Applied Psycholinguistics (2023), **44**, 858–888
doi:10.1017/S0142716423000346

ORIGINAL ARTICLE

Second language speech comprehensibility and acceptability in academic settings: Listener perceptions and speech stream influences

Dustin Crowther, Daniel R. Isbell and Hitoshi Nishizawa

Department of Second Language Studies, University of Hawaiʻi at Mānoa, Honolulu, USA
Corresponding author: Dustin Crowther; Email: dcrowth@hawaii.edu

(Received 20 September 2022; revised 30 May 2023; accepted 9 June 2023; first published online 11 August 2023)

Abstract

Ideally, comprehensible second language (L2) speech would be seen as acceptable speech. However, the association between these dimensions is underexplored. To investigate the relationship between comprehensibility and "academic acceptability," defined here as how well a speaker could meet the demands of a given role in an academic setting, 204 university stakeholders judged L2 speech samples elicited from a standardized English test used for university admissions. Four tasks from 100 speakers were coded for 13 speech stream characteristics. Judgments for comprehensibility and acceptability correlated strongly ($r = .93$). Linear mixed-effects models, used to examine judgments across all tasks and separately for each task, indicated that while random intercepts (i.e., speaker ability, listener severity) explained a substantial amount of total variation (32–44%) in listener judgments compared to speech characteristic fixed effects (8–21%), fixed effects did account for variation in speaker random effects (reducing variation compared to intercept-only models by 50–90%). Despite some minimal differences across task types, the influence of speech characteristics across both judgments was mostly similar. While providing evidence that comprehensible speech can indeed be perceived as acceptable, this study also provides evidence that speakers demonstrate both consistent and less consistent performance, in reference to speech stream production, across performances.

Keywords: acceptability; comprehensibility; speech perception; speech production; target language domain

Listeners' perceptual judgments of second language (L2) speech are well-studied, especially in reference to the global speech dimensions of *comprehensibility* (ease of understanding) and *accentedness* (degree of nativelikeness; Derwing & Munro, 2015). Such dimensions are thought to be general in the sense that they account for qualities of speech that are not bound to communicative context or purpose – just the listener – despite an emerging body of evidence that communicative demand influences listener judgments through associated linguistic and temporal cues

(Crowther et al., 2015a, 2018). Less well-studied are listener perceptual judgments that are more explicitly situated in particular communicative contexts and associated demands. In this study, we investigated the dimension of academic acceptability, which can be defined as *acceptability for spoken communication in a tertiary academic setting*. Though Thomson (2018) proposed that, in a perfect world, comprehensible speech should also be acceptable speech, little research has been conducted on the relationship between these two global speech dimensions. In our investigations, we explore the influence of linguistic (phonology, accuracy, complexity) and temporal (i.e., fluency) speech stream characteristics on listener judgments of comprehensibility and academic acceptability across four open-ended speaking performances elicited through the Duolingo English Test, a high-stakes English proficiency exam used for university admissions purposes. In addition to providing greater insight into the association between comprehensibility and academic acceptability, the inclusion of multiple speaking performances allows us to explore the extent to which L2 speakers' speech stream characteristics and associated listener-based dimensions (acceptability, comprehensibility) are consistent across performances.

Literature review

L2 speech and perceptual judgments

Derwing and Munro (2009) argued that listeners' perceptions of L2 speech serve as the gold standard for assessment, as "what listeners perceive is ultimately what matters most" (p. 478). While listener perceptions have a longstanding history in applied linguistics research (e.g., Lambert et al., 1960), L2 pronunciation research has drawn heavily on dimensions established in Munro and Derwing (1995a): *accentedness, comprehensibility,* and *intelligibility*. The key finding of their study, one frequently replicated (e.g., Huensch & Nagle, 2021), was that foreign-accented speech *could* at the same time be understandable. To measure understanding, Munro and Derwing included listener transcriptions as a measure of intelligibility (i.e., the accuracy with which a listener understands an L2 speaker's intended utterance) and subjective scalar ratings as a measure of comprehensibility (or the perceived ease or difficulty of understanding an utterance). Interestingly, accuracy of understanding has been shown to not necessarily entail ease of understanding (see Derwing & Munro, 2015), indicating that the two measures of understanding could be argued to be partially independent dimensions of L2 speech.

Following Levis (2005), L2 speech scholars have consistently advocated for understandable over native-like speech as the target of pronunciation acquisition and pedagogy. Though interest in intelligibility as a measure of understanding remains high (e.g., Kang et al., 2018), there has been a noticeable shift towards comprehensibility as the measure of primary interest (e.g., Crowther et al., 2022). Beyond methodological advantages – scalar ratings common to comprehensibility research allow for more practical (i.e., quick) and reliable collection of listener judgments (Kennedy & Trofimovich, 2019) – Saito (2021) highlighted how such an approach "has strong ecological validity, as it is assumed to reflect the instant and impressionistic judgements made by interlocutors during oral communication in

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

real-life contexts" (p. 86). As an additional reason for considering comprehensibility over intelligibility as a measure of understanding, consider that comprehensible speech is almost always assessed as intelligible (Derwing & Munro, 2015), which indicates that the acquisition of intelligible speech may outpace the acquisition of comprehensible speech. Thus, a pedagogical focus on comprehensible speech may have the added benefit of improving intelligibility (Thomson, 2018).

Comprehensibility

Though not the earliest use of the term comprehensibility (e.g., Varonis & Gass, 1982), Munro and Derwing (1995a) served to establish comprehensibility as it is generally investigated today (Crowther et al., 2022). That is, as the "[p]erceived degree of difficulty experienced by the listener in understanding speech" (Munro & Derwing, 2015, p. 14). By definition, comprehensibility is a perceptual measure (Munro & Derwing, 2020), with listeners' assessments of ease of understanding a reflection of the processing difficulties faced while attending to foreign-accented speech (e.g., Munro & Derwing, 1995b). As a consequence, speech rated as less comprehensible, beyond generally being less intelligible, tends to elicit less favorable emotional reactions and social judgments (e.g., Dragojevic & Giles, 2016; Lev-Ari & Keysar, 2010).

One area of particular interest in L2 comprehensibility research has been to understand how measures of phonology, fluency, syntax, and lexicon influence listeners' comprehensibility judgments, as understanding these influences enriches understanding of the dimension and helps to distinguish it from other listener perceptions. As the majority of comprehensibility studies rely on audio-only stimuli, listeners' judgments are necessarily based on the speech stream itself (though see Trofimovich et al., 2021; Tsunemoto et al., 2022, for recent, interaction-based judgments of comprehensibility). While the strength of a given measure's influence is variable across studies, the general pattern that has emerged is that listeners, when attending to comprehensibility, attend to a range of phonological (e.g., segmental accuracy, word stress accuracy), temporal (e.g., speech rate, pausing phenomenon), and lexicogrammatical (e.g., grammatical accuracy, lexical appropriateness) measures; this is in contrast to their ratings of accentedness, where they primarily attend to measures of phonology when making judgments (e.g., Crowther et al., 2015b, 2018; Trofimovich & Isaacs, 2012; see Saito's, 2021, meta-analysis). Saito et al. (2016) proposed that this difference stems from listeners' need to make use of all available linguistic information to derive meaning from an utterance when judging comprehensibility. Finally, it has been well-documented that the demands of a given speaking task can influence listeners' comprehensibility ratings. Comprehensibility of speech elicited using a more cognitively complex speaking task tends to be lower than that elicited using a less complex task, with different profiles of linguistic measures associated with each set of ratings (e.g., Crowther et al., 2018).

Acceptability

Beyond comprehensibility (and accentedness), a perceptual dimension that has gained increased attention in L2 speech research is acceptability, though it has yet to reach the same level of interest as other dimensions (Thomson, 2018). A consistent definition of acceptability is elusive. Szpyro-Kozłowska (2014) referred to the

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

"amount of irritation caused by a given accent" (p. 83), while Fayer and Krasinski (1987) discussed both distraction (i.e., the extent to which the speech diverts attention from the message) and annoyance (i.e., the extent to which the listener experiences a negative, subjective reaction to the speech) as detrimental to acceptability judgments. However, a more appropriate way to view acceptability may be along two common streams of inquiry, as identified in Isaacs (2018). The first stream considers the acceptability of L2 speech in reference to a specified norm, often one associated with social power (see Levis, 2006). As such, a judgment of acceptability necessitates a comparison to what a speaker believes to sound proper, raising questions as to whether acceptability in this sense is distinct from accentedness (Flege, 1987). We may view this stream as measuring *acceptability compared to what*. The second stream makes reference to speakers' acceptability, based on their speech, to meet the performative demands of a given role, such as serving as an ITA at an English-medium university (e.g., Ballard & Winke, 2017; Kang, 2012). This second stream situates acceptability as a perceptual measure of a speaker's ability to fulfill a functional role within a given target language use (TLU) domain (e.g., English-medium academic study). Here, the emphasis of measurement is on *acceptability for what*. In comparison to the first stream, listeners need not access their belief on what sounds proper (though they may still do so), but instead should focus on what they believe is necessary for performance in a given TLU domain. We note that, in reference to both streams discussed above, researchers have used a range of adjectives, including not only "acceptability" (Ballard & Winke, 2017; Sewell, 2012), but also "suitability," "appropriateness" (Prikhodkine, 2018), and "effectiveness" (Kang, 2012; Kang et al., 2015; Plakans, 1997).

