Bilingualism: Language and Cognition

cambridge.org/bil

The relationship between intelligibility and comprehensibility in second language speech

Dustin Crowther[1], Daniel Richard Isbell[1], Yoonseo Kim[1] and Jieun Kim[2]

[1]Department of Second Language Studies, University of Hawai'i at Mānoa, Honolulu, HI, USA and [2]Department of English Language and Literature, Soongsil University, Seoul, South Korea

Research Notes

Cite this article: Crowther, D., Isbell, D.R., Kim, Y. and Kim, J. (2025). The relationship between intelligibility and comprehensibility in second language speech. *Bilingualism: Language and Cognition* 1–7. https://doi.org/10.1017/S1366728925100606

Received: 17 October 2024
Revised: 26 August 2025
Accepted: 5 September 2025

Keywords:
comprehensibility; intelligibility; pronunciation; second language; speaking

Corresponding author:
Dustin Crowther;
Email: dcrowth@hawaii.edu

Abstract

This study examined the relationship between intelligibility and comprehensibility in second language speech. Four extended speech samples from 50 speakers spanning a wide range of proficiency were drawn from archived test data. These samples were listened to by 570 English users, who provided comprehensibility ratings and transcriptions to measure intelligibility. The relationship between intelligibility and comprehensibility was strong ($r = .81$, $\rho = .88$) and nonlinear. A segmented regression model suggested a breakpoint for intelligibility scores (transcription accuracy) at 64%, below which speakers were perceived as uniformly hard to understand and above which increased intelligibility was strongly associated with higher comprehensibility.

Highlights

- A strong relationship exists between intelligibility and comprehensibility.
- Relationship between intelligibility and comprehensibility is nonlinear.
- High proficiency speakers (CEFR C, B2) were both intelligible and comprehensible.
- Intermediate speakers (B1, A2) were more intelligible than comprehensible.
- Low proficiency speakers (A1) were low in intelligibility and comprehensibility.

1. Literature review

Second language (L2) speech research commonly employs three listener-based global dimensions: *accentedness* (i.e., how distinguishable an L2 utterance is from that of the target community), *comprehensibility* (i.e., perceived ease/difficulty of understanding an L2 utterance) and *intelligibility* (i.e., accuracy of understanding an L2 utterance) (Munro & Derwing, 1995). Accentedness and comprehensibility are commonly assessed through scalar ratings (Derwing & Munro, 2015), while intelligibility is most commonly assessed through transcription (see Kang et al., 2018). Research has provided consistent evidence that L2 speakers can produce speech that is both intelligible and comprehensible, even in the presence of a (strong) foreign accent (see Crowther et al., 2022; Munro & Derwing, 2011). A number of studies have considered the strength of association between accentedness and comprehensibility, finding moderate to (very) strong correlations (e.g., Munro & Derwing, 1995; Trofimovich & Isaacs, 2012). Theoretically, intelligibility should be a precursor to comprehensibility, as speech low in intelligibility is unlikely to be comprehensible (Derwing & Munro, 2015; Thomson, 2018). More simply, speech that is unintelligible will rarely be perceived as anything other than low in comprehensibility (i.e., a listener who does not understand the intended message would perceive the speech as difficult to comprehend), though speech that is low in comprehensibility may still be intelligible (i.e., speech may be understood, though it requires great effort to do so). However, limited empirical attention has been given to this specific relationship to confirm this belief (see also Chau & Huensch, 2025).

Studies that have featured a correlation between intelligibility and comprehensibility have found varying strengths of association. At the listener level (i.e., within-listener comparisons of transcription accuracy and comprehensibility ratings), Munro and Derwing (1995) found that first language (L1) English listeners hearing L2 English speech had a significant correlation, with a mean correlation of $r = .51$. Jułkowska and Cebrian (2015) further reported strong correlations of .67, .80 and .83 among L1 Polish, Spanish and English listeners, respectively, while Hansen-Edwards et al. (2018), working with speakers and listeners of different English varieties, found a nonsignificant correlation of $r = .08$. Studies that focus on variation among speakers are few in number, but have found strong correlations between increasing intelligibility and comprehensibility of around $r = .60$ in L2 Arabic (Ali, 2023) and L2 Spanish (Nagle & Huensch, 2020[1]), though only .28 in L2 English (Gallant, 2023). Chau and Huensch (2025), in a

[1]We used Nagle and Huensch's (2020) open data to calculated the correlation at $r = .61$, aggregated across sentences and all ratings per speaker.

