Second Language Comprehensibility As a Dynamic Construct

Pavel Trofimovich

Concordia University

Charles L. Nagle

Iowa State University

Mary Grantham O'Brien

University of Calgary

Sara Kennedy

Concordia University

Kym Taylor Reid

Concordia University

Lauren Strachan

Concordia University

This study examined longitudinal changes in second language (L2) interlocutors' mutual comprehensibility ratings (perceived ease of understanding speech), targeting comprehensibility as a dynamic, time-varying, interaction-centered construct. In a repeated-measures, within-participants design, 20 pairs of L2 English university students from different language backgrounds engaged in three collaborative and interactive tasks over 17 minutes, rating their partner's comprehensibility at 2–3 minute intervals using 100-millimeter scales (seven ratings per interlocutor). Mutual comprehensibility ratings followed a U-shaped function over time, with comprehensibility (initially perceived to be high) being affected by task complexity but then reaching high levels by the end of the interaction. The interlocutors' ratings also became more similar to each other early on and remained aligned throughout the interaction. These findings demonstrate the dynamic nature of comprehensibility between L2 interlocutors and suggest the need for L2 comprehensibility research to account for the effects of interaction, task, and time on comprehensibility measurements.

Keywords: comprehensibility; pronunciation; dynamic; interaction; processing fluency; second language

1. Introduction

In their seminal work published 25 years ago, Munro and Derwing (1995) showed that intelligibility and comprehensibility of second language (L2) speech were related yet partially independent constructs. In Derwing and Munro's framework, intelligibility is defined as "the extent to which a speaker's message is actually understood by a listener" (Munro & Derwing, 1995a, p. 76), and comprehensibility refers to listeners' "judgments on a rating scale of how

difficult or easy an utterance is to understand" (Derwing & Munro, 1997, p. 2). The constructs are partially independent because listeners who transcribe L2 utterances (which is a typical measure of intelligibility) nearly perfectly may nevertheless rate the same utterances as hard to understand. As a measure of ease or difficulty of understanding speech, comprehensibility has emerged as a key construct in empirical work focusing on linguistic, cognitive, and social variables associated with speech that is understandable to the listener (Derwing & Munro, 2015). However, in nearly all previous research, comprehensibility has been examined through one-time ratings (after speaking is completed) in monologic tasks (such as picture description) and by listeners evaluating audio recordings only (without seeing speakers). To extend prior research on comprehensibility, we set out to provide a time-sensitive comprehensibility profile by focusing on comprehensibility within interaction, through ratings elicited from L2 speakers themselves as they perform communicative tasks.

2. Background Literature

2.1. Comprehensibility: A Measure of Understanding

Typically assessed through listeners' transcriptions of speech content, intelligibility is often regarded as the gold standard for evaluating listener comprehension (Derwing & Munro, 2015). However, scalar ratings of comprehensibility are a useful measure of listener understanding in many contexts. To begin with, comprehensibility ratings are practical and intuitive, and they can be elicited and scored easily using speech samples featuring the same content. In contrast, intelligibility measures require tasks with unique speech content for each instance when intelligibility is measured (to avoid greater intelligibility for content that is repeated to listeners) and comparatively more time for listeners to complete the tasks. Comprehensibility ratings are also reliable across listeners, meaning that listeners generally agree with each other regardless of

how comprehensibility is measured—through Likert-type scales (Munro & Derwing, 1995), sliding scales (Saito et al., 2017), or direct magnitude estimation (Munro, 2018). Most importantly, although intelligibility and comprehensibility are partially independent, comprehensibility ratings provide a reasonable estimate of listeners' actual understanding of speech (Sheppard et al., 2017). For instance, Munro and Derwing (1995) reported substantial overlap between these dimensions, with correlation coefficients approaching .90, although the magnitude of this link might vary for different speakers and listeners (Matsuura et al., 1999).

Besides being a practical measure of understanding, comprehensibility ratings are also shaped by the linguistic content of speech, which makes comprehensibility a useful metric to understand how various linguistic dimensions in the speaker's speech impact the listener. In their initial work, Munro and Derwing (1995) found associations between listeners' comprehensibility ratings and several linguistic measures derived from the speech being evaluated, including phonemic substitutions, intonation accuracy, and morphosyntactic errors. More recent work has revealed two constellations of linguistic dimensions relevant to comprehensibility: pronunciation (individual segments, prosody, fluency) and lexicogrammar (variety and richness of vocabulary, accuracy and complexity of grammar). The exact combinations of linguistic dimensions feeding into listeners' judgments of comprehensibility can depend on the linguistic background of the speaker and on the speaking task (Crowther et al., 2018), but the general finding has been consistent. Many measures at the level of segments, prosody, fluency, grammar, and discourse have been linked to listeners' ratings of L2 comprehensibility in multiple languages (Bergeron & Trofimovich, 2017; Crowther et al., 2015a; O'Brien, 2014).

2.2. Comprehensibility: An Index of Processing Fluency

Comprehensibility ratings can also be conceptualized in a broader sense, as a measure

capturing listeners' processing fluency, which refers to a person's subjective experience of the ease or difficulty with which information is processed (Reber & Greifeneder, 2017). A key aspect of processing fluency which cuts across various social and psychological domains is that people appraise and respond to various situations based on the perceived difficulty they report while processing a stimulus (e.g., text, image, sound), which may or may not reflect their actual experience with that stimulus. For instance, statements attributed to people whose names are harder to pronounce are considered less trustworthy (Newman et al., 2014), regardless of the actual content of the statements. Similarly, readers exposed to a text printed in a difficult to read font react more negatively to the reading than those who read the same text in an easy to read font, despite having similar text comprehension (Sanchez & Jaeger, 2015; Song & Schwarz, 2008). These findings are strikingly similar to Munro and Derwing's (1995) observation that comprehensibility might be rated differently for speech that is perfectly intelligible, implying that listeners' reactions to speech might be linked not to actual understanding (intelligibility) but to comprehensibility.

There is indeed growing evidence that comprehensibility (as a metric of processing fluency) captures important decisions for listeners. For instance, in social-psychological research on listeners' attitudes, speakers whom listeners perceived as hard to understand were downgraded in listeners' affective and attitudinal evaluations. Such speakers were ascribed negative emotions of annoyance and irritation and labelled less intelligent and successful (Dragojevic et al., 2017). Similarly, in a study focusing on online learning, when students evaluated an instructional video narrated by the instructor who was hard to understand, they downgraded the instructor in their evaluations, expressing negative attitudes towards online coursework and evaluating video content as more difficult, even though students' actual

understanding of the video was not compromised (Sanchez & Khan, 2016). In fact, a comprehensibility scale akin to that used in L2 speech research has now been validated as part of a five-item processing fluency measure, and this measure appears to explain various human judgments (truthfulness, preference, perceived risk), all attributed to processing fluency in prior literature (Graf et al., 2018).

2.3. A Dynamic Approach to Comprehensibility

Speaking and listening are dynamic acts whose properties fluctuate over time, yet comprehensibility has rarely been framed as a dynamic, variable process. Speakers generally alternate between periods of fluent and disfluent speech, with such temporal cycles recurring every 10–30 seconds (Pakhomov et al., 2011). And listeners must continuously adapt their comprehension to process varying levels of accuracy, complexity, and fluency to interpret the speaker's message within an emergent discourse structure (Kuperberg & Jaeger, 2016). Conversation is an inherently social process regarded as a joint, co-constructed activity between interlocutors (Brennan et al., 2018). Based on theoretical views that posit tight coordination between interlocutors (Garrod et al., 2018), comprehensibility could be characterized by variability both within and across interlocutors and could involve a continuous, dynamic adaptation of the interlocutors to each other, with comprehensibility sensitive to both global influences (e.g., time on task, task difficulty) and local issues (e.g., dysfluency, error).

