Comprehensible to whom? Examining rater, speaker, and interlocutor perspectives on comprehensibility in an interactive context

Charlie L. Nagle,[1] **Pavel Trofimovich**,[2] **Mary Grantham O'Brien**,[3] and **Sara Kennedy**[2]

[1]*The University of Texas at Austin* [2]*Concordia University* [3]*University of Calgary*

Nagle, C., Trofimovich, P., O'Brien, M. G., & Kennedy, S. (2022). Comprehensible to whom? Examining rater, speaker, and interlocutor perspectives on comprehensibility in an interactive context. *The Modern Language Journal.* Published online 20 November 2022. https://doi.org/10.1111/modl.12809

Comprehensibility has emerged as a useful and intuitive means of globally evaluating second language (L2) speakers in many research and instructional contexts. In most cases, L2 speakers' comprehensibility is assessed by external listeners who do not engage in extensive communication with the speakers, even though the degree to which a speaker is comprehensible is presumably of greatest concern to their interlocutor. If comprehensibility is defined as the ease with which speakers come to understand one another, then interaction-based assessments, which would include self and peer ratings, might provide different insight into interactive comprehensibility compared to assessments by external listeners. To examine this issue, in this study, 20 pairs of L2 English interactants rated themselves and their partner on 7 occasions distributed throughout a 17-minute interaction encompassing 3 communicative tasks, and recordings of the interaction were subsequently presented to external raters for evaluation. Mixed-effects models were used to compare the shape of the comprehensibility curves over time and the self, partner, and rater scores at each rating episode. Results demonstrated that self and partner assessments were always aligned, but raters consistently assigned significantly lower comprehensibility scores to the interactants. These findings have implications for how comprehensibility, and indeed other listener-based constructs, are assessed.

Keywords: interaction; comprehensibility; pronunciation; self-assessment; listener-based ratings

Comprehensible to whom? Examining rater, speaker, and interlocutor perspectives on comprehensibility in an interactive context

Second language (L2) speakers need to be able to communicate successfully in the L2. Successful communication can be construed in terms of both intelligibility—that is, the extent to which the listener understands the speaker regardless of the amount of effort required—and comprehensibility—that is, the perceived ease with which the listener understands the speaker. Intelligibility and comprehensibility are interrelated but conceptually distinct constructs. L2 speech may be perfectly intelligible even if it is not fully comprehensible, in the sense that intelligible speech often shows varying degrees of comprehensibility depending on how much the listener struggles to understand the speaker (Munro & Derwing, 1995; Nagle & Huensch, 2020). However, the relationship between these two constructs and accentedness—or the degree to which speech (especially pronunciation) aligns with a local native variety of the L2—is comparatively weak, which means that speech may be highly intelligible and comprehensible even if it is moderately to strongly accented (Munro & Derwing, 1995; Nagle & Huensch, 2020).

It comes as no surprise, then, that comfortable intelligibility and comprehensibility, as opposed to nativelike pronunciation, have been identified as the goal of language instruction (Levis, 2020). Beyond its intuitive appeal, comprehensibility has been especially popular with researchers and practitioners because it can be evaluated using simple, easy-to-interpret rating scales with straightforward endpoint descriptors (e.g., difficult to understand–easy to understand). In most research and assessment contexts (e.g., Internet-Based Test of English as a Foreign Language [TOEFL iBT]), comprehensibility is assessed through audio recordings, where a speaker's response to a prompt (e.g., question or images) is evaluated by raters. In other

instructional and formative assessment situations, L2 speakers assess their own comprehensibility (e.g., Kissling & O'Donnell, 2015). Yet in other settings (e.g., International English Language Testing System [IELTS], American Council for the Teaching of Foreign Languages [ACTFL] oral proficiency interview), a rater is present when a speaker responds to an interview question or a prompt, but the interaction between them is minimal, involving only preestablished scripts and protocols.

In most cases, L2 speakers' comprehensibility thus appears to be evaluated either by external listeners who do not engage in extensive communication with these speakers or by the speakers themselves, even though the degree to which a speaker is comprehensible is presumably of greatest concern to their interlocutor. This approach to evaluating comprehensibility raises a key question: Who should be evaluating L2 speakers? That is, should L2 speakers be evaluated by an external listener who does not interact with them, should they be assessing their own performance, or should they be evaluated by their interactive partner? This study's goal was to begin addressing this complex issue by examining how the assessment of comprehensibility compares across external listeners, L2 speakers, and their interlocutors, on the assumption that the resulting evaluations might provide different perspectives on comprehensibility.

Comprehensibility From an External Listener's Perspective

Comprehensibility has been extensively studied from the perspective of external listeners, who are typically either trained or naïve raters evaluating L2 speakers' recorded or live performances (Trofimovich et al., 2022). For external listeners, comprehensibility is primarily a speech-centered construct, such that ease of understanding is associated with many linguistic features (Saito et al., 2017a; Isaacs & Trofimovich, 2012), including lexis and grammar (e.g., appropriate and rich vocabulary, accurate and complex grammar) and pronunciation (e.g.,

accurate word stress). Linguistic influences on comprehensibility also vary based on a speaker's L2 proficiency. For example, Huensch and Nagle (2021) showed that the impact of speech rate and prosody on comprehensibility was strongest in low-proficiency speakers, even though these dimensions predicted comprehensibility ratings for speakers of all ability levels.

Comprehensibility is also listener dependent, in that it reflects not only the speaker's linguistic performance but also the listener's profile. L2 speakers' comprehensibility is impacted by many listener characteristics, including listeners' familiarity with the language being evaluated (Munro et al., 2006), their teaching experience and linguistic training (Isaacs & Thomson, 2013; Saito et al., 2017b), their knowledge of other languages (Saito & Shintani, 2016), and their own L2 learning history (Saito et al., 2019). Nevertheless, regardless of individual differences, various types of listeners, such as native speakers, advanced L2 learners, and bilinguals or multilinguals are generally comparable in the quality and consistency with which they evaluate L2 speakers (Crowther et al., 2016; O'Brien, 2014; Saito & Shintani, 2016).

Finally, comprehensibility is time sensitive, meaning that it fluctuates in response to the varying levels of linguistic accuracy and complexity (e.g., in lexis, grammar) that speakers produce as they strive to convey their message. For instance, Nagle et al. (2019) asked external raters to dynamically assess the comprehensibility of L2 speakers and explain their reasons for downgrading or upgrading their ratings. Analyses of ratings alongside rater comments indicated that various lapses in a speaker's use of grammar and lexis (e.g., inappropriate word choice, missing subject–verb agreement) appeared to influence the raters' time-locked assessments, even if the speaker's message was generally coherent. Thus, seemingly minor linguistic missteps can have a negative impact on comprehensibility if they contradict or disrupt the listener's emergent understanding of what the speaker is trying to say.

Comprehensibility From the Speaker's Perspective

A much less examined perspective on L2 comprehensibility comes from speakers themselves. Accurate self-assessment of speaking and pronunciation is important for autonomous learning (Lee & Chang, 2005; Patri, 2002), yet little is known about how L2 speakers assess their own comprehensibility. Trofimovich et al. (2016) found only a weak relationship (r = .18) between 134 L2 English speakers' self-assessments of comprehensibility and the assessments by three expert raters, with most speakers over- or underestimating their comprehensibility. Overconfident (inflated) self-assessments were particularly pronounced for speakers whose pronunciation (in terms of the production of segments, word stress, rhythm, intonation, and optimal speech flow) was judged as the least accurate by raters. Isbell and Lee (2022) reported a moderate relationship (r = .54) between self- and listener-assessed comprehensibility for 198 speakers of L2 Korean, with overconfident self-ratings observed for speakers of higher proficiency and those who expressed greater satisfaction with, and placed greater value on, their pronunciation. Thus, L2 speakers' self-ratings often diverge from those by external raters (Li & Zhang, 2021). However, just as externally assessed comprehensibility is impacted by speaker and listener variables, speakers' self-assessments appear to include the dual perspective of speaker and listener, in the sense that self-assessments are linked to speaker perceptions of various dimensions of their own speech (Trofimovich et al., 2016) and that different people—as individuals evaluating their own performance—might vary in the extent to which they can self-assess their comprehensibility (Isbell & Lee, 2022).

