Toward a Dynamic View of Second Language Comprehensibility

Charles Nagle (cnagle@iastate.edu)

Iowa State University

Department of World Languages and Cultures

3102G Pearson Hall

505 Morrill Drive

Ames, IA 50011

Pavel Trofimovich and Annie Bergeron

Concordia University

Author Note

This study was supported by an Iowa State University Social Sciences Seed Grant to the first author and grants from the Social Sciences and Humanities Research Council of Canada to the second author. We are deeply grateful to Cristina Uribe for her help with data analyses, to Peter MacIntyre for making the Idiodynamic Software available, and to the anonymous reviewers and the editor, Susan Gass, for their insightful comments and suggestions that helped us refine this article. The data and materials for this study are publicly accessible via the IRIS Repository at https://www.iris-database.org and via the Open Science Framework at https://osf.io/97kur.

*Accepted in *Studies in Second Language Acquisition*

This study took a dynamic approach to second language (L2) comprehensibility, examining how listeners construct comprehensibility profiles for L2 Spanish speakers during the listening task and what features enhance or diminish comprehensibility. Listeners were 24 native Spanish speakers who evaluated 2–5 minute audio clips recorded by three university-level L2 Spanish speakers responding to two prompts. Listeners rated comprehensibility dynamically, using Idiodynamic Software to upgrade or downgrade comprehensibility over the course of the listening task. Dynamic ratings for one audio clip were video-captured for stimulated recall, and listeners were interviewed to understand which aspects of L2 speech were associated with enhanced versus diminished comprehensibility. Results indicated that clips that were downgraded more often received lower global ratings but upgrading was not associated with higher ratings. Certain problematic features and individual episodes caused listeners' impressions to converge, though substantial individual variation among listeners was evident.

Keywords second language speech; rating; comprehensibility; Spanish; Dynamic Systems Theory

Introduction

Recent research into second language (L2) speech learning has been characterized by an increased interest in various linguistic, cognitive, and social variables associated with speech that is understandable to the listener (Derwing & Munro, 2015). One construct that is central to this idea is comprehensibility, which refers to listeners' perception of how easy or difficult speech is to understand. Comprehensible speech has been shown to depend on various speaker and listener factors, such as the linguistic content of speech, the speaking task, and the expertise of the listener (Crowther, Trofimovich, Isaacs, & Saito, 2018; Crowther, Trofimovich, Saito, & Isaacs, 2015; Munro, 2018; Thomson, 2018). However, in all prior research, comprehensibility has been studied as a static construct, based on one-time global judgments provided by raters after listening to short samples of L2 speech (Isaacs & Thomson, 2013; Isaacs & Trofimovich, 2012; Munro & Derwing, 1995a). In light of recent views of language learning and use as dynamic, time-variable processes (de Bot, Lowie, & Verspoor, 2007; van Geert, Steenbeek, & van Dijk, 2011), it would be important to investigate comprehensibility from a dynamic perspective, evaluating listeners' assessments of comprehensibility as they are experiencing speech in real time to determine the factors that enhance or compromise comprehensibility. Therefore, the chief objective of this study was to create a time-sensitive profile of comprehensibility by examining how listeners evaluate comprehensibility for L2 speakers and what speech features influence their moment-to-moment comprehensibility judgments. The secondary objective was to understand the extent to which listeners' dynamic ratings (i.e., upgrading or downgrading the speaker) predict global comprehensibility scores.

Background Literature

A Focus on Comprehensibility

Following from Varonis and Gass (1982), who initially described comprehensibility as "how easy it is to interpret the message" (p. 125), recent L2 research has adopted the definition where comprehensibility "refers to judgments on a rating scale of how difficult or easy an utterance is to understand" (Derwing & Munro, 1997, p. 2). Operationalized as a scalar listener-based rating, comprehensibility is thus distinct from intelligibility, which targets actual understanding assessed through listeners' transcriptions of speech content (Munro & Derwing, 1995a), retellings of narratives (Hahn, 2004), or interviews (Zielinski, 2008). A key facet of comprehensibility is that it captures a listener's processing effort, which implies that comprehensibility need not be aligned with intelligibility (Derwing & Munro, 1997; Munro & Derwing, 1995a). Indeed, listeners can misunderstand utterances that they deem highly comprehensible (Munro, 1998), and the subjective experience of processing difficulty might lead to feelings of negativity and frustration for listeners even if they fully understand the utterance (Dragojevic & Giles, 2016).

Comprehensibility is a useful measure compatible with a focus on understanding as the goal for L2 pronunciation learning (Derwing & Munro, 2015; Levis, 2005). First, comprehensibility captures at least some aspects of listeners' experience with speech as it unfolds over time. For example, listener assessments of comprehensibility have been shown to correlate with the time it takes for listeners to process speech content (Ludwig & Mora, 2017; Munro & Derwing, 1995b), with listeners' decisions about how credible the interlocutor sounds (Lev-Ari & Keysar, 2010), and with listeners' emotional and attitudinal responses to the speaker (Dragojevic & Giles, 2016). These results can be readily interpreted within social psychological research on processing fluency, which refers to a person's subjective experience of the ease or difficulty with which information is processed (Alter & Oppenheimer, 2009; Oppenheimer,

2008). This research has shown that a person's subjective experience of processing difficulty—rather than the actual difficulty—often underlies human judgments, decisions, and actions. For instance, raters evaluate the same university admission statements differently depending on whether they are printed in easy versus hard to read fonts (Oppenheimer, 2006). People also believe that sentences that are easier to process are more truthful (Reber & Schwarz, 1999). Even the performance of financial shares for companies trading in stock exchanges appears to vary as a function of how easy (*Barnings*, *Flinks*) or difficult (*Ulymnius*, *Queown*) the relevant company names sound to people (Alter & Oppenheimer, 2006). As an index of a person's processing ease or difficulty, comprehensibility might thus play an important role for the listener.

Compared to measures of intelligibility, scalar ratings of comprehensibility also seem to offer a more practical and user-friendly alternative to measuring understanding, especially because speakers' intelligibility depends on how it is operationalized. For instance, the same speakers can be shown to be more or less intelligible depending on the outcome measure used (Kang, Thomson, & Moran, 2018; Kennedy, 2009), such as understanding of individual words versus the comprehension of ideas. Capturing intelligibility may also be challenging in many teaching and assessment contexts, since teachers might have little expertise or time to objectively assess the extent to which learners are understood by their interlocutors (e.g., through comprehension questions or transcription exercises). In contrast, assigning a scalar rating of comprehensibility is an efficient and intuitive means of capturing listener experience with speech (Isaacs & Trofimovich, 2012). It is therefore unsurprising that several oral proficiency scales (e.g., TOEFL, IELTS) operationalize understanding as comprehensibility and that comprehensibility is targeted in diagnostic assessment tools for teachers (Isaacs, Trofimovich, & Foote, 2018). In sum, comprehensibility emerges as a theoretically relevant and practical

measure of listeners' experience with L2 speech, thus motivating further research on this construct, especially from a dynamic perspective.

Linguistic Dimensions of L2 Comprehensibility

Because one important source of processing difficulty for the listener stems from the linguistic content of L2 speech, researchers have examined various dimensions linked to comprehensible speech, with the goal of creating a linguistic profile of comprehensibility. For instance, research into L2 English has revealed two broad dimensions—pronunciation (individual segments, prosody, fluency) and lexicogrammar (varied/appropriate use of words and accurate/complex grammar)—which are associated with listeners' comprehensibility ratings for different speaker groups, including native (L1) speakers of French, Farsi, Hindi, Mandarin, and Japanese (e.g., Crowther et al., 2015; Saito, Trofimovich, & Isaacs, 2017). For target languages other than English, speakers' comprehensibility has also been linked to multiple linguistic dimensions, such as pronunciation, lexis, morphology, and fluency in L2 German (O'Brien, 2014), pronunciation, fluency, richness of lexis, and complexity of grammar in L2 French (Bergeron & Trofimovich, 2017), and fluency, appropriateness and variation of lexis, and placement of pitch in L2 Japanese (Saito & Akiyama, 2016). In sum, in evaluating ease or difficulty of understanding across various languages, listeners seem to rely on many linguistic features, not just on speakers' pronunciation and fluency.

However, nearly all current evidence about the linguistic aspects of L2 comprehensibility is correlational, based on associations between comprehensibility ratings and coded or rated measures of L2 speech. In fact, there have been few studies that investigate listeners' reasons for their rating decisions. Isaacs and Trofimovich (2012) asked three experienced ESL teachers to provide comprehensibility ratings for 40 French speakers of L2 English and then to describe the

basis for their judgments. The teachers commented on multiple variables, such as pronunciation, fluency, grammar, vocabulary, discourse structure (use of cohesive devices), and mentioned the availability of context and familiarity with the speaker's L1 as contributing factors (see also Kennedy, Foote, & Dos Santos Buss, 2015). Isaacs and Thomson (2013) compared the comments of 20 experienced ESL teachers and 20 novice raters (with no language teaching background or experience) in response to several scalar ratings of L2 speech, including comprehensibility. The main differences between the raters pertained to experienced raters producing longer comments explaining their decisions, attributing some of their ratings to prior experience with L2 speech, and having access to terminology to describe the linguistic content of speech. Finally, Crowther, Trofimovich, and Isaacs (2016) examined how listeners' knowledge of the speakers' L1 influences their speech ratings, showing that L1 French raters attributed their comprehensibility ratings of L2 French speakers to a broader range of linguistic dimensions, compared to L1 Mandarin raters (see also Foote & Trofimovich, 2018). Whereas these findings generally confirm that comprehensibility is tied to multiple linguistic variables for the listener, there is no research examining listeners' decision-making at various points during their experience with speech. Such data would provide a time-sensitive view of how various linguistic factors might contribute to the ease or difficulty with which listeners understand L2 speech.