Acceptability and the TLU domain (English-medium academic study)

Acceptability in reference to TLU domain task fulfillment has primarily emphasized English-medium academic study (hereafter *academic acceptability*, to reflect the TLU domain which these judgments reference). Even within this TLU domain, research has remained primarily constrained to the rating of ITAs by undergraduate students. For example, Kang (2012) asked 70 native and nonnative undergraduates to rate the instructional competence of 11 ITAs based on a 5-minute audio-only excerpt from a course lecture. She found that while ratings of ITAs' instructional competence were informed by prosodic and fluency measures of speech (e.g., speech rate, pausing), additional variance in ratings came from background variables specific to the undergraduate listeners (e.g., teaching experience, time spent engaging with nonnative speakers). In another example, Dalman and Kang (2023) considered the TOEFL iBT elicited speech of 20 high-proficiency English speakers. Despite near ceiling TOEFL iBT performance, 55 native-speaking undergraduates rarely judged speakers as being perfectly comprehensible nor highly acceptable for either university-level teaching or classroom-based group work. Of additional note is that in multiple linear regression analyses, the global dimensions of comprehensibility and accentedness were found to have only a moderate association with listeners' assessments of acceptability for teaching (adjusted $R^2 = 0.47$) and acceptability for group work (adjusted $R^2 = 0.29$), though listeners did prioritize comprehensibility over accentedness when assigning their ratings. A moderate association between comprehensibility and acceptability is similarly seen in Hosoda and Stone-Romero (2010), who investigated college students' speech perceptions of French- and Japanese-accented job applicants across four jobs (customer service representative, manager trainee, underwriter, data entry). The correlations between "understandability" and suitability ratings ranged from moderate ($r = 0.56$) to non-existent ($r = 0.00$). So, while Thomson (2018) has proposed that comprehensible speech should be seen as acceptable speech, it appears that such an ideal is not fully substantiated in existing research.

An emphasis primarily on ITAs' instructional competence overlooks the fact that L2 English users at English-medium universities interact with a much wider range of stakeholders while engaging in a number of different tasks. Non-instructional roles that may occur on any given day may include engaging with a peer (or two or three) during course-based group discussion or post-class/office hours interactions with instructors. Both stakeholder groups (student peers, instructors) likely hold specific expectations regarding what qualifies as acceptable English for these given interactions. Extending inquiry into academic acceptability beyond the instructional competence of ITAs will allow for a more in-depth understanding of what constitutes acceptable language during English-medium academic study, and how such perceptions do, or do not, align with more commonly employed perceptual judgments of L2 speech (i.e., comprehensibility). Examining the influence of speech stream characteristics will also help determine the degree of overlap between academic acceptability and comprehensibility. Aside from Dalman and Kang (2023), who examined three fluency-related variables and overall pitch range, we know little about how speech stream characteristics influence acceptability judgments.

Consistencies in speech across encounters

The dimensions of comprehensibility and acceptability are defined and operationalized with respect to listener perceptual judgments of speech. In this sense, it has become common to point out that comprehensibility, as a measure of understanding, is co-constructed in the encounter between speaker and listener (Levis, 2005, 2020). Listener ratings are a quantification of these encounters. An open question, however, is whether and to what degree underlying characteristics of individual speakers' speech and listeners' reactions are consistent across encounters. In other words, can some portion of the dimensions of comprehensibility and acceptability be seen as relatively stable individual traits that have consistent effects across encounters? Methodologically, speaker-listener encounters can be seen as interactions between two types of sampled units, speech samples and listeners, with speech samples being isomorphic to or potentially nested within sampled speakers (Barr, 2018). In research on comprehensibility and other listener-based global speech qualities, it is quite common for listeners to react to multiple speakers, and it is in turn common (and uncontroversial) to discuss differences among listeners in terms of judgment severity (e.g., some listeners make more generous judgments than others *in general*, across speakers). It is less common, though, to see research where (a) multiple speech samples (especially more than 2) are collected from speakers and (b) listener judgments of those samples are analyzed simultaneously, both of which would allow for a more *general* comprehensibility of speakers, across speech samples, to be described. While limited in number of studies and number of speech samples per speaker analyzed, studies of L2 Korean (Isbell et al., 2019) and Spanish (Huensch & Nagle, 2021) have demonstrated that variation associated with speakers meaningfully accounts for comprehensibility ratings across samples (i.e., by-speaker random effects contributed to conditional variance explained in mixed-effects models).

The current study

The current study draws upon speech elicited through the Duolingo English Test (DET). Despite ongoing debate regarding to what extent the DET truly reflects academic practices (Wagner, 2020), the DET has developed into a commonly used assessment tool for university admissions (https://englishtest.duolingo.com/institutions). In this sense, speech elicited through the DET, though not *in situ* academic performances, can serve as a barometer of L2 users' preparedness for study in the TLU domain of English-medium academic study. Using speech samples from admissions tests facilitates the inclusion of individuals with varied first languages (L1s) and lower levels of academic language proficiency, in turn allowing for a more comprehensive examination of the dimensions of interest.

The 204 listeners in the current study represented stakeholders within the TLU domain of English-medium academic study. Each listener judged the comprehensibility and academic acceptability of speech elicited from 200 DET test takers. Test takers' speech, across four speaking performances, was additionally coded for 13 speech stream characteristics, spanning phonology, complexity, accuracy, and fluency considerations. Through analyses of this data, we set out to investigate the following research questions:

> RQ1: What is the relationship between judgments of academic acceptability and comprehensibility, in terms of (a) association between judgments and (b) influences of speech stream characteristics on judgments?
>
> RQ2: To what extent do judgments of academic acceptability and comprehensibility, and the influence of speech stream characteristics, vary across speaking performances?

Method

We adopted a cross-sectional associative design involving speech samples elicited in the DET, a high-stakes English proficiency test, and ratings of speech provided by layperson listeners recruited specifically for the study.

Replication package

Research materials and data that can be made publicly available are accessible at https://osf.io/2ujfa/. Some study materials cannot be publicly shared: Original speech files, provided by Duolingo, and our transcriptions of those files, because of test taker privacy and test security concerns. Some research materials, including rating task instructions, questionnaires, are available in the same package as Online Supplement S1 and Online Supplement S2. Data used in statistical analyses are available in the file accept_comp_open_data.csv (with the codebook available in Comp_Accept_Codebook.csv). For the data we share, we ensured that all test taker information has been anonymized. The format of the "Speaker" variable used in the current study to delineate participants neither resembles nor correlates with any type of test taker identification system used by the DET. Our coding schema cannot be used to identify a given DET test taker. Instructions and code required to reproduce all analyses are available in the files 01_primary analyses.R and 02_primary analyses_no af.R.

Participants

Speakers

Speakers consisted of 100 DET test takers provided by Duolingo for the purposes of this study. The speakers' language backgrounds reflected DET test taker demographic trends: Mandarin Chinese ($n = 30$), Arabic ($n = 21$), Spanish ($n = 20$), French ($n = 14$), Persian ($n = 12$), and English[1] ($n = 3$). Speakers reported taking the test for either graduate (N = 49) or undergraduate (N = 51) admissions. Official DET scores provide a standardized measure of speakers' English proficiency. The 100 speakers had a mean DET score of 108 (out of 160; SD = 17.30, median = 110, range = 70–145),[2] closely approximating Duolingo's global test taker score distributions (see Cardwell et al., 2022). Figure 1 illustrates the sample distribution of speaker proficiency.

Listeners

Listeners included four sets of stakeholders from an English-medium university in the United States. All listeners completed a background questionnaire. Among students, 58 graduate students (34 female; mean age = 30.20, *SD* = 7.69) and 58 undergraduates (13 female; mean age = 23.00, *SD* = 5.58) agreed to participate, with a large proportion of each pursuing studies in engineering. A total of 47 faculty members (23 female; mean age = 43.0, *SD* = 11.00) spread across a range of academic disciplines agreed to participate. Finally, 41 administrative staff (28 female; mean age = 45.2, *SD* = 11.10) who were mostly involved in academic department support and student services also agreed to participate. The majority of listeners across groups were native users of English (76.6–87.8%), and all listener groups indicated similar levels of familiarity with foreign accents (mean = 5.51–6.26; 1 = not familiar at all, 9 = very familiar).

Materials – speech samples

Each test taker responded to four speaking prompts which comprised the extended speaking portion of the DET: *Picture-Speak*, *Listen-Speak* (x2), and *Read-Speak*, which were completed in a set order (see Table 1). For each DET speaking task type,

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

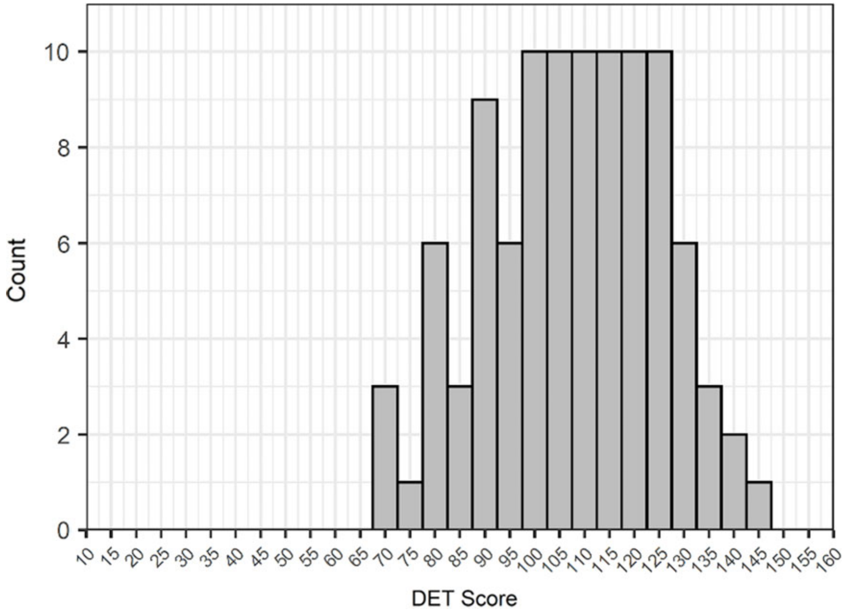


Figure 1. DET score histogram.

test takers are directed to speak for at least 30s and may speak for up to 90s. As evident by the task labels, Picture-Speak required test takers to describe a picture, while Listen-Speak and Read-Speak required test takers to respond to an audio or written prompt, respectively. Example prompts for each task type can be found in *Official Guide for Test Takers* (Duolingo, 2022) and the DET practice test (https://englishtest.duolingo.com; registration required). Given the large item pool developed by the DET (Cardwell et al., 2022), a variety of unique Picture-Speak ($n = 90$), Listen-Speak ($n = 142$), and Read-Speak ($n = 70$) prompts were included in our dataset. Each prompt for Listen-Speak and Read-Speak were additionally classified internally by the DET according to a communicative function (i.e., argument, explanation, description); prompts for Picture-Speak all shared a common function. All four speaking performances from each of the 100 speakers were provided by Duolingo (for a total of 400 speech files). Though we did not trim speech files for length (mean = 69.3s, $SD = 20.7$), we did remove any identifiable information (e.g., full name) and scaled speech files with low intensity to 70 dB.