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

meta-analysis of studies focused on global dimensions of L2 speech, reported a correlation of .57 between intelligibility and comprehensibility across 17 identified studies. While these studies are informative, research designs have typically included smaller numbers of listeners and/or speakers, speakers from limited ranges of L2 proficiency and L1 backgrounds and transcriptions of only short, decontextualized speech excerpts, which may limit our understanding of the full scope of the relationship between intelligibility and comprehensibility. Although seemingly intuitive that intelligibility precedes comprehensibility, the extent to which this holds true and whether such a relationship is completely linear remains in need of empirical support.

To address these concerns, we draw from extended, extemporaneous speech from the speaking portion of Aptis General, a standardized English proficiency exam developed by the British Council (O'Sullivan et al., 2020). Study data included test takers from a range of proficiency levels and L1 backgrounds. For these test takers, we measured intelligibility via transcriptions and comprehensibility via ratings elicited from a large number of L1 English layperson listeners. Working off the premise that intelligibility is a precursor to comprehensibility (i.e., the message of the speaker must be received before speech can be easily understood), we sought to answer the following research question:

To what extent do listener measurements of intelligibility predict listener measurements of comprehensibility?

Given the range in correlation strength in the previously highlighted research, we make no specific prediction here on the extent intelligibility predicts comprehensibility. As such, the study reported next can be considered exploratory in nature.

2. Methodology

2.1. Participants

2.1.1. Aptis test takers and speech samples

We obtained archived speaking performances, courtesy of the British Council, from 50 Aptis general test takers (hereafter speakers) spread across 19 operational forms of the test, which included four samples per speaker (one per Aptis speaking task: short response, picture description, picture comparison, long response). The sample of speakers were stratified by overall Common European Framework of Reference (CEFR) Speaking level (based on Aptis speaking score), with 10 at each of the A1–C levels and featured seven countries of residence/test location (Vietnam = 15, Spain = 10, Albania = 5, Azerbaijan = 5, Colombia = 5, Mexico = 5, Saudi Arabia = 5) that resulted in the inclusion of L1 Albanian, Arabic, Azer, Spanish and Vietnamese speakers. Their Aptis speaking section scores ranged from 5 to 48 (out of 50; mean = 31.18, SD = 13.81). In total, 198 speaking samples were included in analyses, as two separate A1-level speakers did not produce audible speech in one of their responses (one for Task 1, one for Task 4).

2.1.2. Listeners

We recruited 570 listeners via the online research participant pool *Prolific* (https://www.prolific.co), which included 340 US-based (female = 156, male = 184) and 230 UK-based (female = 102, male = 128) listeners. The eligibility criteria for listeners required them to be aged 18–64 (mean = 36.50, SD = 10.60, min = 18, max = 64), use English as their primary language, have no reported hearing or language impairments, and be capable of touch typing. Listeners were recruited from both the United States and United Kingdom, as these two locations are common destinations for Aptis test takers (i.e., the speaking sample in this study). Listeners indicated they had lived in a predominantly English-speaking country for 33.80 years (SD = 12.60) and used English 97% of the time (SD = 12%) in their current profession. They assessed their familiarity with non-native English speech as 5.92 (SD = 2.35) on a 9-point scale (1 = "not at all familiar," 9 = "very familiar").

2.2. Procedure

2.2.1. Speech sample preparation

All speech samples were edited using Audacity 3.3.3 (https://www.audacityteam.org/) to enhance audio clarity and minimize background noise. Extended silences at the outset and conclusion were edited out of each sample, as was any personally identifying information. Each audio file was initially transcribed using Amazon Transcribe (Amazon Web Services, n.d.). Manual corrections were initially performed by the first author, with subsequent corrections made by the other authors, as needed. Speech tokens deemed as unintelligible after review by all authors and after reference to task prompts were recorded as /XXXX/.