The Case for a Dynamic Approach

Speaking and listening are dynamic acts whose properties fluctuate over time. In the case of speaking, for example, L1 speakers generally appear to alternate between periods of fluent and disfluent speech, and these temporal cycles occur on a time scale of 10–30 seconds (e.g., see Pakhomov, Kaiser, Boley, Marino, Knopman, & Birnbaum, 2011, and references therein). In addition to demonstrating various (dis)fluency markers, speakers (and especially L2 users) also

receive feedback on and engage in repair of unintended linguistic errors that may affect comprehensibility (e.g., Mackey, Park, & Tagarelli, 2016). As L2 speakers produce varying levels of accuracy, complexity, and fluency over time, listeners must continuously process this variability to interpret the intended message within an emergent discourse structure, suggesting that a speaker's comprehensibility is likely a dynamic, time-sensitive construct for the listener. For example, it could be that comprehensibility is particularly low at the outset of listening, given that contextual clues related to the topic are not yet available to the listener. Likewise, it could be that pausing, using an incorrect lexical item, or making a morphosyntactic error at a particular point in the discourse could produce a state of low comprehensibility for listeners, as they try to work out what the speaker intended or revise their understanding of the overall message based on new information. Put simply, a dynamic approach to comprehensibility anchored in time-aligned ratings has the potential to provide information on how the timing of different linguistic features of speech affects comprehensibility, conceptualized not as a single rating, but as a dynamic curve that unfolds over time.

Such a listener-centric, dynamic conceptualization of comprehensibility is compatible with views of language learning and use as dynamic, variable processes (de Bot et al., 2007; van Geert et al., 2011). Within such views, speaking and listening would be characterized by variability both within and across individuals, and comprehensibility can be seen as a continuous, dynamic adaptation of the listener to the speaker, for instance, in terms of the listener's processing of the linguistic content in the speaker's utterance. A dynamic focus on comprehensibility also aligns well with recent studies investigating motivation (Dörnyei & Tseng, 2009; MacIntyre & Serroul, 2015), foreign language anxiety (Gregersen, MacIntyre, & Meza, 2014), L2 self (Mercer, 2015), and willingness to communicate (MacIntyre & Legatto,

2011) as variable, time-sensitive constructs. Last but not least, a dynamic look at comprehensibility ratings would extend methodological research on listener-rated L2 speech constructs, such as accentedness and comprehensibility, which to date have only been measured at a single time using Likert-type scales (e.g., Munro & Derwing, 1995a; Southwood & Flege, 1999), sliding scales (e.g., Flege, 1988; Saito et al., 2017), or through direct magnitude estimation by comparing the target speech sample with a reference item (e.g., Brennan, Ryan, & Dawson, 1975; Munro, 2018).

The Current Study

L2 speakers' comprehensibility has emerged as a complex, multidimensional construct, associated with multiple aspects of L2 speech for the listener. However, in all previous research, comprehensibility has been measured as a global judgment by the listener, evaluated at a single time after a brief exposure to a sample of L2 speech (typically 20–30 seconds). The main objectives of this exploratory study were therefore (a) to investigate comprehensibility as a dynamic construct on a longer timescale, evaluating it from the listener's point of view in real time, as he or she experiences L2 speech, and (b) to examine dynamic ratings of comprehensibility in relation to various linguistic dimensions of L2 speech that the listener considers important for comprehensibility. To accomplish these objectives, 24 native Spanish listeners evaluated the speech of three intermediate-level English speakers of L2 Spanish providing narratives in response to two prompts (childhood memory and university studies). The listeners assessed comprehensibility during the entire speech sample (approximately 150–290 seconds), using Idiodynamic Software designed specifically for recording ongoing, time-locked ratings (MacIntyre, 2012). To allow for comparisons with prior research, the listeners also provided a single rating of comprehensibility using a 9-point Likert scale (e.g., Munro &

Derwing, 1995a), so that these global ratings could be compared with dynamic assessments. Upon completion of the ratings, the listeners were shown a video capture of their rating behaviors over time, which was played alongside the original speaker audio, and were asked to comment on potential reasons underlying each of their rating decisions. This study was informed by the following research questions:

1. Can L2 comprehensibility be modeled as a dynamic construct?
2. Which linguistic dimensions of speech are associated with dynamic changes in comprehensibility across time as the listener experiences L2 speech? And how do those linguistic dimensions relate to upgrading or downgrading the speaker?
3. What is the relationship between the dynamic, time-locked ratings and listeners' global assessment of comprehensibility?

Method

Speakers

The speakers included in this study were selected from a larger pool of potential speakers, all of whom were native speakers of English enrolled in fourth- and sixth-semester Spanish courses at a large public university in the United States. Speakers recorded responses to a variety of personally relevant prompts that were modeled on the ACTFL Can-Do Statements (ACTFL, 2015). Clips were normalized for peak intensity and presented to eight native Spanish speakers who were pursuing a graduate degree in a field other than linguistics at the same university. These raters, who represented various dialects of Latin American Spanish, rated fluency and accentedness using 9-point Likert scales, with higher scores indicating better performance (i.e., for fluency, 1 = "very disfluent" and 9 = "very fluent"; for accentedness, 1 = "very strong foreign accent" and 9 = "no foreign accent"). Every effort was made to match

speakers as closely as possible in terms of the following characteristics: (a) gender, (b) instructional level, (c) proficiency, which was estimated using an elicited imitation task, (d) speaking prompt, and (e) prerated fluency and accentedness. Taking these factors into consideration, three female speakers (M_{age} = 19.33 years, *range* = 18–20) were selected. They had begun learning L2 Spanish at age 12 (*range* = 8–15), were enrolled in their sixth semester (third year) of Spanish coursework at the time of recording, and achieved a mean elicited imitation score of 73.33 (*range* = 66–87) out of 120 possible points. As the speakers had predominantly learned Spanish through classroom instruction, they had taken courses with native and near-native Spanish speakers representing a range of dialects, and none of them had spent significant time abroad for the purpose of language learning. Consequently, none of the speakers spoke with a discernible regional accent that could be associated with any single dialect of Spanish.

All three speakers responded to the following two prompts: (a) describe what you are studying, the classes you are taking, and your favorite class; and (b) describe a memorable childhood experience, providing as much detail as possible. The average fluency and accentedness ratings for each speaker, along with other characteristics of the recorded prompts, are included in Table 1. The two prompts that each speaker provided were then combined into a single audio file suitable for the Idiodynamic Software used in this study.

Table 1. *Speaker and File Characteristics*

Variable	Speaker 1	Speaker 2	Speaker 3
Age (years)	20	20	18
Age of learning (years)	15	8	13
Elicited imitation score (0–120)	66	67	87
Prerated fluency (1–9)	6.88 (1.25)	7.38 (1.51)	7.75 (1.58)
Prerated accentedness (1–9)	6.13 (1.89)	5.25 (1.49)	7.75 (1.04)
File length (s)	287	153	128

Note. Age of learning = age at which speakers began learning Spanish; higher scores for fluency and accentedness indicate better performance (i.e., more fluent speech and less foreign accent).

Raters

The raters included 24 native speakers of Spanish (13 males, 11 females) from Colombia (12), Cuba (3), Venezuela, Mexico, Argentina (2 each), Chile, Ecuador, and Paraguay (1 each), all residents of Montréal, Québec, at the time of the study.[1] The choice of Montréal for rater recruitment was largely a matter of convenience, because the research team had access to many Spanish speakers in this large, multilingual urban center. Raters were asked to estimate their ability to speak, listen, write, and read French and Spanish using a 9-point scale (1 = "extremely poor," 9 = "extremely proficient") and to report on their patterns of language use, including their familiarity with L2 Spanish speech (1 = "not at all familiar," 9 = "extremely familiar") and the frequency with which they interacted with L2 Spanish speakers.

All raters (M_{age} = 36.92 years, *range* = 30–46) completed their primary and secondary education in Spanish in their home countries. As reported in Table 2, the majority of raters were first exposed to French later in life (M_{age} = 27.25 years, *range* = 7–41), and rated themselves as

having intermediate to high proficiency in French (*M* = 6.88, *range* = 4–9). In contrast, raters evaluated their English skills as generally weaker (*M* = 5.89, *range* = 1–9), despite being first exposed to English at a younger age (M_{age} = 11.71 years, *range* = 6–35). As residents of Montréal, a predominantly French-speaking city, the raters reported using Spanish (*M* = 41.08%, *range* = 0–50) and French (*M* = 42.88%, *range* = 5–80) to a similar extent for daily interaction, in addition to English (*M* = 15.63%, *range* = 0–50) and Guarani (*M* = 0.37%). The raters estimated themselves to be moderately familiar with L2 Spanish (*M* = 5.25, *range* = 1–9) and varied in their frequency of interaction with L2 Spanish speakers, with 10 raters reporting daily or weekly interaction and 14 claiming to communicate once per month or not at all. Most raters (15) had received some linguistic training (as part of their degree or their L2 French courses), and 10 reported language teaching experience. The rater group thus included multilingual speakers for whom Spanish was the language learned from birth, who used both Spanish and French daily, but who varied in their exposure to L2 speakers of Spanish, linguistic training, and teaching experience.