Nagle et al. (2019) recently explored whether comprehensibility can be construed as dynamic, examining how raters assign ratings in real time. In this study, 24 Spanish-speaking raters evaluated 3-minute speech samples recorded by L2 Spanish speakers responding to personally relevant prompts. The raters first used a computer interface which allowed them to increase or decrease the comprehensibility rating as the speech unfolded. The raters then

completed a stimulated recall interview, commenting on their thoughts while watching a video capture of their rating. Three distinct rater profiles emerged. Non-dynamic raters (the majority) increased or decreased comprehensibility ratings infrequently. Semi-dynamic raters increased or decreased ratings at a high frequency, but the magnitude of change was small. The two dynamic raters also changed ratings at a high frequency, with a high magnitude of change that was generally in the direction of lower comprehensibility. Most raters reported that their ratings moved towards greater comprehensibility either within the same sample or from one sample to another. Over half of the comments about increasing comprehensibility ratings pertained to discourse. These findings implied that comprehensibility ratings—from the perspective of the listener—are dynamic yet highly variable across raters and that these ratings might ultimately reflect discourse (meaning-making) aspects of interaction.

3. The Present Study

As discussed previously, comprehensibility ratings provide a good measure of understanding sensitive to the linguistic profile of speech; they also offer a useful metric of processing fluency relevant to human judgment. To understand the role of comprehensibility in interactive language use, it would be important to understand whether comprehensibility is a stable phenomenon or whether it fluctuates over time. The raters in Nagle et al.'s (2019) study had completed a one-way listening task, with no possibility to interact with a speaker. However, interactive speech, where interlocutors are reacting to one another in real time, is not only an authentic context of language use but also one that is likely amenable to potential changes in comprehensibility. Therefore, this study's goal was to provide a conversation-centric, time-sensitive view of comprehensibility for both interlocutors in a conversation.

To address this goal, we paired L2 English speakers from different language backgrounds,

completing three interactive tasks and rating each other's comprehensibility at approximately 2.5-minute intervals for a total of seven ratings. We examined how the speakers' judgments of each other's comprehensibility changed over time, for each speaker separately and for both conversation partners together in relation to each other's ratings. We also debriefed each speaker to clarify how their interaction and their comprehensibility ratings may have changed over time. Because the raters in Nagle et al.'s (2019) study (although quite variable in their judgments) showed improved ratings within and across the rated speech samples, we expected that comprehensibility ratings would vary across speakers but might show an upward trend as conversation progressed. Based on prior work on speaker–listener adaptation in dialogue (Garrod et al., 2018), we also expected that the two conversation partners might converge on common comprehensibility ratings. Because of the exploratory nature of this study, we made no additional predictions regarding the timing and extent as well as the sources (e.g., task difficulty, language errors) of potential changes in ratings. We asked two broad questions:

1. How do L2 speakers rate each other's comprehensibility over time?
2. Do speakers' ratings of their partners converge or significantly change over time?

4. Method

4.1. Participants

The speakers (M_{age} = 25.85 years, SD = 2.89) included 40 international graduate students (14 women, 26 men) from eight academic disciplines at an English-language university in Canada. The speakers, who reported speaking 17 home languages, had started learning English at a mean age of 8.18 years (SD = 4.58) and had received all primary and secondary schooling in their home countries. As part of university admission requirements, the speakers took standardized language tests and reported IELTS (31) or TOEFL (9) scores. The TOEFL scores

were converted to equivalent IELTS bands using validated conversion metrics (ETS, 2017), with the resulting IELTS scores ranging between 5.5 and 8.0 (M = 6.84, SD = 0.62) for the speaking component and between 6.0 and 9.0 (M = 7.60, SD = 0.95) for the listening component. As students at a university with a large international enrolment, the speakers indicated that they regularly spoke English (M = 56.75% daily, SD = 19.79) and rated themselves as being familiar with accented English (M = 6.33, SD = 1.67) on a 9-point scale. Each speaker was randomly paired with a previously unknown partner from a different language background (resulting in 20 pairs), with the constraint that speakers of related languages (e.g., Hindi and Urdu) were paired with partners from other backgrounds (see Appendix A for background information on the speakers' home languages, genders, and ages).

4.2. Tasks

The speakers engaged in three interactive tasks, administered in a fixed order. The first task (3 minutes) was a warm-up task, with the goal of discovering three things the speakers had in common (e.g., a similar hobby), as a way of helping them become familiar with each other. The second task (7 minutes) was a picture story completion task (Galindo Ochoa, 2017). Each speaker had seven scrambled images from a 14-panel picture story. They could not see each other's pictures and had to share their descriptions to produce a common narrative. The story depicted a man who, after winning a large sum of money and purchasing a new house and a car, experienced several unfortunate events, including a car accident and a robbery; the man eventually realized that the money did not make him happy. The final task (7 minutes) was a problem-solving task, where the speakers identified a common set of solutions to challenges experienced by international students. The speakers were encouraged to share their challenges (e.g., long delays in obtaining visas) before proposing common solutions.

4.3. Repeated Assessments

During approximately 17 minutes of interaction, the speakers evaluated themselves and their partner for comprehensibility seven times. The speakers also evaluated their own and their partner's communicative anxiety and collaborativeness, but these assessments will not be discussed further. The rating episodes were fairly equally spaced: one at the end of each task (Time 1, 4, and 7) and two additional ratings approximately 2.5 minutes and 5 minutes after the beginning of Task 2 (Time 2 and 3) and after the beginning of Task 3 (Time 5 and 6). The rating scales (100-millimeter lines) were printed next to each rated dimension, one labeled "me" for self-rating and the other labeled "my partner" for the rating of the speaker's partner. The scales contained no markings besides labeled endpoints (*difficult to understand–easy to understand*), and the speakers were asked to mark the point on the line which reflected their judgment. Comprehensibility, introduced to the speakers before the tasks, was defined as a judgment of how much effort it takes to understand what someone is saying. Because this report focuses only on peer-ratings (speakers' evaluations of their partners), self-ratings are not discussed further.

4.4. Procedure

Each pair of speakers was tested individually, and the entire session was audio-recorded. The speakers first completed a background questionnaire. Then, a research assistant (RA) described each rated dimension and explained how to complete the ratings, using practice scales. The RA also advised the speakers that they would complete the scales several times, evaluating the immediately preceding 2–3 minutes of interaction, that they would be stopped periodically during Tasks 2 and 3, and that the ratings were private. The speakers then received the testing booklet, with instructions for each task and seven sets of rating scales printed on separate pages. Each task was introduced immediately before the speakers engaged in it: They first read the

printed instructions, then summarized task directions to the RA, and (when applicable) asked clarification questions. The speakers were reminded that the task would stop after the required time had elapsed (3 minutes for Task 1, 7 minutes for Tasks 2 and 3), even if they did not complete their discussion. The RA, who was present in the room during the entire task sequence, stayed away from the speakers during each segment of interaction, using a timer to ensure that task length was comparable across all pairs and that the ratings occurred at evenly spaced intervals. Although the speakers may have felt monitored to some extent, they generally appeared to be focused on completing the tasks. After completing the tasks, both speakers met a different RA in separate rooms and filled out a debrief questionnaire, rating their reactions to the session (100-millimeter scales). Each speaker was then briefly interviewed using guiding questions focusing on their experience during the session.