For comprehensibility ratings, the 198 full speech samples had a mean duration of 106.67 s (SD = 30.25 s, range = 21.58–135.24 s). For intelligibility ratings, our research team segmented each of the 198 samples into AS units. An AS unit is 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). In cases where AS-unit-based segments exceeded 24 words, we identified a logical break in the segment (e.g., clause boundary) to allow for additional segmentation. Initial segmentation was conducted by the first author and then reviewed by the second author. In total, segment length ranged from 2 to 24 words, with a mean length of 9.31 words. The resulting procedure yielded 2846 files for intelligibility transcriptions (mean # of segments = 14.23 per speaker per task, range = 2–33). The number and duration of segments by task are detailed in Table 1.

2.2.2. Data collection

All data collection procedures were carried out remotely using listeners' personal computers via the online experiment platform Gorilla (Anwyl-Irvine et al., 2020). After providing consent and filling in a background questionnaire, listeners completed the experiment. As a means to mitigate potential listener biases (e.g., Kang & Rubin, 2014), listeners knew only that the speech they would hear was from a set of L2 speakers. Given the large number of speech samples to be rated and transcribed, a sparse rating design was employed (i.e., not all listeners heard all files; see Isbell, 2018). First, the 198 speech samples were divided into 4 blocks of 50 files

Table 1. Number and duration of segments in intelligibility speech samples

	No. of segments			Duration of segments (milliseconds)		
	Mean	*SD*	Min–max	Mean	*SD*	Min–max
Task 1	12.7	5.43	3–23	5881	3290	1149–26,352
Task 2	15.8	7.13	3–33	6961	4137	1248–34,824
Task 3	15.8	6.47	3–31	7270	3922	1019–25,176
Task 4	13.1	6.50	2–26	7231	4265	816–32,448
Average	14.4	6.53	2–33	6872.82	3972.58	816–34,824

https://doi.org/10.1017/S1366728925100606 Published online by Cambridge University Press

each (hereafter labeled A, B, C, D). As adequate overlap between listeners is necessary within a sparse rating design, each of the four blocks were then combined to form four supersets of speech samples (AB, BC, CD, DA; i.e., each individual block existed in 2 supersets, with each superset thus comprising 100 files). Each superset was finally divided so that the 100 files were grouped into 25 subsets of 4 (i.e., across the 4 supersets there existed 100 subsets). Each subset was balanced for task, test taker's CEFR level and test taker's L1. Each listener was then assigned to two supersets (e.g., AB and CD) and completed one subset within each (e.g., one subset within AB for comprehensibility ratings, one subset within CD for intelligibility transcriptions). More simply, each listener assessed four speech samples for comprehensibility and transcribed four speech sample sets for intelligibility. This design minimized, but did not eliminate, the chance of a listener being exposed to responses to the same prompt across their comprehensibility and intelligibility assessments. Aggregated at the speaker level (across all four tasks), there was an average of 45.3 judgments (SD = 2.02, min = 41, max = 49) per speaker.

Comprehensibility was assessed on a 9-point scale (1 = "hard to understand," 9 = "easy to understand"). Listeners could select (and change) their rating at any time while hearing each sample, but were required to listen to the entirety of each file. Prior to beginning the comprehensibility section, listeners were asked to rate three practice speech samples, which were chosen to represent speakers with low (1–3), moderate (4–6) and high (7–9) comprehensibility.

Intelligibility was measured via transcription accuracy. Listeners transcribed complete Aptis speaking task responses presented as a series of consecutive audio segments. For each segment, listeners could begin typing the moment the sound file began to play and manually advance to the next segment once the audio finished and they were satisfied with their transcription of the segment. They were only permitted to hear each segment once. During training, listeners were informed that they could ignore punctuation, capitalization and fillers (e.g., uh, eh). In total, listeners transcribed the speech of four speakers (one task each). Prior to beginning the intelligibility section, listeners completed a typing test as a warm up and were then asked to transcribe the speech of two practice speakers.

2.3. Analyses

To investigate the relationship between intelligibility and comprehensibility, correlational (Pearson and Spearman; the former selected in advance for comparability with previous studies and the latter run due to apparent non-linearity of relationships) and regression analyses (linear and segmented) were run in R, including the following packages: *correlation* (v.0.8.3, Makowski et al., 2019) and *segmented* (v. 2.0-4, Muggeo, 2008). Correlation assumptions were assessed graphically and regression model assumptions were tested using the *performance* package (v.0.12.3, Lüdecke et al., 2021). Prior to running analyses, several steps were taken to ensure the data quality.

