integrated into phase three of Harding, Alderson, and Brunfaut's (2015) proposed "ideal" diagnostic process, enabling EAP teachers to check hypotheses based on their prior observations or informal assessments of students' oral language needs relative to overall impressions of their comprehensibility.

As discussed above, the perceptual salience of L1 transfer features in L2 pronunciation make it perhaps the most difficult linguistic component to model in scales that cater to speakers from different L1 backgrounds. This also compounds the challenge of generating a universal pronunciation scale applicable to all world target languages, which will necessarily be generic precisely as a result of the imperative of not being target language specific. This L2 comprehensibility scale development effort, which focused on one target language (English) and for which pronunciation is integral, could inform future instrument validation research to do with L2 proficiency and speaking constructs more generally. This includes projects linking the CEFR levels to both "criterial" English language features (which could include pronunciation) as part of the English Profile Programme (Hawkins & Filipović, 2012) and to high-stakes tests (e.g., Cambridge English Exams), since guidance from the CEFR, including the Phonological Control scale, is limited (Galaczi, ffrench, Hubbard, & Green, 2011). Clearly, there is much more work to be done.

Acknowledgements

We are grateful to Garrett Byrne, Matt Kedzierski, Carla Pastorino, June Ruivivar, Gabriel Smith, and Helen Tan for their help with data collection. We also acknowledge David Collett, Dustin Crowther, Randall Halter, Sara Kennedy, Kazuya Saito, and Ron Thomson for assisting with practicalities in relation to this study or for feeding into its conceptual development. Finally, we thank our EAP teacher participants, whose input and astute reflections shaped the development of the scale.

Author's Note

The rating scale developed in this study along with instructions for use are publicly available on the IRIS Digital Repository (http://www.iris-database.org), the Open Science Framework (https://osf.io/), and the TESOL Resource Center (https://www.tesol.org/connect/tesol-resource-center).

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by an FP7 European Commission Marie Curie Career Integration Grant (grant number PCIG10-GA-2011-303413) and a Social Sciences and Humanities Research Council of Canada Insight Development Grant (grant number 430-2011-0341).

Note

1. It should be noted that there is no commonly agreed definition of intelligibility and comprehensibility in applied linguistics research. In the broad sense of the term, intelligibility and comprehensibility are used interchangeably to mean ease of understanding of L2 speech in general and without reference to how understanding is being measured (Levis, 2006). Derwing and Munro's (2015) definitional distinction is pervasive in research contexts where it is necessary to specify how understanding is being operationalized. Conversely, when the terms are used in rating scales used by human raters, it is already implied that listeners' *perceptions* of understanding are being captured through ratings, making Derwing and Munro's narrow definitional distinction moot. Thus, the use of either term in rating scales can always be interpreted in the broad sense. Comprehensibility applies in the narrow sense when human listeners (as opposed to scoring algorithms) are performing the rating and regardless of the term that is used by the test developer.

References

Alderson, J., Brunfaut, T., & Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics*, *36*, 236–260.

Andrade, M. S. (2006). International students in English-speaking universities: Adjustment factors. *Journal of Research in International Education*, *5*, 131–154.

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.

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. (2015a). Does speaking task affect second language comprehensibility? *The Modern Language Journal*, *99*, 80–95.

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

Derwing, T. M. (2008). Curriculum issues in teaching pronunciation to second language learners. In J. G. Hansen Edwards & M. Zampini (Eds.), *Phonology and second language ac*quisition (pp. 347–369). Amsterdam: John Benjamins.

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

ETS. (2009). The official guide to the TOEFL test (3rd ed.). New York: McGraw-Hill.

ETS. (2011). *TOEFL iBT® research insight: TOEFL® program history*. Princeton, NJ: Educational Testing Service.

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

Foote, J. A., Holtby, A., & Derwing, T. M. (2011). 2010 survey of pronunciation teaching in adult ESL programs in Canada. *TESL Canada Journal*, *29*, 1–22.

Galaczi, E. D., ffrench, A., Hubbard, C., & Green, A. (2011). Developing assessment scales for large-scale speaking tests: A multiple-method approach. *Assessment in Education*, *18*, 217–237.