5. Data Analysis

5.1. Coding

The speakers' ratings of each other's comprehensibility were converted to numerical values, defined as the distance (in millimeters) between the left endpoint and the speaker's mark on the scale (out of 100 points). The speakers' rated responses to the debrief questions were also expressed as numeric values (out of 100 points). The recordings of the speakers' interaction were transcribed and then verified by trained RAs to enable a lexical analysis of each speaker's output. Finally, the speakers' interviews were transcribed, with analysis focusing on the speakers' responses to the two most relevant questions, namely, how their interaction changed over time and which aspects of their partners' speech were most difficult to understand. The speakers' comments were coded thematically, following an iterative process, with response categories derived from the content of the transcripts (Gibson & Brown, 2009). The first author initially

derived codes for themes and subthemes, then a co-author reviewed the coding and suggested modifications to it, until there was full consensus on all coding decisions.

5.2. Identification of Covariates

We identified variables associated with the speakers' comprehensibility ratings so that these variables could be included as covariates in statistical modeling. We first examined the speakers' debrief ratings, on the assumption that the speakers' individual experiences might have impacted their comprehensibility ratings. The speakers generally found the interaction successful (M = 87.5, SD = 9.6) and enjoyable (M = 91.3, SD = 11.4); they considered themselves involved in the tasks (M = 92.2, SD = 9.3) and found their partners pleasant (M = 91.6, SD = 10.4); they also expressed interest in speaking to their partners again (M = 92.3, SD = 12.1) and were satisfied with their performance (M = 81.6, SD = 13.6). None of the six debrief ratings were associated with comprehensibility (all correlations were below .30), so no debrief category was included in subsequent modeling.

We then targeted the speakers' speaking and listening proficiency, as comprehensibility ratings might reflect each speaker's L2 skill level. Across the 20 pairs, the two paired speakers differed (in absolute values) on average by 0.56 points on the IELTS speaking scale (SD = 0.59) and by 1.20 points on the IELTS listening scale (SD = 0.70). Although small, these differences could not be regarded as trivial; therefore, both IELTS speaking and listening scores for each speaker were entered as control covariates in subsequent statistical modeling.

We then focused on the speakers' output in each task using lexical profiling (Cobb, 2019) because comprehensibility might reflect each partner's contribution to the dialogue, in terms of its quantity (tokens) and content richness (types). Although token and type frequencies are basic measures of lexical content, they capture substantial amounts of shared variance (38–61%) in

listener judgments of L2 speech (Trofimovich & Isaacs, 2012). Because type and token frequencies were highly correlated (r = .94), implying their non-independence, only type frequency was used as a covariate, with type values computed separately for each speaker within each segment preceding the rating episode (i.e., before Time 1, between Time 1 and 2, and so on).

In the final check, we examined whether the 20 pairs varied in the amount of time they spent on tasks and in the timing of rating episodes, assuming that comprehensibility might reflect time-on-task differences. On average, the pairs spent 2 minutes and 46 seconds on Task 1, with some completing this task faster than others (01:04–03:14). However, because few pairs completed Tasks 2 and 3 within the time limit, using (nearly) all of the allotted 7 minutes, their time on task was comparable. The pairs spent on average 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). The repeated ratings also occurred at similar intervals, with the rating episodes spaced about 2.5 minutes apart (02:46–02:37–02:32–02:02–02:35–02:34–02:00). Although time on task was generally consistent across pairs and rating episodes, all statistical models were also re-run with a timing covariate that tracked each pair's deviation from the intended rating time, given that the speakers who rated earlier or later than intended may have evaluated each other differently. Finally, model fit was also evaluated using raw timing (actual time of each speaker's assessment) instead of treating time as an equal-interval variable (seven rating episodes).

5.3. Statistical Modeling

The speakers' ratings of each other's comprehensibility were examined through mixed-effects modeling using the lme4 package (Bates et al., 2015) in R version 3.6.1. (R Core Team, 2019). In each set of models, the relevant rating served as the dependent variable, with time

(seven rating episodes) as a fixed factor and random intercepts for pairs and for speakers. Model fit was evaluated by performing likelihood ratio tests on pairs of nested models using the ANOVA function, with a more complex model adopted only when it improved fit. For all model parameters, 95% confidence intervals were derived to determine the statistical significance of each parameter (interval does not cross zero). All models included four fixed effects as control covariates: (a) speakers' IELTS speaking score, (b) speakers' IELTS listening score, (c) lexical type frequency in each speaker's output preceding each rating episode, and (d) a speaker-specific time deviation variable capturing whether a rating episode occurred before or after the intended time. All covariates were centered, such that the sample mean was set to 0 and negative values indicated performance below the mean and positive values performance above the mean.

6. Results

6.1. Comprehensibility Across Time

The first research question asked whether speakers' ratings of their partners' comprehensibility changed over time. Figure 1 illustrates the 40 individual speakers' ratings of their partners' comprehensibility across the seven rating episodes (Time 1–7) with speakers in the same pair shown in the same color. Although different speakers (as rated by their conversation partners) showed varying comprehensibility trajectories, the speakers generally rated each other's comprehensibility high after Task 1 ($M_{\text{Time 1}}$ = 90.69, *SD* = 11.56), reduced their ratings during Task 2 ($M_{\text{Time 2}}$ = 82.14, *SD* = 18.08, $M_{\text{Time 3}}$ = 79.78, *SD* = 17.35), and gradually increased their ratings to approximately the same high initial level by the end of Task 3 ($M_{\text{Time 7}}$ = 92.31, *SD* = 9.17). Moreover, Task 2 tended to yield the most variable comprehensibility ratings, with a U-shaped pattern evident for many speakers.

