Applied Psycholinguistics (2023), **44**, 131–156
doi:10.1017/S0142716423000073

CAMBRIDGE UNIVERSITY PRESS

ORIGINAL ARTICLE

Framing second language comprehensibility: Do interlocutors' ratings predict their perceived communicative experience?

Charlie Nagle[1,*], Pavel Trofimovich[2], Oguzhan Tekin[2] and Kim McDonough[2]

[1]The University of Texas at Austin, Austin, USA and [2]Concordia University, Montreal, Canada
*Corresponding author. Email: cnagle@austin.utexas.edu

(Received 12 May 2022; revised 29 September 2022; accepted 13 December 2022)

Abstract

Comprehensibility has risen to the forefront of second language (L2) speech research. To date, research has focused on identifying the linguistic, behavioral, and affective correlates of comprehensibility, how it develops over time, and how it evolves over the course of an interaction. In all these approaches, comprehensibility is the dependent measure, but comprehensibility can also be construed as a predictor of other communicative outcomes. In this study, we examined the extent to which comprehensibility predicted interlocutors' overall impression of their interaction. We analyzed data from 90 paired interactions encompassing three communicative tasks. Interactive partners were L2 English speakers who did not share the same native language. After each task, they provided self- and partner-ratings of comprehensibility, collaboration, and anxiety, and at the end of the interaction, they provided exit ratings of their overall experience in the interaction, communication success, and comfort interacting with their partner. We fit mixed-effects models to the self- and partner-ratings to investigate if those ratings changed over time, and we used the results to derive model-estimated predictors to be incorporated into regression models of the exit ratings. Only the self-ratings, including self-comprehensibility, were significantly associated with the exit ratings, suggesting a speaker-centric view of L2 interaction.

Keywords: adult second language acquisition; language production; spoken language comprehension; interaction

Comprehensibility refers to a listener's subjective experience of difficulty when processing second language (L2) speech. Although comprehensibility is often aligned with intelligibility, or actual understanding, the two are distinct, in that intelligible speech often shows varying levels of comprehensibility depending on the amount of effort that the listener must invest (Munro & Derwing, 1995; Nagle & Huensch, 2020). Pronunciation experts have long advocated for an intelligibility- and comprehensibility-focused approach to L2 learning and teaching

(Levis, 2020) given that the presence of an L2 accent does not automatically trigger intelligibility or comprehensibility issues (Huensch & Nagle, 2021; Munro & Derwing, 1995, 2020). Comprehensibility also holds intuitive appeal because it can be captured using scalar ratings, which can be adapted to many research and teaching contexts. Therefore, it comes as no surprise that comprehensibility has come to dominate the L2 speech research landscape (for review, see Trofimovich et al., 2022). However, aside from rare exceptions, comprehensibility has to date been investigated as the target and end goal of research, where researchers explore various speaker and listener influences on comprehensibility. Our goal in this study was instead to frame comprehensibility in the context of L2 interaction as an explanatory predictor of interlocutors' perceptions of their communicative experience. To do so, we analyzed fluctuations in self- and partner-comprehensibility ratings over three interactive tasks and included both measures as predictors of interlocutors' summative evaluations of their interactive experience. Although our primary focus in this study was comprehensibility, we also assessed self- and partner-ratings of anxiety and collaboration to provide a more complete picture of how linguistic, social, and affective variables affect interlocutors' perception of interaction.

Background literature

Comprehensibility: The (Brief) story so far

Comprehensibility is one of the most studied constructs in L2 speech research. Practically speaking, it offers a useful metric of listener understanding (Sheppard et al., 2017). Unlike the labor-intensive, content-specific tasks employed to assess intelligibility (e.g., orthographic transcriptions, comprehension questions), scalar comprehensibility ratings (e.g., on a 1–9 scale, where 1 = *hard to understand* and 9 = *easy to understand*) are easy to elicit and analyze in many teaching, assessment, and workplace settings through prerecorded materials or live speaker performances. Whereas intelligibility measures fluctuate depending on whether they focus on listeners' understanding of individual words versus their comprehension of discourse-level content (Kang et al., 2018; Kennedy, 2021), comprehensibility ratings remain consistent across listeners (Nagle, 2019). Lastly, although intelligibility is a sensible end goal for L2 teaching and learning (Levis, 2020), most speakers want to develop speech that causes little difficulty for the listener. Comprehensible speech can indeed be taught and learned, as shown through research conducted in naturalistic and instructed contexts (Lee et al., 2015; Saito, 2021).

As shown in a recent meta-analysis (Saito, 2021), comprehensible L2 speech is associated with multiple linguistic dimensions, such as a speaker's segmental production (e.g., accuracy of individual vowels and consonants), prosody (e.g., word stress placement, intonation accuracy), and temporal fluency (e.g., speech rate, pausing), in addition to various properties of lexis, grammar, and discourse, including vocabulary richness, grammar complexity, and discourse organization. This finding is robust, insofar as it is attested across various L2s (Bergeron & Trofimovich, 2017; O'Brien, 2014; Saito et al., 2016), but the relative importance of each dimension is dictated by the demands of the speaking task (Crowther et al., 2018). The bottom line for language teachers and learners is that comprehensibility is not only about pronunciation but can be achieved through an instructional focus on other aspects of

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

language, such as grammar and vocabulary. Comprehensibility also depends on who is evaluating speech, in that ratings vary as a function of listeners' linguistic training, language teaching and learning experience, familiarity with the target language, and status as monolingual versus multilingual speakers (Isaacs & Thomson, 2013; Saito et al., 2017, 2019; Saito & Shintani, 2016). However, several minor differences aside, recent meta-analytic evidence (Saito, 2021) has shown that novice listeners and L2 speakers generally assign comprehensibility ratings that are similar to those given by experts and native speakers, which should be welcome news for language researchers and teachers.

More recently, comprehensibility has been studied as a dynamic construct that fluctuates as listeners process speech over time. For instance, in Nagle et al. (2019), 24 listeners rated three extended L2 Spanish speech samples, using idiodynamic software to upgrade or downgrade comprehensibility in real time. They also participated in a stimulated interview to explore the rationale behind their decisions. Results demonstrated that listeners were able to overcome lapses in speakers' language use as long as ideas were presented in a logical manner, but even minor language issues, such as grammar errors, triggered major shifts in comprehensibility if the errors interfered with listeners' understanding of the discourse. Researchers have also begun to acknowledge multivariate influences on comprehensibility in interactive contexts, where listeners (as interlocutors) might draw on various verbal and nonverbal cues to continuously shape what they consider to be comprehensible. In Trofimovich et al. (2020), L2 speakers assessed their own and each other's comprehensibility seven times over a 17-minute interaction encompassing three tasks. Although the ratings varied by task, the interlocutors' ratings of one another's comprehensibility became more similar over the course of the interaction, suggesting convergence in the ease with which partners understood one another as the conversation progressed.

In a study examining the same participants as Trofimovich et al. (2020), Nagle et al. (2022) explored the extent to which speakers' ratings of each other's comprehensibility were related to their own and their partner's perceived anxiety and collaboration, on the assumption that comprehensibility might reflect interpersonal dimensions of interaction, such as affect (nervousness experienced during conversation) and behavior (engagement with an interlocutor). Comprehensibility was predicted by how anxious speakers were judged to be by their interlocutors, with greater comprehensibility associated with less perceived anxiety, and by how collaborative they seemed to their interlocutors, such that more comprehensible speech was linked to greater perceived collaboration. Moreover, speakers' self-rated collaboration also positively predicted how comprehensible they sounded to their interlocutor, suggesting that comprehensibility also reflected speakers' own perception of how much or how little they contributed to conversation. Comprehensibility in interaction thus seems to be multidimensional, co-constructed, and dynamic, subject to various influences stemming from both interlocutors.

Comprehensibility: A useful explanatory variable?

In most research to date, comprehensibility has been treated as an outcome variable, with researchers examining various speaker and listener factors that predict it,

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

investigating its development in various contexts, and exploring moment-to-moment and short-term fluctuations in ratings as listeners process speech over time or interact with one another. However, the value of comprehensibility as a construct might also rest on its own predictive strength. Put differently, comprehensibility might help account for aspects of human decision-making or behavior, such as whether interlocutors continue interacting with speakers they find difficult to understand or whether university students drop out of courses delivered by instructors whose speech they consider hard to process. When conceptualized in this way, comprehensibility as an explanatory variable has an intuitive appeal with relevance beyond L2 speech research.