2.3.1. Data quality checks

For comprehensibility ratings, we examined data quality by looking for uniform response patterns in the training task and the primary task. Ultimately, we decided to exclude two listeners' data from the comprehensibility analyses. These two listeners demonstrated uniform response patterns (all comprehensibility ratings in the primary rating task were the same, i.e., all 4s, all 6s). Three other listeners had no variability in their practice judgments, but exhibited variation in their ratings for the primary task and were retained in analyses. In total, 567 listeners were included in comprehensibility analyses. For intelligibility, seven listeners were excluded due to one or both of the following reasons: (1) a large proportion of blank transcriptions (no transcription for over 50% of assigned segments) or (2) an average transcription length that was very short (<20 characters). One of the two listeners who were identified as contributing poor-quality data to the comprehensibility ratings was also flagged and removed from intelligibility analyses; the other listener was found to have demonstrated adequate effort/completion of transcriptions and was maintained. In total, 562 listeners were included in intelligibility analyses.

2.3.2. Scoring intelligibility responses

We scored transcriptions by calculating the proportion (expressed as a percentage) of words in a listener's transcription which were in a criterion transcription. From original verbatim transcriptions (see *Speech sample preparation*), a criterion transcript which excluded extraneous elements, such as filled pauses (e.g., *um*), repetitions (e.g., *the the*) and self-corrections (with the correction being preserved), was created to have the evaluation of listener transcriptions best reflect the message a speaker intended to convey. For each segment, a listener's transcription was searched for a match with each word of the criterion transcription and an aggregate accuracy rate of a listener's transcription was calculated based on all segments in a sample.

2.3.3. Aggregating comprehensibility and intelligibility judgments

To arrive at aggregate measures of intelligibility and comprehensibility that accounted for differences in listeners' ability to decode and transcribe speech (intelligibility) and severity of judgment (comprehensibility), we used many-facet Rasch measurement (MFRM) to analyze the listener response data in *Facets* (Linacre, 2021). Sparse rating designs, as used in this study, are also easily accommodated in MFRM provided that elements are adequately linked (Isbell, 2018; Myford & Wolfe, 2000). We constructed three-facet models which included facets for speakers, listeners and Aptis speaking tasks to produce measures of ability, severity and difficulty, respectively. For the comprehensibility model, a 9-point rating scale was specified, but for intelligibility we modeled the percentages using a scale based on 100 binomial trials. MFRM ability measures for speakers can be represented on the original scale as *fair averages*, which are adjusted to account for differences in other facets (e.g., listener severity). Thus, for intelligibility we transformed Rasch-based estimates of ability to percentages (ranging from 0% to 100%) and comprehensibility ability estimates were transformed to range between 1 and 9 points.

To generate fair averages for each task performance, we estimated models using data only from a given task. Because there were not adequate links in the data when each task was analyzed separately, we anchored the listener measures from the aggregated models, allowing us to generate task-level speaker estimates based on known levels of listener ability/severity.

3. Results

Table 2 provides a descriptive summary of all study variables. The relationship between aggregated intelligibility and comprehensibility (based on speech from across all four tasks) among the 50 speakers

https://doi.org/10.1017/S1366728925100606 Published online by Cambridge University Press

Table 2. Summary of study variables

	N	Mean	*SD*	Median	Min	Max
Comprehensibility (1–9)	50	4.54	2.15	3.90	1.19	8.60
Intelligibility (%)	50	72.06	15.13	75.65	20.06	89.85

Note: Comprehensibility and intelligibility values are computed as Rasch fair averages.

was strong: Pearson's r = .81 [.69, .89], p < .001; Spearman's ρ = .88 [.80, .93], p < .001. As shown in Figure 1 (upper panel), this relationship did not appear to be linear; the relationship between intelligibility and comprehensibility became stronger above roughly 60% intelligibility. Notably, speakers at higher CEFR levels for speaking (based on their Aptis speaking scores) tended to have higher intelligibility and comprehensibility, with A1 speakers making up the majority of speakers below 60%. The relationship between aggregate intelligibility and comprehensibility measures was largely consistent across the four Aptis speaking tasks (r = .77–.83, ρ = .80–.89, see Figure 1, lower panels and Online Supplementary Materials). As such, we focus on aggregate measures.