Although self- and other assessments of comprehensibility tend to be misaligned, self-assessments appear to show time-sensitive properties. For instance, when L2 speakers performed two speaking tasks, where the tasks shared the same procedure but differed in content, their self-

assessed comprehensibility was more closely aligned with external raters' assessments after the second task than after the first one (Strachan et al., 2019). In a longitudinal study, by the end of an academic term, Japanese speakers of L2 English had become more aligned in their self-assessments of comprehensibility with the ratings provided by five trained listeners (Saito et al., 2020). Finally, at the end of 15-week instruction that included activities where speakers of L2 French developed and used criteria for rating comprehensibility and engaged in peer assessment, they provided self-ratings that were closer to external listeners' ratings (Tsunemoto et al., 2022). These findings imply that, similar to external listener evaluations, self-assessments of comprehensibility demonstrate dynamic qualities, though on a coarser grained timescale.

Comprehensibility From the Interlocutor's Perspective

The least explored perspective on comprehensibility is the one from a speaker's interlocutor (i.e., a person communicating with the speaker). In a recent study, Trofimovich et al. (2020) engaged pairs of L2 English speakers in three tasks, asking them to evaluate each other's comprehensibility on seven occasions (2.5 minutes apart) during approximately 17 minutes of interaction. Results showed that comprehensibility ratings followed a U-shaped function, where the speakers rated their partners' comprehensibility as high during the first task, after which the ratings dropped during the second task (likely as a function of the task's difficulty) and then slowly increased during the final task (presumably in response to the speakers' increased familiarity with their partners). The ratings also became progressively aligned within speaker pairs over the course of interaction. To explain their assessments, the speakers cited various pronunciation, content, and discourse issues as reasons for difficulties in understanding their partners. In a follow-up analysis of the same dataset, Nagle et al. (2022) further showed that comprehensibility in an interactive context has behavioral and affective components; speakers

who perceived themselves and their partners to be more collaborative and less anxious also rated their partners to be more comprehensible.

This pattern of comprehensibility assessment in dialogue is generally consistent with the phenomenon of interactive alignment, which refers to the tendency for interlocutors to converge on common language patterns through the social forces of accommodation and psychological mechanisms of priming (Garrod et al., 2018; Giles & Ogay, 2007). Given that comprehensibility is associated for listeners with multiple features of speech (Saito et al., 2017a; Isaacs & Trofimovich, 2012), an alignment in comprehensibility would be expected if interlocutors indeed appropriated and reused each other's language patterns, such as lexical expressions, grammar structures, phonetic realizations of segments and words, utterance length, and pausing frequency (Garrod et al., 2018). An upward trajectory for comprehensibility ratings is also compatible with the notion that listeners' perceptual categories are adaptive to recent experience (Baese–Berk, 2018), where listeners improve in comprehension of unfamiliar L2 speakers in a matter of minutes (Clarke & Garrett, 2004; Xie et al., 2018). Finally, the links between comprehensibility and interlocutors' anxiety and collaboration might reflect the notion that conversation is a form of joint action (Pickering & Garrod, 2021) where, in addition to using language, interlocutors work together to achieve a common goal by reacting to one another's affective states, such as nervousness or joy (Parkinson, 2011), and by coordinating one another's behaviors, such as backchanneling, turn-taking, providing or withholding feedback, and using gesture (Paxton et al., 2016). Thus, speakers' comprehensibility in interaction appears to be shaped by their linguistic (e.g., speech content and quality), behavioral (e.g., degree of perceived collaborativeness), and affective (e.g., extent of perceived anxiety) contributions to the dialogue. Comprehensibility also appears to be dynamic, in that it is co-constructed and displays change over time.

The Present Study

Comprehensibility has emerged as a useful and intuitive means of globally evaluating L2 speakers in many research and instructional contexts (e.g., Isaacs et al., 2017). To date, however, most researchers have elicited L2 speakers' self-assessments or have collected comprehensibility ratings from external listeners who do not interact with L2 speakers, even though a speaker's comprehensibility ostensibly matters most to their interlocutor. Although comprehensibility ratings generally seem to reflect various linguistic properties of the speech being evaluated, to depend on the individual profile of the person providing the rating, and to show various degrees of change over time, assessments by external raters, speakers, and their interlocutors would likely yield different perspectives. For instance, during conversation, speakers may become more familiar with each other's speech patterns, so they will require less effort to understand each other as they continue speaking (Trofimovich et al., 2020). The social coordination that arises as speakers work together to accomplish a common task could manifest itself through enhanced collaboration and decreased anxiety, which may further bolster their mutual comprehensibility (Nagle et al., 2022). Similarly, considering that L2 speakers' self-assessments often diverge from the ratings provided by external listeners (Isbell & Lee, 2022), speakers may be better at self-assessing their comprehensibility in interactions where they have access to verbal and visual cues that signal fluctuations in their own and their partner's comprehensibility. If comprehensibility is defined as the ease with which speakers come to understand each other, then interaction-based assessments, which would include self- and peer ratings, might provide different insight into interactive comprehensibility compared to the assessments by external listeners.

The goal of this study was therefore to compare the assessment of comprehensibility from the perspectives of external listeners, speakers, and their interaction partners. To address

this goal, we revisited the aforementioned dataset that involved L2 speakers interacting with a conversation partner in three communicative tasks (17 minutes per pair), where the interactants rated themselves and their partner on seven occasions (2.5 minutes apart) using a variety of scales, including comprehensibility. In our initial study (Trofimovich et al., 2020), we compared the speakers' comprehensibility ratings of each other over time, examining whether the ratings converged or diverged. In a subsequent publication (Nagle et al., 2022), we explored the speakers' comprehensibility ratings in relation to their self- and peer assessments of collaborativeness and anxiety. For this final report, we analyzed previously unpublished data targeting the speakers' self-rated comprehensibility and external raters' assessments of those speakers, all in relation to how the speakers rated their partners. Of key interest was the shape of the self, partner, and rater comprehensibility curves over the 17-minute interaction and the degree of calibration among the three sets of ratings: (a) self- versus partner assessments, (b) self- versus rater assessments, and (c) partner versus rater assessments.

Given the exploratory nature of this report, we did not formulate specific hypotheses. Nevertheless, drawing on theoretical perspectives that view interaction as socially and linguistically coordinated action (Garrod et al., 2018), where comprehensibility would involve a dynamic adaptation of the interlocutors to each other, we anticipated that self- and partner assessments of comprehensibility might be closely aligned. Based on our previous analyses of this dataset, where partner ratings of comprehensibility tended to improve over time (Trofimovich et al., 2020), we also expected that the speakers' self-ratings would demonstrate a similar upward trend. In light of consistent gaps between L2 speakers' self-assessed comprehensibility and external listeners' assessments (Isbell & Lee, 2022), we also anticipated that the speakers' self-ratings would differ from rater assessments. Finally, given the absence of

prior work comparing the interlocutor versus external listener perspective on speaker comprehensibility, we had no firm predictions regarding this relationship, beyond anticipating potential differences, on the assumption that active participants of interaction would have a different perspective on comprehensibility from those observing it. This study was thus guided by the following exploratory question:

RQ. What is the relationship between the assessments of comprehensibility by external listeners, L2 speakers, and their conversation partners in an interactive setting?

Method

L2 Speakers

The L2 speaker data came from the same corpus of 20 L2–L2 interactions between university-level students analyzed previously (Nagle et al., 2022; Trofimovich et al., 2020). The 40 speakers, who were on average 26 years old (*SD* = 2.89), included 14 women and 26 men, all recently admitted to various degree programs at an English-medium university in Canada (see the Appendix for background information on the speakers' home languages, genders, and ages). They had begun learning English from approximately the age of 8.18 (*SD* = 4.58), mostly through classroom-based instruction in their home countries, and reported 17 different first languages, the most frequent being Farsi (9), Hindi (7), Mandarin (4), and Tamil (3). They indicated a fairly high daily use of English with other L2 speakers (*M* = 64.25%, *SD* = 18.80, where 100% meant *all the time*) and estimated being reasonably familiar with L2-accented English (*M* = 6.33, *SD* = 1.67, where 9 meant *very familiar*), which reflected the multilingual, multicultural context of the university and the city where it is located. The speakers' IELTS scores, submitted to the university as part of admission requirements, were on average 6.84 for speaking (*SD* = 0.62) and 7.60 for listening (*SD* = 0.95), which roughly correspond to the C1

band in the Common European Framework of Reference for Languages (CEFR).