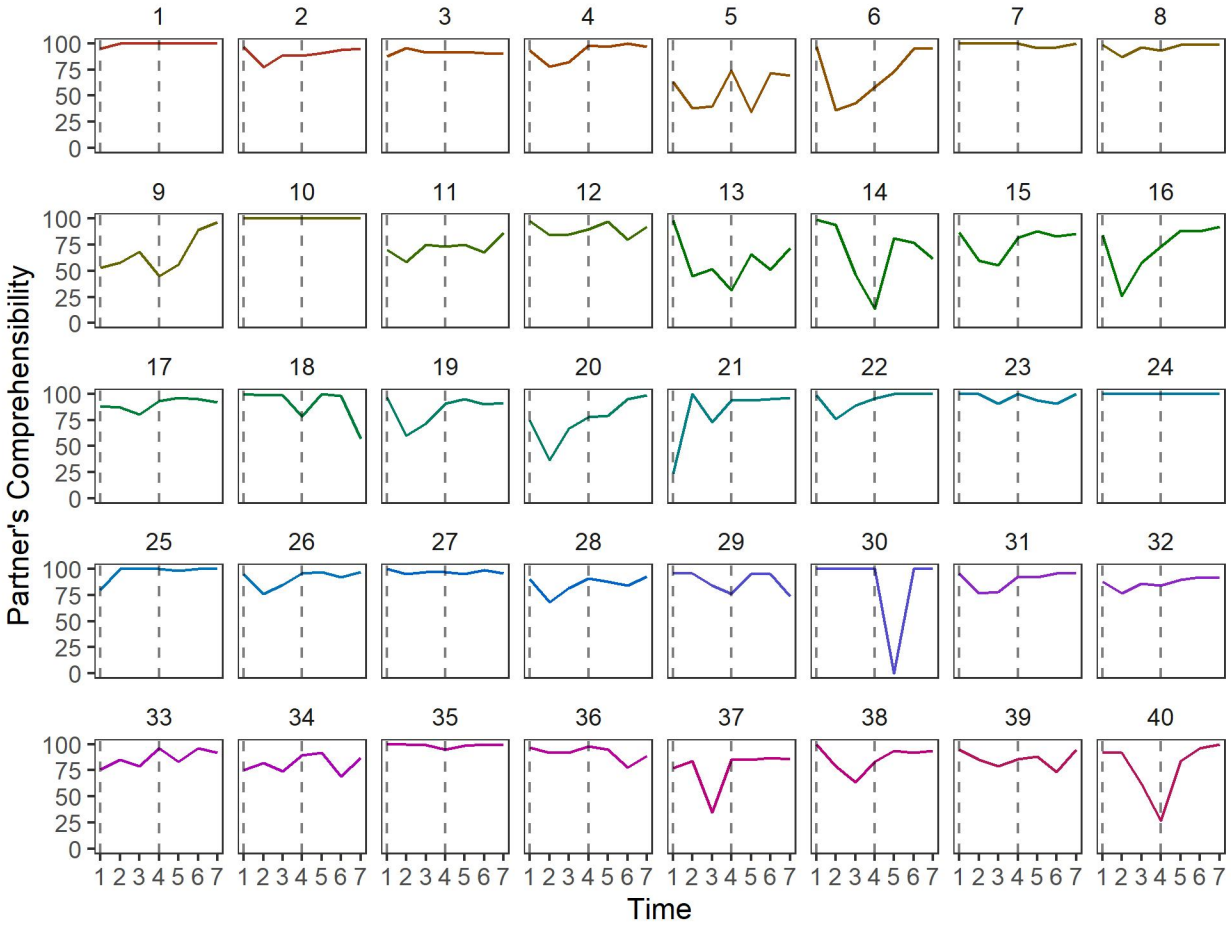

Figure 1. Individual comprehensibility rating trajectories across the seven rating episodes. The vertical dashed lines indicate the three tasks (Task 1: 1, Task 2: 2–4, Task 3: 5–7). Speakers in the same pair are shown in the same color (e.g., 1 and 2, 3 and 4, …, 39 and 40).

To explore the effect of time, we fit four polynomial change models: a null (intercept) model and linear, quadratic, and cubic growth models. Because ratings fluctuated during Task 2, we also fit a piecewise growth model, with time recoded into two dummy variables representing rate of change over Time 1–4 (Tasks 1 and 2) and Time 5–7 (Task 3). In the piecewise model, we estimated linear and quadratic rates of change over Task 2 only, based on the observation that ratings fluctuated most substantially and non-linearly for most participants over that period. This

model was equivalent to the cubic growth model in complexity (i.e., had the same number of terms), but the estimated trajectory was slightly different, insofar as the quadratic (U-shaped) function was limited to the first few datapoints. Polynomial time predictors were fit using the poly function to generate orthogonal terms, preventing autocorrelation among linear, quadratic, and cubic slopes.

With respect to the polynomial models, including a higher-order time function significantly improved model fit: null vs linear, $\chi^2(1) = 6.93$, $p = .008$; linear vs. quadratic, $\chi^2(1) = 10.70$, $p = .001$; quadratic vs. cubic, $\chi^2(1) = 7.87$, $p = .005$. Direct comparison of the cubic and piecewise models using likelihood ratio tests was not possible since the models were not nested. We therefore used the Akaike and Bayesian information criteria to select the best-fitting model because these criteria can be used to compare non-nested models fit to the same dataset (Singer & Willett, 2003). The criterion values for the piecewise model were smaller, suggesting that it was a superior fit to the data. By-speaker random slopes were tested for the time terms, but piecewise models including those effects either did not converge or were singular, suggesting overfit. Therefore, only by-speaker and by-pair random intercepts were retained. When we inspected the model residuals, we identified and removed eight datapoints with standardized residuals greater than 2.5 *SD*s (2.86% of the data) and then refit the model. Table 1 summarizes this model, which accounted for 59% of the variance in comprehensibility ratings (marginal $R^2 = .12$, conditional $R^2 = .59$). Figure 2 shows the model-estimated trajectory (dashed line) and observed individual trajectories (solid lines), which display a great amount of variability.

Table 1. Summary of final mixed-effects model for comprehensibility

Fixed effects	Estimate	*SE*	*t*	95% CI	*p*
Intercept	85.64	2.19	39.02	[81.39, 89.89]	< .001
Tasks 1 and 2					
Time linear	–5.11	12.03	–0.42	[–28.57, 18.36]	.67
Time quadratic	52.78	10.77	4.90	[31.61, 73.66]	< .001
Task 3					
Time linear	1.83	0.68	2.68	[0.50, 3.17]	.008
Covariates					
IELTS Speaking	1.80	1.65	1.09	[–1.53, 4.96]	.28
IELTS Listening	1.04	1.51	27.97	[–1.86, 3.99]	.50
Type frequency	0.63	0.75	0.84	[–0.81, 2.12]	.40
Time deviation	–3.25	3.31	–0.98	[–9.66, 3.23]	.33
Random intercepts		*SD*			
Pair		7.80			
Speaker		6.60			

Note. Tasks 1 and 2 linear and quadratic predictors were orthogonal; they should not be interpreted on the original time scale.

As reported in Table 1, the significant coefficient for the orthogonal Task 1 and 2 quadratic slope shows that changes in comprehensibility were not linear over those datapoints, but rather U-shaped, and the significant coefficient for the Task 3 linear slope shows that comprehensibility increased steadily over Task 3. Comprehensibility was independent of the speakers' lexical contribution or their speaking or listening proficiency; these variables did not explain any

additional model variance. In addition, the speaker-specific time deviation variable, which captured whether a rating episode occurred before or after the intended time, missed significance. Examining the distribution of model residuals revealed heavy tails.[1]

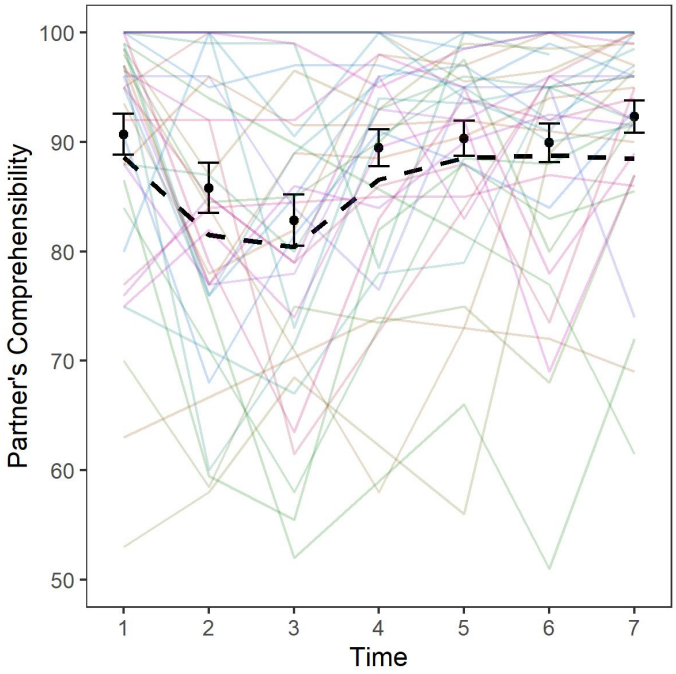

Figure 2. Model-estimated partner comprehensibility trajectory (dashed line) and observed individual trajectories (solid lines). Solid dots designate group mean, and error bars enclose the 95% confidence interval.