Theoretically, intelligibility is a precursor to comprehensibility (i.e., speech that is unintelligible will rarely be perceived as comprehensible, though speech low in comprehensibility may still be intelligible). Accordingly, we fit several regression models in which intelligibility predicted comprehensibility: a linear model, two higher order polynomial models (a model with squared intelligibility and a model with a linear and squared intelligibility term) and a segmented model with a single breakpoint (Table 3). All nonlinear models explained more variance than the linear model, with the linear + squared polynomial model and the segmented model explaining similar proportions of variance (~76%). The latter model provides a useful point of reference: When aggregate intelligibility across the four tasks was less than 63.9%, the slope for intelligibility was very small and not statistically different from zero, but for intelligibility greater than 63.9% there was a large, positive, statistically significant slope. Past 63.9% intelligibility, each 10-percentage-point increase in intelligibility was associated with a 1.5-point gain in comprehensibility.

4. Discussion

The relationship between intelligibility and comprehensibility was strong (ρ = .88). Although not uniform across studies, a similar positive relationship between these two global L2 speech measures has been found elsewhere (e.g., Chau & Huensch, 2025; Munro & Derwing, 1995) and indicates that listeners' comprehensibility ratings reflect, to a degree, their actual understanding of a given utterance, based on their ability to accurately transcribe what was said. A novel and important contribution of this study was the observation that the relationship between intelligibility and comprehensibility was nonlinear. Working off the premise that intelligibility is a precursor to comprehensibility, a segmented regression model indicated that when intelligibility was < 64% (i.e., listeners correctly transcribed less than 64% of words heard), intelligibility was not predictive of comprehensibility. Past this threshold, intelligibility did indeed predict comprehensibility, with a 10 percentage-point increase in intelligibility predictive of a comprehensibility gain of 1.5 points on the 9-point scale. Such a relationship may be indicative that any speech < 64% intelligible is likely to be assessed as low in comprehensibility and it is only beyond this threshold that listeners begin to experience a reduced strain in understanding.

When considering intelligibility and comprehensibility across proficiency levels, the majority of A1 speakers produced speech near or below the 64% threshold. In all but one case, these A1 speakers additionally were rated as low in comprehensibility (≤ 3, on the 9-point scale). When considering intelligibility > 64%, speakers tended to increase in both intelligibility and comprehensibility, with C-level speakers generally outperforming B2, B2 outperforming B1, and B1 outperforming A2. Essentially, in terms of L2 speaking ability, those with high proficiency (C, B2) tended to produce both intelligible and comprehensible speech, those with intermediate proficiency (B1, A2) were more intelligible than comprehensible, and those with low proficiency (A1) produced speech low in intelligibility, which in turn did not allow for anything but low comprehensibility.

Of note, intelligibility levels (~72% accuracy) observed in this study are lower than in some previous research on L2 speech (e.g., Huensch & Nagle, 2021, reported most utterances were transcribed with near perfect accuracy). There are (at least) three factors which might explain this finding: (1) sampling from a wide proficiency range, (2) drawing on extended, spontaneous speech samples in their entirety and (3) adverse listening conditions (e.g., excessive background noise, unexpected interruptions). Any of these three factors are worth exploring, though we focus here specifically on (1) and (2). First, we purposefully included speakers representing a full range of proficiency (A1 through C). Previous intelligibility research has frequently worked with a more constrained proficiency sample (e.g., Munro & Derwing, 1995, featured speakers at B2 or above). It is possible that our lower intelligibility scores are a result of including lower proficiency speakers. Second, to our knowledge, no previous study has attempted to analyze intelligibility through the use of extended, spontaneous speech. In most intelligibility research, listeners are provided with decontextualized utterances drawn from a larger response (see Kang et al., 2018). The use of contextualized samples in our study meant that utterances were full of repetitions, false starts and filler, frequently featured ungrammatical structures and ranged from relatively short (2 words) to relatively long (24 words) segments. As segment length has been shown to predict intelligibility accuracy (e.g., Nagle & Huensch, 2020), this may be another reason for the lower intelligibility scores. As one final comment, listeners were not provided with prompt knowledge in advance (as these prompts were not made available due to test security reasons). Previous studies have frequently made such prompts available to ensure equality across an entire dataset; in this study, a lack of initial familiarity may have contributed to lower intelligibility scores. How speech stream characteristics (related to contextualized speech) impact intelligibility ratings and subsequent comprehensibility ratings of the same speech is an area prime for additional research. Clearly, while chronological, contextualized segments may be more reflective of everyday speech, their use as a measure of intelligibility is in need of further methodological investigation. Finally, intelligibility, as measured in this study, focuses on a surface-level understanding in which the words spoken are assumed to match a speaker's intent. In real-life listening, the possibility exists that listeners may assume an understanding of intent, which is reflected in more positive comprehensibility judgments, though such an understanding may not align with that intended by the speaker (which Smith & Nelson, 1985, referred to as interpretability of an utterance).

