Examining Rater Perception of Holds as a Visual Cue of Listener Nonunderstanding

Kim McDonough, Rachael Lindberg, & Pavel Trofimovich

Concordia University

McDonough, K., Lindberg, R., & Trofimovich, P. (2022). Examining rater perception of holds as a visual cue of listener nonunderstanding. *Studies in Second Language Acquisition*. Published online 14 February 2022. https://doi.org/10.1017/S0272263122000018

Abstract

This study examined whether university students perceive holds (i.e., a listener's temporary cessation of dynamic movement) as a visual cue of nonunderstanding. Conversations between English second language (L2) university students were sampled to extract episodes of other-initiated repair through open clarification requests (e.g., *what?*, *sorry?*). Brief, silent video clips were presented to 60 raters across two experiments who assessed the listener's comprehension, which was their perception about how well the listener had understood the speaker. Experiment 1 tested whether raters can differentiate between the onset and release of listener holds while Experiment 2 examined whether they are sensitive to the sequential organization of holds. Results indicated that raters clearly differentiated between hold onsets and releases and were sensitive to the temporal position of holds in the entire repair sequence. Taken together, these findings suggest that holds are a reliable signal of nonunderstanding with potential implications for L2 teaching and assessment.

Examining Rater Perception of Holds as a Visual Cue of Listener Nonunderstanding

The goal of interaction is to communicate successfully, which entails delivering messages that can be understood by an interlocutor as well as correctly perceiving an interlocutor's intended meaning. Remarkably, the vast majority of interaction occurs without any disruptions to the communication of meaning. However, sometimes a listener fails to understand a speaker's utterance and chooses to seek clarification, which is a type of communication breakdown called nonunderstanding. Having a repertoire of methods for seeking clarification is an important component of interactional competence, which refers to a speaker's ability to access, deploy, and adapt resources for the achievement of mutual understanding in a given interactional context (Roever & Kasper, 2018). In addition to verbal means of expression, interactionally competent interlocutors also deploy a wide range of nonverbal behaviors, such as eye contact, gestures, facial expressions, and posture. The importance of nonverbal behaviors for interactional competence has been recognized in second language (L2) assessment research. For example, prior research has shown that test-takers who were rated as linguistically weak but used nonverbal behaviors associated with active listening (e.g., head nodding and backchannel cues) were viewed as interactionally competent (Jenkins & Parra, 2003). In addition, rater perception studies have shown that nonverbal features of communication, such as eye gaze, facial expressions, gestures, and body language, contribute to authentic interaction and rater evaluations (Ducasse & Brown, 2009; May, 2011). In light of the importance of nonverbal behaviors within interactional competence, the current study examines the visual component of nonunderstanding episodes with clarification requests.

When nonunderstanding occurs, its resolution is locally accomplished through the collaboration and co-construction of meaning by interlocutors (Firth, 1996; Wagner, 1996).

Nonunderstanding has been studied through the focus on repair, which includes practices for interrupting ongoing conversation to deal with problems in speaking, hearing, or understanding (e.g., Schegloff, 1997, 2007; Schegloff et al., 1977), to examine how interlocutors use both verbal messages and nonverbal behaviors to remediate problems. An example of nonunderstanding in which the listener initiates repair through a clarification request is provided in Example 1.

Example 1. Nonunderstanding episode

P61: I'm assuming you're a little older?

P62: Sorry?

P61: How old are you?

P62: I'm twenty-nine.

When the listener (P62) failed to understand the speaker's (P61) initial utterance, she initiated repair through a general or open clarification request. Within the repair sequence, the listener's verbal repair cue (*sorry?*) serves as the first part of an adjacency pair that initiates action, while the speaker's response (*how old are you?*) is the second part that carries out the repair. The resolution of nonunderstanding is demonstrated when the listener provides her age in the final turn. The fact that P61 reformulates her initial question in the third turn indicates that P62's request for repair was understood as such, which exemplifies the next-turn proof procedure for providing evidence of participant understanding of repair practices (Edwards, 2004; Sidnell, 2014).

As defined by Schegloff and Sacks (1973), adjacency pairs, such as the request for clarification and response illustrated in Example 1, consist of two turns produced by different speakers that are adjacent and ordered so that the first part necessarily precedes the second part.

The two parts are related such that certain types of responses are expected, such as an invitation followed by either acceptance or refusal. In Example 1, the first adjacency pair began with P61's query about P62's age in Turn 1. However, the listener was not able to complete that pair with an answer about her age because she failed to understand the question. This nonunderstanding triggered the insertion of an expansion adjacency pair in the form of repair, which is sequentially ordered with the request for clarification (Turn 2) followed by a reformulation of the speaker's question (Turn 3). The second part of the original adjacency pair (i.e., answering the question) is only given in Turn 4 after the inserted repair sequence was complete. As described by Stivers (2013), analyzing the sequential organization of conversation, such as adjacency pairs, is a key tenet of conversation analysis distinguishing it from other approaches to interaction that examine utterances in isolation. When identifying conversational practices, such as repair practices, a key goal is to identify features that have distinctive characteristics, appear in specific locations in a turn or sequence, and serve meaningful actions (Heritage, 2011).