There are only a handful of studies that have centered on comprehensibility as an explanatory variable, exploring the potential consequences of speech that varies in degree of listener-assessed processing difficulty. In an early study, Varonis and Gass (1982) examined how native-speaking English interlocutors responded to a simple request from native and L2 speakers for directions to a well-known location. They observed that the interlocutors approached by L2 speakers tended to repeat the request (often with a rising intonation) and showed reluctance to get involved in a conversation, often accompanied by a pause, a sigh, or a filler like *oh geez*. According to Varonis and Gass, these behaviors arose as a direct reaction to the speaker's comprehensibility. Even though the interlocutors fully understood the speaker's message, their experience was effortful, and the likelihood of a future nonunderstanding was real, so they expressed uncertainty and stalled for time.

Another set of studies where comprehensibility has served as an explanatory variable comes from social-psychological research on processing fluency. Briefly, processing fluency is a metacognitive construct capturing people's perception of the ease or difficulty with which they process information (Schwarz, 2018). A key tenet of processing fluency is that people can readily access subjective perceptions of ease or difficulty when they engage in various tasks, for instance, reading texts, viewing images, hearing speech, or solving math problems. These judgments of ease or difficulty (hence, processing fluency or dysfluency) tend to predict people's reactions arising from their experience in a given task, such that, for example, struggling to decipher the content of a faint photocopy of a job applicant's CV (i.e., experiencing processing difficulty) might be associated with a negative disposition toward the candidate's job suitability, regardless of the actual CV content (Graf et al., 2018). In one processing fluency study, for instance, Sanchez and Khan (2016) asked students to assess e-learning materials narrated by instructors who were either easy or hard to understand. The students exposed to the less comprehensible speaker downgraded this speaker in their evaluations, expressed less favorable attitudes about coursework, and evaluated the materials as more difficult, even though there was no evidence that their actual understanding of the materials suffered. Similarly, in Dragojevic et al. (2017), when listeners evaluated Punjabi- and Mandarin-accented speakers whom they found difficult to understand, they downgraded these speakers in ratings of competence, intelligence, and success, and also attributed more feelings of annoyance and irritation to them, relative to speakers who were easier to understand.

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

The current study

In sum, comprehensibility has been treated largely as an outcome variable, where researchers examine different factors that explain comprehensibility ratings and their time-sensitive properties. Furthermore, comprehensibility is typically framed in terms of other evaluations, where, for example, an external rater or judge assigns a comprehensibility score to speech provided by a speaker with whom the rater does not interact. Apart from processing fluency research, whose goal is largely to explain listeners' affective and attitudinal reactions to speakers as a function of listening effort, comprehensibility has rarely been framed as a predictor of L2 speakers' perceptions about each other and their communicative experience (Varonis & Gass, 1982). Considering that the communicative and social implications of comprehensibility remain under-researched, in this study, we therefore extended work on social and affective dimensions of comprehensibility (Nagle et al., 2022) to examine whether L2 speakers' comprehensibility predicts appraisals of their overall interactive experience.

We sampled the data for this study from the CELFI corpus (McDonough & Trofimovich, 2019). For this corpus, university-level L2 speakers were audio- and video-recorded communicating with an interlocutor from a different language background in three interactive tasks. Their eye gaze was tracked, and their skin conductance was monitored. They also completed a battery of questionnaires (anxiety, motivation, social networks, and acculturation), a working memory task, rating scales after each task (motivation, anxiety, flow, comprehensibility, collaboration), a stimulated recall session about the final task, and a debriefing interview eliciting explanations for their task ratings. Because our data were drawn from an existing corpus, our choice of the outcome variables to be explored in relation to comprehensibility was inevitably constrained. Nonetheless, all speakers in the corpus provided comprehensibility ratings and additionally completed several exit scales with summative evaluations of their conversation, which enabled us to conduct an exploratory analysis of comprehensibility as a predictor of perceived communicative experience.

Underlying this analysis is the idea that increased comprehensibility brings with it communicative and social advantages which can be captured through interlocutors' subjective satisfaction metrics. For example, beyond being able to understand the interlocutor more readily, speakers may perceive their overall experience more favorably, judge their interaction as more successful, and feel more at ease during interaction if they judge their interlocutor as more comprehensible. Based on these considerations, we therefore selected three exit measures provided by the speakers in the corpus as our outcome variables: the speakers' evaluation of their overall experience, their appraisal of communicative success, and their perceived comfort interacting with the partner. Broadly speaking, these metrics captured various facets of speakers' perceived interactional outcomes, such as their overall satisfaction, attainment of communicative goals, and interpersonal rapport. Because recent work has revealed contributions of affective (anxiety) and behavioral (collaboration) dimensions of interaction to interlocutor-perceived comprehensibility (Nagle et al., 2022), we also selected the speakers' ratings of perceived anxiety and collaboration for inclusion in our analyses. The measures of perceived anxiety and collaboration were

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

considered a posteriori, in the sense that their inclusion in the study was determined by their availability in the corpus, not through hypothesizing prior to data collection. Therefore, in this manuscript, we purposefully avoid an extensive conceptual discussion of these factors in relation to the study goals (but see Nagle et al., 2022, for further insights), to reflect the research process as it unfolded. Finally, in light of Gluszek and Dovidio's (2010) suggestion that interlocutors' self-assessments are as important as their assessments of the interlocutor in determining perceived conversational outcomes, we included both self- and partner-perceptions of comprehensibility, anxiety, and collaboration (all available in the corpus).

Because this study relied on existing corpus data and because the conceptual framing of this work was retrospective, we did not generate a priori hypotheses, apart from having a research-informed exploration-driven expectation that speakers' comprehensibility, as a subjective index of the effort they invest in their interactive experience, might have a bearing on their appraisal of the interaction (Dragojevic et al., 2017; Sanchez & Khan, 2016), insofar as greater self- and partner-assessed comprehensibility could be associated with higher exit ratings. Following from prior work (Nagle et al., 2022), we also anticipated that perceived anxiety and collaboration might factor into speakers' outcome ratings, such that less anxiety and more collaboration might be associated with higher exit ratings. This exploratory study, which to our knowledge is the first to examine how L2 speakers' mutual evaluations throughout conversation affect their appraisal of the interaction, was guided by the following question: Do L2 speakers' self- and partner-assessments of comprehensibility, anxiety, and collaboration predict their overall perceived communication experience, communicative success, and comfort interacting with the partner?

Method

The materials, data, and code used for the present study are available at https://osf.io/tzye2/.

Paired interactions

The data set included 90 paired interactions sampled from the Corpus of English as a Lingua Franca Interaction (McDonough & Trofimovich, 2019) which comprises audio recordings of 224 pairs of L2 English speakers enrolled in Canadian English-medium universities. In a 30-minute session, the speakers carried out three, 10-minute tasks: moving to Montréal, close call story, and academic discussion. For the moving to Montréal task, speakers discussed their personal challenges upon arrival in Montréal and brainstormed possible solutions to these challenges. In the close call story task, speakers took turns narrating individual stories about a narrow escape from trouble or danger. For the academic discussion task, speakers first selected one of four topics they wished to discuss (medical ethics, nature vs. nurture, pros and cons of advertising, or motivation for language learning) and then read different short research reports on that topic. After about 5 minutes of reading the individual reports, which included approximately 2 minutes allotted for speakers to prepare their report summary, they compared information and exchanged opinions.

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

The paired interactions were carefully selected, first, to ensure that speakers performed the tasks in different orders to minimize the likelihood that their assessments in each task would be specific to a given task order. Because the first 150 conversation partners in the corpus had completed the tasks in a fixed order (i.e., abc, where a, b, and c correspond to moving to Montréal, close call story, and academic discussion, respectively) before alternate task orders were implemented in equal proportions, the sampled interactions included all available data featuring these alternate orders (acb = 15 pairs, bac = 15 pairs, bca = 16 pairs, cab = 15 pairs, cba = 14 pairs), plus an equivalent random selection of paired performances with the most frequent order (abc = 15 pairs), for a total of 90 paired interactions. The selected interactions were roughly balanced in gender composition (38 female–male, 28 female–female, 23 male–male, 1 male–other). The sample also included different discussion topics selected by speakers for the academic discussion (medical ethics = 14 pairs, nature vs. nurture = 14 pairs, advertising = 30 pairs, motivation = 32 pairs), ensuring that student communication was not specific to a particular topic. Finally, across all sampled interactions, task performances were also matched in duration, with moving to Montréal lasting on average 10.95 minutes (*SD* = 0.80, *range* = 8.48–12.68), close call story lasting 10.86 minutes (*SD* = 0.82, *range* = 8.33–13.93), and academic discussion lasting 11.23 minutes (*SD* = 1.32, *range* = 6.78–14.77) of the total speaking time (i.e., excluding the time spent reading the report and preparing its summary).