Linguistic coding

The 400 speech files were first transcribed via Amazon Transcribe (Amazon Web Services, n.d.), with manual corrections performed by the first and third author. A second set of pruned transcriptions was created, with false starts, filled pauses, and repairs removed to facilitate calculation of some complexity and speed fluency measures. Each speech file was then coded to derive a set of pronunciation, accuracy, complexity, and fluency measures (N = 13). Given that a vast array of speech measures have been shown to associate with L2 speech ratings to some

Table 1. DET extended speaking task types

Order	Description	Communicative Functions Associated with Prompts
1. Listen-Speak 1	20s are provided to listen to a prompt (up to two times) before speaking.	argument, description, explanation
2. Picture-Speak	20s are provided to look at a picture before speaking; the picture remains visible while speaking.	picture description
3. Read-Speak	20s are provided to read a prompt before speaking. The prompt remains visible while speaking.	argument, description, explanation
4. Listen-Speak 2	20s are provided to listen to a prompt (up to two times) before speaking.	argument, description, explanation

extent, we have admittedly included only a small handful of measures, chosen for theoretical, analytical, and practical reasons. Theoretically, we drew upon Suzuki and Kormos (2020), who considered a range of speech measures as predictors of listener ratings of comprehensibility and perceived fluency, as an initial guide to identify potential measures to include in our analyses. Analytically, to guard against overfitting and maintain interpretability, we limited the number of predictors included. This meant that several measures included in a preliminary analysis were removed (see descriptions below). Finally, given the volume of speech coded (>400 minutes, cf. ~99 minutes in Suzuki & Kormos, 2020, and 24 minutes in Trofimovich & Isaacs, 2012), certain coding decisions were made to best make use of the available time of the research team members.

Pronunciation. Following Suzuki and Kormos (2020), we included three measures of pronunciation: substitution rate, syllable structure error rate, and word stress error rate. Due to our extensive number of samples, we coded both substitution rate and syllable structure error rate at the word level, rather than the phoneme or syllable level. With this analytical choice, we sought to strike a balance between efficiency and fidelity for a relatively large amount of elicited speech, falling somewhere between previous approaches involving subjective scalar ratings of overall segmental quality (e.g., Crowther et al., 2015a) and analytic coding at the level of individual segments (e.g., Trofimovich & Isaacs, 2012). This decision was also informed by Munro and Derwing (2006), which found that for sentence-length utterances the presence of a single phonemic substitution had a notable effect on perceptions of comprehensibility, but additional errors had less impact. Though coding at the word level may not capture the total number of errors that speakers may produce, it can still reveal a general pattern of errors, as we would expect a given speaker to make relatively similar errors throughout their speech. We also note that while Suzuki and Kormos (2020) included a measure of speech rhythm in their study, this measure was not predictive of either comprehensibility or perceived fluency in multiple regression analyses. Combined with the extensive time such coding would have required, we do not include rhythm as a measure in the current analyses.

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

Table 2. Intercoder agreement for hand-coded speech variables

	Scope	Percent Agreement (%)	Gwet's AC1
Segmental Errors	Words	94	.94
Syllable Structure Errors	Words	98	.98
Word Stress	Words	99	.99
Lexical Errors	Words	97	.97
Morphosyntactic Errors	Words	98	.98
Syntactic Errors	Clauses	91	.89

1. *Substitution rate*: number of words with a phonemic substitution divided by total number of words produced;
2. *Syllable structure error rate*: number of words with an added/deleted syllable divided by total number of words produced;
3. *Word stress error rate*: number of polysyllabic words with missing/misplaced primary stress divided by total number of words produced.

The first and second authors initially manually coded 10% of all samples for all three measures (see Table 2). Overall percent agreement was high (≥94%) and Gwet's AC_1, a measure of inter-rater reliability argued to be more stable than either Pi (π) or kappa (κ) in the presence of high agreement and/or high prevalence (Gwet, 2008), similarly indicated high inter-rater reliability for each of substitution rate ($AC_1 = .94$), syllable structure error rate ($AC_1 = .98$), and word stress error rate ($AC_1 = .99$).[3] Both authors subsequently collaboratively reviewed all files to identify and resolve all discrepancies in coding before the first author coded the remaining files.

Accuracy. Following Suzuki and Kormos (2020), we included three measures of accuracy.

4. *Lexical error rate*: number of words deemed inappropriate for the context, malformed, or drawn from the speaker's first language divided by the total number of words produced (excluding nonlinguistic filler);
5. *Morphological error rate*: number of words containing a deviation from standard academic English divided by total number of words produced (excluding nonlinguistic filler). Examples of deviations included errors in plural marking, subject-verb agreement, pronoun choice, and articles/determiners;
6. *Syntactic error rate*: number of clauses containing a deviation from standard academic English divided by total number of clauses produced. Examples of deviations included word order violations, missing words/constituents, misapplication of tense or aspect, and subordination errors.

As with the pronunciation measures, the first and second authors manually coded 10% of the speech files for all three measures, with percent agreement again

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

high (≥91%; see Table 2). Gwet's AC_1 indicated high inter-rater reliability for lexical error rate ($AC_1 = .97$), morphological error rate ($AC_1 = .98$), and syntactic error rate ($AC_1 = .89$). After reviewing all files to resolve discrepancies in coding, the second author completed all remaining coding.

Complexity. In Suzuki and Kormos (2020), two measures of syntactic complexity, mean length of AS-units and mean number of clauses per AS-unit, were found to associate strongly with comprehensibility ($rs > .60$). We chose to include only one, Clauses/AS-unit, as it more directly addresses the notion of syntactic complexity in spoken language (Biber et al., 2011). For lexical complexity, we drew on two measures similar to those used in Suzuki and Kormos (2020) that capture two aspects of lexical complexity: lexical sophistication and lexical diversity.

7. *Clauses/AS-unit*: number of clauses divided by number of AS-units per sample. An AS-unit was defined as "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (Foster et al., 2000, p. 365). For example, in the single AS-unit *the boy saw that the girl was holding a bag*, there are a total of 2 clauses (*the boy saw ___, the girl was holding a bag*). Both measures were manually coded by the third author, and verified by the second author during the coding of accuracy measures. In total, only 9.5% of files (N = 38) required adjustment.
8. *COCA spoken mean log frequency*: A log-transformed measure of average word frequency, based on the raw tokens of all words in a corpus. Calculated using the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle et al., 2018);
9. *Measure of textual lexical diversity (MTLD)*: A measure of the range of words used in a given sample, calculated "as the mean length of sequential word strings in a text that maintain a given TTR [type-token ratio] value" (McCarthy & Jarvis, 2010; p. 384). Measured using the Tool for the Automatic Analysis of Lexical Diversity (TAALED; Kyle et al., 2021).

Fluency. Following Suzuki and Kormos (2020), we included measures of speed, breakdown, and repair fluency. Though we began with three measures of speed fluency, preliminary analyses indicated high correlations ($rs > .83$) between speech rate, articulation rate, and mean length of run. As such, we include only one measure, speech rate, in our current analyses.[4] Though Suzuki and Kormos (2020) included five measures of breakdown fluency, as previously stated, we wished to keep our overall number of speech stream measures down in order to allow for greater model interpretability and guard against overfitting. As such, we included only two measures of breakdown fluency, filled pauses per minute and mid-clause silent pauses per minute. A focus on mid-clause, rather than end clause, pauses was chosen, since it was silent pauses in this location that predicted both comprehensibility and perceived fluency ratings in Suzuki and Kormos (2020). Finally, we included a composite measure of repair fluency. All fluency measures were frequency, as opposed to duration, based.

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

10. *Speech rate*: number of syllables divided by total duration of sample;
11. *Mid-clause silent pauses/minute*: number of silent pauses occurring mid-clause divided by 60 seconds;
12. *Filled pauses/minute*: number of filled pauses (e.g., uh, um) divided by 60 seconds;
13. *Repairs/minute*: number of repairs divided by 60 seconds. Repairs included self-corrections, false starts and reformulations, and partial or complete repetitions.

Fluency coding was conducted by the third author, who made use of both hand coding and automated coding of pauses using Praat (Boersma & Weenink, 2022). For speech rate and articulation rate, number of syllables was calculated automatically using the pruned transcripts (i.e., syllable counts did not include filled pauses, false starts, and hesitations). Following standard conventions (e.g., Suzuki & Kormos, 2020), silent pauses consisted of any silence ≥250ms and were automatically identified using de Jong and Wempe's (2009) Praat script. All silence boundaries, initially identified using Praat, were then manually reviewed and adjusted as needed, with filled pauses and pause location additionally coded (based on analyses of unpruned transcripts). Repair measures were derived from those items removed for the pruned transcripts.

Listener procedures and judgment scales

Listeners received an email invitation to complete an online experiment implemented in Gorilla (Anwyl-Irvine et al., 2020), which they could access at home on a personal computer. Given the large number of speech samples, our experiment utilized a sparse rating design (see Isbell, 2018). We divided the 400 DET speech samples into 40 blocks of 10 files each, with each block composed of speakers representing at least four different L1 backgrounds and roughly one speaker per DET score decile. At least two different speaking task types (among Picture, Listen-Speak, and Read-Speak) were represented in each block, with most blocks featuring all three. Each of a speaker's four speaking performances was assigned to different blocks, which were then divided into two sets, A and B. Sets were designed to ensure no speaker had a response in both sets. Each listener was assigned to one block from Set A and one block from Set B, with half of the listeners starting with a block from Set A and vice versa. As such, no listener heard the same speaker more than once. Ultimately, due to randomization and a technical error related to block composition that was remedied partway through data collection (see Online Supplement S2 for details on block assignment and composition), a varied number of listener ratings for each of the 400 audio files were collected, with a mean of 10.40 per file (SD = 7.03, min = 2, max = 37).