Self- and Peer Assessments

The speakers were randomly assigned to dyads, with the constraint that they came from different first language backgrounds so that the only shared language was English (see the Appendix). They completed three interactive tasks, all in the same order. The goal of the first task (3 minutes) was for the speakers to discover three commonalities (e.g., a favorite sport), as a way of getting to know each other. For the second task (7 minutes), the speakers had to complete a coherent, joint story from seven scrambled images distributed to each partner. The story depicted a man who had won a large lottery prize but then experienced misfortune, which made him realize that being rich did not equal happiness (Galindo Ochoa, 2017). The speakers needed to provide each other with verbal descriptions of the images to produce a common narrative. The third task (7 minutes) required the speakers to find common solutions to the problems experienced by international students as they arrive in a new country. They first shared their challenges (e.g., finding housing, accessing healthcare) and then articulated common solutions.

During the 17-minute interaction, the speakers provided seven ratings, evaluating themselves and their respective partners for comprehensibility (among other dimensions not discussed here). The ratings occurred at similar intervals: at the end of each task (Times 1, 4, and 7) and approximately 2.5 minutes and 5 minutes after the beginning of Task 2 (Time 2 and 3) and Task 3 (Times 5 and 6). Comprehensibility was defined for the speakers as a rating of how much effort it takes to understand what someone is saying, and the speakers used continuous scales (100-millimeter lines) printed on paper, one labeled “me” for the self-rating and the other labeled “my partner” for the partner rating. The scales contained only anchor descriptors (*difficult to understand*, *easy to understand*), and the speakers indicated their rating by marking a

cross on the line.

During the interaction, which was audio-recorded for later analysis and presentation to external raters, the two speakers were seated opposite each other with a low barrier preventing them from seeing each other's materials but allowing for an unobstructed view of each other's visual cues. Although both speakers were aware that they were providing self- and peer assessments, they could not see each other's ratings and were not allowed to discuss them. The speakers first heard the researcher define each rated dimension and explain the use of the rating booklet containing instructions for each task and the seven sets of scales (one per page). Each task was introduced separately, always in the same manner, with the speakers first reading printed instructions, then summarizing them to the researcher as a comprehension check, and finally discussing any remaining issues or questions. The speakers were told that Task 1 would last for a maximum of 3 minutes while Tasks 2 and 3 would end after a total of 7 minutes even if they were not completed. The speakers were also informed that they would be evaluating themselves and each other on seven occasions, focusing on the immediately preceding 2–3 minutes of interaction, and that their conversation would be interrupted twice during Tasks 2 and 3 to complete midtask assessments.

External Raters

To obtain external raters' assessments of L2 speaker comprehensibility, 20 expert raters (14 females, 6 males) were recruited from the same university. Consistent with this study's focus on L2 lingua-franca communication, all raters were advanced-level L2 speakers of English with ongoing or completed graduate degrees in linguistics, applied linguistics, education, or related disciplines. The raters, who were on average 32 years old (SD = 4.58), reported speaking 12 different first languages, with Spanish (4), Farsi (4), French (2), and Russian (2) being the most

frequent. They had an average of 7 years (*SD* = 3.81) of prior experience teaching L2 English, both in Canada and abroad, and all had received instruction in phonetics or phonology, with the majority (13) also completing coursework focusing on the teaching of L2 pronunciation. Using the same 9-point scale as the speakers (where 9 meant *very familiar*), the raters estimated their familiarity with L2-accented English as high (*M* = 7.60, *SD* = 0.99). The recruited external raters therefore represented a typical population of trained professionals (all L2 speakers themselves) who might be tasked with evaluating speaking performances by L2 speakers as part of formative, diagnostic, and proficiency assessment, including high-stakes testing (e.g., IELTS, TOEFL iBT). As individuals with phonetics or phonology training and teaching experience, the external raters also illustrated typical profiles of assessors who are often more experienced, trained, and aware than the L2 speakers being assessed. From this standpoint, our decision to recruit trained raters, all with language teaching experience, is ecologically valid, inasmuch as it reflects the general practice of external assessment.

External Assessments

Given the relatively large number of interactions as well as the time and concentration required to evaluate them, the raters were randomly assigned to evaluate five interactions each, such that each interaction was ultimately evaluated by five unique raters. The raters assessed both speakers in each audio-recorded interaction for comprehensibility, along with several other dimensions (not reported here). The raters followed the same timeline as the speakers, providing seven assessments at the same timepoints where the speakers had evaluated themselves and their respective partners in each interaction. Comprehensibility was introduced to the raters using the same definition, and the raters used identical scales (100-millimeter straight lines with the same anchor descriptors), one labelled “Speaker A” for the rating of one speaker and the other labelled

"Speaker B" for the rating of the other speaker.

The raters provided their assessments in an individual session. They first heard the researcher define each rated dimension and then read task descriptions and summarized them for the researcher as a comprehension check. The raters were told that they would be evaluating each speaker on seven occasions, focusing on the immediately preceding 2–3 minutes, and that the timing of each assessment would be announced in the recording (through a prerecorded prompt: "Pause the recording and rate the speakers"). So that the raters could recognize the voices and assign their ratings to the correct speaker, each recording started from a short clip where the speakers introduced themselves by stating their unique participant number (from 1 through 40), their table side (A or B), and their chosen pseudonym, with participant numbers and Speaker A/B designations assigned randomly at the time when the dyadic interaction was originally recorded. To remind the raters of each speaker's identity, the same brief clip with speaker introductions was repeated before each task on the recording.

Before proceeding to evaluate the five recorded interactions, each rater completed a brief rating practice by using a 3-minute sample recording from Task 1 (with a clip introducing the speakers) featuring two additional interactants. To illustrate various timings of the repeated assessments, the practice recording elicited two sets of ratings: once mid-way through the recording and once at its end (both prompted through a prerecorded announcement, as in the target recordings). The five 17-minute audio recordings were presented to each rater for evaluation in a unique random order through a MATLAB interface (Yao et al., 2013). However, to make the rating procedure as similar as possible to the one experienced by the speakers, the raters used a paper booklet to record their ratings. The rating procedure was self-paced, insofar as each rater controlled the playback and pausing of each recording (e.g., starting the recording,

pausing it for rating at each prompted time, then restarting it) and could take breaks between each of the five audio-recorded interactions. However, stopping between assessment episodes or replaying of interaction segments was not allowed. During a short debrief interview at the end of the session, no rater reported any confusion with distinguishing the speakers' voices or using the scales. Each rating session lasted about 2.5 hours.

We used an average-measure, consistency intraclass correlation coefficient to estimate the reliability of the ratings at each of the seven rating episodes. These coefficients, which represent the consistency rather than absolute agreement of ratings averaged over the five individual raters, ranged from .69 to .80, indicating moderate to good reliability at each data point. We therefore averaged scores for the five raters, yielding a single rater comprehensibility score per speaker at each rating episode.[1]

Data Analysis

Target Measures

The three target measures included each speaker's ratings of their own and their partner's comprehensibility (self and partner comprehensibility) and rater assessments of each speaker's comprehensibility (rater comprehensibility). The ratings were expressed numerically (out of 100) by measuring the distance with a ruler (to the nearest millimeter) between the left anchor point and the speaker's or rater's mark on the 100-millimeter scale. In total, there were three sets of ratings per speaker at each of the seven rating episodes corresponding to three perspectives on each speaker's comprehensibility: one self-rating (how the speaker evaluated themselves), one partner rating (how the partner evaluated the speaker), and one (mean) external rating by the five raters who evaluated the same interaction (how external listeners evaluated the speaker). Across the 20 recorded interactions, the speakers spent an average of 2 minutes and 46 seconds on Task

1 (01:04–03:14), 7 minutes and 11 seconds on Task 2 (06:58–07:17), and 7 minutes and 8 seconds on Task 3 (06:23–07:17), with the rating episodes spaced about 2.5 minutes apart (02:46; 02:37; 02:32; 02:02; 02:35; 02:34; 02:00), suggesting that the duration of interactions and the timing of assessments were comparable.