In studying the practices of repair, conversation analysis researchers have pointed out that repair sequences like the one illustrated in Example 1 often occur with nonverbal behaviors that also follow a sequential organization. For example, Seo and Koshik (2010) analyzed tutoring sessions between university students for whom English was either their first language (L1) or their L2, reporting that listeners used two types of head movements when initiating repair: (a) a sharp head tilt or turn to the side with eye gaze and (b) a head poke (i.e., extending the head forward) accompanied by a forward lean. The listener initiated the movements after the speaker completed the utterance with the problematic feature and held the position until the problem had been resolved. Although these held movements most often co-occurred with verbal repair initiators (e.g., *huh*? *sorry*?), there were episodes in which the visual cue initiated repair in

isolation. Also focusing on English interaction, Kendrick (2015) similarly described two visual-only repair sequences in which either a lateral head tilt or a frown was sufficient to initiate repair, although those visual cues more typically co-occurred with a verbal repair initiator.

Turning to nonverbal components of repair in other languages, Floyd et al. (2016) found that listeners in Northern Italian, Cha'palaa, and Argentine sign language who initiate verbal other-repair often temporarily hold a dynamic movement static, which the researchers refer to as holds. For the two spoken languages specifically, the behavior held static was most often eye gaze, followed by head direction (left/right), upper body lean, eyebrow position, and head position (up/down). Their analysis of the sequential organization of the holds found that listeners initiated a hold (i.e., hold onset) and maintained it through the end of their clarification request, and they disengaged the hold (i.e., hold release) during or shortly after the speaker performed the repair. Forward leans have also been found in Mandarin other-initiated repair in the form of intervening questions (i.e., repair initiated during rather than after the speaker's problematic utterance), with listeners leaning forward and holding their lean until a response is provided (Li, 2014). These cross-linguistic findings are similar to the role of head movements and forward leans in English repair sequences identified previously (Kendrick, 2015; Seo & Koshik, 2010) and provide further evidence that the onset and release of the held movements signal the beginning and resolution of nonunderstanding, respectively.

Additional studies have provided evidence for the nonverbal component of repair across languages. For example, in Swiss German sign language, turn-final holds are released when the listener has understood the speaker or the speaker has acknowledged the listener's request (Groeber & Pochon-Berger, 2014), which provides additional evidence that releasing a temporarily static movement is a signal of resumed understanding. The cessation of movement

during repair initiation has also been found in Argentinian sign language (Manrique, 2016) and Yélî Dnye (Levinson, 2015) in the form of a freeze look. In these nonunderstanding episodes, the listener initiates repair by staring at the speaker without moving and maintains the freeze until the problem has been resolved or the listener pursues repair verbally. In sum, although repair initiation typically occurs through both verbal and nonverbal components, researchers acknowledge that some repair initiation utilizes primarily nonverbal resources (Dingemanse, 2015; Levinson, 2015; Manrique, 2016; Seo & Koshik, 2010).

The extensive conversation analysis research has provided valuable insight into the nonverbal behaviors associated with repair practices, specifically the types of movements that are held static during listener holds and their sequential organization as onsets and releases. By identifying the nonverbal signals of repair practiced by multiple speakers of different languages in diverse conversational settings, these researchers have demonstrated generality in repair practices, which can be understood as the extent to which practices are organized in the same way across contexts (see Chenail, 2010, for discussion of generalizability and related constructs in qualitative research). Inspired by this line of research, we were also interested in generality and carried out a series of studies that examined whether nonverbal aspects of repair practices, specifically clarification requests (McDonough et al., 2019, 2021) and recasts (McDonough et al., 2015, 2020a, 2020b), were organized similarly in conversations between university students.

Besides providing evidence of generality, however, we were also interested in exploring whether these nonverbal behaviors are distinctive characteristics of repair practices. If specific visual cues (such as a head poke or forward lean) are uniquely associated with nonunderstanding, then they should not occur when a listener has understood the speaker. In such understanding

episodes, a listener might ask a follow-up question rather than initiate repair, as illustrated in Turn 2 of Example 2.

Example 2. Understanding episode

P230: Yeah it's good for me now but yeah

P229: Did you like French?

P230: It's really hard. It is harder than English

P229: Yes

Unlike the first part of an adjacency sequence in Example 1, the listener's follow-up question in Example 2 (*did you like French?*) does not initiate a repair sequence. The speaker's response in Turn 3 completes the adjacency pair by providing an answer to the question, which indicates that the follow-up question was understood as a request for additional information as opposed to clarification. Since there was no breakdown in the communication of meaning, the listener is unlikely to deploy a hold during the follow-up question if holds are uniquely associated with nonunderstanding. By comparing the listener visual cues that occur during both understanding and nonunderstanding episodes, we aimed to identify whether holds and other visual cues previously identified in repair sequences are distinctive (Heritage, 2011), in the sense that they are uniquely and reliably associated with nonunderstanding.