https://doi.org/10.1017/S1366728925100606 Published online by Cambridge University Press

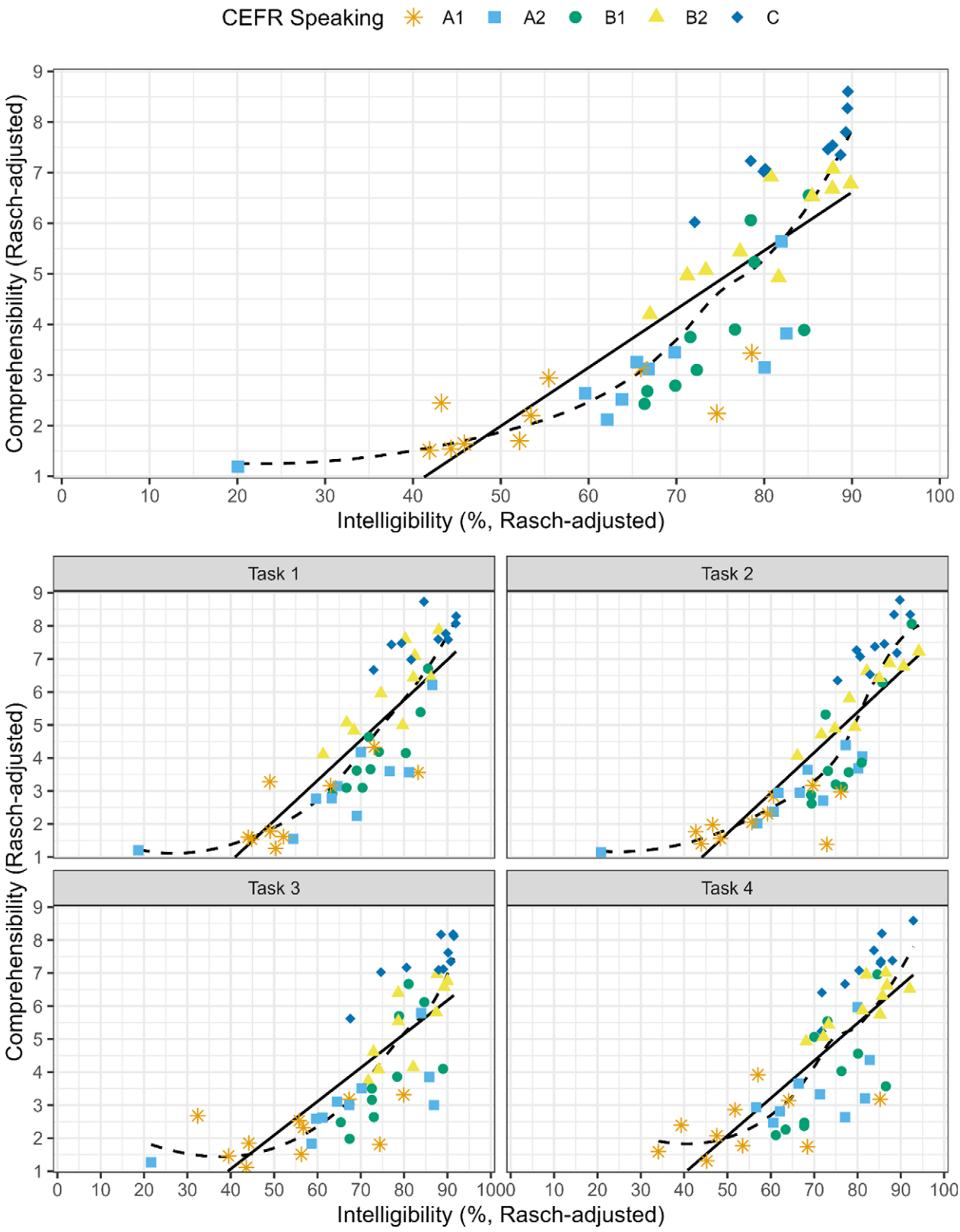

Figure 1. Scatterplots of intelligibility and comprehensibility measures.