Statistical Modeling

We used mixed-effects models to analyze self, partner, and rater comprehensibility scores over seven rating episodes, examining alignment between the three sets of ratings, that is, the extent to which the self and partner comprehensibility ratings mirrored each other during the interaction, as well as the extent to which the self and rater and the partner and rater comprehensibility ratings aligned with (or diverged from) each other over time. We fit models in R (Version 4.0.4; R Core Team, 2021) using the lme4 package (Bates et al., 2015). Time, rating source (three levels: self, partner, and rater), and the Time × Rating Source interaction were the primary fixed effects of interest, but following our previous work (Nagle et al., 2022; Trofimovich et al., 2020), we also included the following standardized covariates: (a) speakers' IELTS speaking and listening scores and (b) type frequency. The speaking and listening scores were included to control the effect of the speakers' proficiency on their comprehensibility assessments, considering that across the 20 pairs, the interlocutors diverged from each other (in absolute terms) on average by 0.56 points on the IELTS speaking scale (SD = 0.59) and by 1.20 points on the listening scale (SD = 0.70). Type frequency, which we derived by lexically profiling each segment preceding the rating episode, was used to account for the possibility that comprehensibility ratings might reflect the amount of content produced by each speaker before each assessment. Because type and token frequencies were highly correlated in our dataset (r = .94), implying their nonindependence, we included only type frequency as a lexical covariate.

Also, in line with our previous work, we split the data into two datasets corresponding to the first two tasks (four data points) and the third task (three data points), to maintain a comparable number of ratings across the two datasets. All models included by-speaker and by-pair random intercepts. We used QQ plots to inspect model residuals and plotted fitted values against residuals to check for linearity. We also used the DHARMa package (Version 0.3.3.0; Hartig, 2020) to simulate residuals as an additional check for normality of residuals, to examine the dispersion of model residuals, and to check for outliers.

We undertook two conceptually distinct analyses. First, treating time as a continuous predictor, we compared rate of change for the three sets of ratings through a series of pairwise contrasts: self versus partner, self versus rater, and rater versus partner. Second, treating time as a categorical predictor, we compared the ratings at each rating episode, using the same pairwise comparisons, to determine whether the three sets of ratings were statistically different from one another at each timepoint. That is, for one analysis, we estimated the rate and shape of change over time, and for the other, we examined the difference in self, partner, and rater comprehensibility scores at each timepoint. In both cases, we used the emmeans package (Version 1.5.2-1; Lenth, 2020) to conduct pairwise comparisons.

Results

We first computed means and standard deviations for the self, partner, and rater comprehensibility scores. As reported in Table 1, for the most part, the self and partner scores exceeded 80 (except partner scores at Times 2 and 3, which were slightly below that threshold), with most scores in the 80–90-point range on the 100-point scale. In contrast, all rater averages were below 80. Thus, at least descriptively, self and partner scores exceeded rater scores at all timepoints.

Table 1

Descriptive Statistics for Self, Partner, and Rater Comprehensibility Scores

Time	Self		Partner		Rater	
	M	*SD*	*M*	*SD*	*M*	*SD*
1	87.74	14.32	89.02	15.54	73.30	14.51
2	82.79	18.11	79.60	20.92	73.82	12.59
3	80.88	16.16	78.66	18.53	75.61	12.83
4	84.75	18.99	83.45	20.94	76.76	13.07
5	86.79	17.57	86.66	19.22	76.92	13.48
6	90.18	12.52	89.93	11.18	76.41	13.84
7	93.08	7.85	91.43	10.63	78.75	11.86

Comprehensibility in Tasks 1 and 2

We fit a series of preliminary models to the Task 1 and 2 data. However, the residuals of those models showed a substantial deviation from normality. To bring them closer to a normal distribution, we applied a Box-Cox transformation to the original 100-point comprehensibility outcome measure. After this transformation, model diagnostics showed that residuals followed an approximately normal distribution, and outlier and dispersion tests applied to the simulated residuals revealed no significant issues. However, the plot of fitted and residual values was slightly fanlike, suggesting that the linearity assumption may not have been completely upheld and that results should be interpreted with caution.

As shown in Figure 1, the self and partner comprehensibility ratings were curvilinear over the first two tasks. To capture that curvilinearity, we fit linear and quadratic functions for time, using the poly function to compute orthogonal polynomials (thereby avoiding autocorrelation among the two time functions). Pairwise comparisons (summarized in Table 2) showed no significant differences in linear and quadratic rates of change for the self versus partner comprehensibility scores and for the self versus rater comprehensibility scores. However, both the linear and quadratic partner-versus-rater comparisons approached significance, suggesting that the partner and rater comprehensibility scores showed a distinct pattern of change over the first two tasks. Visual inspection confirms this difference (see Figure 1): The partner scores showed the greatest curvature, whereas the rater scores showed the least amount of curvature over time.

Figure 1

Model-Estimated Comprehensibility (With 95% Confidence Intervals) by Rating Source in Tasks 1 and 2

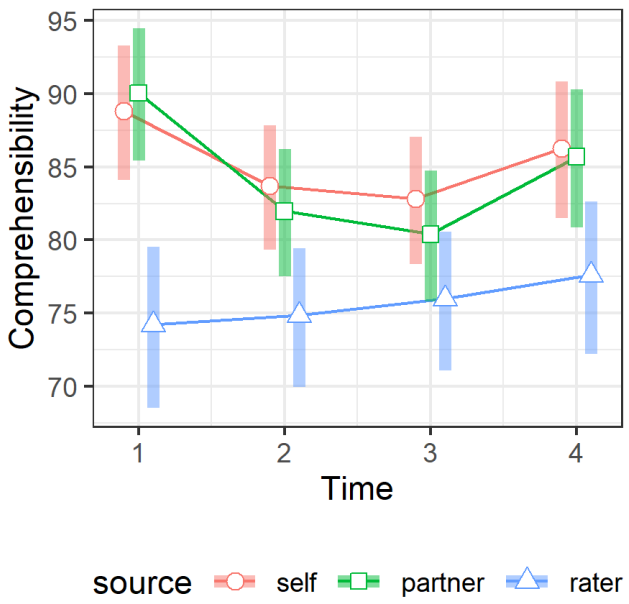

Table 2

Pairwise Comparisons for Linear and Quadratic Slopes in Tasks 1 and 2

Comparison	*b*	*SE*	95% CI	*t*	*p*
Linear slope					
Self–Partner	0.70	1.20	[–2.19, 3.59]	0.59	.829
Self–Rater	–2.02	1.24	[–5.01, 0.96]	–1.63	.241
Partner–Rater	–2.72	1.26	[–5.75, 0.31]	–2.17	.086
Quadratic slope					
Self–Partner	–1.30	1.40	[–4.67, 2.07]	–0.93	.623
Self–Rater	1.95	1.41	[–1.45, 5.35]	1.39	.356
Partner–Rater	3.25	1.44	[–0.23, 6.73]	2.25	.072

Note. Pairwise least-square slope comparisons for the three rating sources; *p* values were adjusted using the Tukey method to account for three comparisons.

In the second analysis, we treated time as a categorical variable, which allowed us to locate significant pairwise differences in comprehensibility scores at each rating episode. As reported in Table 3, the self-versus-partner comparisons never reached significance, demonstrating that self and partner ratings were aligned throughout the interaction (in all cases, the difference in self and partner ratings was less than 3 points). In contrast, the self-versus-rater and partner-versus-rater comparisons were statistically significant at three of the four time points; only the Time 3 comparisons did not reach significance. Across the board, the raters assigned consistently lower comprehensibility scores than the individuals who were involved in the interaction. The difference was greatest at the first timepoint, where the raters assigned scores

that were approximately 15 points lower than the interactants' own scores. Pairwise differences at the other timepoints were smaller, on the order of 5–10 points. Overall, then, the raters provided ratings indicating that they had to invest significantly more effort to understand the interactions than the conversation partners did.

Table 3

Pairwise Comparisons at Each Rating Episode in Tasks 1 and 2

	Self–Partner			Self–Rater			Partner–Rater		
Time	*b*	*t*	*p*	*b*	*t*	*p*	*b*	*t*	*p*
1	–1.47	–0.60	.820	14.20	5.19	< .001	15.67	5.76	< .001
2	2.46	0.94	.620	9.84	3.55	.001	7.38	2.63	.027
3	1.67	0.62	.809	5.77	2.09	.099	4.10	1.47	.312
4	0.83	0.33	.943	8.96	3.34	.004	8.14	3.02	.009

Note. Pairwise least-square mean comparisons for the three rating sources; *p* values adjusted using the Tukey method to account for three comparisons.

Comprehensibility in Task 3

We followed the same procedure to estimate differences in self, partner, and rater comprehensibility scores in the third task. We applied the same Box-Cox transformation to the comprehensibility outcome measure to bring Task 3 model residuals closer to normality. In Task 3, comprehensibility trajectories were linear (Figure 2), so we tested differences only in linear slopes. The results of the Task 3 linear slope model, reported in Table 4, showed that there were no significant pairwise differences in linear rate of change.