The pairs were composed of 180 L2 English speakers with a mean age of 23.50 years (*SD* = 3.97, *range* = 18–45) who were pursuing various undergraduate (92) and graduate (88) degrees. There was generally a balanced breakdown of pair composition by student status, where 25 pairs were composed of undergraduate students, 24 pairs consisted of graduate students, and 39 pairs included graduate–undergraduate pairings (with two students in two separate dyads failing to provide their degree status). As part of university admission requirements, the speakers took standardized language tests and reported IELTS (85) or TOEFL (38) scores, whereas the remaining speakers provided no test score. To enable comparisons, the TOEFL scores were first converted to equivalent IELTS bands through established conversion metrics (ETS, 2017), and the resulting total IELTS scores (where each represents a speaker's performance in listening, reading, writing, and speaking) ranged between 5.5 (modest user) and 9.0 (expert user), with the median score of 7.0 corresponding to good user (*M* = 7.17, *SD* = 0.67). The speakers were assigned to pairs without consideration of their proficiency; however, interlocutor differences in proficiency were small in terms of absolute IELTS scores (M_{diff} = 0.70, *SD* = .59), such that the two interacting partners' IELTS scores generally fell within one band value (Mdn_{diff} = 0.50). The speakers had formally studied English for a mean of 13.25 years (*SD* = 5.98, *range* = 1–30) and resided in Montréal for about 2.19 years (*SD* = 3.13, *range* = 2 weeks–20 years). They came from 35 language backgrounds, the largest being Mandarin (28), French (23), Arabic (22), Hindi (17), Farsi (13), and Tamil (12). Using a 100-point scale (0 = *never*, 100 = *all the time*), they indicated that they regularly spoke English at university (*M* = 85.82, *SD* = 19.30) and rated themselves as fairly high in English speaking (*M* = 7.19, *SD* = 1.23) on a 9-point scale (1 = *poor*, 9 = *excellent*). The two interacting

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

partners were randomly assigned to dyads with the constraint that they had only English as a common language.

Data collection occurred in a university lab, where the speakers were seated at a table across from each other, with all task instructions provided by a research assistant (RA). After the speakers signed the consent form, the RA explained the target rating scales and provided the definitions of the terms (see below). Next, the RA introduced each of the three tasks, first by giving task instructions orally and then providing the same information on a handout. To minimize any observer-related effects, the RA left the room, giving speakers approximately 10 minutes to complete each task. After each task, the two speakers were given the rating scales and evaluated themselves and each other. At the end of the session (i.e., after completing all three tasks), the speakers provided several exit ratings and completed background questionnaires.

Target ratings

Immediately after each task, speakers rated themselves and their partner in terms of linguistic, socio-affective, and behavioral dimensions. The linguistic dimension included ratings of comprehensibility (ease of understanding). The socio-affective dimension involved perceptions of speaker anxiety (degree of stress, worry, or nervousness). The behavioral dimension was defined through perceived collaboration (working with someone to produce or create something). Speakers also provided ratings of speech flow and motivation, but these data are not analyzed here. All ratings were operationalized through continuous scales, which were 100-milimeter lines printed on paper with no markings aside from endpoints. The left endpoint was equivalent to the rating of 0 while the right endpoint corresponded to the rating of 100, and there were two scales per rated construct, with the first targeting each student's own performance (labeled "me") and the other targeting their conversation partner (labeled "my partner"): comprehensibility (*difficult to understand–easy to understand*), anxiety (*high level of anxiety–low level of anxiety*), and collaboration (*I did not work well with my partner/my partner did not work well with me–I worked well with my partner/my partner worked well with me*). Before starting each task, speakers were given definitions for each construct (see Appendix A), and they could clarify any remaining questions before carrying out the tasks. The same definitions were available alongside the scales so speakers could consult them as they rated themselves and their partner.

After completing all three communicative tasks, speakers provided several individual exit ratings capturing various facets of their interactive experience. Of key importance are three exit measures, all rated through similar 100-millimeter scales with the left endpoint equivalent to the rating of 0 and the right endpoint equivalent to the rating of 100: (a) appraisal of the overall communicative experience (*my experience during this session was very negative–my experience during this session was very positive*), (b) evaluation of communication success in terms of how productively the interaction unfolded (*my conversation partner and I were not successful at communicating–my conversation partner and I were very successful at communicating*), and (c) perception of the ease or comfort interacting with the conversation partner (*I felt very uncomfortable interacting with my conversation partner–I felt very comfortable interacting with my conversation partner*). For both self- and

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

partner-evaluations following each task and the exit measures at the end of the session, speakers indicated their rating by putting a cross on each line corresponding to their evaluation.

Data analysis

Speakers' self- and partner evaluations following each task as well as their exit ratings were converted to numerical values by measuring the distance in millimeters between the left scalar endpoint and the speaker's mark on each scale (out of 100 points). Because interlocutors' perceptions might be influenced by the range of content produced by each speaker, such as when one interlocutor engages in repetitive talk relying on the same limited word set whereas another contributes diverse lexical content to the discussion, we included the number of content word types as a control covariate. To derive this measure, we computed the total number of distinct content words (i.e., unique nouns, adjectives, verbs, and adverbs) from each speaker's total lexical output in each task. In our own prior work, the number of word types (a basic measure of lexical diversity) was significantly related to interlocutors' comprehensibility ratings (Nagle et al., 2022; Trofimovich et al., 2020), so it was important to control variation in the range of content produced by each speaker. The relationship between speakers' self- and partner-ratings of comprehensibility, anxiety, and collaboration and the three outcome measures was examined (as described below) in R version 4.2.1 (R Core Team, 2022). The data set and R markdown used for analysis are publicly available at https://osf.io/tzye2/.

Results

Descriptive analyses

We first computed descriptive statistics and correlations for the three outcome ratings. At the end of their interactive session, speakers provided high evaluations of their overall perceived experience ($M = 91.53$, $SD = 8.92$, *range* = 50–100), communicative success ($M = 87.46$, $SD = 16.09$, *range* = 4–100), and comfort interacting with the partner ($M = 88.53$, $SD = 13.71$, *range* = 4–100). However, despite feeling generally positive, they expressed a range of judgments from relatively low to high. According to field-specific benchmarks (Plonsky & Oswald, 2014; $r = .25$, small; $r = .40$, medium; $r = .60$, large), overall experience showed a medium association with communicative success ($r = .49, p < .001$) and with comfort ($r = .47, p < .001$), whereas communicative success and comfort had a weaker association ($r = .35$, $p < .001$). With 12–24% variance in common, the three outcome measures therefore seemed to capture sufficiently distinct evaluative dimensions of L2 interaction.

We then computed descriptive statistics for the self- and partner-ratings of comprehensibility, anxiety, and collaboration at the end of each task. As shown in Table 1, the ratings were overall high, indicating that speakers perceived themselves and their partners to be comprehensible and collaborative. The high ratings for anxiety, where higher scores indicate lower anxiety, further demonstrate that speakers were not very anxious throughout the interaction. The self-ratings tended to decrease slightly over time, whereas the partner-ratings were mostly stable, save

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

Table 1. Means and standard deviations for self- and partner-ratings over time

Rating	First task	Second task	Third task
Self-comprehensibility	82.85 (14.27)	80.48 (16.39)	79.39 (17.24)
Self-anxiety	81.33 (20.87)	79.43 (21.91)	77.23 (24.13)
Self-collaboration	88.57 (11.21)	87.32 (15.01)	86.96 (13.75)
Partner comprehensibility	81.88 (16.84)	80.78 (17.69)	77.52 (19.71)
Partner anxiety	81.83 (20.59)	80.83 (21.67)	81.16 (20.42)
Partner-collaboration	88.00 (12.47)	89.02 (12.28)	87.62 (14.06)

Table 2. Pearson correlations between predictor variables (Mean values across all tasks)

Rating	1	2	3	4	5
1 Self-comprehensibility					
2 Self-anxiety	.36*				
3 Self-collaboration	.49*	.39*			
4 Partner-comprehensibility	.13	–.06	.03		
5 Partner-anxiety	.01	.01	–.02	.31*	
6 Partner-collaboration	–.02	–.01	.01	.58*	.50*

Note. *$p < .01$ (two-tailed).

partner-comprehensibility, which like the self-ratings decreased slightly as speakers moved through the three communicative tasks.[1]

In terms of potential relationships between the predictors, to obtain a general sense of intercorrelations, we computed Pearson coefficients between self- and partner-ratings pooled over the three tasks (see Appendices B and C for time- and task-specific correlations). As shown in Table 2, self- and partner-ratings of comprehensibility, anxiety, and collaboration showed little association ($r = –.06–.13$), suggesting that there was little agreement between the speakers in how they viewed themselves and each other across six different task orders and three tasks.[2] Within each set of self- versus partner-ratings, however, the three predictors showed weak-to-medium relationships, where comprehensibility was more strongly associated with collaboration ($r = .49–.58$) than with anxiety ($r = .31–.36$), which provides general support to Nagle et al.'s (2022) conclusion that comprehensibility in interaction is differentially tied to perceived collaboration and anxiety.