Table 2. *Rater Characteristics* (*n* = 24)

Background variable	*M*	*SD*	Range
Age	36.92	3.84	30–46
Age of exposure: L2 French	27.25	7.52	7–41
Self-rated French proficiency	6.88	1.19	4–9
Age of exposure: L2 English	11.71	5.77	6–35
Self-rated English proficiency	5.89	2.52	1–9
Percent daily Spanish use	41.08	16.98	0–50

Percent daily French use	42.88	20.23	5–80
Percent daily English use	15.63	17.24	0–50
Familiarity with L2 Spanish speech	5.25	2.64	1–9

Rating Procedure

Individual experimental sessions, which took place in a quiet location and lasted between 60 and 90 minutes, were conducted by the third author (a native speaker of French and a near-native speaker of Spanish), with each rater allowed to use Spanish, French, or both languages during the session. All printed materials were presented in the raters' L1 (Spanish). The raters first completed a questionnaire eliciting information about their language background and then followed the researcher's oral instructions using a short booklet introducing the study, providing the definition of the rated construct with examples, and illustrating the rating interface. Comprehensibility was defined as the amount of effort that it takes to understand what someone is saying.

Dynamic ratings of comprehensibility were collected using Idiodynamic Software (MacIntyre, 2012). The software, which is freely available from http://faculty.cbu.ca/pmacintyre, allows users to record time-locked ratings (in 1 second increments) by clicking to raise or lower the level of the rated construct to values between ±5, relative to the baseline (marked by a straight line crossing 0). The raters were instructed to click the button labeled "Increase comprehensibility" when they felt that the speaker became easier to understand and to click the button labeled "Decrease comprehensibility" when they felt that the speaker became more difficult to understand. The raters were also told that each successive click of the mouse corresponded to an additional increase or decrease in their rating—which would appear as an upward or downward block on a color bar graph—and were encouraged to keep clicking the

button if they perceived the speaker's speech to become increasingly easier or harder to understand. In the absence of any rating activity from the user, the software engages a built-in auto-zero function, returning the rating to the baseline at the rate of one point (click) per second (MacIntyre, 2012), and the raters were made aware of this function (see Appendix A for a screenshot of the interface).

Before rating the target audio clips, the raters practiced using the interface with an additional clip (63 seconds) recorded by a near-native female L2 Spanish speaker and featuring a response to a different prompt (describe your personality). Once all raters confirmed that the task was clear, they used a high-quality headset to listen to the audio clips and rated them using the onscreen interface. The raters evaluated the three speakers' clips using six randomized orders (e.g., 1–2–3, 3–2–1), with four raters randomly assigned to each order. Each speaker's responses to the two prompts (childhood memory, university studies) were played back to back, but with an equal number of raters assigned to each of the two prompt orders. The raters were not allowed to take notes or pause the clips because the intention was to capture comprehensibility ratings in real time. Before the last speaker's audio clips were loaded, the researcher set up video capture to record the rater's interaction with the rating interface so that the last rating could be used for a stimulated recall procedure (henceforth, stimulated interview). This was done to ensure that the raters were fully familiar with the rating interface and its use before they commented on their thought processes.

Immediately after the rating, stimulated interview was carried out with each rater, using the video capture file (MP4) containing the ratings for the last speaker. The video file included the speaker's audio, a visual bar graph showing the direction of the ratings (positive or negative), and the mouse clicks/movements from the rater. Each session was recorded using a digital voice

recorder (VN-8100PC). The raters were told that they could stop the recording at any time to share their comments, but that the researcher might also follow up with specific questions. They were instructed to focus on what they were thinking at the time when they clicked upward or downward to indicate their rating and to comment about what helped them make their decisions. The researcher waited for listeners to stop the video and provide a comment, intervening with questions when obvious spikes and dips in the ratings were left without comment (e.g., Can you tell me what you were thinking here?) or when there were long stretches with no click activity (e.g., What made you keep your rating here?). The stimulated interview sessions lasted between about 5 and 17 minutes (*M* = 10 minutes 45 seconds) depending on the number of times the raters upgraded or downgraded the final speaker and their propensity to comment on the dynamic ratings they provided.

After the interview session, the raters completed a short questionnaire, evaluating their understanding of comprehensibility as a rating category and their comfort when rating it, the perceived difficulty of the rating task, and their rating confidence by placing a cross at the location corresponding to their assessment on a 100-milimeter continuous scale. These scales were followed by six open-ended debriefing questions, and the raters' responses to these questions were audio recorded for later analysis. The questions assessed the raters' previous experience with rating speech, their thoughts about what they considered to be the easiest and the most difficult aspects of the rating, their perception about how their ratings may have changed within and across speakers, and their comments about the speech features that they found easiest and most difficult to understand (see Appendix B for a copy of the debriefing questionnaire).

At the end of the testing, the raters heard the same audio clips again (practice file followed by the three speakers' audio clips, all played in the same order used for the dynamic

ratings), but this time providing a single global comprehensibility rating using a 9-point scale (1 = “difficult to understand,” 9 = “easy to understand”). Although the speech content was already familiar to the raters, which might have influenced the global rating, it was important to minimize familiarity effects on dynamic assessments by eliciting them first, in keeping with the focus on raters’ dynamic behaviors. Again, because raters were familiar with the speech content through dynamic assessments, they were permitted to listen to as much or as little as necessary to make their global rating decision, and the researcher, who was controlling the audio playback, recorded the time needed by each rater to make his or her decision for each audio clip. This methodological decision, which departs from the usual practice whereby listeners wait until the end of a speech sample to provide a global rating decision (e.g., Munro & Derwing, 1995a), made it possible to make comparisons of the timing of the global rating and the (first) dynamic assessment directly. Despite the utility of this approach, it is important to bear in mind that each listener had a different level of re-exposure to the productions of the L2 speakers. Thus, some listeners may have been relying more heavily on memory rather than on the actual speech at the time of the global ratings.

Data Analysis

In terms of raters’ understanding of comprehensibility as the target dimension, initial evaluation of the debriefing questionnaires revealed that raters appeared to understand comprehensibility as a rating category (*M* = 92.96, *range* = 74–100), felt comfortable rating it (*M* = 85.88, *range* = 61–100), perceived the rating task as being easy (*M* = 86.83, *range* = 49–100), and showed strong confidence in their ratings (*M* = 89.04, *range* = 49–100). For quantitative analyses, we extracted the timing, magnitude, and direction (upgrade, downgrade) of the dynamic rating activity from the Idiodynamic Software’s data output file for the second and third

(final) speakers and tabulated corresponding global comprehensibility ratings. Because the key feature of the Idiodynamic Software is that it allows users to provide ratings conceptualized as deviations from the baseline, we reasoned that each rater needed sufficient time to become familiar with the characteristics of the intermediate-level Spanish samples. Consequently, we excluded data from the first speaker rated. In other words, the near-native sample was used to familiarize listeners with the ratings interface, the first intermediate-level sample was considered a calibration trial, and the remaining two speakers that were rated were experimental trials included in data analysis. Additionally, ratings were screen-captured during the final clip, yielding a dataset that integrated the timing, magnitude, and direction of ratings over time with the listener's stated reasons for these ratings for the last speaker.

For the global ratings (collected for comparison with the dynamic assessments, which were the focus of this research), interrater reliability was estimated using a two-way, average-measure, consistency intraclass correlation (ICC) coefficient and Cronbach's alpha. The consistency ICC reached .92 (p = .03) for the entire group of listeners, but Cronbach's alpha was substantially lower (α = .63). Inspection of the individual by-rater statistics indicated that two listeners' scores did not align with the scores provided by the others. Once these individuals were removed from the dataset, the consistency ICC decreased (r = .79, p < .001), but Cronbach's alpha improved substantially (α = .79).[2] Although the content of the clips was already familiar to raters, to provide a global rating, on average they listened to 40.57 seconds of speech (0–153) per prompt, with no difference in timing between the prompts, before providing their assessment (see Table 3 for a summary of global ratings).

Table 3. *Descriptive Statistics for Global Comprehensibility Ratings (1–9 Scale)*

	Speaker 1		Speaker 2		Speaker 3		Time to rating (s)		
Prompt	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	Range
Studies	6.75	1.53	6.94	1.44	6.38	1.41	39.71	30.11	0–135
Childhood	6.13	1.93	6.19	1.56	6.56	1.21	41.42	32.39	3–153

The audio recordings of stimulated interviews were transcribed by the third author. Thematic coding of the listeners' comments for each question was done through empirical coding drawn from the content of the transcripts (Gibson & Brown, 2009). In an iterative process, the first author derived codes for themes and subthemes from the transcribed comments. The third author reviewed the coding, suggesting modifications to the coding of certain themes and subthemes, which was then commented on by the second author. This process continued until there was full consensus on the themes and subthemes and on coding decisions. The entire set of transcripts, along with detailed descriptions and examples of each coded category, were then given to a new coder (a French–Spanish bilingual with a graduate degree in L2 pedagogy), who independently recoded all stimulated interview comments. Intercoder reliability ($\kappa = .67$) was within the range of substantial agreement (.61–.80) for Fleiss' kappa (Landis & Koch, 1977). The third author and the new coder then reviewed cases where the coding was different to come to an agreement. The responses to open-ended debriefing questions were transcribed by the third author and were tagged for major themes based on the response content. Because these responses were straightforward (i.e., they did not require much interpretation), these coding decisions were not subjected to second coding.