Figure 2

Model-Estimated Comprehensibility (With 95% Confidence Intervals) by Rating Source in Task 3

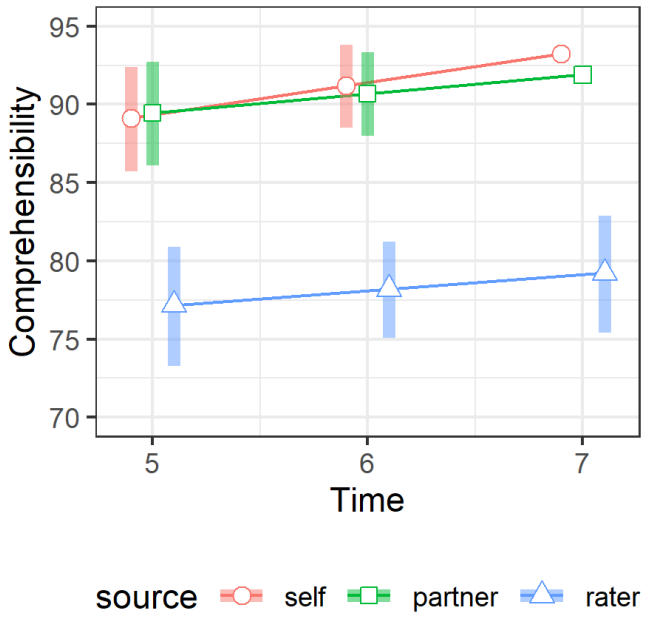

Table 4

Pairwise Linear Slope Comparisons in Task 3

Comparison	*b*	*SE*	95% CI	*t*	*p*
Self–Partner	0.84	1.37	[–2.52, 4.20]	0.61	.814
Self–Rater	1.03	1.48	[–2.59, 4.66]	0.70	.766
Partner–Rater	0.19	1.48	[–3.44, 3.83]	0.13	.991

Note. Pairwise least-square slope comparisons for the three rating sources; *p* values adjusted using the Tukey method to account for three comparisons.

As reported in Table 5, there were no significant differences in self and partner ratings in Task 3, which aligns with the findings for the first two tasks, suggesting that self- and partner

ratings were aligned for the speakers throughout the entire interaction (again, in all cases, the difference between the two ratings was less than 2 points). However, the significant self-versus-rater and partner-versus-rater differences that were observed in the first two tasks persisted in Task 3 and were greater in magnitude. For Task 3, the raters assigned comprehensibility scores that were 10–15 points lower than those provided by interactants, which indicates that they had to invest more effort to understand the interaction than the interactants did.

Table 5

Pairwise Comparisons at Each Rating Episode in Task 3

	Self–Partner			Self–Rater			Partner–Rater		
Time	*b*	*t*	*p*	*b*	*t*	*p*	*b*	*t*	*p*
5	–0.31	–0.16	.987	11.68	5.48	< .001	11.99	5.63	< .001
6	0.42	0.22	.975	13.42	6.35	< .001	13.00	6.14	< .001
7	1.37	0.72	.750	13.73	6.70	< .001	12.37	6.00	< .001

Note. Pairwise least-square mean comparisons for the three rating sources; *p* values adjusted using the Tukey method to account for three comparisons.

Discussion

Considering that external raters, L2 speakers, and their interactive partners might provide different perspectives on comprehensible L2 speech, we examined L2 speakers' self- and partner assessments of comprehensibility, relative to external raters' evaluations. Although self, partner, and rater comprehensibility scores showed a similar (upward) trajectory over time, there were differences in how individuals involved in the interaction assessed comprehensibility compared to external raters who did not take part in it. Notably, the self- and partner assessments were always aligned. In contrast, the external raters' evaluations were significantly lower than the self-

and partner assessments at six of the seven rating episodes (except Time 3, which was in the middle of Task 2), meaning that the external raters generally found it more effortful to understand the speakers than the interactants themselves.

The Speaker Versus the Interlocutor Perspective

The present analysis, which appears to be the first to examine self-rated comprehensibility in interaction, extends prior work (Nagle et al., 2022; Trofimovich et al., 2020) by showing that L2 speakers self-rate their comprehensibility similarly to how it is rated by their interaction partners. Regardless of the task or the timing of self- and partner-assessments, the two interlocutors' ratings were on average within 2–3 points of each other (on a 100-point scale), implying that the interactants had a fairly accurate view of how effortful it was for their interlocutors to understand their speech. A tight coordination between self- and partner assessments is predicted by a view that considers interaction to be shaped by such forces as social accommodation to an interlocutor (Giles & Ogay, 2007), long-term sociocognitive adaptation (Pickering & Gambi, 2018), and spontaneous mimicry (Arnold & Winkielman, 2020), whereby interlocutors, including L2 speakers, converge in verbal and nonverbal behaviors, such as speech rate, pause frequency, phonetic production, and gesture use (Garrod et al., 2018). Because interactants act as speakers and listeners, they might be especially attuned to the verbal and nonverbal feedback provided by each other, for example, in the form of clarification requests or facial cues (e.g., raised eyebrows) to indicate difficulty in understanding (Mauranen, 2006; Seo & Koshik, 2010). This feedback might have helped the interactants adjust their self-perception of comprehensibility against the actual evidence provided by their interlocutors (Kleinschmidt & Jaeger, 2015). Alternatively, speakers might be actively engaged in predictive processing (Pickering & Gambi, 2018), meaning that they are not only assembling their own

utterances but are also covertly simulating (anticipating) their interlocutor's next conversational moves. By engaging in self- and other prediction (Corps et al., 2018), the interactants might have developed heightened awareness of how comprehensible they sounded to their interlocutors, resulting in calibrated self-assessments in dialogue.

Although the speakers' self- and partner assessments were aligned, they were clearly subject to task and time effects (see Figures 1–2). The aligned self- and partner assessments followed a U-shaped trajectory, where they were initially high after Task 1 and then declined by about 10 points during Task 2, followed by an upward trend throughout Task 3. In terms of task effects, the initial task (which functioned as a warm-up) was relatively easy because speakers discussed everyday topics, with an unlimited number of possible commonalities to consider. The second task, which was ranked the hardest by the speakers and which was not completed by any interacting pair within the allotted 7 minutes (Trofimovich et al., 2020), was more complex due to the requirement for the speakers to exchange unique, nonshared information across 14 scrambled images. The final task was again less challenging because it elicited the speakers' shared experiences as recently arrived international students. Assuming that all conversation partners shared at least some lived experiences needed for this task (e.g., obtaining health insurance, registering for coursework, applying for a part-time job), they could complete the task by co-constructing agreed-upon solutions to the challenges they had in common. Considering that tasks of different complexity impose varying demands on the speaker and thus increase or decrease processing effort for the listener (Crowther et al., 2015, 2018), the speakers likely had more opportunity to experience communication breakdown in the second task, which would explain the U-shaped comprehensibility function shown in this dataset.

In terms of time effects, an upward trend for self- and partner assessments throughout the

speakers' interactive experience would be predicted both by a micro perspective, which assumes that listeners' perceptual categories are malleable, with the consequence that listeners rapidly adapt to an unfamiliar speaker's speech (Clarke & Garrett, 2004; Xie et al., 2018), and by a macro perspective, which suggests that listeners with greater exposure to speech assign higher comprehensibility ratings (Kang, 2012; Saito & Shintani, 2016). Nevertheless, it would be premature to draw definitive conclusions about the separate roles of task versus time in comprehensibility ratings because these variables were confounded in this dataset. Indeed, all interaction partners completed the tasks in the same sequence, which makes it impossible to determine whether the obtained U-shaped function for the self- and partner-assessments was determined by task complexity, which created varying levels of linguistic challenge for the speakers, leading to fluctuations in their processing difficulty, or whether an upward rating trajectory was driven by time, as speakers gained experience communicating with each other, regardless of the specifics of each task. Most likely, both forces were at play, in the sense that individual task demands and cumulative speaking time mattered for how the speakers evaluated their own and their partners' comprehensibility. In future research, researchers might want to replicate and extend these findings in an investigation where interlocutors provide self- and partner ratings of comprehensibility for several tasks, as long as they are performed by different interacting pairs in different orders. Alternatively, researchers might choose to engage different interacting pairs in repeated tasks or tasks that feature highly familiar content, so that they could isolate time effects while minimizing task influences on comprehensibility. These concerns notwithstanding, regardless of potential task and time effects, which may have impacted how the speakers rated their comprehensibility in the present dataset, the self- and partner assessments were always aligned while the external raters' evaluations were consistently more severe than

both those assessments at six of the seven rating episodes.