6.2. Convergence or Divergence in Comprehensibility

The second research question asked whether the speakers' ratings of each other's comprehensibility became more aligned during interaction. Table 2 summarizes descriptive statistics for comprehensibility ratings at each of the seven rating episodes, separately for the two speakers across the 20 pairs (i.e., for Speaker A vs. Speaker B). The two speakers in each pair were designated as A or B in a random fashion, determined by seat assignment (at opposite sides

of a table) upon a speaker's arrival in a testing room.

Table 2. Summary of comprehensibility ratings for Speaker A (as rated by Speaker B) and Speaker B (as rated by Speaker A) across the seven rating episodes

	Speaker A			Speaker B		
Rating	*M*	*SD*	*Range*	*M*	*SD*	*Range*
Time 1	87.32	13.92	53–100	93.90	7.83	75–100
Time 2	81.20	20.50	38–100	83.19	15.47	36–100
Time 3	79.63	17.34	40–100	79.93	17.81	43–100
Time 4	85.18	18.27	32–100	88.56	11.23	58–100
Time 5	86.03	16.58	35–100	91.89	7.95	73–100
Time 6	88.40	12.94	51–100	91.45	9.19	69–100
Time 7	90.88	9.51	69–100	93.82	8.80	62–100

As shown in Table 2, on average, the two speakers across all pairs appeared the most divergent in each other's comprehensibility ratings during the first rating episode, after Task 1 ($M_{\text{Speaker A}} = 87.32$ vs. $M_{\text{Speaker B}} = 93.90$), such that one speaker in a pair was perceived as being more comprehensible than the other. However, the two speakers generally converged on a common rating approximately 5 minutes into the interaction at Time 2 ($M_{\text{Speaker A}} = 81.20$ vs. $M_{\text{Speaker B}} = 83.19$), and remained fairly aligned after that, except perhaps at Time 5 ($M_{\text{Speaker A}} = 86.03$ vs. $M_{\text{Speaker B}} = 91.89$). Illustrated graphically in Figure 3, the speakers' comprehensibility ratings of their respective partners generally followed the same U-shaped trajectories, but the ratings were substantially mismatched only at the outset of the interaction.

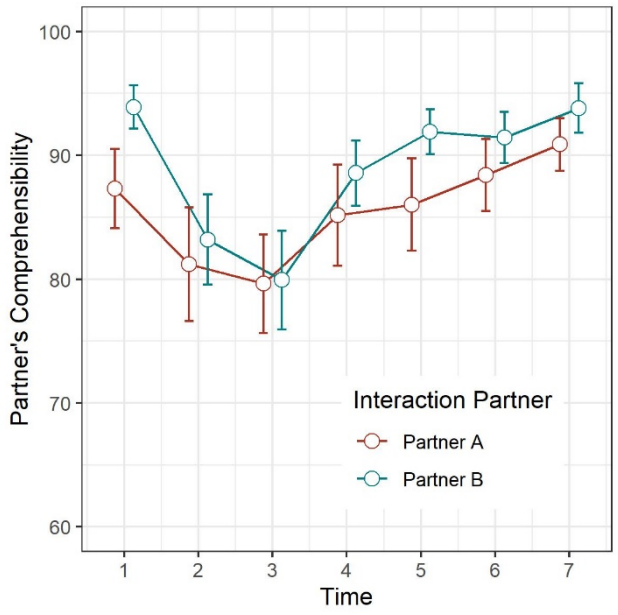

Figure 3. Mean comprehensibility for both speakers in each pair across the seven rating episodes. Vertical bars encompass 95% confidence intervals around the mean values. Speaker A and B designations are random within each pair.

Preliminary plotting of group and individual data for within-pairs differences in comprehensibility did not suggest a definitive pattern for change over time; instead, for some pairs, differences in partner comprehensibility diminished over ratings, whereas for others, ratings were most dissimilar near the end of the interaction. Considering this variability, we fit exploratory polynomial models to the absolute value of the within-pair difference in comprehensibility, comparing each model to the baseline (intercept) model. None of these models significantly improved fit over the intercept model. In a follow-up exploratory analysis, which is conceptually similar to the piecewise model reported above, we split the dataset into separate subsets corresponding to Tasks 2 and 3, each with three datapoints, and examined change in alignment in each subset independently. Because the Task 2 and 3 subsets contained

only three datapoints, we could only estimate linear and quadratic rates of change.

For Task 2, neither the linear nor quadratic model significantly improved fit over the intercept model. However, for Task 3, the linear model improved fit over the intercept model, albeit marginally, $\chi^2(1) = 4.15$, $p = .04$. Including by-pair random slopes for linear time resulted in singular fit, so the model reported in Table 3 contained only by-pair random intercepts. As before, between-speaker differences in comprehensibility were unrelated to speakers' proficiency and the timing of their ratings. However, lexical characteristics were marginally related to comprehensibility; speakers producing more word types were rated as less comprehensible. Residuals for both final models were normally distributed, with only minor excursions at the upper tail.

Table 3. Summary of fixed effects for between-speaker differences in comprehensibility

Fixed effects	Estimate	*SE*	*t*	95% CI	*p*
Task 2					
Intercept	13.20	2.49	5.31	[8.64, 17.75]	< .001
IELTS Speaking	–0.33	2.84	–0.12	[–5.54, 4.85]	.91
IELTS Listening	2.14	2.07	1.03	[–1.64, 5.92]	.32
Type frequency	–2.44	2.35	–1.04	[–6.99, 1.97]	.30
Time deviation	2.40	4.24	0.57	[–5.35, 10.21]	.58
Task 3					
Intercept	17.70	3.23	5.49	[11.83, 23.89]	< .001
Time linear	–4.45	2.25	–1.97	[–8.91, –0.18]	.06
IELTS Speaking	1.60	2.43	0.66	[–2.79, 6.01]	.52
IELTS Listening	0.58	1.83	0.32	[–2.73, 3.90]	.76
Type frequency	–3.96	1.93	–2.05	[–7.84, –0.45]	.05
Time deviation	5.17	3.77	1.37	[–1.64, 12.00]	.19

6.3. Interview Responses

To clarify individual rating patterns, we examined the speakers' interview responses to two questions: how their interaction changed, and which aspects of their partners' speech were most difficult to understand. As shown in Table 4, to explain change over time, the speakers made 58 comments, most of which (44 or 76%) encompassed four categories. In three such categories (33 or 57%), the speakers attributed change to (a) reduced anxiety and increased confidence and comfort, (b) improved or sustained collaboration, or (c) enhanced knowledge of their partner:

- I think maybe at the beginning, we were a bit stressful since we just began and it was

like conversation, but then we were more relaxed (S22);

- I think from the first to the last, the collaboration, the sense of collaboration is increased and cooperate more (S9);
- The first activity was about… finding the common things, when you find common things, then it was easier to communicate… so it went easy and easy as time (S1).