Gilbert, J. (2012). *Clear speech student's book: Pronunciation and listening comprehension in North American English*. Cambridge: Cambridge University Press.

Ginther, A., & Elder, C. (2014). *A comparative investigation into understandings and uses of TOEFL iBT Test, the International English Language Testing Service (Academic) Test, and the Pearson Test of English for graduate admissions in the United States and Australia: A case study of two university contexts*. ETS Research Report, RR–14–44. Princeton, NJ: ETS.

Harding, L. (2017). What do raters need in a pronunciation scale? The users' view. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment* (pp. 12–34). Bristol, UK: Multilingual Matters.

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, *32*, 317–336.

Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the common European framework*. Cambridge: Cambridge University Press.

IELTS. (2009). *Official IELTS practice materials*. Los Angeles, CA: IELTS International.

IELTS. (2015). *IELTS Guide for teachers: Test format, scoring and preparing students for the test*. Retrieved March 17, 2017, from www.ielts.org/-/media/publications/guide-for-teachers/ielts-guide-for-teachers-2015-uk.ashx

Isaacs, T., & Trofimovich, P. (2011). *International students at Canadian universities: Validating a pedagogically-oriented pronunciation scale*. Unpublished corpus of second language speech.

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., Yu, G., & Chereau, B. M. (2015). *Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale*. IELTS research reports online series, 4.

Jenkins, J. (2014). *English as a lingua franca in the international university: The politics of academic English language policy*. Abingdon, UK: Routledge.

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

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: Praeger.

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.

Li, A., & Post, B. (2014). L2 acquisition of prosodic properties of speech rhythm. *Studies in Second Language Acquisition*, *36*, 223–255.

LTRC. (2014). *The 36th Language Testing Research Colloquium call for papers: Towards a universal framework*. Retrieved June 22, 2016, from http://ltrc2014.nl/call-for-papers.html

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Peter Lang.

Major, R. C. (2012). Foreign accent. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Hoboken, NJ: Wiley.

Mora, J. C., & Darcy, I. (2017). The relationship between cognitive control and pronunciation in a second language. In T. Isaacs & P. Trofimovich (Eds.), *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, UK: Cambridge University Press.
North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang.
Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, *12*, 237–241.
Pearson (2012). *PTE Academic score guide*. (n.p.): Pearson Education. Available at: http://pearsonpte.com/wp-content/uploads/2014/07/PTEA_Score_Guide.pdf
Rogerson-Revell, P. (2011). *English phonology and pronunciation teaching*. London: Continuum.
Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, *37*, 217–240.
Swan, M., & Smith, B. (Eds.). (2001). *Learner English: A teacher's guide to interference* (2nd ed.). Cambridge: Cambridge University Press.
Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition*, *19*, 122–140.
Turner, C. E. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *Canadian Modern Language Review*, *56*, 555–584.
Zhang, J., & Goodson, P. (2011). Predictors of international students' psychosocial adjustment to life in the United States: A systematic review. *International Journal of Intercultural Relations*, *35*, 139–162.
Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, *36*, 69–84.

Appendix. L2 English comprehensibility global and analytic scales.

Comprehensibility level	Overall description of comprehensibility (summary statement)
5	**Speech is effortless to understand** Errors, are rare and do not interfere with the message **Sounding native-like or producing hesitation- or error-free speech is *not* necessary to achieve a level 5 (highest level)
4	**Speech requires little effort to understand** Errors minimally interfere with the message
3	**Speech requires some effort to understand** Errors somewhat interfere with the message
2	**Speech is effortful to understand** Errors are detrimental to the message
1	**Speech is painstakingly effortful to understand or indecipherable** Errors are debilitating to the message **Not enough *comprehensible* language is generated for coherent communication, confining the speaker to level 1
UR	**Unable to Rate the speech** No assessable speech sample is produced (e.g., unresponsive to the task, no articulation of English-like sounds)

Overall description of comprehensibility (1 = low comprehensibility; 5 = high comprehensibility).