Change over time in self- and partner-ratings

Analysis plan

Our primary interest was to investigate the relationship between the self- and partner-ratings of comprehensibility, anxiety, and collaboration and speakers' perception of their overall experience, communicative success, and comfort interacting

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

with their partner, which they evaluated at the end of the interaction. We were specifically interested in a time-sensitive analysis, considering each predictor averaged over the three tasks as well as change in each predictor. As a first step, following Gries (2021), we inspected the distribution of self- and partner-ratings. All variables showed a limited range, with most ratings concentrated at the upper end of the scale, as might be expected for advanced L2 speakers enrolled in an L2-medium university. To improve the normality of the predictors and, in doing so, prevent problems when fitting and evaluating the statistical models, we applied a Box-Cox transformation to each variable (Gries, 2021).

After inspecting the distribution of the data and applying the Box–Cox transformation to make distributions more normal, we fit a series of mixed-effects models to explore the extent to which the self- and partner-predictors showed significant change over time. The goal of this step was (a) to identify the model (intercept or growth) that was the best fit to each of the measures and (b) to use that model to generate a new set of model-estimated, person-specific values to be included as predictors of the exit ratings. To examine potential changes in self- and partner-ratings over time, we fit a null (intercept-only) model and a linear growth model to each construct. We compared the models using a likelihood ratio test, and if the linear growth model was a significantly better representation of the data (i.e., if the chi-square test was significant), we tested by-speaker random slopes for time to determine if there was significant between-speaker variability in linear rate of change. All models contained by-speaker random intercepts and the word type control covariate. If the intercept-only model provided the best fit to the data, we extracted the by-speaker random intercepts, which represent each speaker's overall estimate on the predictor. If the linear growth model with by-speaker random slopes for time provided the best fit, we extracted by-speaker random intercepts and slopes to capture individual variation in both parameters. In that case, the intercepts continue to represent the speaker's overall estimate for each predictor, and the slopes represent the estimated linear change trajectory for each speaker. We reasoned that both types of variables (overall average and change trajectory) could be related to speakers' overall impression of the interaction. In this study, we therefore took a data-driven approach, wherein we first modeled the structure of the ratings over time and subsequently used that model-based information to generate the speaker-level estimates that would serve as predictors for our ultimate analytic goal of investigating how speakers' perception of their own and their partner's comprehensibility, anxiety, and collaboration affected their exit ratings. Using model-estimated variables is preferable to using simple averages because the model-estimated scores are computed in relation to other significant effects while accounting for the nested data structure.

Modeling change over time

The results of the first step, modeling change in self- and partner-ratings over the interaction, are displayed in Table 3. The best fit to the self- and partner-ratings of comprehensibility was a linear growth model with by-speaker random intercepts (a unique intercept for each speaker). When we tested by-speaker random slopes for time (a unique rate of change for each speaker), models were singular, suggesting

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

Table 3. Summary of models fit and fit statistics for null and linear growth models

	Self				Partner			
Model	Deviance	χ^2	df	*p*	Deviance	χ^2	df	*p*
Comprehensibility								
Intercept	13937				13938			
Linear growth	13932	5.12	1	.024	13929	8.61	1	.003
Linear growth RS	Singular				Singular			
Collaboration								
Intercept	13617				13535			
Linear growth	13616	0.90	1	.343	13535	0.30	1	.584
Anxiety								
Intercept	14199				14018			
Linear growth	14196	3.16	1	.075	14018	0.01	1	.917
Linear growth RS	Singular							

Note. If the linear growth model was significant, a second model with by-speaker random slopes (RS) for time was fit to examine if speakers displayed different rates of change.

overfit. For self- and partner-ratings of anxiety and collaboration, the intercept model provided the best fit to the data, which demonstrates that there was no statistically significant group-level change in the anxiety and collaboration ratings. Thus, the self- and partner-ratings of comprehensibility changed significantly over time (although modeling did not support estimating individual change trajectories for each speaker), whereas the self- and partner-ratings of anxiety and collaboration did not.

Relationship between self- and partner-ratings and outcome measures

Analysis plan

Based on the findings of the initial change analysis, we generated six new variables representing model estimated by speaker intercepts for each construct to be included as predictors of the exit ratings: self-comprehensibility intercepts, partner-comprehensibility intercepts, self-collaboration intercepts, partner-collaboration intercepts, self-anxiety intercepts, and partner-anxiety intercepts. In all cases, these new variables represented the relevant global model-estimated rating for each speaker. We then included these six model-estimated variables as predictors of speakers' ratings of their overall experience, communicative success, and comfort interacting with their partner, which they evaluated at the end of the interaction. The goal of this step was to examine if speakers' self- and partner-ratings, which we assessed repeatedly throughout the interaction and modeled to derive a model-estimated mean score for each individual, were significantly related to their exit ratings, which reflected their global appraisal of their interactive experience.

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

Because speakers provided only one exit rating for each construct, we used multiple regression rather than a mixed-effects model. First, we fit a maximal model, including all effects of interest (all model-estimated self- and partner-intercepts). We then backward-tested fixed effects, removing predictors one by one until we reached the minimally adequate model. After establishing that model, we followed Field et al.'s (2012) recommendations to check model assumptions. We inspected model residuals for large values, including influential datapoints, which we determined using Cook's distance and leverage values. We also checked residuals for normality and linearity. Lastly, we used the Durbin–Watson test to check the assumption of independence and checked for multicollinearity by computing variance inflation factors. All predictors were standardized.

Predicting overall experience

After backward-testing fixed effects, we arrived at a final model for overall experience with self-comprehensibility and self-collaboration; removing the other predictors, including all partner-ratings, did not significantly alter model fit. When we checked the overall experience model, 7 of 180 observations (3.89% of the data) had large residuals (over |2|), which is below the 5% threshold considered as a cutoff. None of the cases with large residuals appeared to be influential based on Cook's distance (<1) and leverage scores (<.04 for this dataset). The Durbin–Watson test was significant ($p = .026$), which would suggest that the assumption of independence was violated. However, the statistic was close to 2 (2.26), which provides confidence in the model. Plotting model residuals revealed some issues with linearity, so we used the sandwich package (Zeileis, 2004; Zeileis et al., 2020) to compute robust standard errors (SEs) and compare them to the original model. For self-comprehensibility, the robust SE was 0.74, which is comparable to the 0.68 estimated in the original model. For self-collaboration, the robust and original SEs were the same (0.68). We adopted the robust SEs and used them to recompute 95% confidence intervals (CIs) and p values. Doing so did not alter our findings because both predictors remained statistically significant, despite slight adjustments to p values. For the sake of completeness and transparency, both original and robust estimates are reported in Table 4. Self-collaboration showed a far stronger relationship to speakers' appraisal of their overall experience than self-comprehensibility, given that the coefficient for collaboration (*estimate* = 3.62) was approximately twice as large as the coefficient for comprehensibility (*estimate* = 1.82). Both predictors were positive, meaning that speakers who rated themselves as more collaborative and comprehensible tended to have a more positive impression of their overall experience.