The fourth category (11 or 19%) included largely negative comments pertaining to speakers' difficulty with a task, which was exclusively Task 2, or to other methodological issues:

- The second task was a bit difficult to comprehend because of lack of clarity, I wouldn't blame [partner] for [it] because he tried his best to show me the real picture like he was having (S2);
- But with the interruptions, this is something that breaks you… and then you have to rate again, but even with that it's super easy to continue dealing with that (S35).

Table 4. Frequency of comments (*k*) and number of pairs (out of 20) contributing comments

	Change over time			Understanding difficulty		
Coded category	*k*	%	Pairs	*k*	%	Pairs
Anxiety, comfort, confidence	14	24.1	11	1	2.3	1
Task effects	11	19.0	10	4	9.3	4
Enhanced knowledge of partner	11	19.0	10			
Increased or sustained collaboration	8	13.8	6			
Accent familiarity	5	8.6	5	6	14.0	6
Shared experience and knowledge	4	6.9	3			
No change, no issue with understanding	3	5.2	3	10	23.3	9
General improvement	2	3.4	2			
Grammar				1	2.3	1
Vocabulary				2	4.7	2
Fluency				2	4.7	2
Voice quality				2	4.7	2
Sufficient explanation and details				4	9.3	4
Pronunciation				11	25.6	10
Total	58	100		43	100	

To explain understanding difficulty, the speakers cited pronunciation issues, which made up a quarter (11 or 26%) of the 43 comments produced. Pronunciation issues included generally unclear accent and problems with specific sounds, words, and intonation; vocabulary, grammar, or fluency were rarely mentioned as barriers to understanding, which is unsurprising given that for most speakers the term “accent” encompasses various language issues, including lexical

choice, grammatical appropriateness, and issues of fluency or flow:

- I think the accent, his accent was difficult for me (S32);
- Some letters were not pronounced clearly, like… when he was saying "thief" if I heard the "teef" and I have to ask him to repeat it to understand (S6);
- Intonation and pronunciation, I think, she didn't have good intonation (S26).

In another set of comments (11 or 23.3%), the speakers cited no difficulty in understanding each other, largely explained through both partners sharing a cultural background or partners' joint teamwork:

- Because of the community we belong, like it's easy for us to understand… what he is actually going to talk about (S19);
- When he stop speaking, then I speak; sometimes when I stop speaking, he speak… we cooperate well (S25).

Accounting for 9–14% of the comments, other reasons for understanding difficulty included task-specific issues (again limited to Task 2), familiarity with partners' accent, and partners' ability to express ideas clearly or provide sufficient detail:

- Accent a little bit different, but I get used to this… it's like I understand it's not perfect British English that I learned at school (S22);
- He's not explaining the part of the picture… he's giving only one two three pictures scenarios (S31).

More importantly, the speakers' individual comprehensibility ratings (plotted in Figure 1) did not appear to unambiguously map onto the stated reasons for how their communication changed or which issues contributed to difficulty in understanding. For example, of the seven speakers rated consistently as being highly comprehensible (1, 3, 7, 10, 24, 27, and 35 in Figure

1), there were only two cases where the partner cited no problem with understanding the speaker. Similarly, across the speakers whose comprehensibility was rated as changeable (dynamic trajectories in Figure 1), 11 partners reported no problems contributing to difficulty in understanding these speakers or reported no change to communication over time.

7. Discussion

This exploratory study's goal was to examine whether comprehensibility could be viewed as a dynamic, time-sensitive construct for both interlocutors in a conversation and to explore whether comprehensibility ratings might be co-dependent on both interlocutors and thus subject to convergence or divergence effects over time. We found evidence for a dynamic change in comprehensibility consistent with an exponential (U-shaped) trendline which was independent of speakers' proficiency, lexical contribution, or the timing of a rating episode. Although the speakers' comprehensibility judgments displayed a great amount of inter-individual variability, they rated their partners' comprehensibility as generally high after Task 1, their ratings then dropped during Task 2 and increased gradually throughout Task 3. In terms of the relationship between interlocutors' ratings, although the best-fitting model showed no significant time effect, the absolute differences in mutual ratings seemed to diminish over time and tasks, approximating a linear function, especially during Task 3, suggesting that the speakers' ratings showed more similarity over time. Based on interview comments, the most frequent changes to communication patterns were decreased anxiety and increased confidence, improved collaboration, and enhanced knowledge of the partner. The most cited issues leading to difficulties in understanding were various pronunciation issues, task-specific influences, and partner's ability to provide sufficient content detail.

7.1. Time- and Task-Sensitive View of Comprehensibility

In their study exploring how raters' comprehensibility ratings evolved over time, Nagle et al. (2019) provided a micro perspective on comprehensibility as a time-sensitive construct, arguing that comprehensibility judgments displayed several properties of dynamic systems (de Bot et al., 2007), including change over time and nonlinearity. For instance, comprehensibility judgments in that study were variable both within and across the raters and displayed nonuniformity, in the sense that different types of linguistic issues (e.g., phonemic errors, lexical substitutions), with their particular timing and location in the evolving narrative, elicited different reactions from different raters. To complement this micro-level view of comprehensibility, the present findings offer a global, macro-level perspective, demonstrating that comprehensibility ratings for both speakers in a conversation, while overall highly variable, seem to fluctuate in tandem in extended communication. The two macro variables emerging from this dataset with relevance to comprehensibility are time and task.

That comprehensibility ratings are sensitive to time (understood broadly as listeners' cumulative experience with a speaker's speech) is unsurprising. To begin with, time might have a negative influence on comprehensibility, so that speech evokes more effortful processing for the listener, at least early in a conversation. For example, raters sometimes assign harsher ratings when they evaluate the same sentence-length speech samples again, because raters might become increasingly aware of how the speakers' output differs from the language expected by the rater (Flege & Fletcher, 1992). Similarly, there seems to be little consistency between raters' evaluations of separate short sentences by the same speaker, suggesting that ratings of shorter speech samples might not be representative of ratings of longer discourse produced by the same speaker (Munro, 2018).

Time might also impact comprehensibility positively, such that, as interaction proceeds, speech becomes less effortful for the listener. For instance, raters with greater linguistic exposure or experience (language teachers, multilinguals) generally assign higher ratings than those with less exposure or experience (Kang, 2012; Saito & Shintani, 2016). An upward trend in comprehensibility would also be compatible with the notion that listeners' perceptual categories are highly adaptive to recent experience (Baese-Berk, 2018). In the end, both negative and positive time-bound forces may have been at play here, yielding a U-shaped comprehensibility function, with negative influences operating early in the interaction, until a certain temporal threshold was reached, and positive influences acting as comprehensibility attractors later on.

Comprehensibility ratings also seemed to depend on the communicative task performed by interlocutors, on the assumption that different tasks impose greater or lesser demands on the speaker and thus increase or decrease processing effort for the listener. Increased task difficulty likely elicits more sophisticated language from speakers, while also increasing opportunities for them to make errors or experience a communication breakdown (Robinson, 2005), which may explain why raters experience greater processing effort in evaluating more complex tasks (Crowther et al., 2015b). In this study, the dip in comprehensibility following the first rating (illustrated by a quadratic time function for Tasks 1–2) was likely due to higher cognitive demands in the second task, with ratings continuing to rise as speakers moved through an easier task (shown by a positive linear time function for Task 3).