Results

Dynamic Profiles

To respond to the first research question relating to the raters' dynamic profiles while

rating comprehensibility, we plotted and inspected individual rater data to determine the extent to which the raters adopted a dynamic approach and if their approach changed from the second to the third speaker rated. We classified raters as dynamic, semi-dynamic, or non-dynamic based on the frequency and magnitude of click activity (see Table 4). Dynamic raters displayed high click frequency and magnitude, semi-dynamic raters high frequency but lower magnitude, and non-dynamic raters low frequency and magnitude. Although click frequencies for semi- and non-dynamic raters partially overlapped, the semi-dynamic raters always utilized a larger portion of the scale. Dynamic raters (n = 2) continuously evaluated comprehensibility over the clip and seemed to establish a high comprehensibility benchmark for speakers, downgrading speakers to a lower positive value or allowing them to fall away from positive scores (i.e., allowing the auto-zero function of the software to reduce the comprehensibility rating) instead of using the negative portion of the scale. Consequently, comprehensibility curves were characterized by high peaks and deep valleys on the positive side of the scale for these two individuals. The semi-dynamic raters (n = 4) displayed the same pattern of continuous ratings, but the magnitude of the click activity was less pronounced than what was observed for the dynamic group. Moreover, ratings for this group were typically centered on the baseline (i.e., on zero), alternating between a narrow band of positive and negative values. Finally, the non-dynamic raters (n = 18) upgraded or downgraded the speakers far less frequently, with some raters only clicking a single time across the entire clip. For this group, ratings of ± 1 were common, resulting in relatively flat comprehensibility curves. Table 4 reports descriptive statistics for click activity for each listener group, and Figure 1 displays plots for dynamic and semi-dynamic raters and a representative sample of the non-dynamic raters for the second (panel a, top) and third (panel b, bottom) speaker rated. As is evident from the plots, raters' approach to the ratings was consistent across

speakers. Thus, in response to the first question, nearly all raters registered shifts in comprehensibility as they listened to the clips, but only about 25% (6 out of 24) displayed evidence of frequent shifts in comprehensibility that would suggest an unambiguously dynamic profile.

Table 4. *Number and Timing of Clicks for Dynamic Comprehensibility Ratings*

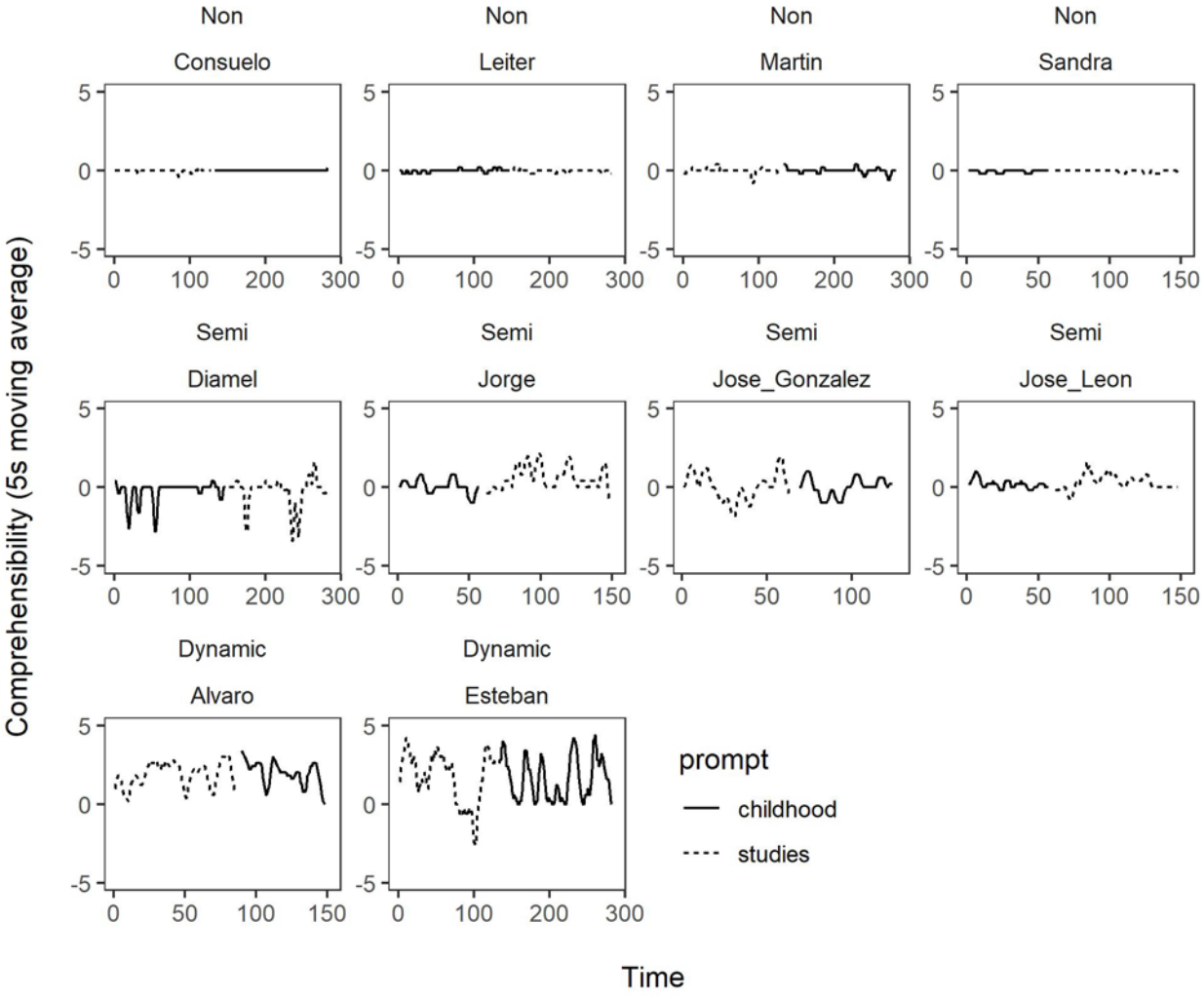
	Positive clicks (*k*)			Negative clicks (*k*)			Time to first click (s)		
Listeners	*M*	*SD*	Range	*M*	*SD*	Range	*M*	*SD*	Range
Dynamic (2)	419.00	85.85	299–503	13.25	8.06	5–24	5.25	0.96	4–6
Semi-dynamic (4)	46.38	21.74	11–70	30.50	28.82	7–98	7.75	3.15	5–14
Non-dynamic (18)	3.86	5.77	0–28	6.56	7.26	0–33	37.47	53.97	5–252

In terms of raters' experience of engaging in dynamic assessments, 10 raters commented in their open-ended debriefing responses that the subjective aspect of the assessment was the most difficult aspect of the task for them, whereas 14 noted that listening and assessing speech in their native language was relatively easy. Compared to two raters who thought that their rating had become stricter, six raters commented that they had become more lenient toward each speaker over time, as illustrated in this comment:

- But 30 seconds later, it starts to… it's more fluent, we understand it very well. Maybe at the beginning we are stricter with comprehensibility but after…. (Yanet, non-dynamic rater)[3]

As many as 13 raters admitted to becoming progressively more lenient as they proceeded from one clip to the next across the three speakers, as illustrated in the following comment:

- The last [speaker] seemed… clearer to me. At the beginning, [with] the first [speaker], I was stricter. (Consuelo, non-dynamic rater)