Conversation analysis researchers would likely adopt a micro-analytic approach to address the question of distinctiveness, such as by comparing the nonverbal behaviors that occur during repair sequences and follow-up questions. However, as primarily quantitative researchers, we adopted an alternative approach that elicited the perceptions of naïve observers to determine whether they can differentiate between understanding and nonunderstanding episodes. Clearly, visual cues of nonunderstanding are "real" because interlocutors respond to them by

reformulating their prior utterances, and conversation analysts have used next-turn proof procedures to document their occurrence. Our question was whether these cues are sufficiently distinctive that they can be perceived by external observers from the same speech community as the interlocutors (henceforth, raters), which in this case was university students. As pointed out by Toerin (2014), quantitative research that applies the findings of conversation analysis typically explores the association between specific interactional practices and other aspects of the social world. Reflecting this orientation, our work explores whether the nonverbal behaviors of nonunderstanding have implications for L2 teaching and assessment by first demonstrating that these behaviors can be perceived. If members of a speech community can neither detect a nonverbal behavior nor associate it with a distinct interactional meaning, then this would raise doubts about its potential relevance or application to broader issues in L2 learning.

Adopting this methodological orientation, McDonough et al. (2019) compared listener visual cues and rater perceptions of understanding and nonunderstanding episodes from conversations between L2 English speakers (N = 21) and a French–English bilingual listener who had been instructed to provide feedback as appropriate. Analysis of video-recorded conversations revealed that nonunderstanding episodes contained more listener holds and head nods than the understanding episodes, which provided evidence of the generality of nonverbal repair practices. Next, students (N = 66) from the same university were randomly assigned to rating conditions that manipulated access to the speaker's voice (clear or distorted) and the listener's face (clear or blurred) to rate speaker comprehensibility (*Hard for me to understand* and *Easy for me to understand*) and listener understanding (*He understood 0%* and *He understood 100%)* using a 100-millimeter scale. Although decontextualizing and manipulating the videos poses challenges to the ecological validity of the interactions, the experimental control

allows for the identification of the unique contribution of nonverbal behaviors (i.e., what the visual component adds to the verbal repair signal). The ratings showed that raters with access to the listener's face rated listener comprehension lower during nonunderstanding episodes than raters who only heard the speaker's voice. Put simply, seeing the listener's face provided the raters with visual information that helped them determine when the listener had trouble understanding the speaker. However, as an exploratory study, the findings were based on the behavior of a single listener who had been asked to provide feedback, which limited the generalizability of the study's findings.

To confirm the association between holds and nonunderstanding and explore the salience of visual cues to raters, McDonough et al. (2021) carried out a replication study drawing on a corpus of conversations between L2 English university students. Analyzing the transcripts, they identified 79 nonunderstanding episodes of the same type tested in the initial study. They then analyzed the video-recordings to determine whether those episodes contained holds and other visual cues identified in the initial study (e.g., head nods, blinks). They selected a subset of those episodes (n = 35) for rating and paired them with an understanding episode (n = 35) from the same interlocutors. Students at the same university (N = 90) rated the 70 episodes in terms of speaker comprehensibility and listener comprehension using the same sliding scales, with raters randomly assigned to conditions that manipulated access to the speaker's voice and face as in the initial study. Both the analysis of the 79 episodes and the ratings of the 35 matched episodes confirmed the association between holds and nonunderstanding reported in conversation analysis studies and the initial exploratory study. New analysis to classify holds based on the type of held movements, where some holds involved a single movement while others had multiple movements, revealed that 67% of the holds included a head movement (e.g., tilts, pokes, turns)

while 40% had an open mouth, and 37% had a forward lean. Although raters clearly recognized that listeners had comprehension difficulties for nonunderstanding episodes, they could differentiate between understanding and nonunderstanding episodes equally well through access to the speaker's voice or the listener's face or both, which raises questions about any additive benefits for visual cues when assessing listener comprehension.

Taken together, the findings of the two rating studies with university students (McDonough et al., 2019, 2021) confirm that L2 English university students clearly recognize their peers' holds and associate them with listener nonunderstanding, which confirms the observations of conversation analysts. However, the extent to which those holds provide additional meaningful information beyond the listener's verbal repair initiators remains unclear due to the conflicting findings for rating conditions. Although holds were uniquely associated with nonunderstanding, their occurrence did not consistently aid observers in detecting challenges with listener comprehension. Previous studies demonstrated that some repair initiation occurs visually only (Dingemanse, 2015; Levinson, 2015; Manrique, 2016; Seo & Koshik, 2010), which suggests that the nonverbal cues of nonunderstanding in isolation are meaningful enough to elicit repair between interlocutors. It is unknown, however, if such signals are sufficiently useful for identifying nonunderstanding to warrant pedagogical interventions to raise L2 speakers' awareness of nonverbal components of repair practices.

In summary, previous conversation analysis research has provided rich information about the occurrence and sequential organization of holds and other nonverbal features of repair initiation, and subsequent quantitative studies have confirmed that L2 English university students uniquely associate holds with problems in nonunderstanding. Although visual only repair initiation (i.e., holds and freeze looks) has been shown to occur during conversation, previous

research has not specifically examined if external observers can recognize them as signals of nonunderstanding when presented without any speech. If holds communicate meaning visually, then the nonunderstanding that they convey should be detectable even in the absence of the speaker's initial utterance or the listener's clarification request. To test this possibility, the current study presents silent videos showing holds during other-initiated repair with clarification requests (e.g., *sorry*, *huh*) and elicits rater perceptions about the listener's comprehension (i.e., to what extent the listener appeared to understand the speaker) in two experiments. Reflecting the sequential organization of holds, Experiment 1 tests raters' ability to differentiate between hold onsets that signal a problem versus hold releases that indicate a return to understanding. To further test the association between holds and nonunderstanding and the importance of their sequential organization, Experiment 2 tests raters' ability to differentiate among understanding episodes and to distinguish holds presented in their naturally-occurring four-turn sequence and those presented in reversed order. If the meaningfulness of holds is linked to the sequential order of onsets and releases, then raters should be more successful at identifying problems with listener comprehension when the holds appear in their naturally-occurring sequence. Based on prior research that elicited rater perceptions (McDonough et al., 2019, 2021), we predicted that perceived listener comprehension would be lower for hold onsets (as compared to hold releases) and lower in naturally-occurring hold episodes (as opposed to understanding episodes or reversed hold episodes).