5. Conclusion

We identified a nonlinear relationship between intelligibility and comprehensibility, with L2 speech transcribed accurately below a breakpoint of 64% uniformly low in comprehensibility, whereas speech above this breakpoint demonstrated increased intelligibility strongly associated with higher comprehensibility. While a novel finding, some limitations should be considered. First, a larger speaker sample would increase the precision and generalizability of our findings. In addition, we recruited listeners through Prolific and all procedures were conducted online. While online research has

https://doi.org/10.1017/S1366728925100606 Published online by Cambridge University Press

Table 3. Regressions modeling the relationship between comprehensibility and intelligibility

	Linear			Squared			Linear + squared			Segmented		
	Est.	95% CI	*p*	Est.	95% CI	*p*	Est.	95% CI	*p*	Est.	95% CI	*p*
Intercept	−3.78	[−5.54, −2.03]	<0.001	−0.59	[−1.55, 0.36]	0.218	3.99	[0.16, 7.82]	0.041	0.44	[−2.36, 3.24]	0.752
Intelligibility	0.12	[0.09, 0.14]	<0.001				−0.15	[−0.27, −0.03]	0.017			
Intelligibility^2				0.00	[0.00, 0.00]	<0.001	0.00	[0.00, 0.00]	<0.001			
Intelligibility < 63.9%										0.03	[−0.02, 0.09]	0.243
Intelligibility > 63.9%										0.15	[0.08, 0.23]	<0.001
R^2/R^2 adjusted	0.664/0.656			0.732/0.726			0.763/0.753			0.765/0.750		

Note: The *p* values associated with terms after a breakpoint are calculated in reference to a difference from the pre-breakpoint term. R model syntax: Linear: lm(com ~ int, data = d); Squared: lm(com ~ int_2, data = d); Linear + Squared: lm(com ~ int + int_2, data = d); Segmented: segmented(Linear, seg.Z = ~int).

become common in L2 speech research, it must be noted that researcher control over the rating environment is limited and listener attention may be a concern (Nagle & Rehman, 2021). Finally, this study focused on L2 English speech, a trend in L2 pronunciation research (Crowther & Isbell, 2023; Levis, 2021); further consideration of non-English target speech is warranted. Notwithstanding these limitations, this study provides initial evidence that the relationship between L2 speech intelligibility and comprehensibility may be nonlinear, with L2 speakers needing to reach a threshold of intelligibility prior to producing speech that listeners find easier to understand. Although continued research is needed, recognizing the existence of a nonlinear relationship, on top of the potential use of extended, conceptualized segments of speech across a range of proficiencies, should inform future investigations.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S1366728925100606.

Data availability statement. The speech samples used in this study were provided by the British Council and drawn from operational Aptis language test data. Access to these data is restricted for test security and privacy reasons. Requests for data access can be made through the British Council's English Language Research program.

Acknowledgments. This study was funded by a British Council Assessment Research Grant awarded to Dustin Crowther and Daniel R. Isbell.

Competing interests. None.

References

Ali, M. M. (2023). The foreign-accentedness, comprehensibility, and intelligibility of L2 Arabic speech. *Language Teaching Research* https://doi.org/10.1177/13621688231158787.

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.

Chau, T., & **Huensch, A.** (2025). The relationships among L2 fluency, intelligibility, comprehensibility, and accentedness: A meta-analysis. *Studies in Second Language Acquisition* Published online 6 February 2025. https://doi.org/10.1017/S0272263125000014.

Crowther, D., **Holden, D.**, & **Urada, K.** (2022). Second language speech comprehensibility. *Language Teaching*, **55** (4), 470–489. https://doi.org/10.1017/S0261444821000537.