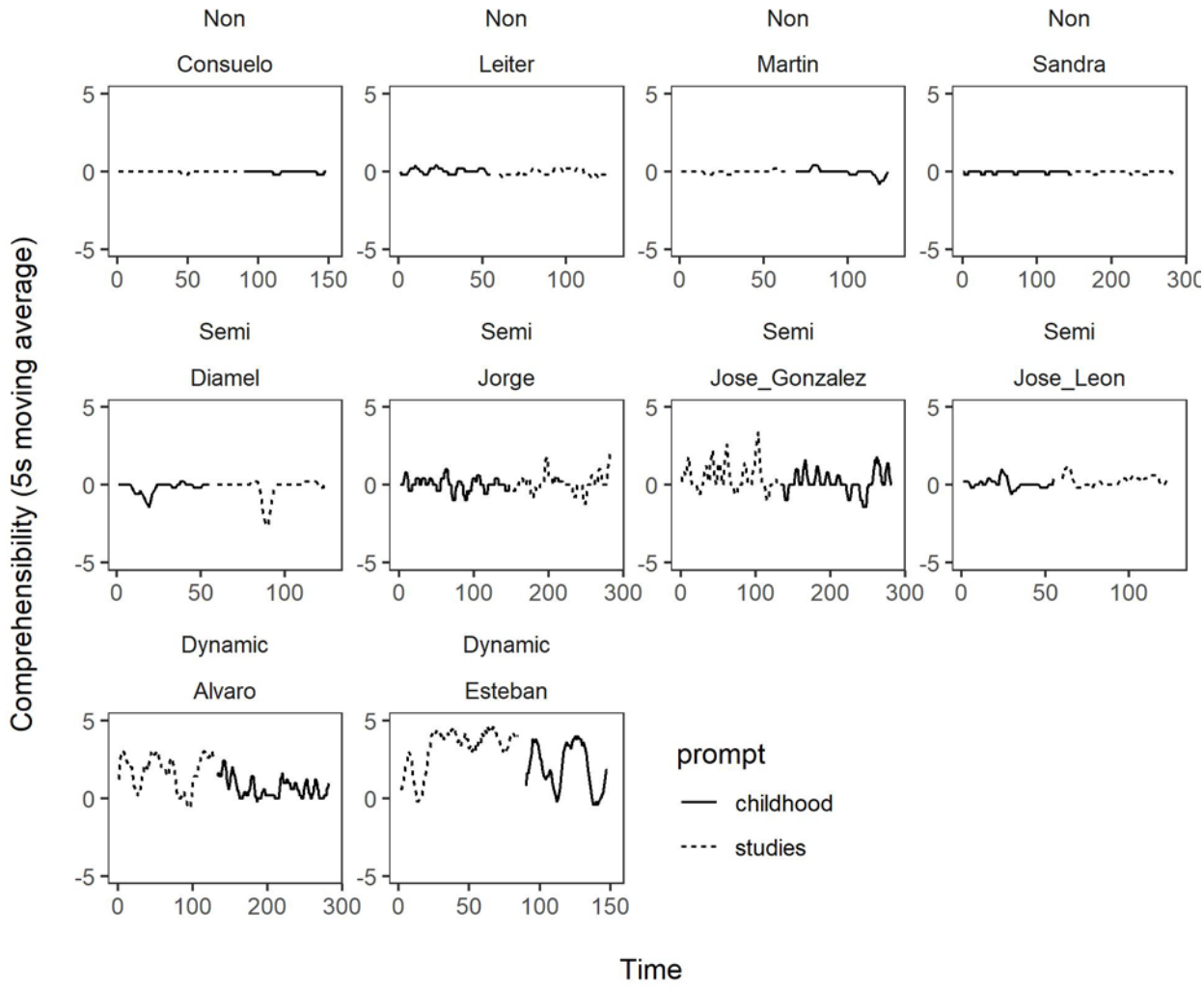


Figure 1. Five-second moving window comprehensibility plots for non-dynamic, semi-dynamic, and dynamic listener groups for the second (panel a, top) and the third (panel b, bottom) speaker rated. The four non-dynamic listeners plotted are a representative subset of the larger group (n = 18). Scales for the x-axes vary depending on the speaker order to which the listeners were assigned. Comprehensibility curves reflect the order in which the two prompts were combined for each rater (childhood memory prompt–studies prompt or studies prompt–childhood memory prompt).

Linguistic Dimensions Underlying Dynamic Ratings

To answer the second research question concerning the linguistic dimensions associated with the dynamic ratings, stimulated interview comments were analyzed focusing first on general patterns and then on the relative importance of specific linguistic issues to upgrading and

downgrading the speaker. A detailed description of each coded category with 1–2 representative examples appears in Appendix C.

General patterns. Table 5 summarizes the number and percentage of comments provided per coded category, as well as the number of raters who made at least one comment pertaining to the category during their stimulated interview; a breakdown of comments by rater group (dynamic, semi-dynamic, non-dynamic) is provided in Appendix D. The comments categorized as negative were associated with raters downgrading a speaker's comprehensibility; positive comments were linked to raters upgrading a speaker's comprehensibility; neutral comments were associated with no rating (clicking) activity and were thus cited as reasons for no change in ratings. In total, raters made 434 comments during the stimulated interview, most often in reference to discourse structure (31%). This category encompassed both sequencing (e.g., "I think it was the sequence of events there, it wasn't coherent") and completeness issues (e.g., "She didn't finish the idea, the idea wasn't finished for me"). Comments pertaining to the lexis (17%) and grammar (16%) categories revealed that raters were sensitive to a variety of errors, including accuracy issues that arguably would not affect the intended meaning of the utterance. For instance, one rater mentioned downgrading a speaker for using the incorrect grammatical gender: "[She said] *todas las días*. It's *todos los días* ('everyday,' *día* is a masculine noun in Spanish)," and another rater commented on an incorrect verb conjugation: "[She said] *fuimos* ('we went'), it was the way she conjugated the verb…it wasn't [correct]." In this case, the speaker was discussing a trip her family had taken, and instead of conjugating the verb in the third person singular to agree with "family," a singular noun in Spanish, she used the first person plural, which seemed to confuse or at least distract the rater.

Comments for fluency (9%) and pronunciation (6%) patterned similarly. For example,

one rater indicated that he understood what the speaker intended to say even though he perceived a different word due to a mispronunciation: "It was clear that [the word] was *favorita* ('favorite'), but I heard *ahorita* ('right now')." In this case, downgraded comprehensibility could be attributed to a mismatch between top-down, contextually driven processing (i.e., the word that the rater anticipated based on previous information) and bottom-up, perceptual processing (i.e., the word that the rater perceived). Referring to the same stretch of speech, another rater remarked on a vowel quality issue: "*Cose, cose favorita* ('favorite, favorite thing,' but reducing the final vowel of *cosa* to schwa). I'm not sure, there isn't a pure vowel there, but I know that in English, there are two vowels that can be /a/, it's something like that. It wasn't *cosa*, it was *cose*." In this example, it seems that vowel reduction may have distracted the rater, increasing processing demands and reducing comprehensibility.

Table 5. *Frequency of Coded Comments (k) and Number of Raters (Out of 24) Contributing Comments Through Stimulated Interviews*

	Downgrade		Upgrade		No clicking		Total		
Coded category	*k*	Raters	*k*	Raters	*k*	Raters	*k*	%	Raters
Discourse	26	13	85	16	25	15	136	31	22
Lexis	46	13	13	8	14	9	73	17	23
Grammar	43	14	23	8	5	4	71	16	16
Fluency	13	9	20	7	5	5	38	9	15
Incomprehensible	27	15	0	0	2	2	29	7	15
Pronunciation	16	8	9	7	3	3	28	6	14
Listener experience	7	6	13	7	4	4	24	6	13

No comment	1	1	16	9	6	4	23	5	11
Forgot reason	4	3	2	2	1	1	7	2	5
Software use	3	3	1	1	1	1	5	1	4
Total	186		182		66		434	100	

Note. No comment = instances when a rater provided no reason for rating. Forgot reason = instances when a rater could not recall or articulate a specific reason.

Relevance of specific dimensions to comprehensibility. For ease of referencing, we will refer to individual raters by pseudonyms in this section. Although raters commented on multiple linguistic dimensions relevant to comprehensibility (see Table 5), in many instances, they overlooked micro-level issues with language use (e.g., incorrect verb conjugation, inaccurate or imprecise vocabulary, etc.) in favor of macro-level content and discursive issues so long as the main ideas were presented in a logical order that facilitated comprehension. This was particularly true when raters felt that a speaker's comprehensibility improved, with nearly half of reasons for upgrading the speaker (47%) related to discourse, as illustrated by the following comments:

- The conjugation isn't the best, but the main idea is there. (José León, semi-dynamic rater, discussing Speaker 3: childhood)
- The word order isn't right, but the ideas are logically sequenced, and that makes [the response] understandable even if the words aren't placed where they should be. (Juan Fernando, non-dynamic rater, discussing Speaker 1: studies)
- I'm going to stop here. Even though she has gender agreement errors, the idea is pretty easy to understand. (Diamel, semi-dynamic rater, discussing Speaker 3: childhood)

Both José León and Juan Fernando upgraded the speaker (+1 in both cases), which is indicative of the overall trend of upgrading the speaker when raters perceived the content to be logical,

coherent, and well organized. Although Diamel did not upgrade or downgrade the speaker at that precise moment, at the end of the file, she upgraded the speaker significantly (+5), stating that she did so because "everything was very easy to understand at the end."

Even though upgrading the speaker was more common for the discourse category, raters also downgraded the clip or alternated between upgrading and downgrading when they struggled to comprehend the relationship between events, or when they were waiting for the speaker to complete a thought, as shown in the following sequence from Esteban, a dynamic rater (discussing Speaker 2: studies):

- She was saying that it was very interesting for [her], but I was still lost because she was saying that biology was interesting for her… but there was a lapse, a gap in the construction of the conversation because I didn't know what she was studying. (–1 at 0:18 and –2 at 0:19)
- There I recovered, I began to tell myself, "Well, I'm following you, let's go, I'm not going to stop. Maybe, in context, I'll know what you're studying, at least I know it's something related, something related to biology." (+2 at 0:20)
- I saw the bar [the comprehension bar graph of the software] going up and down. Then I said to myself, "If she takes a long time to convey to me what she wants to say, during that time, there is no comprehension, right? No, there's not." I'm simply waiting for her to say something, but she takes a while to put the phrase together, and when she does, that creates a gap in the conversation, in the relationship that we have. But the two words that she said before hooked me back into the conversation [even though I was waiting] and helped me concentrate on what she was saying at that moment. (+1 at 0:21, +1 at 0:22, and + 3 at 0:23)

- I said to myself, “I’ve got it. Biology, chemistry, everything had, everything was coming together, the sciences.” (significant upgrading in the +4 to +5 range from 0:25 to 0:30)

At the same time, raters were not always able to recover the intended message, especially when the speaker abruptly changed topics. In this case, the rater was more likely to downgrade the speaker like Yanet did (–1):