Listeners made several judgments about files related to comprehensibility and academic acceptability. For comprehensibility, listeners provided three judgments using 6-point scales drawn from Schmidgall and Powers (2021), with prompts as follows:

- How *certain* are you that you understood the speaker?
- How *easy* was it to understand the speaker?
- How *comprehensible* was the speaker?

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

Though less common than the use of a single scale, the use of a multi-item scale for comprehensibility is not unheard of. For example, Kang et al. (2010) made use of a five-item scale (easy/hard to understand, incomprehensible/highly comprehensible, needed little effort/lots of effort to understand, unclear/clear, and simple/difficult to grasp the meaning), with reliability across items high (Cronbach $\alpha = .94$). In Isbell et al. (2023), we investigated the relationship between stakeholders' speech perceptions and DET speaking performance using the same speech samples analyzed here. As Schmidgall and Powers' (2021) reported a similar analysis focused on TOEIC speaking performance, we made use of their three-item scale as a measure of comprehensibility.

Listeners subsequently provided four judgments on 6-point scales that targeted the acceptability of a given speaker's English across different academic roles. Given that acceptability is a less well-established dimension in the literature, we drew inspiration from several previous studies. The first acceptability item targeted acceptability for university teaching, as this has been the primary focus of academic acceptability research (e.g., Dalman & Kang, 2023; Kang, 2012). The second item, following Dalman and Kang (2023), focused on group work in classes. The final two items, acceptability for undergraduate and graduate study, were included since the DET serves as a tool for assessing academic preparedness for university study. The four final prompts were as follows:

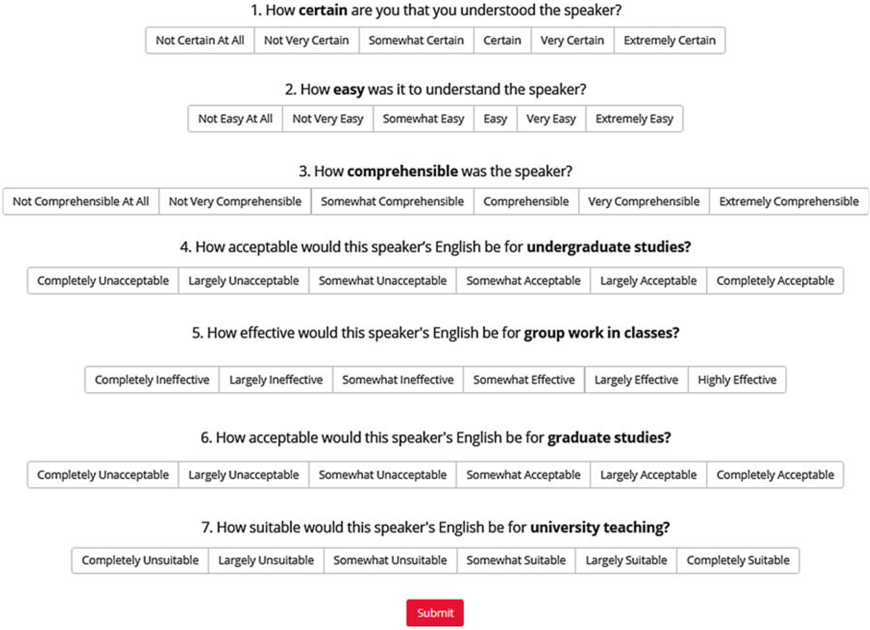
- How acceptable would this speaker's English be for *undergraduate studies*?
- How effective would this speaker's English be for *group work in classes*?
- How acceptable would this speaker's English be for *graduate studies*?
- How suitable would this speaker's English be for *university teaching*?

Figure 2 illustrates how these judgment scales were displayed in the online experiment platform Gorilla, which includes the descriptors for each point of the seven scales. Listeners received training on each target dimension and completed two practice ratings prior to beginning the experiment. Two attention checks, which required listeners to click on a specified rating on three rating scales, were included in the experiment.

To get a sense of how well the judgment items worked together as measures of the overarching construct, we aggregated judgments across listeners for each file and examined inter-item correlations (Table 3). The items for each scale were highly intercorrelated, indicating coherence of the dimensions being measured. Though the strongest correlations among items were generally those within the same scale (e.g., the correlations among the three comprehensibility items), it is also worth noting the high correlations among items across the two scales; we return to the relationship between comprehensibility and acceptability later in the manuscript. Additionally, in Isbell et al. (2023), many-facet Rasch models yielded strong evidence of unidimensional measurement and good fit of judgment items for both comprehensibility and acceptability.

Reliability of each judgment scale and for listener averages across all scales for comprehensibility and acceptability were estimated using 2-way random effects intraclass correlations (ICC, McGraw & Wong, 1996) using the irrNA package

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press



1. How **certain** are you that you understood the speaker?

2. How **easy** was it to understand the speaker?

3. How **comprehensible** was the speaker?

4. How acceptable would this speaker's English be for **undergraduate studies?**

5. How effective would this speaker's English be for **group work in classes?**

6. How acceptable would this speaker's English be for **graduate studies?**

7. How suitable would this speaker's English be for **university teaching?**

Figure 2. Speech judgment questions and interface.

(v.0.2.2, Brueckl & Heuer, 2021) in R to accommodate missingness in a sparse rating design. As shown in Table 4, absolute agreement among listeners' single scores was low (.27 – .34), indicating variability in listener scale use (i.e., some listeners were more lenient in judgments while others were stricter, see Isbell et al., 2023). However, the degree of consistency for average ratings from *k* randomly selected raters was considerably higher (.79 – .84). Our primary analyses are based on the average of all judgments for each dimension from each listener, and the ICCs for raters' aggregated comprehensibility and acceptability scores were both .84 and .85, respectively.

Analyses

An initial check of listener judgment quality appeared satisfactory, suggesting that most listeners were paying attention throughout the experiment. Overall, 1160/1230 attention checks were passed (94%); 192 listeners responded correctly to 4/6 checks (94%), and 176 of these listeners responded correctly to every attention check (86%). Of concern, however, were 12 listeners who answered ≤3/6 attention checks correctly. A closer inspection revealed no clear aberrant response patterns (e.g., uniform responses, such as selecting the most positive category for all judgments). Parallel analyses to those presented below without these 12 listeners differed little (see Online Supplement S3: Tables S3.1–S3.5). As such, we report our analyses inclusive of all 204 listeners but make notes of minor differences when relevant.

Based on Isbell et al. (2023), only minor differences existed between academic listening groups (undergraduate students, graduates, faculty, and administrative staff) in

Table 3. Pearson's correlations among comprehensibility and acceptability items (scores averaged across listeners)

	Comprehensibility			Acceptability		
	1	2	3	1	2	3
Comprehensibility						
1. Certainty						
2. Ease	.95					
3. Comprehensibility	.93	.95				
Acceptability						
1. Undergraduate Study	.88	.88	.91			
2. Group Work	.89	.90	.91	.94		
3. Graduate Study	.86	.87	.89	.95	.94	
4. Course Instruction	.85	.89	.89	.88	.91	.94

Note. N = 400 speech files, all *p*-values < .001.

Table 4. Intraclass correlations of speech judgments

	2-way ICC	
	A, 1	C, k
Comprehensibility (Average)	.34	.84
1. Certainty	.28	.80
2. Ease	.32	.84
3. Comprehensibility	.28	.80
Acceptability (Average)	.34	.85
1. Undergraduate Study	.27	.80
2. Group Work	.31	.83
3. Graduate Study	.30	.82
4. Course Instruction	.34	.85

Note. A, 1 = agreement of single scores assigned to a file. C, k = consistency of scores averaged across raters for files (equivalent to Cronbach's α, see McGraw & Wong, 1996).

responses to judgment questions. While faculty tended to be most lenient in their ratings, and administrative staff most harsh, the magnitudes of differences between group average scores were small (less than half a point in most cases). Importantly, the four groups used the scales similarly, and assessed both comprehensibility (certainty > comprehensibility > ease) and acceptability (undergraduate > group work > graduate > teaching) items following the same hierarchy. As such, and in line with our research questions in the present study, we treated listeners as a single group and made use of composite scores from each listener, averaged across judgment questions, as dependent variables for both comprehensibility and acceptability models.

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

For all analyses involving inferential statistics, we adopted an alpha of $p < .05$. When interpreting the magnitudes of correlations, we drew on Plonsky and Oswald's (2014) guidelines (small > .25, medium > .40, and large > .60). For mixed-effects models, we considered the standardized coefficient estimates, overall model R^2 (including marginal and conditional, with Plonsky & Ghanbar's, 2018, guidelines informing interpretation of the former), and furthermore considered the degree to which speech stream fixed effects reduced the amount of random intercept variation associated with speakers (i.e., the degree to which Level 1 fixed effects reduced Level 2 random effect variation compared to a "null" random effects only model; Bryk & Raudenbush, 1992). To check assumptions of linear mixed-effects regression models, we examined bivariate correlations among predictors to screen for collinearity and plotted model residuals (histograms, Q-Q plots) to assess normality.

Results

Descriptive statistics for study variables are provided in Table 5. These descriptive statistics are aggregated across all 400 spoken performances, with listener judgments aggregated across rating questions and listeners (descriptive statistics for each of the 4 task types are available in Online Supplement S4). The ICC characterizes how similar values of each variable were across speakers' sets of performances. Notably, averaged listener judgments associated with a speaker were moderately similar across performances, as were segmental pronunciation characteristics (phonemic substitutions and syllable structure errors) and disfluencies (mid-clause pauses and repairs/repetitions). The speech stream characteristics most similar within speakers across all performances were speech rate and filled pauses.