Predicting communication success

Like the final model for overall experience, the final model for communication success contained two predictors: self-comprehensibility and partner-anxiety. This model passed all checks except for linearity. The nonsignificant Durbin–Watson test ($p = .228$) showed that the assumption of independence was upheld. Seven cases (3.89% of the data set) had large residuals, but their Cook's distance and leverage values were all within prescribed thresholds, which suggests that none of those cases

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

Table 4. Summary of Best-Fitting Model for the Exit Ratings

	Standard				Robust		
	Estimate	***SE***	95% CI	***p***	***SE***	CI	***p***
Overall experience							
Self-comprehensibility	1.82	0.68	[0.43, 3.17]	.008	0.74	[0.37, 3.27]	.014
Self-collaboration	3.62	0.68	[2.27, 4.96]	<.001	0.68	[2.28, 4.95]	<.001
Communication success							
Self-comprehensibility	5.53	1.13	[3.29, 7.77]	<.001	1.07	[3.43, 7.64]	<.001
Partner-anxiety	0.97	1.13	[−1.27, 3.20]	.396	1.40	[−1.78, 3.71]	.490
Comfort							
Self-collaboration	4.62	1.09	[2.47, 6.77]	<.001	1.00	[2.66, 6.57]	<.001
Self-anxiety	1.78	1.09	[−0.37, 3.93]	.104	1.05	[−0.28, 3.84]	.090

had an undue influence on the model. Model residuals were normally distributed, but the plot depicting fitted versus residual values showed some deviation from linearity; notably, like the overall experience model, there seemed to be a narrowing of residual variance at higher values. Thus, we adopted the same technique to compute robust SEs, 95% CIs, and adjusted *p* values. As reported in Table 4, the robust and original SEs for self-comprehensibility were similar (1.07 vs. 1.13), but the robust SE for partner-anxiety indicated that variability in the original estimate had been underestimated (1.40 vs. 1.13). The self-comprehensibility predictor was highly significant in both cases, whereas the partner-anxiety predictor was not. The large positive coefficient for self-comprehensibility (*estimate* = 5.53) demonstrates that speakers who rated themselves as more comprehensible also tended to perceive their communication with the partner to be more successful.

Predicting comfort interacting with partner

Lastly, for the model of comfort interacting with the partner, backward testing of fixed effects showed that self-collaboration and self-anxiety significantly improved model fit. This model also passed all checks except for linearity. The Durbin–Watson test was nonsignificant ($p = .800$), and model residuals were normally distributed. There were nine cases (5%) with large residuals, which is within the 5% threshold, and as in the other models, those cases did not appear to exert an undue influence on the model based on their Cook's distance and leverage values. When we computed robust SEs, we found that they were nearly identical to the SEs of the original model (1.00 vs. 1.09 for self-collaboration; 1.05 vs. 1.09 for self-anxiety). Although these small differences were unlikely to change the outcome of the analysis, for the sake of parity with the other models, we nonetheless recomputed 95% CIs and *p* values (Table 4). The results remained the same, in that there was a significant relationship between self-collaboration and speakers' appraisal of comfort, whereas the relationship between self-anxiety and comfort missed significance. The positive

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

coefficient suggests that speakers who rated themselves as more collaborative felt more comfortable interacting with their partner.

Discussion

The goal of this study was to examine if L2 speakers' self- and partner-assessments of comprehensibility (in addition to their ratings of anxiety and collaboration) predicted perceived interactional outcomes, defined as speakers' ratings of their overall experience, communicative success, and comfort interacting with the partner. Self-comprehensibility was the only predictor of communicative success and an additional predictor of perceived overall experience. In contrast, self-collaboration was the primary predictor of perceived overall experience and the only predictor of comfort. These effects were positive, such that speakers who rated themselves as more comprehensible and collaborative evaluated the interaction more favorably. Outside early research on comprehensibility (Varonis & Gass, 1982) and recent work on processing fluency (Dragojevic et al., 2017), this appears to be the first demonstration that comprehensibility has predictive strength in L2 speakers' appraisals of their interactive experience. In the following sections, we first discuss the primary finding that only the self-ratings predicted interactive outcomes before discussing each of the predictors in detail.

A speaker-centric view of L2 interaction?

A striking finding of this study is that all measures relevant to speakers' post-interaction ratings included their self- rather than partner-assessments (see Table 4). Against this backdrop, an interesting question to consider is why speakers' evaluations of their partner's comprehensibility, anxiety, and collaboration did not factor into their summative evaluations. Several explanations could account for this finding. For one, speakers induced in happy moods tend to engage in egocentric linguistic behaviors (Kempe et al., 2013). Given that all interactions in this data set were largely positive, where the overall experience was judged at a mean of 91.53 (with a 50–100 range) and no speakers provided negative comments in a post-interaction debrief, (happy) speakers may have been especially attuned to their own comprehensibility, anxiety, and collaboration. Another potential explanation relates to listener expectation. Speakers may hold certain expectations about the linguistic competence of their interlocutor, and these expectations can lead to increased lexical competition and poor memory for the content of conversations (Lev-Ari et al., 2018), which might make speakers less sensitive to their interlocutor's linguistic and nonlinguistic behavior. Because all speakers were well aware that they were conversing with an L2 interlocutor, their real or imagined communication difficulty may have resulted in an additional processing load for them (Horton & Gerrig, 2005), leading them to pay less attention to their partner's comprehensibility, anxiety, and collaboration states.

A third possibility is that speakers simply approached their conversations as additional L2 practice, with the consequence that the interactional behaviors of their partner (e.g., clarification requests, backchannelling, and eye contact) were primarily seen as positive or negative feedback targeting the speaker, rather than as

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

goal-oriented joint action relevant to both participants in a conversation (Brennan et al., 2018; Clark, 1996). In fact, in prior analyses of the data drawn from the present corpus, interlocutors' visual and interactive behaviors were attested frequently across all three tasks. For example, hand gestures and head movements (e.g., nods, tilts) occurred frequently in the moving to Montréal task (McDonough et al., 2022); breaking eye contact (looking away), blinking, and hand gesturing were attested most frequently in the close call story task (Tsunemoto et al., 2022a); and cases of backchanneling, nodding, and responsiveness (i.e., completing or elaborating on the interlocutor's utterance) emerged as most common interactive behaviors in the academic discussion task (Trofimovich et al., 2021). From this standpoint, speakers may have adopted an egocentric perspective on their performance because they interpreted interlocutor visual and interactive behaviors as reflecting the peaks and troughs in their own comprehensibility, anxiety, and collaboration. Finally, an increased role of self-assessments in speakers' evaluations of their interactive experience may reflect the halo effect, whereby speakers project a positive image of themselves on the conversation. Regardless of their potential origins, these findings suggest that the success of L2 interaction, as evaluated by its participants, seems to depend on a speaker's own (perceived) contributions to it.

(Self) predictors of the interactive experience

Turning to the predictors themselves, self-comprehensibility was associated with two of the three outcome measures, predicting speakers' ratings of their communicative success and overall experience but not their ratings of comfort. This finding implies that self-comprehensibility has a functional rather than interpersonal association with perceived interactional outcomes. In this sense, our findings align with prior work on processing fluency (Dragojevic et al., 2017; Jensen & Thøgersen, 2020; Sanchez & Khan, 2016), where listeners' processing effort is more strongly tied to their judgments of the content of speech (e.g., in terms of its informativeness, learning potential) and the speaker's competence (e.g., education, intelligence) than to their attitudes toward the speaker's personality (e.g., warmth, friendliness). In essence, as they create discourse that varies in comprehensibility, speakers presumably track their processing effort, using these perceptions to guide their judgment about how well they achieve their communicative goals in terms of exchanging distinct task-relevant ideas or reaching consensus. In contrast, processing effort appears to contribute little to how speakers evaluate their comfort level, which can be broadly understood through the constructs of interpersonal cohesion and rapport. These aspects of interaction may be best explained through nonlinguistic behaviors, such as interlocutors' gestures, body postures, mannerisms, facial expressions, and displays of emotion (Duffy & Chartrand, 2015). Needless to say, it remains for future work to determine which aspects of interaction, including comprehensibility, are more or less important to how interlocutors appraise both their experience and their partner in terms of functional (e.g., task completion, goal attainment) and interpersonal (e.g., rapport, comfort) value.