- She didn’t finish her thought, and then I couldn’t figure out how it was related to what she said afterwards. I mean, [I didn’t know] where she was going, there was something that got lost in the middle, and the first idea, she wasn’t able to express what she wanted to say, and she didn’t know how to complete her thought. She abandoned that idea and started another. (Yanet, non-dynamic rater, discussing Speaker 2: childhood)

As is evident in these examples, raters were committed to understanding the speaker, trying to piece together an intended message that was sometimes disparate, both in terms of its content and the time within which it was delivered. Relatedly, over half of the raters recognized their own status as L2 learner-users and leveraged that perspective to process speakers’ responses. Like the discourse category, the rater’s experience category was oftentimes, but not always, associated with upgrading. For instance, Yeny responded positively (+1) to Speaker 2’s use of pause words:

- She used a pause word just like we [Spanish speakers who are learning French] do! When you’re presenting something, you start saying, “Uhm, well, let’s see.” That’s what she’s doing, and that’s exactly what you should do! It happens because we’re still learning [the language]. (Yeny, non-dynamic rater, discussing Speaker 2: childhood)

On the other hand, Aleli identified with Speaker 2 but nevertheless downgraded her due to the increased processing effort that was required:

- She [Speaker 2] didn't know the word, but I imagined, I put myself in her place and was able to understand her, but I still lowered her score because I had to think about what she wanted to say. (Aleli, non-dynamic rater, discussing Speaker 2: childhood)

Lastly, in one particularly insightful comment pertaining to Speaker 2, Yanet reflected on how she arrived at her ratings, even though she chose not to upgrade or downgrade comprehensibility:

- I asked myself, "Does that affect comprehension?" And then I thought, "It depends who's listening." In my case, it didn't bother me, because I knew she said "sand" [in English] because she was looking for the word in Spanish but couldn't find it. (Yanet, non-dynamic rater, discussing Speaker 2: childhood)

Despite raters' efforts to understand the speakers—and their general propensity to relate to the speakers as fellow L2 users—there were times when word choice, pronunciation, or grammatical constructions compromised their comprehension, and these categories were slightly weighted toward downgrading (see Table 5; a further example contrasting raters' comments in response to the same speaker's audio clip appears in Appendix E). For example, when Speaker 3 was describing her classes, she mentioned *gimia* (pronounced [himiə]), instead of *química*, "chemistry." Six of the eight raters who completed the stimulated interview targeting this speaker mentioned this particular issue, downgrading the speaker 1 to 3 points (n = 5 for –1, and n = 1 for –3). One rater interpreted the word to be a shortened form of *gimnasia* ([xim.ˈna.s̪ia]), "gym class/physical education," guessing that the speaker was taking a gym class, and one correctly guessed that the speaker was trying to describe a chemistry course. The remaining raters indicated that they simply did not know what she was describing. Interpreting the word was probably made more difficult by the fact that the speaker paused beforehand and said "uhm," and so some raters reported hearing *algimia* (approximately [əhimiə]).

In terms of grammar, speakers sometimes alternated between different tenses, confused verb endings, or both, making certain portions of their speech difficult to process, as was the case for Juan Fernando:

- It's the conjugation. She's not recounting what happened in the past, rather, it's as if she had gone to Hawaii many times, as if she had spent every Christmas in Hawaii. It's not just one thing, it's everything, the way she's describing it and the conjugations. (Juan Fernando, non-dynamic rater, describing Speaker 1: childhood, –1)

In some cases, it was not so much that comprehensibility was significantly impacted, but rather that language use errors were distracting:

- It's the changes in the verb tenses… they distract me even though I understand. (Mariet, non-dynamic rater, discussing Speaker 1: childhood, –1)

In sum, raters seemed to rely on discursive elements (i.e., the overarching organization and logic of ideas) to interpret speakers' responses, at times aided by their own experience as L2 speakers, but they did not hesitate to downgrade comprehensibility in the presence of salient and/or persistent grammatical, lexical, and pronunciation errors. In terms of the relevance of specific linguistic dimensions to enhanced or decreased comprehensibility, as summarized in Table 5, clearly organized ideas were likely to enhance comprehensibility, but the absence of discourse clarity did not necessarily cause comprehensibility to decline. In contrast, lexis and grammar were mostly cited to explain negative click activity, suggesting that raters were more apt to notice and penalize an error than they were to praise speakers for accurate or sophisticated language use. Pronunciation and fluency were cited relatively proportionately to explain positive and negative clicks.

Relationship Between Dynamic and Global Ratings

In response to the third question, which asked how dynamic and global assessments of comprehensibility related to each other, relationships between raters' dynamic and global ratings were explored using linear mixed-effects modeling. As a first step, the following listener background characteristics were integrated into separate models for the global ratings (overall scalar rating) and for the dynamic assessments (total number of clicks, total number of positive clicks, and total number of negative clicks): (a) familiarity with L2 Spanish speech (Familiarity, continuous, 9-point scale); (b) frequency of interaction with L2 Spanish speakers (Interaction, categorical, two levels: low = less than once a month vs. high = once a week or more); (c) teaching experience (Teaching, categorical, two levels: no vs. yes); and (d) previous training in linguistics (Linguistics, categorical, two levels: no vs. yes). Prompt was included as a fixed effect, and all models contained by-rater random intercepts. The only significant effect that emerged from these initial analyses was prompt. On average, the university studies prompt was rated as more comprehensible (estimate = 0.40, *SE* = 0.18, $p = .03$) and was upgraded more frequently (estimate = 9.81, *SE* = 3.89, $p = .01$). There was no statistically significant difference in the frequency with which the prompts were downgraded (i.e., in the number of negative clicks each prompt received).

Total click activity, positive click activity, and negative click activity were then incorporated into three separate models to examine the extent to which they predicted global comprehensibility scores. These more complex models were compared against the baseline global rating model described above by performing a Chi-square test on the difference in their deviance statistics. Neither total click activity ($\chi^2(1) = 0.001$, $p = .97$) nor positive click activity ($\chi^2(1) = 0.21$, $p = .65$) improved model fit. However, incorporating negative click activity marginally enhanced the model ($\chi^2(1) = 3.70$, $p = .06$). As reported in Table 6, according to

model estimates, downgrading the clip was associated with a .03 unit decrease in global comprehensibility. This effect is scalar, such that more frequent downgrading would be associated with an increasingly lower global rating.

Table 6. *Mixed-Effects Model Parameters for Global Comprehensibility Ratings*

Fixed effects	Estimate	*SE*	*t*	95% CI	*p*
Intercept	6.19	0.67	9.27	[4.83, 7.55]	< .001
Familiarity	–0.05	0.12	–0.44	[–0.29, 0.19]	.67
Interaction	0.52	0.61	0.87	[–0.72, 1.78]	.40
Teaching	–0.16	0.59	–0.27	[–1.36, 1.04]	.79
Linguistics	0.58	0.59	0.99	[–0.61, 1.78]	.33
Prompt (studies)	0.42	0.18	2.38	[0.07, 0.77]	.02
Negative clicks	–0.03	0.02	–1.94	[–0.07, 0.001]	.06
Random effects		*SD*			
Raters (intercept)		1.13			

Because the effect of downgrading just missed significance ($p = .06$), we undertook a follow-up analysis using a flattened dataset, taking into account only the directionality of listeners' clicks. For example, downgrading scores of –1, –3, and –5 (i.e., listeners who opted to downgrade the speaker once, three times, or five times at a given point in time) were converted to scores of –1, indicating that all speakers had downgraded the listener even though the magnitude of their response varied. We hypothesized that there might be a stronger relationship between downgrading, irrespective of the perceived magnitude of the drop in comprehensibility, and the global ratings. This hypothesis was supported: The model integrating the negative click

behavior significantly improved fit over the baseline ratings model ($\chi^2(1) = 4.92$, $p = .03$), and the coefficient for negative clicks representing the relationship between downgrading the speaker and global comprehensibility rating was stronger (estimate = –0.07, *SE* = 0.03, $p = .03$, 95% CI = [–0.14, –0.01]). As in previous analyses, the total number of clicks and the number of positive clicks were not significantly related to the global comprehensibility ratings in the flattened dataset.

Lastly, we examined in both datasets whether this effect—that is, whether the relationship between downgrading frequency and the global comprehensibility ratings—varied as a function of the speaker or the order in which the speakers were rated (e.g., if a stronger relationship was evident for the third than for the second speaker rated). Integrating those two variables into the model as interaction terms with the downgrading predictor did not significantly improve model fit over the simpler models, which suggests that the observed effect cannot be attributed to the specific speech characteristics of a particular speaker, nor to raters' experience with this type of rating paradigm as they moved through the speakers and prompts. In other words, it is not the case that raters' dynamic ratings became more strongly associated with their global comprehensibility ratings as they progressed through the rating task.

Discussion

This study sought to examine L2 comprehensibility as a dynamic construct and to clarify the extent to which different linguistic dimensions of speech are associated with changes in comprehensibility across time as the listener experiences L2 speech. Twenty-four native speakers of Spanish provided moment-to-moment comprehensibility ratings while they were listening to samples from three native English speakers (intermediate learners of L2 Spanish). As raters were evaluating the final speaker, the rating interface was video recorded, which served as the basis

for a subsequent stimulated interview to determine why raters chose to upgrade or downgrade the speaker at a particular moment. After the dynamic rating and stimulated interview, raters provided a global comprehensibility rating for each speaker and were debriefed.