The Interactant Versus the External Rater Perspective

A key finding of this study is that L2 speakers' comprehensibility was evaluated differently by active participants in the interaction than by external listeners assessing it from the sidelines. The speakers tended to rate their own and their partners' comprehensibility on average 10–15 points higher than the raters who listened to the interaction. For partner-rated comprehensibility, this finding is novel, as no prior research to date has directly compared L2 speakers' assessments of comprehensibility between external raters and L2 speakers themselves. For self-rated comprehensibility, the demonstrated gap between self- and rater assessments is consistent with prior work on comprehensibility (e.g., Isbell & Lee, 2022; Saito et al., 2020) and other L2 speaking skills (e.g., Lee & Chang, 2005; Patri, 2002) showing that L2 speakers' self-assessments often diverge from external raters' evaluations, with most L2 speakers prone to overconfident (more lenient) self-assessment.

There are several plausible reasons for the obtained difference between interactant- and rater-assessed comprehensibility. First, because lingua franca speakers sometimes do not signal problems of understanding—employing a "let it pass" strategy, especially if a problematic utterance is not crucial to the overall message (Firth, 1996)—it is possible that the L2 speakers in this study similarly overlooked at least some of their processing difficulty, upgrading their own and their partners' comprehensibility relative to the assessments by external raters, who were cognizant of this processing difficulty. Second, because the speakers were actively engaged in communication to complete task objectives (e.g., reconstructing a joint story from images), they may have lacked sufficient cognitive resources to develop or articulate a refined understanding of their own and their partners' comprehensibility. In contrast, the external raters were listening

to the interaction, not actively participating in it, so they could allocate greater attentional resources to (i.e., focus on) assessing each speaker's comprehensibility. Alternatively, the speakers may have only reflected on their own and their partner's comprehensibility retrospectively, at each rating episode, while otherwise paying little attention to comprehensibility issues during conversation. The external raters, instead, may have focused on the speakers' comprehensibility throughout the recording, since their task (listening, then rating) was different from that of interactants (completing an interactive task, then rating). The obtained difference in interactant- versus rater-assessed comprehensibility may have thus reflected various degrees of cognitive workload experienced by the person providing the assessment.

Third, the speakers evaluated their live performances while the external raters listened to an audio recording, which lacked the immediacy of live conversations and may have also been less successful at replicating the auditory quality of in-person speaking. In this sense, potential differences in self- and partner assessments versus the assessments by external raters may have stemmed from the medium in which the performance was presented, given that live or video evaluations of the same speakers often elicit more generous evaluations from raters (Nakatsuhara et al., 2021; Nambiar & Goon, 2016; Neu, 1990). Fourth, the external raters in this study were L2 speakers with training in linguistics and phonetics, including L2 pronunciation. Although trained and untrained listeners often provide similar assessments (Huang, 2013; Isaacs & Thomson, 2013), expert raters (including teachers) might also demonstrate more severity in their evaluations compared to raters with no teaching background (Galloway, 1980), such as the L2 interactants in this study. Indeed, the external raters not only received more pronunciation-specific training but they were also more proficient as L2 speakers, compared to the interactants. Although these differences in language training, awareness, and proficiency are ecologically

valid—in that our external raters and L2 interactants illustrated typical profiles of external assessors and L2 speakers being assessed—the obtained differences in self- and partner assessments versus external ratings may have reflected the distinct experiential and linguistic profiles of the participant groups.

Finally, the interactants and the external raters differed in their access to visual cues available for rating comprehensibility. Whereas the speakers had full access to each other's facial expressions, gestures, body postures, and displays of emotion (e.g., smiling, irritation), which have been linked to understanding in dialogue (Floyd et al., 2016; Seo & Koshik, 2010), the external raters could not readily avail themselves of these cues because they were listening to the interaction, not observing it. There is emerging evidence that visual cues such as averting eye gaze (Tsunemoto et al., 2021) and nodding (Trofimovich et al., 2021) are associated with comprehensibility when ratings take place in a video format. Thus, the interactants in this study may have taken advantage of various visual cues that they produced and observed, which was not possible for the external raters.

Comprehensibility From Inside and Outside Interaction

Regardless of which factors explain the observed interactant–rater differences, the present findings imply a crucial distinction in how comprehensibility is assessed by active participants of interaction versus those who evaluate it from the sidelines. This distinction is akin to many similar dissociations observed in various aspects of human behavior, including language learning and use. In L2 learning, for example, Mackey (1999) showed that only active participants in interaction (i.e., those who heard the target language forms negotiated, repeated, and otherwise highlighted by their interlocutors in response to communication breakdowns) demonstrated language-learning gains, whereas those who observed the same interaction showed

no benefit. In assessment, according to a meta-analysis of 69 studies across various subject fields (Li et al., 2016), the assessments by teachers (i.e., external assessors) and those by peers (i.e., individuals performing target assessment tasks) often diverge, showing a moderate association (r = .63), with even weaker associations reported for specific subject domains (i.e., medical) and modes of assessment (i.e., computer assisted). In language development, American 9-month-old infants exposed to speakers of Mandarin Chinese as part of live interaction with a speaker learned to discriminate between Mandarin Chinese sounds, but the infants exposed to similar video- and audio-recorded input (i.e., through observing a speaker) showed no evidence of learning (Kuhl et al., 2003). And in neuroscience, individuals listening to speech that they believed to be produced by a live social partner showed increased brain activity in the regions linked to mentalizing, which refers to understanding the intentional states of a person, compared to listening to speech that they believed to be prerecorded (Rice & Redcay, 2016). As these examples imply, people's experience in interactive social situations differs from their experience observing or imagining those situations (Li & Jeong, 2020; Redcay & Schilbach, 2019).

Although definitive conclusions based on the present findings are clearly premature, the interactant–observer distinction in assessments of L2 comprehensibility implies divergent perspectives for speakers versus external raters on what comprehensible speech entails. First, a view of comprehensibility from within interaction reflects each interactant's dual role as a speaker and a listener, where an interactant can both initiate a communicative action and respond to one. This dual role would entail various co-constructed processes of speech planning, prediction, imitation, and short- and long-term adaptation (Arnold & Winkielman, 2020; Corps et al., 2018; Pickering & Gambi, 2018), which may be less pronounced or absent in observers of interaction. Second, a perspective of comprehensibility from within interaction is anchored in

interactants' shared experience in a social activity often driven by a common goal, which may be comprehended but not necessarily experienced by an outside observer in a similar way. Third, a view of comprehensibility within interaction is grounded in interactants' codependent affective and emotional states (Nagle et al., 2022). Even more so, active interaction—through overt or covert imitation of one's interactive partner—leads not only to increased affect, such as empathy, liking, and rapport, but also to enhanced prosocial behaviors, such as generosity and helpfulness, decreased prejudice, improved performance in cognitive and motor tasks, and improved ability to exercise self-control (Duffy & Chartrand, 2017). With no opportunity to engage in interaction, an external observer not only would experience interactants' emotions differently but also would not benefit from affective or behavioral benefits stemming from interaction. In a nutshell, what makes an L2 speaker's speech easy or difficult to understand, and to what degree, would necessarily be different from the perspective of the speaker's interactive partner versus an external observer.

Pedagogical Implications

The results of the current study have implications for L2 speech assessment. If the ultimate goal is to determine the effort required for those outside of an interaction to understand L2 speakers' speech, then instructors are justified in requiring students to carry out monologic tasks that are subsequently evaluated by instructors themselves or by other external raters. If, however, instructors are interested in the extent to which L2 speakers are comprehensible in interaction, then requiring them to carry out interactive tasks may be more appropriate. Finally, if the ultimate goal is to assess speaker comprehensibility in interaction, we recommend that instructors themselves take part in interactive tasks or, at the very least, imagine that they are taking part in the interaction that they assess. Indeed, given that communication is a two-way

street, with both speakers and listeners contributing to understanding, it seems advantageous to operationalize assessments from an interactive perspective.