Appendix. (Continued)

Comp	Pronunciation	Fluency	vocabulary	Grammar
5	• Pronunciation is **effortless to understand** • Errors do not interfere with the message • Pitch variation may make the speech sound lively or engaging • Sounding native-like is *not* expected	• Fluent speech, which is optimally paced, is **effortless to understand** • Hesitation markers are used at appropriate junctures or strategically to sustain listener attention	• Fluent speech, which is optimally paced, is **effortless to understand** • Hesitation markers are used at appropriate junctures or strategically to sustain listener attention	• Grammatical use conveys precise meaning or nuance, resulting in speech that is **effortless to understand** • Errors do not interfere with the message • Complex sentences may be used, depending on the task
4	• Pronunciation requires little **effort to understand** • Errors minimally interfere with the message • Speech may be characterized by too many or too few variations in pitch, sounding disjointed or monotone	• Mostly fluent speech, which may be slightly too fast or slow, requires **little effort to understand** • Hesitation markers are generally used at appropriate junctures	• Sufficient lexical choice mostly relevant to the task requires **little effort to understand** • Errors minimally interfere with the message • Unusual or less familiar lexical expressions may be used	• Grammatical use mostly conveys precise meaning, resulting in speech that requires **little effort to understand** • Errors minimally interfere with the message • A mix of simple and complex sentences are used
3	• Pronunciation requires some **effort to understand** • Errors somewhat interfere with the message (e.g., misplaced word stress, sound substitutions, not stressing important words in a sentence)	• Somewhat fluent speech, which is too fast or slow, requires **some effort to understand** • Hesitation markers are occasionally used at inappropriate junctures	• Simple lexical choice requires **some effort to understand** • Errors somewhat interfere with the message • Occasional gaps in vocabulary make the speech somewhat labored, although meaning is still roughly conveyed	• Grammatical use conveys general meaning, resulting in speech that requires **some effort to understand** • Errors somewhat interfere with the message • Simpler sentences are used instead of more complex ones

(Continued)

Appendix. (Continued)

Comp	Pronunciation	Fluency	vocabulary	Grammar
2	• Pronunciation is **effortful to understand** • Errors are detrimental to the message (e.g., misplaced word stress, sound substitutions, not stressing important words in a sentence) • Production difficulties may obscure the meaning of a few words	• Speech, which is markedly dysfluent or too fast, is **effortful to understand** • Hesitation markers are frequently used at inappropriate junctures • Compensatory strategies are used to offset gaps in fluency (e.g., ideas are described in a roundabout way, self-correction)	• Limited lexical choice and frequent lexical errors are **effortful to understand** • Errors are detrimental to the message • Frequent gaps in vocabulary may make the speech labored or unelaborated • Lexical chunks may be used to compensate for limited vocabulary	• Grammatical use may obscure meaning, resulting in speech that is **effortful to understand** • Errors are detrimental to the message • Only basic sentence structures are used
1	• Pronunciation is painstakingly **effortful to understand** • Errors are debilitating to the message (e.g., misplaced word stress, sound substitutions, not stressing important words in a sentence) • Production difficulties may make words sound slurred or indistinct	• Speech, which is extremely dysfluent or much too fast, is **painstakingly effortful to understand** • Hesitation markers are very frequently used at inappropriate junctures, leading to halting or "broken" speech • No compensatory strategies are used to offset gaps in fluency	• Extremely simplistic or limited lexical choice and very frequent lexical errors make the speech **painstakingly effortful to understand** • Errors are debilitating to the message • Frequent gaps in vocabulary make the speech unelaborated or indecipherable • No lexical chunks are used to compensate for limited vocabulary	• Grammatical use obscures meaning, making the speech **painstakingly effortful to understand** • Errors are debilitating to the message • Only very basic or fragmented sentences are used
UR	Unable to Rate. Speaker does not produce an assessable sample of speech (e.g., unresponsive to the task, no articulation of English-like sounds)			

1 = low comprehensibility; 5 = high comprehensibility.
Note: The pronunciation and fluency criteria may weigh more heavily in assessments of comprehensibility than the vocabulary and grammar criteria.