L2 Comprehensibility and Its Linguistics Dimensions

Plotting comprehensibility curves for each rater revealed that response patterns were not necessarily driven by either the speaker or the prompt. Instead, curves appeared to depend on the particular response strategy that the rater adopted. Dynamic raters evaluated the speaker continuously and typically made use of the positive side of the continuum (i.e., +1 to +5), reserving the lower end of this band (i.e., +1 and +2) for decreasing comprehensibility and the upper end (e.g., +3 to +5) for increasing comprehensibility. Like their dynamic counterparts, the semi-dynamic raters frequently evaluated the speakers, but the range of their ratings was narrower, limited to ± 1 in most cases. The vast majority of raters, however, fell into the non-dynamic group. They evaluated comprehensibility far less frequently, often upgrading or downgrading the speaker only once or twice over the entire listening experience.

With respect to raters' explanations for their click activity, multiple categories were cited as reasons for upgrading or downgrading comprehensibility, including discourse organization, lexis and grammar, pronunciation, fluency, and raters' experience (i.e., in some cases, raters referenced their own status as L2 users and seemed to approach ratings from that particular perspective). On one hand, these results align with findings of previous research demonstrating that various aspects of discourse structure, lexis and grammar, and pronunciation all influence global comprehensibility judgments (e.g., Crowther et al., 2018). On the other hand, when click behavior was combined with the qualitative comments provided by raters during the stimulated interview, a richer picture of category use with respect to upgrading and downgrading the

speaker emerged. For example, whereas discourse was frequently cited as a reason for upgrading comprehensibility, lexis and grammar were more frequently associated with downgrading. In fact, raters were oftentimes willing and able to overlook lapses in language use (i.e., word choice, subject-verb agreement, verb aspect, etc.) when the overall flow of information was coherent, that is, when they perceived the discourse to be well organized. At the same time, certain language use issues were difficult to overcome, such as when the speaker used a word that the raters could not process (e.g., *gimia* vs. *química*, "chemistry"). In these instances, the error seemed to interrupt or interfere with raters' understanding of the overall response—for example, raters were unable to ascertain the relationship between *gimia* and the other science courses that the speaker was describing. In other words, the timing and gravity of the error conjointly influenced raters' response to the speech, such that in some cases, multiple raters converged with respect to downgrading the speech.

When dynamic click activity was integrated into mixed-effects models of the global comprehensibility ratings while controlling for rater background (i.e., raters' familiarity with L2 Spanish speech, their frequency of interaction with L2 Spanish speakers, and previous linguistic training and teaching experience), a negative relationship emerged between negative clicks (i.e., downgrading the speaker) and global comprehensibility score. Raters who downgraded the speaker more often tended to rate that speaker as less comprehensible. This was evident in datasets encoding (a) the direction and magnitude of click behavior (where a marginal effect emerged) and (b) only the direction of click behavior (where a statistically reliable effect obtained). Furthermore, models incorporating an interaction term with the speaker and order of speakers did not significantly improve fit, which suggests that the effect of downgrading the speaker did not vary depending on the particular speech characteristics of the speaker or the

order in which the speakers were rated. In contrast to these findings, relationships between the total number of clicks and global ratings and between positive clicks and global ratings did not reach significance. On the basis of these results, it seems likely that raters weigh periods of low comprehensibility more heavily in their global comprehensibility judgments, irrespective of their tendency to reward the same speaker for speech that is easy to understand. This novel finding aligns well with results of prior work showing that listeners tend to comment negatively rather than positively in relation to comprehensibility, particularly when discussing the performance of L2 speakers from linguistic backgrounds other than their own (Foote & Trofimovich, 2018) and evaluating comprehensibility of lower-proficiency L2 speakers (e.g., Kennedy et al., 2015).

Comprehensibility as a Dynamic System

If comprehensibility is a dynamic construct, then it should display some of the core properties of dynamic systems, including change over time, interconnectedness of elements, self-organization into preferred and dispreferred states, and non-linearity or threshold effects (de Bot et al., 2007; de Bot, Lowie, Thorne, & Verspoor, 2013). The dynamic approach (as applied to the current study) suggests that comprehensibility certainly changes over time and appears to display nonlinearity, insofar as the timing and the location of the error might produce a variable response in different listeners. This point was particularly salient in relation to specific errors, such as dysfluencies or lexical substitutions (e.g., *gimia*), which might not compromise comprehensibility in certain discourse contexts (e.g., early in a response) for some listeners but in other contexts might cause other listeners to question their understanding of the entire response. Similarly, as illustrated in Appendix E, where the ratings of two listeners were compared in response to the same clip, listeners may downgrade a speaker for a consonant or vowel substitution, leading to a case of local unintelligibility, but at the same time might

disregard major fluency issues, such as persistent pausing or numerous hesitations. By contrast, other listeners might overlook segmental substitutions but may lose focus because of ongoing disfluency.

With respect to the interconnectedness of elements, it seems that the strength and relationship of the linguistic dimensions of comprehensibility to one another (at least to the extent that listeners could comment on these dimensions through stimulated interview) also appears to vary from one listener to another, depending on how the individual has construed the listening task. Self-organization was evident for some listeners, in that their comprehensibility curves displayed longer plateaus, indicating that ratings gravitated toward a certain level of comprehensibility. For instance, for the dynamic raters, comprehensibility drifted toward the upper end of the scale, often resting at ceiling (+5) until an error caused a disruption. The tentative portrait that emerges from this exploratory analysis is dynamic. However, it must be acknowledged that only six of the 24 raters in this study showed a clearly dynamic rating pattern, while the remaining raters likely approached the task more holistically, upgrading or downgrading comprehensibility only rarely. In future research, it would be important to understand whether specific rating behaviors are associated with a particular strategy adopted by the rater or whether these behaviors reflect real-time processing demands for the rater (see Ludwig & Mora, 2017; Munro & Derwing, 1995b), such as the need to attend to multiple linguistic dimensions in a speaker's speech while providing its ongoing assessment. Future research will also need to clarify whether and how quickly listeners settle on optimal or non-optimal comprehensibility states and the extent to which comprehensibility is subject to nonlinearity by examining the aggregated effect of constellations of factors.

In the present study, certain linguistic categories and individual episodes appeared to act

as attractor states, in that multiple listeners became attuned to the same feature and responded similarly. For example, discourse structure was frequently cited as a reason for upgrading comprehensibility even for listeners who adopted different rating strategies, which extends prior research establishing links between discourse structure and L2 comprehensibility (e.g., Tyler & Bro, 1992). To that point, comparing the non-dynamic and dynamic raters (see Appendix E) revealed that they both responded positively to the list of university courses that the speaker described near the beginning of the clip in response to the studies prompt, and indicated that they upgraded comprehensibility because of the coherence of that particular stretch of speech. Just as discourse predominantly operated as a positive attractor, lexical and grammatical issues routinely elicited a negative response, and could be construed as comprehensibility repellers, in line with work showing links between grammar and lexis and L2 comprehensibility (e.g., Isaacs & Trofimovich, 2012; Munro & Derwing, 1995a). Yet, even individual episodes that were troublesome for groups of 3–5 individuals did not seem to pose much of a problem for the remaining listeners, which suggests that using inappropriate vocabulary or misconjugating a verb should be considered minor repellers whose contouring of the comprehensibility landscape may not induce an equal response from all interlocutors.

Establishing a Comprehensibility Baseline

One important conceptual issue that arises out of this research is how listeners establish a comprehensibility baseline. Instructions made it clear to raters that they should upgrade or downgrade speakers when they perceived a change in comprehensibility. Thus, raters may have established a somewhat unique interpretation of the overall or absolute level of comprehensibility that served as the baseline for their ratings. For instance, some raters may have hypothesized that speakers would be relatively difficult to understand, lowering their

comprehensibility threshold, whereas others may have made the opposite assumption. These perspectives could have had a cascade effect on ratings, affecting the frequency and magnitude of dynamic click activity. In their debrief comments, two raters noted that they had become stricter as they were moving from one prompt to the next for the same speaker, compared to six raters who indicated becoming more lenient. Likewise, moving from one speaker to the next, two raters noted that they became stricter and 13 more lenient. At the same time, reported frequency of interaction with L2 Spanish speakers and some training in linguistics did not seem to influence either the dynamic or global comprehensibility ratings. Taken together, these findings suggest that it is not prior experience with L2 speech in general that shapes listeners' perception of comprehensibility, but rather experience with a specific speaker's speech or possibly experience with speech from speakers of similar ability. Consequently, even though nearly half of raters reported daily or weekly interactions with non-native Spanish speakers, those interactions may have involved advanced L2 speakers using the language for daily communication, such that their production would differ both quantitatively and qualitatively from the intermediate-level learner speech that was evaluated in this study.

A second and related question that deserves attention is how much experience listeners need to establish a reliable baseline for dynamic ratings to proceed, and for that matter, for global ratings to be dependable (see Munro, 2018, for a similar argument). In the present study, average time to first click during dynamic ratings was 30 seconds, suggesting that about 30 seconds of speech may be needed to make an initial judgment, and it took raters, on average, about 40 seconds of listening to an audio clip (notably, even after having heard and evaluated the same clip previously) to provide a global, static rating. Although the timing of the global judgment varied widely across raters (0–153 seconds), because they were free to listen to as much or as

little as needed, this finding implies that, at minimum, a 30-second speech sample might be necessary for listeners to provide a judgment (at least for intermediate-level speakers exposed to the L2 in a classroom context), while more dependable ratings might require longer samples. Ultimately, more research is needed to understand how listeners construct an appropriate baseline when evaluating L2 speech. In particular, it would be useful to have raters evaluate comprehensibility on absolute and relative scales (i.e., overall level of comprehensibility vs. change in comprehensibility) to arrive at a more complete picture of listeners' perception of L2 speech over time.