When choosing speech assessment tasks (formative or summative), instructors are encouraged to provide L2 speakers with less constrained interactive tasks like those in Tasks 1 and 3, as these may be less cognitively demanding and may encourage more fluent—and ultimately more comprehensible—L2 speech. More linguistically constrained tasks like Task 2 in the current study may impose a greater cognitive burden that may, in turn, have a negative impact on comprehensibility assessments (Crowther, 2020). Alternatively, tasks can be sequenced according to speakers' proficiency. For instance, in line with ACTFL guidelines (2012), tasks that allow learners to discuss personally relevant and tangible information may be more suitable at the intermediate level, whereas more advanced speakers—who should be able to use abstract language—could be asked to engage in tasks that require a higher proportion of external referents and/or unfamiliar scenarios. This type of scaffolding could help L2 speakers self-assess comprehensibility in a level-appropriate manner.

Limitations and Future Work

Based on the present findings, a key goal of future work would be to clarify how an L2 speaker's performance shapes listener perception of that speaker's comprehensibility when the listener is an active participant of interaction versus an external observer. To address this goal, researchers might need to examine how various linguistic dimensions of the speaker's speech interact with different characteristics of the listener's profile (e.g., amount of training, knowledge of multiple languages) on a dynamic timescale, as the listener repeatedly assesses the same speaker as an external rater or engages in interaction with the speaker as a conversation partner. In future work, it would also be important to understand how quickly and to what degree L2

speakers align their self-assessments with those of their interlocutors in interactive settings, for example, through engaging both interactive partners in interviews or concurrent or retrospective recall protocols focusing on reasons for why they consider their own and their partner's speech comprehensible at various points in interaction. Similarly, researchers might wish to supplement interlocutor ratings of comprehensibility in interaction with online measures, using eye-tracking (Godfroid et al., 2020) or pupillometry (Schmidtke, 2018) to understand interlocutors' processing depth, workload, or engagement in dialogue. Finally, future research should address the various limitations of this study. For example, it would be important to examine L2 speakers' comprehensibility as they interact with native speakers, which might yield implications for L2 speaker assessment by native-speaking raters, and to separate time and task effects by engaging speakers in the same tasks presented in different orders. Moreover, it would be critical to examine whether and how interactant versus rater assessments differ when the cognitive workloads are matched (e.g., both the external rater and the interactant must complete similar task objectives) rather than mismatched (e.g., the external rater observes an L2 speaker interacting to complete a task objective), when external raters do or do not have access to the speakers' visual cues, and when the interactants and external raters are comparable in their linguistic and experiential profiles (e.g., as L2-speaking university students).

Conclusion

Comprehensibility ratings provide researchers with valuable insights into a wide range of linguistic and nonlinguistic factors that affect the effort required for understanding. Our goal was to determine if it matters whether those who assess comprehensibility are active participants in a conversational exchange. Whereas previous studies have primarily relied upon raters who are external to the speech event, we have demonstrated the value of comparing the insights provided

by the interlocutors themselves to those provided by external raters. We have demonstrated that taking part in an interaction may encourage both conversation partners to make use of everything they have at their disposal (e.g., giving and receiving verbal and non-verbal cues to understanding, adjusting speech rate and the use of pauses, and predicting forthcoming utterances) to ensure mutual understanding. External listeners, who are not invested in communicative success, do not have access to this wide range of resources as they assess the amount of effort required to understand the speech. While we fall short of suggesting one type of assessment over another, we do recommend that researchers consider the goals of the assessment as well as the benefits and drawbacks of each type of rater as they determine how to appropriately assess comprehensibility and, indeed, other listener-based constructs.

Note

[1] Materials, data, and analysis code for this project can be accessed at https://osf.io/c3g7s/?view_only=0c53aec4343e45908b1b9317790392db

References

American Council on the Teaching of Foreign Languages. (2012). *ACTFL proficiency guidelines*. https://www.actfl.org/resources/actfl-proficiency-guidelines-2012

Arnold, A. J., & Winkielman, P. (2020). The mimicry among us: Intra- and inter-personal mechanisms of spontaneous mimicry. *Journal of Nonverbal Behavior*, *44*, 195–212. https://doi.org/10.1007/s10919-019-00324-z

Baese–Berk, M. (2018). Perceptual learning for native and non-native speech. In K. D. Federmeier & D. G. Watson (Eds.), *The psychology of learning and motivation: Current*

topics in language (pp. 1–29). Academic Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America*, *116*, 3647–3658. https://doi.org/10.1121/1.1815131

Corps, R. E., Gambi, C., & Pickering, M. J. (2018). Coordinating utterances during turn-taking: The role of prediction, response preparation, and articulation. *Discourse Processes*, *55*, 230–240. https://doi.org/10.1080/0163853X.2017.1330031

Crowther, D. (2020). Rating L2 speaker comprehensibility on monologic vs. interactive tasks: What is the effect of speaking task type. *Journal of Second Language Pronunciation*, *6*, 96–121. https://doi.org/10.1075/jslp.19019.cro

Crowther, D., Trofimovich, P., & Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility: Nonnative listeners' perspectives. *Journal of Second Language Pronunciation*, *2*, 160–182. https://doi.org/10.1075/jslp.2.2.02cro

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility? *Modern Language Journal*, *99*, 80–95. https://doi.org/10.1111/modl.12185

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2018). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition*, *40*, 443–457. https://doi.org/10.1017/S027226311700016X

Duffy, K. A., & Chartrand, T. L. (2017). *From mimicry to morality: The role of prosociality.* In W. Sinnott-Armstrong & C. B. Miller (Eds.), *Moral psychology: Virtue and character*

(pp. 439–464). Boston Review.

Firth, A. (1996). The discursive accomplishments of normality: On 'lingua franca' English and conversation analysis. *Journal of Pragmatics*, *26*, 237–259. https://doi.org/10.1016/0378-2166(96)00014-8

Floyd, S., Manrique, E., Rossi, G., & Francisco, T. (2016). Timing of visual bodily behavior in repair sequences: Evidence from three languages. *Discourse Processes*, *53*, 175–204. https://doi.org/10.1080/0163853X.2014.992680

Galindo Ochoa, J. A. (2017). *The effect of task repetition on Colombian EFL students' accuracy and fluency* (Unpublished master's thesis). Concordia University, Montreal, Canada.

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal*, *64*, 428–433. https://doi.org/10.1111/j.1540-4781.1980.tb05218.x

Garrod, S., Tosi, A., & Pickering, M. J. (2018). Alignment during interaction. In. S.–A. Rueschemeyer & M. G. Gaskell (Eds.), *The Oxford handbook of psycholinguistics* (pp. 575–593). Oxford University Press.

Giles, H., & Ogay, T. (2007). Communication accommodation theory. In B. B. Whaley & W. Santer (Eds.), *Explaining communication: Contemporary theories and exemplars* (pp. 293–309). Lawrence Erlbaum.

Godfroid, A., Winke, P., & Conklin, K. (2020). Exploring the depths of second language processing with eye tracking: An introduction. *Second Language Research*, *36*, 243–255. https://doi.org/10.1177/0267658320922578

Hartig, F. (2020). DHARMa: Residual diagnostics for hierachical (multi-level/mixed) regression models. R package version 0.3.3. https://CRAN.R-project.org/package=DHARMa

Huang, B. H. (2013). The effects of accent familiarity and language teaching experience on raters' judgments of non-native speech. *System*, *41*, 770–785. http://doi.org/10.1016/j.system.2013.07.009

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*. Advance online publication. https://doi.org/10.1111/lang.12451

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*, 135–159. https://doi.org/10.1080/15434303.2013.769545

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility. *Studies in Second Language Acquisition*, *34*, 475–505. https://doi.org/10.1017/s0272263112000150

Isaacs, T., Trofimovich, P., & Foote, J. A. (2017). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*, *35*, 193–216. https://doi.org/10.1177/0265532217703433

Isbell, D. R., & Lee, J. (2022). Self-assessment of comprehensibility and accentedness in second language Korean. *Language Learning*. Advance online publication. https://doi.org/10.1111/lang.12497

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly*, *9*, 249–269. https://doi.org/10.1080/15434303.2011.642631

Kissling, E. M., & O'Donnell, M. E. (2015). Increasing language awareness and self-efficacy of FL students using self-assessment and the ACTFL proficiency guidelines. *Language*

Awareness, 24, 283–302. https://doi.org/10.1080/09658416.2015.1099659

Kleinschmidt D. F., & Jaeger T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*, 148–203. https://doi.org/10.1037/a0038695

Kuhl, P. K., Tsao, F. M., & Liu, H. M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences, 100*, 9096–9101. https://doi.org/10.1073/pnas.1532872100

Lee, S. K., & Chang, S. (2005). Learner involvement in self- and peer-assessment of task-based oral performance. *Language Research, 11*, 711–735.