Experiment 1

Conversation Corpus Overview

The videos rated in the current study were drawn from the Corpus of English as a Lingua Franca Interaction (CELFI), which consists of 224 paired conversations between L2 English

students at Montreal-area universities (McDonough & Trofimovich, 2019) with most of them studying at Concordia University (67%). As students, they had met a minimum English proficiency requirement to be admitted to their universities, which was a TOEFL iBT score of 75 or equivalent plus university EAP language courses. When asked to report their latest standardized proficiency test results, 62% of the students reported scores from the TOEFL iBT (*Mdn* = 110, *IQR* = 21) or IELTS (*Mdn* = 7, *IQR* = 1) tests. Based on the minimum requirement and the range of reported proficiency test scores, the students in the CELFI corpus range from the B2 to C2 levels in the Common European Framework of Reference. Students were randomly assigned to carry out three communicative tasks (posing solutions to problems encountered when moving to a new city, a close-call narrative, and an academic discussion task) with someone from a different L1 background. The self-reported gender composition of the pairs was controlled so that there were approximately the same number of male–male, female–female, and female–male dyads. The students' interaction while carrying out the three tasks was audio- and video-recorded, 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. These data were collected as part of CELFI, but only transcripts of the audio-recordings and video extracts from their conversations were used for the two experiments reported here.

Sampling Nonunderstanding Episodes from CELFI

All 224 CELFI transcripts were analyzed for nonunderstanding episodes that consisted of the four-turn sequence: (a) the speaker's initial utterance (Turn 1), (b) the listener's nonspecific,

open clarification request, such as *sorry*, *pardon*, *what*, or *huh* (Turn 2), (c) the speaker's repair (Turn 3), and (d) the listener's response showing understanding (Turn 4). Example 3 illustrates a nonunderstanding episode in which the listener requests clarification of the speaker's initial question in Turn 2 (*sorry?*) after which the speaker rephrases the question in Turn 3 and the listener answers the question in Turn 4.

Example 3: Nonunderstanding episode

P294: yeah... do you need to take course in your master?

P293: sorry?

P294: do you need to take courses or you only do research?

P293: no I–mine is course based masters

This analysis identified 139 listeners in the corpus (139/448 or 31%) who produced at least one clarification request of this type.

To ensure comparability across the listeners' nonunderstanding episodes to be used in this experiment, the following inclusion criteria were applied: (a) the speaker's initial utterance contained at least three words; (b) there was minimal speaker–listener overlap between turns; (c) the hold onset occurred in Turn 2; (d) the hold release occurred in Turn 4; and (e) the hold movement (e.g., head tilt or forward lean) was controlled. Application of the inclusion criteria led to the selection of 25 nonunderstanding episodes with four types of holds: forward lean only (5), forward lean with raised eyebrows or smile (7), head poke only (6), and head poke with raised eyebrows or smile (7). In terms of their background information, the listeners for these hold videos (14 women and 11 men) were students in undergraduate (56%) and graduate (44%) degree programs and spoke 13 different L1s with the most frequent being Mandarin (28%), Tamil (16%), Farsi (12%), and Bengali (8%). They ranged in age from 18 to 29 with a mean age

of 22.6 years (*SD* = 2.8). They had been living in Canada for a mean of 2.9 years (*SD* = 4.6) and had studied English for a mean of 14.2 years (*SD* = 4.8). Their reported proficiency test scores were similar to the median values for the students in the larger corpus.

Rating Stimuli

To create the hold onset and release videos, the four-turn nonunderstanding episodes were extracted using video editing software (VideoPad) into two clips. The 25 hold onset videos showed the listener from the last second of Turn 1, the hold onset in Turn 2, and the first second of Turn 3 when the hold was maintained. The 25 hold release videos showed the listener from the last second of Turn 3 with the hold, the hold release, and the remainder of Turn 4. As the fourth turn varied in length across episodes, it was cut at a natural speaking point to be the same length for all release videos (~2 seconds). On average, the hold onset videos were 3.76 seconds long (*SD* = 0.91), and the hold release videos were 3.08 seconds long (*SD* = 0.74). As the video clips were short, a 3-second countdown was added to the beginning of each one to allow raters time to prepare for the start of the video. The videos were presented without sound so that no verbal contributions from the speaker or listener could influence raters' judgements of listener comprehension. If raters heard the listeners request clarification (e.g., *sorry? what? pardon me?*), then it would clearly indicate that the listener had not understood, and the raters could give low comprehension scores without considering the listeners' visual cues. Without sound, however, the raters could only orient to the listeners' nonverbal behavior when assessing their degree of comprehension.