Conclusion

Focusing on comprehensibility as a dynamic construct, this study sought to ascertain how listeners arrive at a global comprehensibility judgment and which dimensions of L2 speech listeners associate with moment-to-moment shifts in comprehensibility during the listening task. Findings demonstrate that negative click behavior during the dynamic ratings was negatively associated with global scores, suggesting that the incidence of low comprehensibility might be the primary determinant of global ratings regardless of the frequency with which the same speaker was rewarded for high comprehensibility. Many linguistic features were associated predominantly with one response pattern: Discourse (and fluency) were cited as reasons for upgrading comprehensibility while lexical and grammatical errors were mentioned as reasons for downgrading the speaker. Nevertheless, there was substantial individual variation in how listeners approached the rating task and in the speech features that elicited a response, both in terms of rating activity and comments provided during the stimulated interview. Overall, these results underscore the need to unpack the dynamic properties of comprehensibility, and possibly other dimensions of L2 speech. It would also be worthwhile to adopt a broader definition of

processing fluency (cf. Oppenheimer, 2008), which encompasses the construct of comprehensibility, to include desire or motivation to continue listening, an operationalization that could take into account both the linguistic features of speech and its content. Adopting such a definition could shed light on whether and to what extent dynamic ratings are associated with various communicative consequences for L2 interlocutors.

Notes

1. One reviewer pointed out that the dialect of the raters could have influenced their perception of the speech. We acknowledge this point and believe it would be advantageous for future research to explore this possibility. For the sake of this study, we tried to recruit raters representing a range of Spanish dialects to mirror the characteristics of the L2 speakers, all of whom had learned Spanish in the classroom and had themselves been exposed to multiple varities of Spanish through their instructors and textbooks.

2. One reason for reliability indices being lower compared to those typically reported in previous research (e.g., Derwing & Munro, 2015) could be that the experience of evaluating speech dynamically may have interfered with listeners' holistic judgments of each speaker's comprehensibility. Because this study focused on dynamic ratings (with static ratings used for comparison purposes), all listeners' data were included in subsequent analyses.

3. All direct quotes were translated from Spanish or French (reflecting the language in which the listener chose to provide comments).

References

Alter, A. L., & Oppenheimer, D. M. (2006). Predicting short-term stock fluctuations by using processing fluency. *PNAS, 103*(24), 9369–9372. doi:10.1073/pnas.0601071103

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*(3), 219–235. doi:10.1177/1088868309341564

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. doi:10.1111/flan.12285

Brennan, E. M., Ryan, E. B., & Dawson, W. E. (1975). Scaling of apparent accentedness by magnitude estimation and sensory modality matching. *Journal of Psycholinguistic Research, 4*(1), 27–36. doi:10.1007/BF01066988

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

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. doi:10.1017/S027226311700016X

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

de Bot, K., Lowie, W., Thorne, S. L., & Verspoor, M. (2013). Dynamic Systems Theory as a comprehensive theory of second language development. In M. D. P. García Mayo, M. J.

Gutierrez Mangado, & M. Martínez Adrián (Eds.), *Contemporary approaches to second language acquisition* (pp. 199–221). Philadelphia, PA: John Benjamins.

de Bot, K., Lowie, W., & Verspoor, M. (2007). A Dynamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition, 10*(01), 7–21. doi: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), 1–16.

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

Dörnyei, Z., & Tseng, W.-T. (2009). Motivational processing in interactional tasks. In A. Mackey & C. Polio (Eds.), *Multiple perspectives on interaction: Second language research in honor of Susan M. Gass* (pp. 117–134). London: Routledge.

Dragojevic, M., & Giles, H. (2016). I don't like you because you're hard to understand: The role of processing fluency in the language attitude process. *Human Communication Research, 42*(3), 396–420. doi:10.1111/hcre.12079

Flege, J. E. (1988). Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America, 84*(1), 70–79. doi:10.1121/1.396876

Foote, J., & Trofimovich, P. (2018). Is it because of my language background? A study of language background influence on comprehensibility judgments. *Canadian Modern Language Review*, *74*, 253–278. doi:10.3138/cmlr.2017-0011

Gibson, W., & Brown, A. (2009). *Working with qualitative data*. Thousand Oaks, CA: Sage Publications.

Gregersen, T., MacIntyre, P. D., & Meza, M. D. (2014). The motion of emotion: Idiodynamic

case studies of learners' foreign language anxiety. *The Modern Language Journal, 98*(2), 574–588. doi:10.1111/modl.12084

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*(2), 201-223. doi:10.2307/3588378

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10*(2), 135–159. doi:10.1080/15434303.2013.769545

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition, 34*, 475–505. doi:10.1017/S0272263112000150

Isaacs, T., Trofimovich, P., & Foote, J. A. (2018). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*, *35*, 193–216. doi:10.1177/0265532217703433

Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, *68*(1), 115–146. doi:10.1111/lang.12270

Kennedy, S. (2009). L2 proficiency: Measuring the intelligibility of words and extended speech. In A. G. Benati (Ed.), *Issues in second language proficiency* (pp. 132–146). London, UK: Continuum.

Kennedy, S., Foote, J. A., & Dos Santos Buss, L. K. (2015). Second language speakers at university: Longitudinal development and rater behaviour. *TESOL Quarterly*, *49*(1), 199–209. doi:10.1002/tesq.212

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical

data. *Biometrics*, *33*(1), 159–174. doi:10.2307/2529310

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology, 46*(6), 1093–1096. doi:10.1016/j.jesp.2010.05.025

Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*(3), 369–377. doi:10.2307/3588485

Ludwig, A., & Mora, J. C. (2017). Processing time and comprehensibility judgments in non-native listeners' perception of L2 speech. *Journal of Second Language Pronunciation, 3*(2), 167–198. doi:10.1075/jslp.3.2.01lud

MacIntyre, P. D. (2012). The idiodynamic method: A closer look at the dynamics of communication traits. *Communication Research Reports, 29*(4), 361–367. doi:10.1080/08824096.2012.723274

MacIntyre, P. D., & Legatto, J. J. (2011). A dynamic system approach to willingness to communicate: Developing an idiodynamic method to capture rapidly changing affect. *Applied Linguistics, 32*(2), 149–171. doi:10.1093/applin/amq037

MacIntyre, P. D., & Serroul, A. (2015). Motivation on a per-second timescale: Examining approach-avoidance motivation during L2 task performance. In Z. Dörnyei, P. D. MacIntyre, & A. Henry (Eds.), *Motivational dynamics in language learning* (pp. 109–138). Tonawanda, NY: Multilingual Matters.

Mackey, A., Park, H. I., & Tagarelli, K. M. (2016). Errors, corrective feedback and repair: Variations and learning outcomes. In G. Hall (Ed.), *The Routledge handbook of English language teaching* (pp. 499–512), New York, NY: Routledge.

Mercer, S. (2015). Dynamics of the self: A multilevel nested approach. In Z. Dörnyei, P. D.

MacIntyre, & A. Henry (Eds.), *Motivational dynamics in language learning* (pp. 139–163). Tonawanda, NY: Multilingual Matters.

Munro, M. J. (1998). The effects of noise on the intelligibility of foreign-accented speech. *Studies in Second Language Acquisition*, *20*(2), 139-154.

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). New York, NY: Routledge.

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

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38*(3), 289–306. doi:10.1177/002383099503800305

O'Brien, M. G. (2014). L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning, 64*(4), 715–748. doi:10.1111/lang.12082

Oppenheimer, D. M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: problems with using long words needlessly. *Applied Cognitive Psychology, 20*(2), 139–156. doi:10.1002/acp.1178

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12*(6), 237–241. doi:10.1016/j.tics.2008.02.014

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*(6), 619–635. doi:10.1016/j.jneuroling.2011.06.002

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition, 8*(3), 338–342. doi:10.1006/ccog.1999.0386

Saito, K., & Akiyama, Y. (2016). Video-based interaction, negotiation for comprehensibility, and second language speech learning: A longitudinal study. *Language Learning, 67*(1), 43–74. doi:10.1111/lang.12184

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. doi:10.1093/applin/amv047

Southwood, M. H., & Flege, J. E. (1999). Scaling foreign accent: direct magnitude estimation versus interval scaling. *Clinical Linguistics & Phonetics, 13*(5), 335–349. doi:10.1080/026992099299013

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

Tyler, A., & Bro, J. (1992). Discourse structure in nonnative English discourse: The effect of ordering and interpretive cues on perceptions of comprehensibility. *Studies in Second Language Acquisition, 14*(1), 71–86. doi:10.1017/S0272263100010470

van Geert, P., Steenbeek, H., & van Dijk, M. (2011). A dynamic model of expert-novice co-adaptation during language learning and acquisition. In M. S. Schmid & W. Lowie (Eds.), *Modeling bilingualism: From structure to chaos* (pp. 235–266). Amsterdam, The Netherlands: John Benjamins.

Varonis, M., & Gass, S. (1982). The comprehensibility of non-native speech. *Studies in Second Language Acquisition, 4*(2), 114–136. doi:10.1017/S027226310000437X

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System, 36*(1), 69–84. doi:10.1016/j.system.2007.11.004