Figure 3 further illustrates the distribution of study variables (density plots on the diagonal) and presents bivariate scatterplot and Pearson correlation coefficients for all pairs of study variables. The averaged listener judgments of comprehensibility and acceptability were strongly correlated at $r = .93$, and several speech stream variables demonstrated moderate to large correlations with each judgment. Notably, speech stream variables tended to have similar correlations with both listener judgments. Correlations among speech stream variables were mostly small.

Linear mixed-effect regression models were used to examine the influence of speech stream characteristics on listener judgments for each speech performance. The dependent variable in each model, comprehensibility or acceptability, was the average of a single listener's ratings for all questions (3 comprehensibility questions or 4 acceptability questions). Fixed effect predictors included functional demands of each speaking task type, pronunciation variables, and complexity, accuracy, and fluency indices. While the full model with all fixed effects was of primary interest, intermediate models that added each group of variables were run to examine changes in variance explained.

Random intercepts were included for speakers, to account for the nesting of spoken performances and judgments, and for listeners, to account for the nesting of judgments within each listener. In more substantive terms, the by-listener random effects account for differing levels of judgment severity. Attempts to include random

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

Table 5. Summary of speech variables

Variable	M	SD	95% CI	Range	ICC*
Listener Judgments					
Comprehensibility†	3.97	0.78	[3.89, 4.05]	1.67–5.89	.62
Acceptability†	4.15	0.77	[4.07, 4.23]	1.91–5.75	.64
Pronunciation					
Substitution Rate	0.05	0.04	[0.05, 0.05]	0.00–0.25	.52
Syllable Structure Error Rate	0.02	0.03	[0.02, 0.02]	0.00–0.22	.58
Word Stress Error Rate	0.01	0.02	[0.01, 0.01]	0.00–0.11	.23
Complexity					
Clauses per AS-unit	1.97	0.69	[1.90, 2.04]	0.75–5.50	.23
MTLD	40.43	11.07	[39.35, 41.51]	17.40–101.64	.16
COCA Mean Log Word Freq.	3.13	0.16	[3.11, 3.15]	2.54–3.54	.31
Accuracy					
Lexical Error Rate	0.03	0.02	[0.03, 0.03]	0.00–0.13	.30
Morphosyntax Error Rate	0.02	0.02	[0.02, 0.02]	0.00–0.14	.21
Syntax Error Rate	0.10	0.11	[0.09, 0.11]	0.00–0.60	.34
Fluency					
Speech Rate	2.23	0.64	[2.17, 2.29]	0.76–4.31	.77
Mid-clause Silent Pauses/minute	12.05	4.87	[11.57, 12.53]	0.00–35.72	.59
Filled Pauses/minute	9.06	7.14	[8.36, 9.76]	0.00–38.50	.87
Repairs/minute	4.33	3.35	[4.00, 4.66]	0.00–16.05	.54

Note. $N = 400$ speaking performances.
*ICC across all 4 speaking tasks.
†Based on the average of all items and all raters.

slopes to account for possible variation across listeners in the influence of speech stream characteristics were unsuccessful, as the inclusion of more than a single random slope (uncorrelated with any intercept) resulted in singular fits or nonconvergence.

Model results are shown in Table 6. As indicated by R^2 values, fixed effects explained relatively little variation in judgments for both models while a substantial amount of variation (~40–45%) was accounted for by-speaker and listener random intercepts. In other words, differences among speaker ability and listener severity were the most powerful influences on judgments across speaking performances. Speaker and listener random intercepts were also highly correlated across the two models, $r_{\text{speakers}} = .96$ ($p < .001$) and $r_{\text{listeners}} = .85$ ($p < .001$).

Closer inspection of speaker random effects and task-level fixed effects indicated that several fixed effects had noteworthy explanatory power. Compared to intercept-only models (with no fixed effects), speaker random intercept variation in final

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

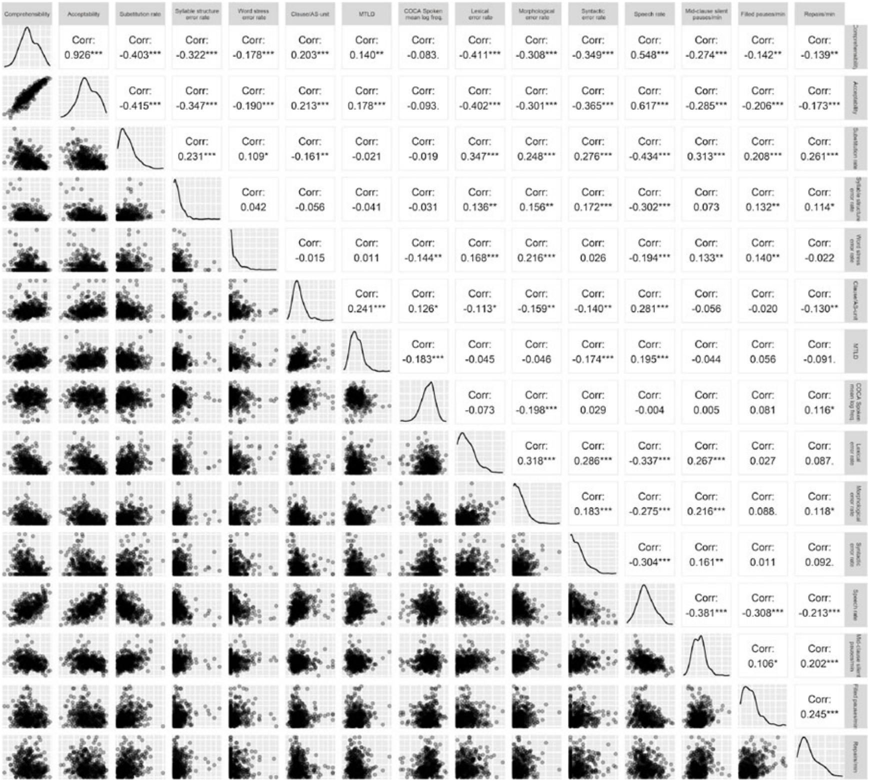


Figure 3. Correlations and distributions of study variables.

models was reduced by 56% for comprehensibility and 65% for acceptability. Thus, Level 1 fixed effect variables associated with spoken performances could account for considerable amounts of variation in Level 2 random effects associated with speakers. Fixed effects influences on judgments were similar across both models, with speech rate, all lexicogrammatical accuracy variables, and phonemic substitutions having statistically significant standardized coefficients of comparable size. Most of these statistically significant coefficients were small and negative (as expected for error rates), though the magnitude of the speech rate coefficient was positive and notably larger. Syllable structure error rate was only statistically significant in the acceptability model. These results are largely consistent with models excluding listeners who passed fewer than 4/6 attention checks (see Online Supplement S3: Table S3.2), but with a lack of statistical significance for substitution rate ($p = .05$) in both models (coefficient magnitudes were nearly identical).

Listener perceptions by speaking task type

To examine how speech stream characteristics might influence listener judgments differently across speaking task types, we conducted linear mixed-effect model analyses separately for each task type: Listen-Speak 1, Picture-Speak, Read-Speak,

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

Table 6. Linear mixed-effect model results for comprehensibility and acceptability

	Comprehensibility					Acceptability				
Predictors	β	CI	*p*	ΔR^2	$\Delta \tau_{\text{Speaker}}$	β	CI	*p*	ΔR^2	$\Delta \tau_{\text{Speaker}}$
(Intercept)	−0.06	[−0.18 – 0.07]	.377	.00	.24	−0.02	[−0.15, 0.10]	.696	.00	.21
Function (Step 1)				.00	−.00				.00	−.00
Description	0.07	[−0.02, 0.17]	.134			0.01	[−0.08, 0.09]	.875		
Explanation	0.06	[−0.04, 0.15]	.246			0.03	[−0.06, 0.12]	.492		
Picture	0.07	[−0.03, 0.17]	.147			0.07	[−0.02, 0.16]	.130		
Pronunciation (Step 2)				.01	−.04				.01	−.04
Substitution Rate	−0.04	[−0.08, −0.01]	**.026**			−0.04	[−0.08, −0.00]	**.026**		
Syllable Structure Error Rate	−0.01	[−0.05, 0.02]	.463			−0.05	[−0.08, −0.01]	**.013**		
Word Stress Error Rate	−0.01	[−0.04, 0.02]	.504			−0.00	[−0.03, 0.03]	.971		
Complexity (Step 3)				−.00	.00				.00	−.01
Clauses per AS-unit	−0.01	[−0.05, 0.02]	.438			0.00	[−0.03, 0.03]	.951		
MTLD	−0.01	[−0.04, 0.02]	.696			0.01	[−0.02, 0.04]	.562		
COCA Mean Log Word Freq.	−0.02	[−0.06, 0.01]	.166			−0.03	[−0.06, 0.00]	.060		
Accuracy (Step 4)				.02	−.05				.02	−.04
Lexical Error Rate	−0.05	[−0.08, −0.01]	**.005**			−0.04	[−0.07, −0.01]	**.005**		
Morphosyntax Error Rate	−0.04	[−0.07, −0.01]	**.015**			−0.04	[−0.07, −0.01]	**.008**		
Syntax Error Rate	−0.07	[−0.10, −0.03]	<**.001**			−0.05	[−0.09, −0.02]	**.001**		

(*Continued*)

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

Table 6. (*Continued*)

	Comprehensibility					Acceptability				
Predictors	β	CI	*p*	ΔR^2	$\Delta\tau_{\text{Speaker}}$	β	CI	*p*	ΔR^2	$\Delta\tau_{\text{Speaker}}$
Fluency (Step 5)				.06	−.05				.07	−.05
Speech Rate	0.22	[0.16, 0.28]	<**.001**			0.23	[0.18, 0.28]	<**.001**		
Mid-clause Silent Pauses/minute	0.03	[−0.01, 0.08]	.100			0.02	[−0.01, 0.06]	.213		
Filled Pauses/minute	0.02	[−0.03, 0.08]	.421			0.03	[−0.02, 0.08]	.305		
Repairs/minute	0.01	[−0.03, 0.05]	.634			0.01	[−0.02, 0.05]	.549		
Random Effects										
σ^2	0.50					0.44				
τ_{00}	0.27 $_{\text{Listener}}$					0.35 $_{\text{Listener}}$				
	0.10 $_{\text{Speaker}}$					0.07 $_{\text{Speaker}}$				
Observations		4143					4143			
Marginal R^2/Conditional R^2	.09/.48					.10/.54				

Note. The ΔR^2 column represents changes in marginal R^2. Boldface indicates *p*-values <0.05.