Lenth, R. (2020). emmeans: Estimated marginal means. R package version 1.3.2. https://CRAN.R-project.org/package=emmeans

Levis, J. (2020). Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation, 6*, 310–328. https://doi.org/10.1075/jslp.20050.lev

Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., Lyu, Y., Chung, K. S., & Suen, H. K. (2016). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*, 245–264. https://doi.org/10.1080/02602938.2014.999746

Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing, 38*, 189–218. https://doi.org/10.1177/0265532220932481

Li, P., & Jeong, H. (2020). The social brain of language: Grounding second language learning in social interaction. *NPJ: Science of Learning, 5*, 8. https://doi.org/10.1038/s41539-020-

0068-7

Mackey, A. (1999). Input, interaction, and second language development: An empirical study of question formation in ESL. *Studies in Second Language Acquisition*, *21*, 557–587. https://doi.org/10.1017/S0272263199004027

Mauranen, A. (2006). Signalling misunderstanding in English as a lingua franca communication. *International Journal of the Sociology of Language*, *177*, 123–150. https://doi.org/10.1515/IJSL.2006.008

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

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, *28*, 113–131. https://doi.org/10.1017/S0272263106060049

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*, 329–351. https://doi.org/10.1075/jslp.20009.nag

Nagle, C. L., Trofimovich, P., O'Brien, M. G., & Kennedy, S. (2022). Beyond linguistic features: Exploring the behavioral and affective correlates of comprehensible second language speech. *Studies in Second Language Acquisition*, *44*, 255–270. https://doi.org/10.1017/S0272263121000073

Nagle, C., Trofimovich, P., & Bergeron, A. (2019). Toward a dynamic view of second language comprehensibility. *Studies in Second Language Acquisition*, *41*, 647–672. https://doi.org/10.1017/s0272263119000044

Nambiar, M. K., & Goon, C. (2016). Assessment of oral skills: A comparison of scores obtained through audio recordings to those obtained through face-to-face evaluation. *RELC Journal*, *24*, 15–31.

Nakatsuhara, F., Inoue, C., & Taylor, L. (2021). Comparing rating modes: Analysing live, audio, and video ratings of IELTS speaking test performances. *Language Assessment Quarterly*, *18*, 83–106. https://doi.org/10.1080/15434303.2020.1799222

Neu, J. (1990). Assessing the role of nonverbal communication in the acquisition of communicative competence in L2. In R. C. Scarcella, E. S. Andersen, & S. D. Krashen (Eds.), *Developing communicative competence in a second language* (pp. 121–138). Newbury House.

O'Brien, M. G. (2014). L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning*, *64*, 715–748. https://doi.org/10.1111/lang.12082

Parkinson, B. (2011). Interpersonal emotion transfer: Contagion and social appraisal. *Personality and Social Psychology Compass*, *5*, 428–439. https://doi.org/10.1111/j.1751-9004.2011.00365.x

Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing*, *19*, 109–131. https://doi.org/10.1191/0265532202lt224oa

Paxton, A., Dale, R., & Richardson, D. C. (2016). Social coordination of verbal and nonverbal behaviors. In P. Passos, K. Davids, & J. Y. Chow (Eds.), *Interpersonal coordination and performance in social systems* (pp. 259–274). Routledge.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*, 1002–1044. https://doi.org/10.1037/bul0000158

Pickering, M. J., & Garrod. S. (2021). *Understanding dialogue: Language use and social interaction*. Cambridge University Press.

R Core Team. (2021). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews: Neuroscience*, *20*, 495–505. https://doi.org/10.1038/s41583-019-0179-4

Rice, K., & Redcay, E. (2016). Interaction matters: A perceived social partner alters the neural processing of human speech. *NeuroImage*, *129*, 480–488. https://doi.org/10.1016/j.neuroimage.2015.11.041

Saito, K., & Shintani, N. (2016). Do native speakers of North American and Singapore English differentially perceive comprehensibility in second language speech? *TESOL Quarterly*, *50*, 421–446. https://doi.org/10.1002/tesq.234

Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (2019). How do L2 listeners perceive the comprehensibility of foreign-accented speech? Roles of L1 profiles, L2 proficiency, age, experience, familiarity and metacognition. *Studies in Second Language Acquisition*, *41*, 1133–1149. https://doi.org/10.1017/S0272263119000226

Saito, K., Trofimovich, P., & Isaacs, T. (2017a). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*, 439–462. https://doi.org/10.1093/applin/amv047

Saito, K., Trofimovich, P., Abe, M., & In'nami, Y. (2020). Dunning–Kruger effect in second language speech learning: How does self-perception align with other-perception over time? *Learning and Individual Differences*, *79*, 1–10.

https://doi.org/10.1016/j.lindif.2020.101849

Saito, K., Trofimovich, P., Isaacs, T., & Webb, S. (2017b). Re-examining phonological and lexical correlates of second language comprehensibility: The role of rater experience. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 141–156). Multilingual Matters.

Schmidtke, J. (2018). Pupillometry in linguistic research: An introduction and review for second language researchers. *Studies in Second Language Acquisition, 40*, 529–549. https://doi.org/10.1017/S0272263117000195

Seo, M. S., & Koshik, I. (2010). A conversation analytic study of gestures that engender repair in ESL conversational tutoring. *Journal of Pragmatics, 42*, 2219–2239. https://doi.org/10.1016/j.pragma.2010.01.021

Strachan, L., Kennedy, S., & Trofimovich, P. (2019). Second language speakers' awareness of their own comprehensibility: Examining task repetition and self-assessment. *Journal of Second Language Pronunciation, 5*, 347–373. https://doi.org/10.1075/jslp.18008.str

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition, 19*, 122–140. https://doi.org/10.1017/s1366728914000832

Trofimovich, P., Isaacs, T., Kennedy, S., & Tsunemoto, A. (2022). Speech comprehensibility. In T. M. Derwing, M. J. Munro, & R. Thomson (Eds.), *The Routledge handbook of second language acquisition and speaking* (pp. 174–187). Routledge.

Trofimovich, P., Nagle, C. L., O'Brien, M. G., Kennedy, S., Taylor Reid, K., & Strachan, L. (2020). Second language comprehensibility as a dynamic construct. *Journal of Second*

Language Pronunciation, 6, 430–457. https://doi.org/10.1075/jslp.20003.tro

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*, 435–461. https://doi.org/10.1075/jslp.21006.tro

Tsunemoto, A., Lindberg, R., Trofimovich, P., & McDonough, K. (2021). Visual cues and rater perceptions of second language comprehensibility, accentedness, and fluency. *Studies in Second Language Acquisition*. Published online 5 August 2021. https://doi.org/10.1017/S0272263121000425.

Tsunemoto, A., Trofimovich, P., Blanchet, J., Bertrand, J., & Kennedy, S. (2022). Effects of benchmarking and peer-assessment on French learners' self-assessments of accentedness, comprehensibility, and fluency. *Foreign Language Annals*, *55*, 135–154. https://doi.org/10.1111/flan.12571

Xie X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *Journal of the Acoustical Society of America*, *143*, 2013–2031. https://doi.org/10.1121/1.5027410

Yao, Z., Saito, K., Trofimovich, P., & Isaacs, T. (2013). *Z-Lab* [Computer software]. https://github.com/ZeshanYao/Z-Lab

Appendix

Background Information for Speaker Pairs

Pair	Speaker A			Speaker B		
	Native language	Gender	Age	Native language	Gender	Age
1	Farsi	male	26	Tamil	male	24
2	Hindi	female	24	Malayalam	male	25
3	Vietnamese	male	31	Arabic	female	25
4	Mandarin	male	24	Farsi	female	26
5	Farsi	male	30	Bengali	male	27
6	Hindi	female	24	Mandarin	female	23
7	Kannada	male	25	Portuguese	male	24
8	Gujarati	female	27	Azeri	male	25
9	Arabic	male	26	Punjabi	female	24
10	Tamil	male	24	Hindi	male	23
11	Hindi	male	23	Russian	female	28
12	Hindi	female	24	Farsi	male	28
13	Mandarin	female	24	Farsi	male	24
14	Nepali	male	23	Tamil	male	22
15	Farsi	male	27	Hindi	female	27

16	Hindi	male	26	Farsi	male	35
17	Tulu	female	25	Farsi	male	29
18	Portuguese	male	32	Farsi	male	30
19	Mandarin	female	23	Bengali	male	29
20	Urdu	male	22	Kannada	female	26