In addition to comprehensibility, we examined measures of interlocutor-rated anxiety and collaboration as predictors of L2 speakers' perceived interactional

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

outcomes, having chosen these measures retrospectively (rather than through a priori hypothesis testing) while being aware of recently reported links between these measures in conversational tasks (Nagle et al., 2022). Self-collaboration emerged as the primary predictor of speakers' perceived overall experience and the only predictor of their comfort interacting with the partner. That collaboration accounted for speakers' post-interaction ratings is hardly surprising, assuming that an optimal conversational experience—both in terms of its overall merit and its interpersonal value—is likely grounded in interlocutors having a sense that they show commitment to the dialog. Collaboration, which can be broadly defined as a speaker's degree of interest, engagement, and participation in an activity (Philp & Duchesne, 2016), can be displayed in multiple ways. For example, speakers might show collaboration through cognitive engagement by devoting sustained attention or effort to performing a task, through behavioral engagement by providing sufficient task-relevant talk, and through social and emotional engagement by demonstrating reciprocity and mutuality in conversation, such as participating in turn-taking or elaborating on the partner's ideas (Duran & Spitzberg, 1995; Galaczi & Taylor, 2018). What was surprising, however, is that only self-collaboration mattered for how speakers appraised their conversational experience and comfort. This is a novel finding, implying that speakers likely evaluated the interaction more favorably and felt more comfortable in conversation if they themselves demonstrated attention to task instructions, showed commitment toward task completion, produced task-relevant content, and managed the discussion without dominating the conversation or abstaining from it (Nakamura et al., 2021; Qiu & Lo, 2017). In the absence of clear behavioral measures of collaboration, these speculative interpretations must be revisited in future work.

There was no evidence in this data set that perceived anxiety mattered for how speakers judged their interactional outcomes, apart from a weak, nonsignificant effect of self-anxiety on the rating of comfort. Broadly defined, anxiety refers to a person's negative emotional reaction experienced during communication (Gardner & MacIntyre, 1993), where increased anxiety might impair language processing, disrupt the flow of interaction, and promote negative attitudes and motivational dispositions (Dewaele, 2010; MacIntyre & Gardner, 1994). For example, if speakers experience anxiety or perceive their interlocutor as being anxious, they might be reluctant to continue speaking or may view their interaction as unsuccessful, engaging in avoidance behaviors, such as abandoning task goals and providing frequent backchannels as a way of evading a turn (Ely, 1986; Steinberg & Horwitz, 1986). However, when anxiety is low, interlocutors might embrace their communicative goals, which would reinforce their goal-oriented positive actions, such as increased participation or helpfulness, resulting in a positive outlook on interaction. There was some evidence for such anxiety–collaboration links (see Table 2), especially for the partner-ratings ($r = .50$). However, perceived anxiety did not seem to factor into speakers' post-interaction ratings (at least independently from collaboration) most likely because most conversations were friendly and relaxed. This, of course, does not mean that perceived anxiety would be irrelevant to speakers' appraisals of their interaction in other situations, such as high-stakes exams or employment interviews or if the speakers had lower or differing levels of language proficiency.

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

Changes in perceptions throughout the interaction

One final point that deserves attention is the extent to which the self- and partner-ratings changed over the three interactive tasks. The only ratings that demonstrated statistically significant change over time were self- and partner-comprehensibility, which decreased slightly throughout the interaction. This result runs counter to that reported by Trofimovich et al. (2020), who found that speakers' comprehensibility assessments increased throughout the interaction. There are two potential explanations. First, the order in which speakers in the present study completed the interactive tasks was rotated and counterbalanced across dyads, whereas in Trofimovich et al. (2020), all dyads completed tasks in a fixed order, making it impossible to disentangle task and time effects. Specifically, in Trofimovich et al. (2020), speakers first completed a warm-up task, then performed a challenging picture narration task, and ended with a less challenging shared experience task, which could explain the U-shaped trend in comprehensibility ratings (i.e., a comprehensibility trough during the challenging task and a peak during the shared experience task, where speakers' comprehensibility assessments may have been boosted by task order insofar as they moved from a more to a less difficult task). And second, the fact that comprehensibility ratings decreased in this study may indicate that speakers gradually gained a more realistic perspective on their own and their partner's comprehensibility as the conversation progressed. Indeed, there is research demonstrating that L2 speakers' self-assessments become more aligned with assessments provided by external raters after a series of tasks (Strachan et al., 2019; Tsunemoto et al., 2022b). Thus, a downward trend may have reflected speakers' increased awareness of their own and their partner's comprehensibility. At the same time, all ratings were high overall and relatively stable throughout the interaction, which is likely a reflection of speakers' status as highly proficient and comprehensible L2 English users. It seems probable that if this study were replicated with speakers of different proficiency and comprehensibility levels, ratings would be more variable over time.

Limitations, future work, and conclusion

During interaction, speakers make various perceptual judgments about the interaction itself and about each other (MacIntyre & Ayers-Glassey, 2020), and these judgments likely have both short-term and long-term consequences. In this study, we explored whether L2 speakers' judgments of one another's comprehensibility, anxiety, and collaboration contributed to how positively they evaluated their conversations. We showed that speakers' perception of an interaction, at least as assessed at the end of a three-task sequence, was mainly linked to their own perceived performance, in terms of comprehensibility and collaboration. Despite their promise, these findings need to be interpreted in light of several limitations. One limitation is our use of existing corpus data, which precluded a priori hypothesis testing. Predictably an existing data set constrained the nature and scope of available measures. In future research, in addition to providing overall appraisals of the quality or success of their interaction, speakers might be asked to make judgments about their future plans to communicate, collaborate on course or work projects, or spend time in social situations with the interlocutor. As noted throughout the discussion,

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

incorporating judgments about both the interaction and the interlocutor would be worthwhile given that these judgments may be driven by distinct sets of predictors. The value of comprehensibility and collaboration as predictive constructs would certainly be enhanced if researchers could show that, for a given speaker, a particular experience with an interlocutor who is more or less comprehensible or collaborative impacts that speaker's future actions. Ideally speaking, impressionistic judgments should be supplemented with observations of actual behavior, where speakers report both the frequency and intensity of their future communicative exchanges. Indeed, it seems probable that any model of communicative outcomes would need to include both subjective measures of interlocutor experience and objective measures of interlocutor behavior.

The use of an existing data set also constrained our choice of speakers and tasks. Presumably issues of comprehensibility might be enhanced for interlocutors whose speaking skills are lower than those expected for university students engaged in L2-medium instruction, or in speaking tasks where attaining a communicative goal carries higher stakes (e.g., oral language exams, job interviews). As noted previously, understanding predictive links between comprehensibility (or collaboration) and speakers' attitudes and behaviors must also be accompanied by measures of their linguistic and interactive performance in order to isolate the specific aspects of interaction linked to speakers' eventual decision-making. Finally, it would be beneficial to consider the degree of calibration between interlocutors (Trofimovich et al., 2020). One interlocutor may view the conversation as highly successful, whereas the other might have a less favorable view of the interaction, in which case it would be unlikely for the interlocutors to interact again in the future. Given that interaction requires active planning and participation from both individuals, the pairs that would be most likely (to want) to interact again are pairs where both members view the interaction in highly positive terms. Thus, an important next step in interaction-based comprehensibility research, including research on communicative outcomes, is developing measures and models that appropriately capture interlocutor calibration.

Broadly speaking, aside from a few notable exceptions (Dragojevic et al., 2017; Gluszek & Dovidio, 2010; MacIntyre & Ayers-Glassey, 2020), there is ample room for theoretical and conceptual thinking to develop data-driven, dynamic, time-sensitive predictive models of how various linguistic, affective, social, and behavioral dimensions of interaction act together to determine the success of a given communicative exchange. For example, it is easy to imagine a self-reinforcing, cyclical process, wherein partners who judge one another to be easy to understand, collaborative, and relaxed during a conversation are more likely to develop a positive view of the interaction, potentially laying the groundwork for future interactions with the same and other similar individuals. This process, articulated over time, could serve as an engine for L2 development, insofar as speakers who are judged as easy to interact with might ultimately have more opportunities for the type of sustained, meaningful L2 interaction that is likely to stimulate learning. A worthwhile next step is therefore for researchers to develop new and accelerate existing work targeting subjective, perceived dimensions of interaction as catalysts of communication-driven language development.

Conflict of interest. The authors declare none.

Replication Package. Replication data and materials for this article can be found at https://osf.io/tzye2/.