Crowther, D., & **Isbell, D. R.** (2023). Second language speech comprehensibility: A research agenda. *Language Teaching*. https://doi.org/10.1017/S026144482300037X.

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

Foster, P., **Tonkyn, A.**, & **Wigglesworth, G.** (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, **21** (3), 354–375. https://doi.org/10.1093/applin/21.3.354.

Gallant, J. (2023). Typed transcription as a simultaneous measure of foreign-accent comprehensibility and intelligibility: An online replication study. *Research Methods in Applied Linguistics*, **2** (2), 100055. https://doi.org/10.1016/j.rmal.2023.100055.

Hansen Edwards, J. G., **Zampini, M. L.**, & **Cunningham, C.** (2018). The accentedness, comprehensibility, and intelligibility of Asian Englishes. *World Englishes*, **37** (4), 538–557. https://doi.org/10.1111/weng.12344.

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. https://doi.org/10.1111/lang.12451.

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

Jułkowska, I. A., & **Cebrian, J.** (2015). Effects of listener factors and stimulus properties on the intelligibility, comprehensibility and accentedness of L2 speech. *Journal of Second Language Pronunciation*, **1** (2), 211–237. https://doi.org/10.1075/jslp.1.2.04jul.

Kang, O., & **Rubin, D.** (2014). Listener expectations, reverse linguistic stereotyping, and individual background factors in social judgments and oral performance assessment. In J. M. Levis & A. Moyer (Eds.), *Social dynamics in second language accent* (pp. 239–254). De Gruyter Mouton.

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**, 115–146. https://doi.org/10.1111/lang.12270.

Levis, J. (2021). L2 pronunciation research and teaching: The importance of many languages. *Journal of Second Language Pronunciation*, **7** (20), 141–153. http://doi.org/10.1075/jslp.21037.lev.

Linacre, J. M. (2021). *A user's guide to FACETS Rasch-model computer programs (program manual 3.83.5)*. Winsteps.com.

Lüdecke, et al. (2021). Performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, **6** (60), 3139. https://doi.org/10.21105/joss.03139.

Makowski, D., **Ben-Shachar, M. S.**, **Patil, I.**, & **Lüdecke, D.** (2019). Methods and algorithms for correlation analysis in R. *Journal of Open Source Software*, **5** (51), 2306. https://doi.org/10.21105/joss.02306.

Muggeo, V.M.R. (2008). Segmented: An R package to fit regression models with broken-line relationships. *R News*, **8**/1, 20–25. https://cran.r-project.org/doc/Rnews/

Munro, M. J., & **Derwing, T. M.** (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, **45** (1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x.

Munro, M. J., & **Derwing, T. M.** (2011). The foundations of accent and intelligibility in pronunciation research. *Language Teaching*, **44** (3), 316–327. https://doi.org/10.1017/S0261444811000103.

Myford, C. M., & **Wolfe, E. W.** (2000). Strengthening the ties that bind: Improving the linking network in sparsely connected rating designs. *ETS Research Report Series*, **2000** (1), i–34. https://doi.org/10.1002/j.2333-8504.2000.tb01832.x.

Nagle, C. L., & **Huensch, A.** (2020). Expanding the scope of L2 intelligibility research: Intelligibility, comprehensibility, and accentedness in L2 Spanish. *Journal of Second Language Pronunciation*, **6** (3), 329–351. https://doi.org/10.1075/jslp.20009.nag.

Nagle, C. L., & **Rehman, I.** (2021). Ng L2 speech research online: Why and how to collect online ratings data. *Studies in Second Language Acquisition*, **43** (4), 916–939. https://doi.org/10.1017/S0272263121000292.

O'Sullivan, B., **Dunlea, J.**, **Spiby, R.**, **Westbrook, C.**, & **Dunn, K.** (2020). *Technical report: Aptis general technical manual* (version 2.2). British Council.

Smith, L. E., & **Nelson, C. L.** (1985). International intelligibility of English: Directions and resources. *World Englishes*, **4** (3), 333–342. https://doi.org/10.1111/j.1467-971X.1985.tb00423.x.

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

Trofimovich, P., & **Isaacs, T.** (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, **15** (4), 905–916. https://doi.org/10.1017/S1366728912000168.

https://doi.org/10.1017/S1366728925100606 Published online by Cambridge University Press