In terms of task difficulty, the initial task had low cognitive demands because speakers had an unlimited range of possible commonalities to consider. The second task was more complex due to the need to exchange nonshared information by identifying and describing referents in 14 scrambled images (a conclusion also supported in interview comments). The final

task was less demanding because partners had shared access to the initial information and could complete the task by co-constructing agreed-upon solutions. Until the effects of time and task are disentangled in future work by rotating task order across speakers, our interim conclusion is that comprehensibility trajectories reflect individual and joint contributions of interlocutors becoming accustomed to each other and to specific tasks and their features over time. A key qualification here is that such interlocutor experiences across tasks and time appear to be subject to vast amounts of inter-individual variability, which also needs to be explained in future work.

7.2. Between-Speaker Alignment in Comprehensibility

Comprehensibility ratings for the two speakers engaged in interaction appeared to be co-dependent, showing a trend for convergence over time. Between-speaker comprehensibility differences were approximately 15–20 points on a 100-point scale and were highly variable. However, these differences generally decreased over time, particularly during Task 3, dropping below a 10-point difference by the end of the interaction. This novel finding extends prior work on interactive alignment (Garrod et al., 2018) to include interlocutor alignment in comprehensibility. Interactive alignment reflects a phenomenon whereby interlocutors converge on common speech patterns, driven by such social forces as accommodation to an interlocutor (Giles & Ogay, 2007) and psychological mechanisms of priming (Garrod et al., 2018), with alignment involving multiple features of speech, including utterance length, speech rate, phonetic realizations of segments and words, volume, and pausing frequencies and lengths (Garrod et al., 2018). Convergence in various speech patterns has also been attested among L2 speakers and has been shown to depend on speech style and speaker proficiency (Berry & Ernestus, 2018). The finding that frequency of word types was negatively associated with speaker convergence in comprehensibility highlights another variable that might modulate alignment, in this case, by

increasing interlocutors' effort in understanding speech with increased lexical content.

The obtained evidence for speaker convergence in comprehensibility is also consistent with prior research on listener adaptation to foreign accent (Baese-Berk, 2018). For instance, listeners rapidly get attuned to the speech of unfamiliar L2 speakers, often requiring just over a minute of experience (Clarke & Garrett, 2004). Xie et al. (2018) recently extended these findings to show that listeners improve quickly (in a matter of minutes) in speed and accuracy of comprehension of unfamiliar L2 speakers, arguing that long- and short-term adaptations to L2 speech might be driven by similar mechanisms. Our finding of a rapid convergence in interlocutors' comprehensibility ratings, which generally occurred within 1–3 minutes of their experience in the initial task (see Figure 3), is suggestive of a parallel phenomenon for comprehensibility. Adaptation to L2 comprehensibility (at least for L2 speakers) might involve interlocutors engaging in a process of adjusting their expectations of the effort involved in understanding their partners, by checking these expectations against the actual linguistic evidence available in discourse (for a potential model, see Kleinschmidt & Jaeger, 2015). Because the discourse in dialogue is usually co-constructed by both interlocutors (e.g., through turn-taking and feedback as part of attaining a common interactive goal) and likely involves interlocutors predicting upcoming content and potential language errors, it is reasonable that L2 interlocutors (especially those at comparable L2 skill levels) would arrive at a shared, conversation-specific (rather than speaker-specific) view of comprehensibility. As shown in Figure 3, once a shared understanding of comprehensibility has been reached (which might require more time for some pairs than for others), this shared rating of comprehensibility is what describes both partners' conversational experience across time and task.

8. Limitations and Future Work

A major limitation of this work, which prevented us from making specific predictions beyond asking exploratory questions, is that tasks were not rotated across speakers. As discussed previously, it is important to examine whether similar U-shaped comprehensibility trajectories would emerge when speakers engage in communicative tasks ordered differently, clarifying how interlocutors' cumulative shared experience impacts their comprehensibility ratings in tasks that increase versus decrease in cognitive difficulty across time. Similarly, the speakers' comments regarding changes in their interaction patterns and reasons for difficulty in understanding their partners did not unambiguously explain the speakers' comprehensibility rating trends. For instance, the speakers may not have been aware of how and why their perceived effort of understanding their partners varied, which would implicate an implicit component to ratings. More likely, however, the speakers did not possess the needed terminology to describe their thought processes and largely resorted to the categories made salient to them (see Table 4), through either the experimental procedure (anxiety, collaboration) or conversation tasks (understanding, getting to know partners). The link between interaction-based comprehensibility ratings and interlocutor awareness of comprehensibility should be investigated further, using different combinations of interlocutors that vary in language proficiency and experience.

With respect to interactive alignment, as suggested by an anonymous reviewer, between-speaker convergence or divergence in comprehensibility could be potentially misleading, in the sense that speakers may have given the impression of convergence or divergence because they approached the rating task using different criteria, showed varying degrees of rating severity, or tended to avoid extreme rating values (thus demonstrating a regression-to-the-mean effect). Future work should therefore revisit the validity of interactive ratings of comprehensibility by

ensuring (at minimum) that raters are trained on the use of the rating scale to the point of calibrated performance. It would also be interesting to explore potential effects of cognitive workload on alignment in comprehensibility ratings. For example, certain interactive tasks might be particularly prone to highlighting partner-specific comprehensibility issues in interaction, preventing or delaying convergence. Similarly, it might be useful to explore long-term effects of interlocutors' extended conversational experience on their perception of comprehensibility, focusing on speakers' judgments of the same and new partners in another instance of interaction, after a delay. In light of the alignment between both partners' comprehensibility scores in extended interaction, it might also be fruitful to examine the validity of a joint (rather than speaker-specific) measure of comprehensibility for both partners in a conversational dyad. Finally, comprehensibility ratings, as useful measures of listener understanding and listener processing fluency, could be examined in relation to such conversation phenomena as speakers' engagement in dialogue, their participation patterns, or their affective response to the task or their partner, to clarify the role of processing effort in interlocutor experience in dialogue.

9. Conclusion

Over the last 25 years, comprehensibility ratings have become a valuable metric that captures various facets of listeners' experience with L2 speech, implicated in multiple social, linguistic, and psychological phenomena. Our goal was to extend prior work on comprehensibility by providing a conversation-centric, dynamic view of this construct in interaction. Our findings imply that listeners' judgements of L2 comprehensibility can change in real time according to listeners' immediate experience, particularly for listeners in interactive speaking tasks, and that such judgments may be subject to convergence effects over time. These initial results call for rigorous future research in order to understand whether

comprehensibility—as a proxy for listeners' effort in understanding speech—could capture many other important real-life dimensions of L2 speakers' performance (communication anxiety collaborativeness, engagement, affective response) as they evolve in interaction in real time.

Note

1. Approximately 19% of the data occurred at the maximum value for comprehensibility (100), suggesting that the dataset was somewhat inflated at the highest range. A zero/one beta regression model was fit to approximate this distribution using the glmmTMB package to account for inflation at either extreme by estimating separate effects for $0 <$ values < 1 and for 1 versus other values. Regression findings confirmed results for the linear model, namely, a significant quadratic trend for Tasks 1–2 (*estimate* = 3.07, *SE* = .95, $z = 3.24$, $p = .001$) and a significant linear trend for Task 3 (*estimate* = .17, *SE* = .06, $z = 2.72$, $p = .007$), except that in this model residuals were normally distributed.