Notes

1. In this study, because we controlled for task order by counterbalancing the order in which speaker dyads completed the tasks, we choose not to examine task and task order effects for several reasons. First, integrating task order into the analyses would have involved incorporating a categorical predictor with several levels and relevant interactions with other predictors, which would have made the models far more complex. Second, we view order effects as a separate research question, worthy of independent investigation. Nevertheless, to enable researchers to generate testable hypotheses for future research focusing on task order effects, we provide descriptive statistics for all target ratings separately for each task order (see Appendix D).
2. Whereas the overall lack of relationships between self- and partner ratings of comprehensibility, anxiety, and collaboration is intriguing, it would be premature to draw definitive conclusions from this findings, given that these associations are based on pooled data across three different tasks, each varying in their individual demands, and across six different task orders. We therefore generated correlation matrices by time (first, second, and third tasks) and task (moving, close call, and academic discussion). As shown in Appendices B and C, the time- and task-specific correlations largely mirror the pattern of results we obtained for the global correlations: moderate correlations for the self-self and partner-partner measures and very weak correlations between the self-partner measures. Nevertheless, as discussed in the previous note, it would be interesting to examine how task order potentially affects the strength of the associations obtained between the self-partner measures in particular.

References

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

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.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

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

Dewaele, J.-M. (2010). Multilingualism and affordances: Variation in self-perceived communicative competence and communicative anxiety in French L1, L2, L3 and L4. *International Review of Applied Linguistics*, **48**, 105–129. https://doi.org/10.1515/iral.2010.006

Dragojevic, M., Giles, H., Beck, A.-C., & Tatum, N. C. (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

Duffy, K. A., & Chartrand, T. L. (2015). The extravert advantage: How and when extraverts build rapport with other people. *Psychological Science*, **26**, 1795–1802. https://doi.org/10.1177/0956797615600890

Duran, R. L., & Spitzberg, B. H. (1995). Toward the development and validation of a measure of cognitive communication competence. *Communication Quarterly*, **43**, 259–275. https://doi.org/10.1080/01463379509369976

Ely, C. M. (1986). An analysis of discomfort, risktaking, sociability, and motivation in the L2 classroom. *Language Learning*, **36**, 1–25. https://doi.org/10.1111/j.1467-1770.1986.tb00366.x

ETS. (2017). *TOEFL iBT® test: Compare scores*. https://www.ets.org/toefl/score-users/ibt/compare-scores.html

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE Publications.

Galaczi, E. D., & Taylor, L. (2018). Interactional competence: Conceptualizations, operationalizations, and outstanding questions. *Language Assessment Quarterly*, **15**, 219–236. https://doi.org/10.1080/15434303.2018.1453816

Gardner, R. C., & MacIntyre, P. D. (1993). On the measurement of affective variables in second language learning. *Language Learning*, **43**, 157–194. https://doi.org/10.1111/j.1467-1770.1992.tb00714.x

Gluszek, A., & Dovidio, J. F. (2010). The way they speak: A social psychological perspective on the stigma of nonnative accents in communication. *Personality and Social Psychology Review*, **14**, 214–237. https://doi.org/10.1177/1088868309359288

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

Gries, S. T. (2021). (Generalized linear) mixed-effects modeling: A learner corpus example. *Language Learning*, **71**, 757–798. https://doi.org/10.1111/lang.12448

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, **96**, 127–142. https://doi.org/10.1016/j.cognition.2004.07.001

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

Jensen, C., & Thøgersen, J. (2020). Comprehensibility, lecture recall and attitudes in EMI. *Journal of English for Academic Purposes*, **48**, 100912. https://doi.org/10.1016/j.jeap.2020.100912

Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, **68**, 115–146. https://doi.org/10.1111/lang.12270

Kempe, V., Rookes, M., & Swarbrigg, L. (2013). Speaker emotion can affect ambiguity production. *Language and Cognitive Processes*, **28**, 1579–1590. https://doi.org/10.1080/01690965.2012.755555

Kennedy, S. (2021). Difficulties understanding L2 speech due to discourse- versus word-level elements. *Journal of Second Language Pronunciation*, **7**, 315–342. https://doi.org/10.1075/jslp.18015.ken

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, **36**, 345–366. https://doi.org/10.1093/applin/amu040

Lev-Ari, S., Ho, E., & Keysar, B. (2018). The unforeseen consequences of interacting with non-native speakers. *Topics in Cognitive Science*, **10**, 835–849. https://doi.org/10.1111/tops.12325

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

MacIntyre, P., & Ayers-Glassey, S. (2020). Competence appraisals: Dynamic judgements of communication competence in real time. In W. Lowie, M. Michel, A. Rousse-Malpat, M. Keijzer, & R. Steinkrauss (Eds.), *Usage-based dynamics in second language development* (pp. 155–175). Multilingual Matters.

MacIntyre, P. D., & Gardner, R. C. (1994). The subtle effects of language anxiety on cognitive processing in the second language. *Language Learning*, **44**, 283–305. https://doi.org/10.1111/j.1467-1770.1994.tb01103.x

McDonough, K., Kim, Y.-L., Uludag, P., Liu, C., & Trofimovich, P (2022). Exploring the relationship between behavior matching and interlocutor perceptions in L2 interaction. *System*, **108**, Article 102865. https://doi.org/10.1016/j.system.2022.102865

McDonough, K., & Trofimovich, P. (2019). *Corpus of English as a Lingua Franca interaction (CELFI)*. Montréal, Canada: Concordia University.

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. (2020). Foreign accent, comprehensibility and intelligibility, redux. *Journal of Second Language Pronunciation*, **6**(3), 283–309. https://doi.org/10.1075/jslp.20038.mun

Nagle, C. (2019). Developing and validating a methodology for crowdsourcing L2 speech ratings in Amazon Mechanical Turk. *Journal of Second Language Pronunciation*, **5**, 294–323. https://doi.org/10.1075/jslp.18016.nag

Nagle, C., & 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., 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

Nagle, C., 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

Nakamura, S., Phung, L., & Reinders, H. (2021). The effect of learner choice on L2 task engagement. *Studies in Second Language Acquisition*, **43**, 428–441. https://doi.org/10.1017/S027226312000042X

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

Philp, J., & Duchesne, S. (2016). Exploring engagement in tasks in the language classroom. *Annual Review of Applied Linguistics*, **36**, 50–72. https://doi.org/10.1017/S0267190515000094

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, **64**, 878–912. https://doi.org/10.1111/lang.12079

Qiu, X., & Lo, Y. Y. (2017). Content familiarity, task repetition and Chinese EFL learners' engagement in second language use. *Language Teaching Research*, **21**, 681–698. https://doi.org/10.1177/1362168816684368

R Core Team. (2022). *R: A language and environment for statistical computing*. Retrieved from https://www.R-project.org

Saito, K. (2021). What characterizes comprehensible and native-like pronunciation among English-as-a-second-language speakers? Meta-analyses of phonological, rater, and instructional factors. *TESOL Quarterly*, **55**, 866–900. https://doi.org/10.1002/tesq.3027

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. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, **37**, 217–240. https://doi.org/10.1017/S0142716414000502

Saito, K., Trofimovich, P., Isaacs, T., & Webb, S. (2017). 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.

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

Schwarz, N. (2018). Of fluency, beauty, and truth: Inferences from metacognitive experiences. In J. Proust & M. Fortier (Eds.), *Metacognitive diversity: An interdisciplinary approach* (pp. 25–46). Oxford University Press.

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

Steinberg, F. S., & Horwitz, E. K. (1986). The effect of induced anxiety on the denotative and interpretive content of second language speech. *TESOL Quarterly*, **20**, 131–136. https://doi.org/10.2307/3586395

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., & 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**(3), 435–461. https://doi.org/10.1075/jslp.21006.tro

Tsunemoto, A., Lindberg, R., Trofimovich, P., & McDonough, K. (2022a). Visual cues and rater perceptions of second language comprehensibility, accentedness, and fluency. *Studies in Second Language Acquisition*, **44**(3), 659–684. https://doi.org/10.1017/S0272263121000425

Tsunemoto, A., Trofimovich, P., Blanchet, J., Bertrand, J., & Kennedy, S. (2022b). Effects of benchmarking and peer-assessment on French learners' self-assessments of accentedness, comprehensibility, and fluency. *Foreign Language Annals*, **55**(1), 135–154. https://doi.org/10.1111/flan.12571

Varonis, E. M., & Gass, S. (1982). The comprehensibility of non-native speech. *Studies in Second Language Acquisition*, **4**, 114–136. https://doi.org/10.1017/S027226310000437X

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, **11**, 1–17. https://doi.org/10.18637/jss.v011.i10

Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software*, **95**, 1–36. https://doi.org/10.18637/jss.v095.i01

Appendix A

Definitions of rated constructs

Term	Explanation
Comprehensibility	This term refers to how much effort it takes to understand what someone is saying. If you can understand with ease, then a speaker is highly comprehensible. However, if you struggle and must listen very carefully, or in fact cannot understand what is being said at all, then a speaker has low comprehensibility.
Anxiety	This term refers to the level of stress, worry, or nervousness that someone is feeling while completing a task. If you are (or you believe that your partner is) experiencing very little worry or stress while completing the task, then the anxiety level is low. If you are (or you believe that your partner is) feeling very worried or nervous about the task, however, the anxiety level is high.
Collaboration	This term refers to the action of working with someone to produce or create something. If you are actively participating and working together as a team more than as an individual, then you are collaborating, or working well together. If you are not actively participating or working together as a team, then you are not collaborating or working well together.