Raters

Raters included 30 students (21 females, 9 males) recruited from the same Montreal universities as the listeners in the videos on the assumption that they would represent the same

student population (i.e., potential peers of the students in the videos). They were undergraduate (67%) and graduate (33%) students between the ages of 19 and 41 (M = 25.03, SD = 5.74). Their L1s included Canadian or World Englishes (11), Portuguese (4), Arabic (3), French (3), Mandarin (2), Spanish (2), Tamil (2), Manipuri, Hebrew, and Danish. The L2 English raters had been studying English on average for 17.44 years (SD = 6.20) and the non-Canadian born raters reported a mean length of residence of 3.8 years (SD = 6.1). As compared to the listeners in the hold videos, the English L2 raters had a similar length of residence in Canada, similar amount of prior English study, and equally diverse L1 backgrounds. The greater proportion of raters from English L2 backgrounds (63%) reflected the distribution of English L2 (52%) and English L1 (48%) students at Concordia University where the majority of the listeners and raters were studying, and the linguistic diversity of Montreal where only 16% of the population reports English as their L1 and only 23% report using English as their predominant home language (Statistics Canada, 2017).

Rating Materials and Procedure

The entire procedure was conducted using the LimeSurvey online interface (https://www.limesurvey.org), where raters first completed the consent form (2 minutes), and then were given instructions for the rating procedure and explanations of the rating criteria (2 minutes). After practicing rating two video clips from listeners whose data were not included in the study (2 minutes), the 50 target video clips were presented to raters in a unique random order. Each video appeared on a separate survey page and played automatically, allowing the rater to only view it once. Below the videos were 100-millimeter slider scales which raters used to evaluate the listener's comprehension (i.e., how much they thought the student in the video understood the speaker), which was the key variable of interest for this experiment. The

endpoints for the comprehension sliding scale were *this student understood 0%* (negative endpoint on the left side) and *this student understood 100%* (positive endpoint on the right side). The initial slider position was set in the middle of the scale. An additional sliding scale was used to elicit the raters' perceptions about how easily the listener seemed to understand the speaker (*extremely difficult* and *extremely easy*), which was intended to capture the listeners' processing effort. A third sliding scale was used to check whether the edited videos looked natural (*extremely unnatural* and *extremely natural*). After the video rating task (20 minutes), the raters filled out a background questionnaire (5 minutes), a personality test (3 minutes), and facial expression recognition test (15 minutes). The analysis focuses on the listener comprehension ratings only because they provide the most direct assessment of whether raters associated listener holds with a lack of understanding.[1] Participants were remunerated $30 for their time.

Results

Prior to addressing the research question, we first examined whether the raters were consistent in their evaluation of listener comprehension by calculating the two-way mixed average-measures intraclass correlation coefficients for the hold and release videos. The coefficient was .93 for both video types, which indicates a high level of agreement across raters. To obtain one listener comprehension score for hold videos and one score for release videos for each rater, we obtained a mean score by summing the ratings and dividing by total videos for the hold and release videos separately. To determine whether raters can recognize the hold onset as a signal of nonunderstanding and the hold release as a signal of resumed understanding, their listener comprehension ratings were compared. Raters assessed listener comprehension lower in the onset videos (M = 34.93, SD = 9.40) than the release videos (M = 68.70, SD = 14.21). A paired-samples t test indicated that the difference was statistically significant, $t(29) = 13.06$, $p <$

.001, $d = 2.80$, with a large effect size ($d \geq 1.40$) based on benchmarks for applied linguistics research (Plonsky & Oswald, 2014).

To explore whether their perceptions about listener comprehension varied by hold type, we compared the raters' hold onset ratings for episodes with leans, head pokes, leans with facial expressions, and head pokes with facial expressions. As shown in Table 1, the raters provided the lowest listener comprehension ratings for lean holds, followed by head poke holds, lean holds with facial expression, and head poke holds with facial expression.

Table 1

Listener Comprehension Ratings by Held Behavior (Out of 100)

	Comprehension	
Hold type	*M*	*SD*
Lean ($k = 5$)	23.31	9.88
Head poke ($k = 6$)	33.71	13.47
Lean with facial expression ($k = 7$)	35.15	11.29
Head poke with facial expression ($k = 7$)	41.64	9.60

A repeated-measures ANOVA (sphericity assumed) indicated that there was a statistically significant difference in perceived listener comprehension ratings, $F(3, 87) = 29.55$, $p < .0001$, partial $\eta^2 = .51$. Post hoc comparisons with a Bonferroni adjustment indicated that there were significant differences ($p \leq .015$, $d \geq 0.88$) for all paired comparisons except for head poke versus lean with facial expression ($p = 1.00$, $d = .12$) (see Table 2).