Listen-Speak 2. In all task-specific models, random intercepts were included for speakers and listeners. As before, we were primarily interested in models with all relevant predictors, but intermediate models were constructed to investigate the incremental contribution of communicative function (except for Picture-Speak, for which communicative function is uniform), pronunciation, complexity, accuracy, and fluency variables.

Table 7 provides a summary of the task-specific models for comprehensibility and acceptability (full results tables, including all coefficients, are available in Online Supplement S4: Appendix B and C). Across all task types, speaker and listener random intercepts, accounted for in the conditional R^2 values, explained more variation in judgments (34–43%) than fixed effects (14–21%). In intercept-only models, the amount of variation in judgments explained by speakers and listeners was greater for acceptability than comprehensibility. For Read-Speak and Listen-Speak 1 and Listen-Speak 2, functional demands associated with prompts accounted for little variance explained in either listener judgment. Speech stream variables accounted for the greatest amount of variation in judgment for the two Listen-Speak performances. The complexity variables appeared to account for more variance in judgments for both Listen-Speak performances than for Picture-Speak and Read-Speak. Notably, the inclusion of fixed effects substantially reduced the amount of variation associated with speaker random effects (τ_S). Most drastically, the speaker random effect variance was reduced to .03 in the Listen-Speak 2 model for acceptability, down nearly 90% from the intercept-only model value of .24. The other acceptability and comprehensibility models with all variables reduced speaker random variation by 50–75%.

Looking at specific speech stream predictors, Fig. 4 compares the magnitudes of standardized coefficients across speakers' performances for comprehensibility and acceptability, with error bars indicating 95% confidence intervals. Speech rate was clearly the most influential and consistent predictor across task types and judgments, with positive, statistically significant values in all models. Lexical error rate also had fairly consistent negative effects across task types and judgments, though not statistically significant in every model. Conversely, some predictors were consistently near zero in magnitude and failed to achieve statistical significance across task types and judgments, including word stress error rate, clauses per AS-unit, MTLD, and all fluency variables aside from speech rate.

Other predictors showed similar magnitudes across judgments, but differed by task type. Substitution rate showed larger effects for both comprehensibility and acceptability in Picture-Speak and Listen-Speak 1, though it was only statistically significant as a predictor of comprehensibility in Listen-Speak 1. Syllable structure error rate had no impact in Picture-Speak, but had notable negative effects in Read-Speak and both Listen-Speak performances for both judgments. Morphosyntactic accuracy predictors also differed by task type. Morphological error rate was a larger negative predictor on Listen-Speak 2 (and a similar trend was seen in Listen-Speak 1), but had no large or reliable effect in Picture-Speak or Read-Speak. Conversely, syntactic error rate was a statistically significant and negative predictor in the Picture-Read and Read-Speak tasks for both judgments, but had smaller, nonsignificant effects in the two Listen-Speak performances.

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

Table 7. Variance explained (R^2) and speaker variance for task-specific comprehensibility and acceptability models

	Intercept-Only		+Functional Demand			+Pronunciation			+Complexity			+Accuracy			+Fluency		
	R^2		R^2			R^2			R^2			R^2			R^2		
Model	Cond.	τ_S	Marg.	Cond.	τ_S	Marg.	Cond.	τ_S	Marg.	Cond.	τ_S	Marg.	Cond.	τ_S	Marg.	Cond.	τ_S
Comprehensibility																	
Picture-Speak	.50	.24	–	–	–	.08	.51	.17	.09	.51	.16	.13	.51	.12	.16	.51	.09
Read-Speak	.47	.22	.00	.48	.22	.09	.48	.15	.10	.49	.14	.13	.49	.12	.14	.49	.11
Listen-Speak 1	.53	.26	.00	.54	.26	.10	.54	.18	.14	.54	.15	.17	.55	.12	.18	.55	.11
Listen-Speak 2	.55	.27	.01	.55	.27	.08	.55	.18	.10	.54	.16	.16	.54	.11	.21	.55	.07
Acceptability																	
Picture-Speak	.56	.21	–	–	–	.08	.57	.14	.09	.57	.14	.12	.57	.10	.15	.57	.07
Read-Speak	.56	.20	.00	.57	.20	.07	.57	.14	.09	.57	.13	.13	.57	.10	.15	.58	.08
Listen-Speak 1	.56	.25	.00	.56	.25	.10	.57	.16	.15	.57	.12	.17	.57	.10	.19	.57	.08
Listen-Speak 2	.60	.24	.01	.60	.24	.08	.60	.16	.11	.60	.12	.16	.60	.07	.21	.60	.03

Note. +Fluency corresponds to the final model with all fixed effects predictors added.
Marg. = Marginal R^2.
Cond. = Conditional R^2.
τ_S = Speaker random effect variance.

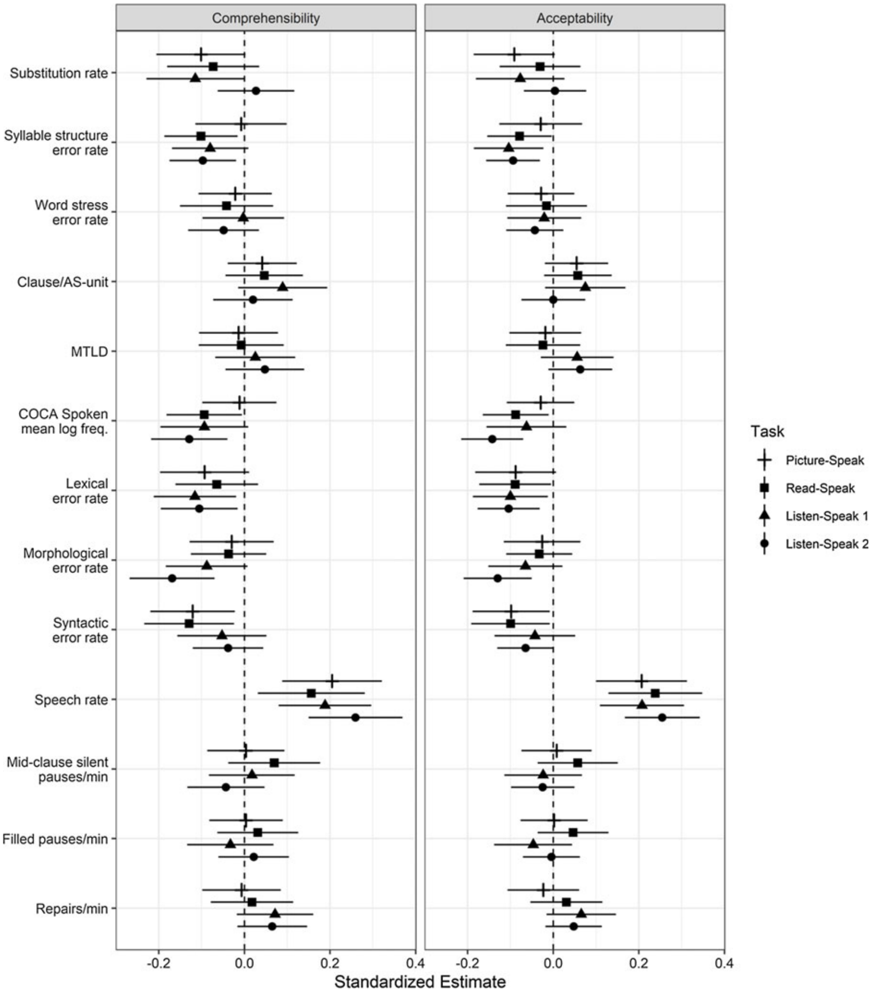


Figure 4. Coefficient plot for task-specific judgment models.

With regard to differences when listeners who passed fewer than 4/6 attention checks were excluded from analyses (see Online Supplement S3: Tables S3.3–S3.5), total amounts of variance accounted for (model R^2 and reductions in speaker random effect variance) were nearly identical. The magnitudes of fixed effect coefficients were also largely similar, but there was one instance where a speech stream predictor failed to achieve statistical significance: substitution rate in the comprehensibility model for Listen-Speak 1.

Discussion

We sought to understand in more depth the relationship between listeners' perceptions of the well-researched speech dimension of comprehensibility and the

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

less-understood dimension of acceptability. Importantly, acceptability in this study was conceptualized as listeners' perception of a speaker's ability to perform a specific role within a specific TLU domain (i.e., English-medium academic study, or academic acceptability). To shed light on this relationship, we not only considered the association between the two dimensions, but the linguistic profile of speech stream characteristics that appeared to influence listeners' judgments of each. This relationship was explored across four speaking performances per speaker, allowing for consideration of speakers' consistency across performances. We discuss the results of the study with respect to our two research questions, before concluding with additional considerations for the dimension of acceptability in L2 speech.

RQ1: the relationship between judgments of academic acceptability and comprehensibility

Given the strong association between the two dimensions ($r = .93$), and the similar speech stream profiles for each, it would seem that our findings constitute support for the idea that acceptability can be (almost) indistinguishable from comprehensibility, at least for this population of listeners judging speech in reference to particular TLU domain needs. The only speech stream characteristic which appeared to differ in the multi-task comprehensibility and acceptability models was syllable structure error rate, which had a statistically significant but small negative effect on acceptability judgments but nearly no effect on comprehensibility judgments. Nonetheless, the bivariate correlations between syllable structure error rate and comprehensibility and acceptability were similar in magnitude ($r = -.32$, $r = -.35$, respectively), which calls into question the meaningfulness of this difference.