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

Appendix B

Pearson correlations between predictor variables by time

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. s comp t1																	
2. s comp t2	.66																
3. s comp t3	.54	.49															
4. s anx t1	.31	.18	.22														
5. s anx t2	.26	.31	.16	.52													
6. s anx t3	.22	.25	.29	.50	.48												
7. s collab t1	.43	.33	.29	.45	.37	.25											
8. s collab t2	.33	.38	.33	.22	.24	.17	.48										
9. s collab t3	.29	.36	.34	.27	.23	.24	.55	.54									
10. p comp t1	.01	.00	−.02	−.13	−.05	−.13	−.14	−.01	.00								
11. p comp t2	.21	.29	.12	−.03	.04	−.02	.07	.11	.11	.61							
12. p comp t3	.04	.08	.12	−.11	−.04	.06	−.06	.04	.00	.58	.53						
13. p anx t1	−.01	.04	−.01	−.01	−.02	−.03	−.04	.06	.03	.29	.17	.19					
14. p anx t2	−.03	.09	−.10	−.05	.08	−.03	−.03	.01	.03	.25	.20	.19	.58				
15. p anx t3	.00	.08	−.02	.01	.04	.10	−.11	−.08	.00	.25	.18	.25	.55	.46			
16. p collab t1	−.11	−.02	−.06	−.06	.02	−.09	−.12	.03	−.01	.54	.29	.32	.39	.31	.34		
17. p collab t2	.02	.00	.01	.03	.04	.00	.02	.15	.10	.50	.46	.39	.32	.39	.37	.54	
18. p collab t3	−.02	.02	.01	−.07	.08	−.02	−.13	.01	−.02	.48	.38	.43	.33	.32	.43	.59	.64

Note. s = self, p = partner, comp = comprehensibility, anx = anxiety, collab = collaboration, t1 = first task, t2 = second task, t3 = third task.

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

Appendix C

Pearson correlations between predictor variables by task

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. s comp m																	
2. s comp c	.56																
3. s comp d	.61	.47															
4. s anx m	.30	.19	.23														
5. s anx c	.21	.28	.20	.46													
6. s anx d	.26	.18	.35	.56	.47												
7. s collab m	.40	.32	.32	.24	.26	.17											
8. s collab c	.30	.36	.29	.28	.34	.24	.49										
9. s collab d	.44	.25	.40	.25	.29	.33	.56	.53									
10. p comp m	.21	.08	.15	.09	−.11	.02	−.01	.10	.14								
11. p comp c	.06	.12	.04	−.05	−.06	−.03	−.05	.01	.05	.52							
12. p comp d	.05	.02	.12	−.07	−.14	−.02	−.10	.04	.02	.57	.58						
13. p anx m	.04	−.03	−.06	.07	−.01	−.04	−.08	.03	−.04	.23	.17	.16					
14. p anx c	.00	.00	−.06	.04	.05	.03	−.08	.04	−.07	.23	.27	.24	.57				
15. p anx d	.06	.06	.04	−.03	−.06	.04	.11	−.05	.03	.21	.23	.24	.50	.54			
16. p collab m	−.04	−.09	.00	.03	−.04	.08	−.04	.01	−.02	.41	.31	.40	.40	.41	.39		
17. p collab c	.02	−.04	.02	.04	−.05	.00	.00	.05	.05	.37	.51	.46	.25	.40	.32	.60	
18. p collab d	.05	−.02	.01	−.02	−.13	−.04	.00	−.02	.03	.42	.37	.50	.27	.38	.41	.55	.64

Note. s = self, p = partner, comp = comprehensibility, anx = anxiety, collab = collaboration, m = moving challenges task, c = close call task, d = academic discussion task.

https://doi.org/10.1017/S0142716423000073 Published online by Cambridge University Press

Appendix D

Means and standard deviations for self- and partner-ratings over time by task order

	Order 1			Order 2			Order 3		
	Task 1 Moving	Task 2 Close Call	Task 3 Discussion	Task 1 Moving	Task 2 Discussion	Task 3 Close Call	Task 1 Close Call	Task 2 Moving	Task 3 Discussion
S Comp.	83.90 (12.05)	85.80 (12.48)	81.60 (15.42)	83.97 (13.22)	85.23 (13.73)	83.63 (15.31)	75.00 (19.51)	80.23 (18.39)	80.03 (19.69)
S Anx.	73.30 (28.55)	77.97 (27.47)	79.53 (27.06)	82.63 (19.96)	82.93 (22.47)	88.43 (10.62)	77.60 (23.73)	87.03 (13.12)	86.30 (19.83)
S Collab.	84.80 (16.03)	86.30 (14.28)	87.07 (15.28)	88.80 (9.27)	85.37 (16.10)	89.37 (10.07)	85.47 (20.93)	90.73 (9.99)	91.53 (11.15)
P Comp.	80.20 (17.25)	82.60 (19.28)	78.57 (19.70)	80.30 (16.79)	80.40 (15.24)	79.37 (17.14)	76.17 (19.35)	77.40 (22.45)	78.13 (23.02)
P Anx.	78.80 (24.97)	79.97 (24.52)	83.33 (23.38)	80.00 (21.66)	83.47 (22.74)	90.90 (9.17)	76.50 (27.51)	85.47 (16.19)	85.90 (16.94)
P Collab.	84.43 (17.62)	85.23 (19.32)	87.17 (14.63)	89.23 (9.04)	88.93 (12.19)	90.97 (8.56)	91.13 (10.29)	89.27 (14.25)	91.07 (13.84)
	Order 4			Order 5			Order 6		
	Task 1 Close Call	Task 2 Discussion	Task 3 Moving	Task 1 Discussion	Task 2 Moving	Task 3 Close Call	Task 1 Discussion	Task 2 Close Call	Task 3 Moving
S Comp.	76.56 (15.69)	80.84 (13.78)	82.91 (12.80)	69.37 (20.96)	83.37 (14.27)	84.20 (15.89)	79.14 (15.77)	77.75 (16.80)	82.71 (14.99)
S Anx.	67.78 (21.77)	76.47 (22.71)	83.25 (18.81)	69.33 (24.85)	80.80 (22.43)	83.87 (20.25)	68.25 (23.79)	81.86 (19.43)	80.79 (18.36)
S Collab.	85.16 (12.50)	84.47 (12.99)	87.72 (10.32)	87.07 (12.92)	91.70 (10.39)	89.17 (18.51)	86.36 (13.54)	88.71 (11.29)	87.68 (9.25)
P Comp.	80.25 (13.64)	77.62 (17.57)	84.19 (10.28)	74.20 (23.53)	84.87 (18.60)	85.33 (19.54)	76.11 (19.15)	81.00 (16.75)	84.36 (13.10)
P Anx.	69.69 (21.58)	81.53 (17.67)	81.47 (14.40)	78.37 (19.46)	82.80 (24.62)	86.67 (20.55)	73.82 (21.15)	82.11 (15.95)	82.50 (20.90)
P Collab.	86.81 (10.46)	82.62 (14.71)	86.41 (11.26)	87.87 (16.06)	91.50 (10.07)	91.23 (10.89)	88.43 (12.04)	88.86 (10.81)	87.21 (10.13)

Note. S Comp. = Self-Comprehensibility; S Anx. = Self-Anxiety; S Collab. = Self-Collaboration; P Comp. = Partner Comprehensibility; P Anx. = Partner Anxiety; P Collab. = Partner Collaboration.

Cite this article: Nagle, C., Trofimovich, P., Tekin, O., and McDonough, K. (2023). Framing second language comprehensibility: Do interlocutors' ratings predict their perceived communicative experience? *Applied Psycholinguistics* **44**, 131–156. https://doi.org/10.1017/S0142716423000073