Table 2

Post hoc Tests for Comprehension Ratings by Hold Type

Hold type	*M* (*SD*)	Hold type	*M* (*SD*)	*t*(29)	*p*	*d*
Head poke	33.71 (13.47)	Lean	23.31 (9.88)	4.87	.001	0.89
		Head poke with facial expression	41.64 (9.60)	–4.36	.001	0.80
		Lean with facial expression	35.15 (11.29)	–0.74	.467	0.13
Lean	23.31 (9.88)	Head poke with facial expression	41.64 (9.60)	–10.95	.001	2.00
		Lean with facial expression	35.15 (11.29)	–5.27	.001	0.96
Head poke with facial expression	41.64 (9.60)	Lean with facial expression	35.15 (11.29)	3.31	.002	0.61

Discussion

To summarize the findings of Experiment 1, these raters clearly perceived hold onsets as a signal of a listener's difficulty comprehending the speaker and hold releases as signalling a return to understanding. Thus, it appears that English L1 and L2 raters from the same university community can interpret L2 English speakers' holds accurately as providing visual signals of both the initiation and resolution of nonunderstanding. The findings support the results of prior rating studies that reported an association between holds and nonunderstanding (McDonough et al., 2019, 2021) and provide evidence that raters associate hold releases with a return to understanding, which has not been tested previously. Furthermore, analysis of the specific hold

types revealed that the raters attributed the lowest comprehension ratings to body lean holds and head poke holds. When these held movements also occurred with held facial expressions, comprehension ratings were higher. It is possible that the facial expressions, such as smiling, might dilute the nonunderstanding signal of the lean or head poke because smiling is often interpreted as a sign of understanding (McDonough et al., 2021). Indeed, when asked to explain what visual cues they based their ratings on, 16 raters mentioned smiling as a signal of understanding as compared to only three raters who stated that it was a sign of nonunderstanding. In addition, smiling might occur with a variety of nuanced expressions that are temporally linked to the verbal utterances in ways that communicate unique meanings, which could be confirmed through micro-analytic techniques.

Raters' ability to differentiate between hold onsets and releases provides a possible explanation for the null findings for rating condition (speaker's voice vs. listener's face) reported in McDonough et al. (2021). By providing raters with a visual image of the listener's face during the speaker's initial utterance only (i.e., before the hold onset), their rating stimuli failed to present the hold onset in isolation or in its naturally-occurring sequence, thereby likely diminishing its impact. It is also possible that holds are more meaningful as a sign of nonunderstanding when raters have access to their entire sequential organization with both the onset and release. Although Experiment 1 demonstrated that raters can differentiate between hold onsets and releases when they are presented in isolation, it is not known whether raters are sensitive to their sequential organization across an entire four-turn sequence. Based on the key premise of conversation analysis that talk is sequentially organized, it seems plausible that raters could differentiate between four-turn visual sequences with a hold (i.e., nonunderstanding episodes) and without a hold (i.e., understanding episodes). Furthermore, as raters clearly

differentiated between the visual cues associated with a hold onset and release, their ability to interpret hold episodes should be greater when those two signals are presented in their naturally-occurring sequence (i.e., in Turn 2 and Turn 4, respectively) as opposed to the opposite order.

To test these possibilities, Experiment 2 compared rater perceptions about listener comprehension for four-turn sequences that showed listeners' visual cues either of understanding episodes (i.e., no hold) or nonunderstanding (with hold). We expected that raters would assign lower comprehension ratings to the nonunderstanding episodes as these included listener holds. To test raters' sensitivity to the sequential organization of holds, we also elicited their perceptions about listener comprehension in nonunderstanding episodes when the hold onset and release are reversed. Because the meaning of a hold is indicated by both its onset and release in that order, we predicted that perceived listener comprehension would be lower when the hold was presented in its naturally-occurring sequence as opposed to the reversed order.

Experiment 2

Sampling Episodes from CELFI

As in Experiment 1, episodes were sampled from the CELFI corpus of L2 English speakers. Because of the narrower focus on the sequential organization of holds, the initial episode pool consisted of 42 transcripts identified during Experiment 1 as having a nonunderstanding episode with a hold onset in Turn 2 and a release in Turn 4. We returned to those 42 transcripts to locate all listeners who engaged in (a) a second nonunderstanding episode and (b) an understanding episode. Example episodes from the same listener (P62) are provided in Table 3. Whereas the listener requested clarification of the speaker's initial utterance in Turn 2 of both nonunderstanding episodes (*sorry?*), she asked a follow-up question in Turn 2 of the understanding episodes (*how they published that?*).

Table 3

Sample Nonunderstanding and Understanding Episodes

Turn	ID	Nonunderstanding 1	Nonunderstanding 2	Understanding
1	P61	I'm assuming you're a little older?	They have dead bodies at Concordia too	All their stuff was published without their consent
2	P62	Sorry?	Sorry?	How they published that?
3	P61	How old are you?	They have dead bodies at Concordia	I don't know. Uh cuz wait, consent isn't required to publish a genome
4	P62	I'm 29.	Oh really?	Oh.

Finally, the videos of the new episodes were analyzed to ensure that (a) the nonunderstanding episodes depicted a hold onset in Turn 2 and hold release in Turn 4 and (b) the understanding episodes did not include holds. This process identified 12 listeners who each contributed one understanding and two nonunderstanding episodes that met the criteria for a total of 36 episodes. Six of these 12 listeners had contributed one nonunderstanding episode to Experiment 1 rating stimuli. In terms of their background information, the listeners (50% women) were students in undergraduate (50%) and graduate (50%) degree programs and spoke six different L1s with the most frequent being French (33%), Mandarin (17%), Tamil (17%), and Farsi (17%). They ranged in age from 18 to 29 with a mean age of 23.1 years (*SD* = 3.6). They had been living in Canada for a mean of 4.0 years (*SD* = 4.6) and had studied English for a mean of 13.6 years (*SD* = 4.0). Their reported proficiency test scores were similar to the median values for the entire corpus.