Compared to previous studies on speech stream predictors of L2 English comprehensibility, the overall amount of variance accounted for by speech stream-based fixed effects in this study appears somewhat low on the surface. In the general models of comprehensibility and acceptability, inclusive of all four speaking performances, speech stream fixed effects accounted for only 8–10% of total variance in judgments. These amounts are comparable in magnitude to Dalman and Kang (2023), which found adjusted R^2 values ranging from .04 to .08 for linear regression models with four predictors for comprehensibility and acceptability. In comparison, and towards the other extreme, a recent study by Suzuki and Kormos (2020) found that a handful (5) of speech stream variables could explain over 92% of the variance in comprehensibility judgments made on argumentative speech elicited from 40 L1 Japanese speakers of English (see also R^2 values of .74–.87 in Crowther et al., 2015a, 2018; .50 in Kang et al., 2010; .86 in Trofimovich & Isaacs, 2012, etc.). We note, however, that fixed effects in this study did explain a substantial amount of variation in by-speaker random intercepts, by 56–65% in the multi-task models. We elaborate on this finding shortly.

RQ2: variation in judgments of academic acceptability, comprehensibility, and speech stream characteristics across speaking task types

The overall effects of speaking task type on listener judgments were minimal. Average judgments of comprehensibility across task types ranged between 3.92 and

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

4.01 and average judgments of acceptability between 4.11 and 4.23. In terms of speech stream characteristics, there were also some broad similarities across task types and dimensions, namely, speech rate and lexical errors had fairly consistent influence on judgments and several other measures consistently had null effects. However, there were several interesting differences in speech stream influences across task types. For one, more variance overall was explained by speech stream fixed effects in both Listen-Speak performances for both dimensions. Furthermore, syllable structure errors had statistically significant effects on Read-Speak and both Listen-Speak performances, but not Picture-Speak. While we did not have access to the prompts speakers responded to (but see the example item in the *Official Guide for Test Takers,* Duolingo, 2022), it may be the case that the pictures used were less likely to elicit multisyllabic words in the first place. There was also an interesting dynamic where morphological errors had more pronounced influence on the Listen-Speak performances while syntactic errors had more influence on Picture-Speak and Read-Listen. While this finding is more difficult to explain and interpret, it nonetheless suggests that task types have some influence on speaker performance and/or ensuing listener judgments. These subtle differences are contrasted by studies such as Crowther et al. (2015a) and Crowther et al. (2018), where more distinct profiles of speech stream influences on comprehensibility emerged in tasks that varied in complexity. In both studies, lexicogrammatical measures appeared more influential for a more complex integrated task, in which speakers were required to employ greater reasoning and perspective-taking, versus less complex long turn and picture narration tasks. However, we note that whereas both Crowther et al. (2015a) and Crowther et al. (2018) compared distinctly different tasks, the tasks included in the current study might be considered more similar than they were different, with the exception of the picture description task. Both Read-Speak and Listen-Speak required speakers to respond to a prompt, with the only task differences being the modality of prompt delivery and the availability of the reading prompt during speakers' response. As such, the largely similar profiles of speech stream influences may be due to speakers drawing on similar processes to respond to prompts.

A key characteristic of the current study was that, given the source of speech elicited (high-stakes English proficiency test [i.e., DET]), it was possible to include speakers representing a wide range of linguistic backgrounds and overall proficiency. That is, our population included both those who would and those who would not receive admission into English-medium university study (see Isbell et al., 2023), ranging in proficiency from roughly CEFR levels B1 to C1. This is in contrast to many existing studies, where the speaker population was more homogenous. For example, Suzuki and Kormos (2020) included only L1 Japanese speakers of English, with most at the CEFR B1-B2 proficiency level. While Crowther et al. (2015a, 2018) did include speakers from several different L1s, they had all achieved at minimum a proficiency level high enough to allow for undergraduate study at an English-medium university. As a final point of comparison, while speakers in the current study all completed the same battery of DET speaking tasks, the prompts they completed were drawn from a large item pool, minimizing the potential effects of prompt familiarity on listeners' judgments. Given the earlier referenced differences in variance explained with previous speech rating research,

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

from the listener's perspective, speech stream characteristics may have a larger influence on judgments when repeatedly hearing speech elicited using the same prompt delivered by speakers of a common L1 background and similar range of proficiency. With a uniform prompt, the topic, content, and even some linguistic features (e.g., topical vocabulary, grammar structures, discourse markers) of responses may overlap more from speaker to speaker. This similarity may reduce listener cognitive efforts on comprehending meaning and allow for closer attention to fluency, disfluencies, and phonological form when judging such speech.

Alignment across speech performances

One notable contribution of this study is the examination of L2 speech dimensions based on listener judgments (comprehensibility, acceptability) as attributes of speakers that are generalizable across several speaking performances. Before accounting for speech stream characteristics, comprehensibility and acceptability scores showed moderate levels of consistency within speakers across performances (i.e., ICC values). We also observed notable consistencies of several speech stream characteristics within speakers, particularly related to fluency and segmental pronunciation phenomena, while other speech characteristics showed more within-speaker variation across performances, like lexicogrammatical accuracy and complexity. Accounting for these features as fixed effects in models showed how characteristics of individual spoken performances systematically explained variance in judgment outcomes within and across speakers, as inclusion of such variables as fixed effects reduced the total amount of speaker random intercept variation in the all-task general models (by 56–65%) and task-specific models (by 50–90%). In other words, both consistent and less consistent aspects of speaker performance have roles to play in determining the outcomes of speaker-listener encounters.

Of course, not all variation in judgments was explained. Remaining random variation associated with speakers may be accounted for by speech stream characteristics and other variables we did not include in this study, such as suprasegmental measures related to prosody (e.g., Trofimovich & Isaacs, 2012), discourse features (Suzuki & Kormos, 2020) or collocation use (Saito, 2020; Saito & Liu, 2022). Similarly, the substantial variation associated with listeners could also likely be explained by a number of factors, such as L1/L2 status of English, academic role, experience interacting with L2 English speakers, and so on (e.g., Isaacs & Thomson, 2013; O'Brien, 2016). Variation unattributed to speakers and listeners is likely to be accounted for, in part, by differences in specific prompts.

Limitations

Due to the nature of the sparse rating design we employed (i.e., not all listeners rated all speech samples), we were limited in our ability to account for the potential effect of listener variables such as L1/L2 status or accent familiarity, both of which have been found to influence listener perception of global speech dimensions (e.g., Ballard & Winke, 2017; Crowther et al., 2016). As referenced, in Isbell et al. (2023), the four listener groups (undergraduates, graduate students, faculty, staff), as a whole, tended to rate similarly across speech dimensions, which may help to alleviate concerns regarding such individual differences, though how such differences may influence ratings of academic acceptability should be considered in future research. An additional methodological limitation is that the simultaneous judging of comprehensibility and acceptability may have inflated the similarity of those judgments. Though O'Brien (2016) found that the rating of three perceptual dimensions (accentedness, comprehensibility, fluency) separately versus simultaneously resulted in only minor differences in listener judgments, the high association between aggregated comprehensibility and acceptability judgments ($r = .93$) found here may suggest that reconsideration and/or further investigation might be necessary. Finally, we acknowledge that the choice of linguistic measures, and the way in which these measures are coded, is variable across studies. Decisions made for the current study were made drawing upon existing research, but also simultaneously accounting for the need to code over 400 minutes of speech data. Refinement of our linguistic coding procedures could potentially lead to variations in our findings.

Conclusion

Findings pertaining to the association between comprehensibility and acceptability and the patterning of speech stream characteristics that influenced judgments of each construct support the idea that acceptability can be (almost) indistinguishable from comprehensibility (Thomson, 2018). This finding, of course, comes with a pair of caveats. For one, the emphasis here was on academic acceptability, where elicited ratings were specific to the academic items presented to listeners. The same speech, presented for judgment to stakeholders representing different TLU domains (e.g., business, tourism), may be viewed quite differently. Any interpretation of acceptability, as operationalized in the current study, should remain within the specific TLU domain of interest. Second, the speech elicited was not representative of *in situ* language use. Thus, listeners' judgments were not based on speech representative of actual academic production. This leaves a question regarding to what extent "acceptable" speech, as deemed in the current study, would predict success in the actual academic domain (see, for example, Bridgeman et al., 2012, or Schmidgall & Powers, 2021). Yet, despite such concerns, judgments across listeners were found to be highly reliable. As all listeners were stakeholders within the TLU domain of interest, this would provide support for a claim that even if speech may not be representative of actual in situ performance, stakeholders with knowledge of the TLU domain share, at least to some extent, a degree of criteria for what constitutes acceptable speech for a given academic role. An additional finding of interest, and one seemingly less explored in L2 speech research, was that speakers' productions generated moderately similar listener perceptions across performances. Whereas prior studies (e.g., Crowther et al., 2015a, 2018) have emphasized how judgments of speech and linguistic/temporal influences on those judgments differ across tasks, we here highlight how consistencies in the characteristics of speakers' performances, and presumably underlying competence, are identifiable and also play a role in predicting individual outcomes of listener judgments. Further research into these cross-task consistencies will greatly enhance our understanding of L2 speaking ability and what determines how listeners make judgments of comprehensibility and acceptability.

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

Data availability statement. All raw data/materials from this study, which includes official Duolingo English Test speaking performances and transcriptions, are not available. Data used in statistical analyses are available.

Replication package. Research materials and data that can be made publicly available are accessible at https://osf.io/2ujfa/.

Acknowledgments. We would like to thank Duolingo English Test personnel for reading and commenting on earlier versions of this article to verify our description of test content and procedures. As all data collection and analysis described in this study were conducted independently, any and all remaining errors are our own.