Funding

This study was supported by grants from the Social Sciences and Humanities Research Council of Canada (SSHRC) to the first, third, and fourth authors.

Acknowledgements and Open Materials and Data Statement

We are grateful to Clinton Hendry for help with data collection, Aki Tsunemoto for help with data analyses, and to the anonymous reviewers and the journal editor for the insightful comments and suggestions that helped us refine this article. All materials and data from this study are publicly accessible through the Open Science Framework at https://osf.io/bpzgm and the IRIS digital repository at http://www.iris-database.org.

References

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

Bergeron, A., & Trofimovich, P. (2017). Linguistic dimensions of accentenedness and comprehensibility: Exploring task and listener effects in second language French. *Foreign Language Annals*, *50*, 547–566. https://doi.org/10.1111/flan.12285

Berry, G. M., & Ernestus, M. (2018). Phonetic alignment in English as a lingua franca: Coming together while splitting apart. *Second Language Research*, *34*, 343–370. https://doi.org/10.1177/0267658317737348

Brennan, S. E., Kuhlen, A. K., & Charoy, J. (2018). Discourse and dialogue. In S. L. Thompson-Schill (Ed.), *The Stevens' handbook of experimental psychology and cognitive neuroscience* (pp. 145–209). Wiley.

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

Cobb, T. (2019). VocabProfilers [computer program]. https://www.lextutor.ca/vp

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

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

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2015a). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly*, *49*, 814–837. https://doi.org/10.1002/tesq.203

de Bot, K., Lowie, W., & Verspoor, M. (2007). A Dynamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition*, *10*, 7–21. https://doi.org/10.1017/s1366728906002732

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*, 1–16. https://doi.org/10.1017/S0272263197001010

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

Dragojevic, M., Giles, H., Beck, A.-C., & Tatum, N. T. (2017). The fluency principle: Why foreign accent strength negatively biases language attitudes. *Communication Monographs*, *84*, 385–405. https://doi.org/10.1080/03637751.2017.1322213

ETS (2017). *TOEFL iBT® and IELTS® academic module scores: Score comparison tool*. http://www.ets.org/toefl/institutions/scores/compare

Flege, J., & Fletcher, K. (1992). Talker and listener effects on the perception of degree of foreign accent. *Journal of the Acoustical Society of America*, *91*, 370–389. https://doi.org/10.1121/1.402780

Galindo Ochoa, J. A. (2017). *The effect of task repetition on Colombian EFL students' accuracy*

and fluency (Unpublished master's thesis). Concordia University, Montreal.

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.

Gibson, W., & Brown, A. (2009). *Working with qualitative data*. Sage.

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.

Graf, L. K. M., Mayer, S., & Landwehr, J. R. (2018). Measuring processing fluency: One versus five items. *Journal of Consumer Psychology*, *28*, 393–411. https://doi.org/10.1002/jcpy.1021

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

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

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition, and Neuroscience*, *31*, 32–59. https://doi.org/10.1080/23273798.2015.1102299

Matsuura, H., Chiba, R., & Fujieda, M. (1999). Intelligibility and comprehensibility of American and Irish Englishes in Japan. *World Englishes*, *18*, 49–62. https://doi.org/10.1111/1467-971X.00121

Munro, M. J. (2018). Dimensions of pronunciation. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 413–431). Routledge.

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

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

Newman, E. J., Sanson, M., Miller, E. K., Quigley-McBride, A., Foster J. L, Bernstein, D. M., & Garry, M. (2014). People with easier to pronounce names promote truthiness of claims. *PLoS ONE*, *9*(2), e88671. https://doi.org/10.1371/journal.pone.0088671

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

Pakhomov, S. V., Kaiser, E. A., Boley, D. L., Marino, S. E., Knopman, D. S., & Birnbaum, A. K. (2011). Effects of age and dementia on temporal cycles in spontaneous speech fluency. *Journal of Neurolinguistics*, *24*, 619–635. https://doi.org/10.1016/j.jneuroling.2011.06.002

R Core Team (2019). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org

Reber, R., & Greifeneder, R. (2017) Processing fluency in education: How metacognitive

feelings shape learning, belief formation, and affect. *Educational Psychologist*, *52*, 84–103. https://doi.org/10.1080/00461520.2016.1258173

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching*, *43*, 1-32. https://doi.org/10.1515/iral.2005.43.1.1

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., Trofimovich, P., & Isaacs, T. (2017). Using listener judgements 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

Sanchez, C. A., & Jaeger, A. J. (2015). If it's hard to read, it changes how long you do it: Reading time as an explanation for perceptual fluency effects on judgment. *Psychonomic Bulletin and Review*, *22*, 206–211. https://doi.org/10.3758/s13423-014-0658-6

Sanchez, C. A., & Khan, S. (2016). Instructor accents in online education and their effect on learning and attitudes. *Journal of Computer Assisted Learning*, *32*, 494–502. https://doi.org/10.1111/jcal.12149

Sheppard, B. E., Elliott, N. C., & Baese-Berk, M. M. (2017). Comprehensibility and intelligibility of international student speech: Comparing perceptions of university EAP instructors and content faculty. *Journal of English for Academic Purposes*, *26*, 42–51. https://doi.org/10.1016/j.jeap.2017.01.006

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis*. Oxford University

Press.

Song, H., & Schwarz, N. (2008). If it's hard to read, it's hard to do: Processing fluency affects effort prediction and motivation. *Psychological Science*, *19*, 986–988. https://doi.org/10.1111/j.1467-9280.2008.02189.x

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

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

Appendix A Background information for speaker pairs

	Speaker A			Speaker B		
Pair	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

Address for correspondence

Pavel Trofimovich

Concordia University (S-FG 5.150)

1455 De Maisonneuve Blvd. W.

Montreal, QC, Canada H3G 1M8

Pavel.Trofimovich@concordia.ca

https://orcid.org/0000-0001-6696-2411

Co-author information

Charles L. Nagle

Iowa State University

Department of World Languages and Cultures

505 Morrill Road, 3102G Pearson Hall

Ames, IA, United States 50011

cnagle@iastate.edu

Mary Grantham O'Brien

University of Calgary

School of Languages, Linguistics, Literatures and Cultures

C216 Craigie Hall, 2500 University Drive NW

Calgary, AB, Canada T2N 1N4

mgobrien@ucalgary.ca

https://orcid.org/0000-0001-5873-4013

Sara Kennedy

Concordia University (FG 5.150)

1455 De Maisonneuve Blvd. W.

Montreal, QC, Canada H3G 1M8

Sara.Kennedy@concordia.ca

https://orcid.org/0000-0001-5484-8048

Kym Taylor Reid

Concordia University (FG 5.150)

1455 De Maisonneuve Blvd. W.

Montreal, QC, Canada H3G 1M8

kym.taylor@concordia.ca

https://orcid.org/0000-0003-3915-3576

Lauren Strachan

Concordia University (FG 5.150)

1455 De Maisonneuve Blvd. W.

Montreal, QC, Canada H3G 1M8

lauren.strachan@concordia.ca

https://orcid.org/0000-0002-5030-7866