Rating Stimuli

One nonunderstanding episode from each listener was randomly assigned to remain in its naturally-occurring sequence with the hold onset in Turn 2 and the hold release in Turn 4. The other nonunderstanding episode was edited to present the turns in the reverse order. The reversed episodes were created using DaVinci Resolve to first split the video clip into four turns and then create an episode with the turns in reverse order (i.e., 4–3–2–1). This order was selected because it showed the release before the hold, which is contrary to their naturally-occurring sequence, but avoided placing the hold onset in the final position, which is a privileged position in memory tasks because of the distinctiveness of the material experienced at the end of sequences or lists (Kelley et al., 2013, 2015).[2] To soften the abruptness of the areas where the video was cut and re-spliced so that the image would not "jump" between turns, the "smooth cut transition" setting was used, which seamlessly blended the re-spliced elements together. The understanding episode videos were not manipulated. The reversed hold videos had a mean length of 5.92 seconds (*SD* = 2.60), the intact hold videos were 6.21 seconds long (*SD* = 8.5), and the videos without holds were slightly longer with a mean of 8.92 seconds (*SD* = 3.28). All videos were silenced, and a 3-second countdown was added to the beginning of each clip.

Raters

The 30 new raters for Experiment 2 (57% women) represented the same speech community as those in Experiment 1 and the CELFI corpus, namely, linguistically-diverse university students in Montreal. They were studying in undergraduate (53%) and graduate (47%) degree programs and ranged in age between 19 and 35 (*M* = 25.17, *SD* = 3.97). Their L1s included Canadian or World Englishes (10), French (5), Bengali (3), Hindi (2), Mandarin (2), Turkish (2), Vietnamese (2), Cantonese, Punjabi, Tamil, and Urdu. The L2 English raters had

been studying English on average for 17.9 years (*SD* = 5.2) and the non-Canadian born raters reported a mean length of residence of 4.8 years (*SD* = 5.1). As compared to the listeners in the hold videos, the English L2 raters had a similar length of residence in Canada, similar amount of prior English study, and equally diverse L1 backgrounds. As in Experiment 1, there was a greater proportion of English L2 raters (67%), which represents the linguistic diversity of both the university student community (which has a relatively equal percentage of L1 and L2 English speakers) and the city of Montreal (where less than 25% of the population uses English as their predominant home language).

Rating Materials and Procedure

Just as in Experiment 1, the entire rating procedure was administered online using LimeSurvey (https://www.limesurvey.org). After completing the consent form (2 minutes), raters were given instructions and explanations of the rating criteria (2 minutes). The same sliding scale from Experiment 1 was used to assess the listener's comprehension (*this student understood 0%* and *this student understood 100%*). The sliding scales for naturalness and ease of understanding from Experiment 1 were also used, but are not reported for the main analysis.[3] After reviewing the scales and definitions, the raters practiced with two video clips from students who did not appear in the main rating task (2 minutes). Finally, for the main rating task, the raters viewed the 36 target video clips in random order with each video appearing one at a time and playing automatically (20 minutes). Raters viewed each video only one time. After the videos, the raters filled out a background questionnaire (5 minutes), personality test (3 minutes), and facial expression recognition test (15 minutes). Participants were remunerated $30 for their time.

Results and Discussion

After confirming the raters' internal consistency using a two-way mixed average-measures intraclass correlation coefficient (.90), the ratings were averaged per rater separately for each episode type (understanding, nonunderstanding intact, nonunderstanding reversed). The goals of Experiment 2 were to clarify whether raters orient to a hold as a nonverbal signal of nonunderstanding and associate that meaningfulness with a hold's naturally-occurring sequential organization. Raters gave the highest listener comprehension ratings to the understanding episodes without holds (M = 74.42, SD = 12.57). Raters assigned the lowest ratings when the holds were presented in their naturally-occurring order with the onset in Turn 2 and the release in Turn 4 (M = 58.29, SD = 14.06). The episodes with holds in reverse order (i.e., 4–3–2–1) elicited scores that fell between the other two episode types (M = 64.17, SD = 11.77). A repeated-measures ANOVA (sphericity assumed) indicated that there was a statistically significant difference for perceived comprehension ratings, $F(2, 58) = 42.99$, $p < .0001$, partial $\eta^2 = .60$. Post hoc comparisons with a Bonferroni adjustment indicated that there were significant differences ($p \leq .001$, $d \geq 0.75$) for all paired comparisons (see Table 4).

Table 4

Post hoc Tests for Comprehension Ratings by Episode Type

Episode type	*M* (*SD*)	Episode type	*M* (*SD*)	*t*(29)	*p*	*d*
Intact (1–2–3–4) hold	58.29 (14.06)	Reversed (4–3–2–1) hold	64.17 (11.77)	–4.13	.001	0.75
		No hold	74.42 (12.57)	–7.81	.001	1.43
Reversed (4–3–2–1) hold	64.17 (11.77)	No hold	74.42 (12.57)	–6.70	.001	1.08

Results of Experiment 2 demonstrate that raters can differentiate between understanding and nonunderstanding episodes regardless of the location of the hold onset and release in the four-turn sequence. However, they rate listener comprehension lower when holds unfold in their natural sequence. Although the presence of a hold onset out of order was sufficient to elicit ratings lower than those provided to the understanding episodes, raters clearly interpreted holds in their intact sequence as the stronger visual cue of listener difficulty understanding the speaker.