Funding statement. Funding from Duolingo awarded to Daniel R. Isbell and Dustin Crowther supported this study.

Competing interests. The authors have no conflicts of interest to disclose.

Ethics approval statement. This study was conducted in accordance with *Applied Psycholinguistics* informed consent guidelines. Study protocols were reviewed and approved by the University of Hawai'i Office of Research Compliance Human Studies Program (Protocol ID 2020-00996).

Notes

[1] Despite reporting English as their L1, all three speakers completed the DET as a measure of English proficiency for the purposes of admissions into an English-medium university.

[2] For comparison's sake, a DET score of 108 may be seen as roughly equivalent to a TOEFL iBT score of 80–85, an IELTS score of 6.5, and a Common European Framework of Reference (CEFR) level of B2 (Cardwell et al., 2022).

[3] Data and R script used to calculate inter-rater reliability for both pronunciation and accuracy measures are available at https://osf.io/2ujfa/.

[4] Speech rate was chosen rather than articulation rate, as speech rate was a) the focus of Munro & Derwing's (2001) investigation into the effects of fluency on comprehensibility ratings, and b) found to have a higher association with speaking proficiency ($r = .77$ vs. $r = .55$) in Yan et al., 2021. Suzuki et al's (2021) meta-analysis similarly found speech rate to correlate more strongly than articulation rate with listeners' perception of fluency.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, **52**(1), 388–407.

Barr, D. J. (2018). Generalizing over encounters: Statistical and theoretical considerations. In S.-A. Rueschemeyer & M. G. Gaskell (Eds.), *The Oxford handbook of psycholinguistics* (pp. 917–929). Oxford University Press.

Ballard, L., & Winke, P. (2017). Students' attitude towards English teachers' accents: The interplay of accent familiarity, comprehensibility, intelligibility, perceived native speaker status, and acceptability as a teacher. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 121–140). Multilingual Matters.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, **41**(1), 5–35.

Boersma, P., & Weenink, D. (2022). *Praat: Doing phonetics by computer [Computer program]*. Version 6.2.14. Retrieved 29 July 2022 from http://www.praat.org/

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing*, **29**(1), 91–108.

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

Brueckl, M., & Heuer, F. (2021). *irrNA: Coefficients of interrater reliability – generalized for randomly incomplete dataset in R.* R package version 0.2.2. Retrieved from https://cran.r-project.org/package=irrNA

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* SAGE.

Cardwell, R., LaFlair, G. T., & Settles, B. (2022). *Duolingo English test: Technical manual.* Duolingo Research Report. Retrieved from https://duolingo-testcenter.s3.amazonaws.com/media/resources/techinical_manual.pdf

Crowther, D., Holden, D., & Urada, K. (2022). Second language speech comprehensibility. *Language Teaching*, **55**(4), 470–489.

Crowther, D., Trofimovich, P., & Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility: Nonnative listeners' perspectives. *Journal of Second Language Pronunciation*, **2**(2), 160–182.

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

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2018). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition*, **40**(2), 443–457.

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

Dalman, M., & Kang, O. (2023). Validity evidence: Undergraduate students' perceptions of TOEFL iBT high score spoken responses. *International Journal of Listening*, **37**(2), 113–126.

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, **41**(2), 385–390.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, **42**(4), 476–490.

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

Dragojevic, M., & Giles, H. (2016). I don't like you because you're hard to understand: The role of processing fluency in the language attitudes process. *Human Communication Research*, **42**(3), 396–420.

Duolingo (2022). *Official guide for test takers.* Duolingo, Inc. Retrieved from https://englishtest.duolingo.com/guide

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, **37**(3), 313–326.

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

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, **21**(3), 354–375.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, **61**(1), 29–48.

Hosoda, M., & Stone-Romero, E. (2010). The effects of foreign accents on employment-related decisions. *Journal of Managerial Psychology*, **25**(2), 113–132.

Huensch, A., & Nagle, C. (2021). The effect of speaker proficiency on intelligibility, comprehensibility, and accentedness in L2 Spanish: A conceptual replication and extension of Munro and Derwing (1995a). *Language Learning*, **71**(3), 626–668.

Isaacs, T. (2018). Shifting sands in second language pronunciation teaching and assessment research and practice. *Language Assessment Quarterly*, **15**(3), 273–293.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, **10**(2), 135–159.

Isbell, D. R. (2018). Assessing pronunciation for research purposes with listener-based numerical scales. In O. Kang and A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 89–111). Routledge.

Isbell, D. R., Crowther, D., & Nishizawa, H. (2023). Speaking performances, stakeholder perceptions, and test scores: Extrapolating from the Duolingo English Test to the university. *Language Testing*. Published online 24 April 2023.

Isbell, D. R., Park, O. S., & Lee, K. (2019). Learning Korean pronunciation: Effects of instruction, proficiency, and L1. *Journal of Second Language Pronunciation*, **5**(1), 13–48.

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly*, **9**(3), 249–269.

Kang, O., Rubin, D., & Lindemann, S. (2015). Mitigating US undergraduates' attitudes toward international teaching assistants. *TESOL Quarterly*, **49**(4), 681–706.

Kang, O., Rubin, D. O. N., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, **94**(4), 554–566.

Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, **68**(1), 115–146.

Kennedy, S., & Trofimovich, P. (2019). Comprehensibility: A useful tool to explore listener understanding. *Canadian Modern Language Review*, **75**(4), 275–284.

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, **50**(3), 1030–1046.

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, **18**(2), 154–170.

Lambert, W. E., Hodgson, R. C., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *The Journal of Abnormal and Social Psychology*, **60**(1), 44–51.

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, **46**(6), 1093–1096.

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

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

Levis, J. (2020). Changes in L2 pronunciation: 25 years of intelligibility, comprehensibility, and accentedness. *Journal of Second Language Pronunciation*, **6**(3), 277–282.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, **42**(2), 381–392.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, **1**(1), 30–46.

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, intelligibility, and comprehensibility in the speech of second language learners. *Language Learning*, **45**(1), 73–97.

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, **38**(3), 289–306.

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech the role of speaking rate. *Studies in Second Language Acquisition*, **23**(4), 451–468.

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, **34**(4), 520–531.

Munro, M. J., & Derwing, T. M. (2015). A prospectus for pronunciation research in the 21st century: A point of view. *Journal of Second Language Pronunciation*, **1**(1), 11–42.

Munro, M. J., & Derwing, T. M. (2020). Foreign accent, comprehensibility, and intelligibility, redux. *Journal of Second Language Pronunciation*, **6**(3), 283–309.

O'Brien, M. G. (2016). Methodological choices in rating speech samples. *Studies in Second Language Acquisition*, **38**(3), 587–605.

Plakans, B. S. (1997). Undergraduates' experiences with and attitudes toward international teaching assistants. *TESOL Quarterly*, **31**(1), 95–119.

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal*, **102**(4), 713–731.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, **64**(4), 878–912.

Prikhodkine, A. (2018). Language regard and sociolinguistic competence of non-native speakers. In B. E. Evans, E. J. Benson, & J. N. Stanford (Eds.), *Language regard: Methods, variation, and change* (pp. 218–238). Cambridge University Press.

Saito, K. (2020). Multi-or single-word units? The role of collocation use in comprehensible and contextually appropriate second language speech. *Language Learning*, **70**(2), 548–588.

https://doi.org/10.1017/S0142716423000346 Published online by Cambridge University Press

Saito, K. (2021). What characterizes comprehensible and native-like pronunciation among English-as-a-second-language speakers? Meta-analyses of phonological, rater, and instructional factors. *TESOL Quarterly*, **55**(3), 866–900.

Saito, K., & Liu, Y. (2022). Roles of collocation in L2 oral proficiency revisited: Different tasks, L1 vs. L2 raters, and cross-sectional vs. longitudinal analyses. *Second Language Research*, **38**(3), 531–554.

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism: Language and Cognition*, **19**(3), 597–609.

Schmidgall, J., & Powers, D. E. (2021). Predicting communicative effectiveness in the international workplace: Support for *TOEIC(R)* Speaking test scores from linguistic laypersons. *Language Testing*, **38**(2), 302–325.

Sewell, A. J. (2012). The Hong Kong English accent: Variation and acceptability. *Hong Kong Journal of Applied Linguistics*, **13**(2), 1–21.

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

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *The Modern Language Journal*, **105**(2), 435–463.

Szpyro-Kozłowska, J. (2014). *Pronunciation in EFL instruction: A research-based approach*. Multilingual Matters.

Thomson, R. I. (2018). Measurement of accentedness, intelligibility and comprehensibility. In O. Kang and A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 11–29). Routlege.

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, **15**(4), 905–916.

Trofimovich, P., Tekin, O., & McDonough, K. (2021). Task engagement and comprehensibility in interaction: Moving from what second language speakers say to what they do. *Journal of Second Language Pronunciation*, **7**(3), 435–461.

Tsunemoto, A., Lindberg, R., Trofimovich, P., & McDonough, K. (2022). Visual cues and rater perceptions of second language comprehensibility, accentedness, and fluency. *Studies in Second Language Acquisition*, **44**(3), 659–684.

Varonis, E. M., & Gass, S. M. (1982). The comprehensibility of nonnative speech. *Studies in Second Language Acquisition*, **4**(2), 114–136.

Wagner, E. (2020). Duolingo English test, revised version July 2019. *Language Assessment Quarterly*, **17**(3), 300–315.

Yan, X., Kim, H. R., & Kim, J. Y. (2021). Dimensionality of speech fluency: Examining the relationships among complexity, accuracy, and fluency (CAF) features of speaking performances on the Aptis test. *Language Testing*, **38**(4), 485–510.

Cite this article: Crowther, D., Isbell, D. R., & Nishizawa, H. (2023). Second language speech comprehensibility and acceptability in academic settings: Listener perceptions and speech stream influences. *Applied Psycholinguistics* **44**, 858–888. https://doi.org/10.1017/S0142716423000346