General Discussion

In this systematic investigation of rater perception of holds as a visual cue of listener nonunderstanding, these university students clearly recognized the difference between fellow students' hold onsets and releases, rating onsets as more closely associated with a problem understanding the speaker (Experiment 1). They also perceived holds in which the L2 English interlocutors held a single movement (body lean or head poke) static to be more strongly tied to nonunderstanding as compared to holds with multiple held movements (Experiment 1). When presented with videos of the entire repair sequence, raters differentiated between nonunderstanding episodes with holds and understanding episodes without holds (Experiment 2). Furthermore, they were sensitive to the sequential organization of holds in that they associated holds in their naturally-occurring sequences with listener comprehension problems to a greater extent than holds in a reversed turn order (Experiment 2). In sum, raters, the majority of whom were English L2 speakers, clearly recognize their peers' holds as a nonverbal cue with unique components for signalling the beginning and resolution of listener comprehension problems.

In terms of their ability to differentiate between understanding and nonunderstanding episodes based on visual cues only (i.e., the presence or absence of a hold), the raters downgraded listener comprehension for nonunderstanding episodes. They also revealed

sensitivity to the moves within a hold sequence by giving lower perceived comprehension ratings to hold onsets (Turn 2) than hold releases (Turn 4). However, the association between perceived comprehension and holds was affected by the type of movement held static. The configurations most clearly associated with lower listener comprehension were individual movements, such as forward body leans and head pokes, whereas combinations of body leans and head pokes with facial expressions, such as raised eyebrows or smiling, elicited higher comprehension ratings, possibly because the cues provided conflicting or ambiguous information. For example, laughing and smiling are subtle markers of communication breakdowns (Matsumoto, 2018; Pitzl, 2010), yet raters tend to comment on these facial expressions as markers of understanding (McDonough et al., 2021). It appears that raters associate holds with nonunderstanding when a single, easily perceptible body movement is held in isolation rather than in combination with other cues. However, this conclusion must be revisited in future work by eliciting qualitative comments from raters and through micro-analytic approaches to identify how facial expressions might enhance or dilute the meaning associated with holds in the form of body leans or head pokes.

A particularly novel contribution of this study is that the sequential organization of holds contributes to its strength as a signal of nonunderstanding. A temporal constraint on the meaningfulness of holds is fully compatible with cross-linguistic evidence in interactive practices, where multimodal behaviors are tightly organized in space and time (Enfield & Levinson, 2006; Enfield et al., 2013; Floyd et al., 2016). Holds mark nonunderstanding more saliently when their natural sequential organization is preserved with the onset in Turn 2 and the release in Turn 4. Although the presence of a hold onset in any position in a four-turn sequence may be sufficient for raters to recognize listener comprehension difficulty, the meaningfulness of holds is greater when they are presented in their natural sequence. An interesting question is

whether interlocutors themselves are aware that holds provide visual signals for the onset of nonunderstanding and for the resumption of understanding. Although it can be difficult to identify holds during real-time conversation and immediately solicit interlocutor perceptions, future research might first approach this question by having a slight time delay between the initial conversation and the presentation of hold videos to the interlocutors. To avoid drawing undue attention to holds, future studies might include episodes of other types, such as the understanding episodes tested in Experiment 2.

For the assessment of interactional competence, the present findings underscore that L2 speakers' ability to initiate and respond to repair is an important skill, particularly when it comes to the assessment of L2 speakers' performance in interactive speaking tests (Galaczi & Taylor, 2018; Roever & Kasper, 2018). Raising L2 speakers' awareness of the visual cues that can signal nonunderstanding (with or without a verbal appeal for repair) might enable speakers to demonstrate interactional competence, either by initiating self-repair before the listener requests clarification or by reformulating their initial utterance rather than simply repeating it in response to the listener's request. Furthermore, greater awareness of the visual cues of nonunderstanding may help L2 speakers engage in active listening by nonverbally signalling speakers that they are having difficulty understanding. Thus, in light of the importance and salience of visual cues as signals of nonunderstanding, assessing L2 speakers' interactional competence would benefit from considering visual aspects of interaction and how both speakers and listeners can deploy visual cues to achieve mutual communicative success. In terms of repair specifically, the ability to initiate a repair as a listener and successfully carry it out as a speaker might be evaluated as positive indicators of interactional competence. At minimum, it would be important that raters and examiners involved in scoring interactive speaking tests consider L2 speakers' ability to

provide and interpret visual signs of nonunderstanding rather than base their assessments on speech only. For pedagogy, the finding that holds can be recognized as a sign of listener nonunderstanding by external observers of conversation opens up the possibility of carrying out various instructional interventions to help L2 speakers build the nonverbal behavior component of their interactional competence. The goal of such interventions would be to explore various pedagogical ways of raising L2 speakers' awareness of visual cues so that they can signal, detect, anticipate, and avert communication breakdowns. These pedagogical interventions might follow the global template for metacognitive training (e.g., Wenden, 1999) which includes raising awareness through communicative practice (e.g., Nakatani, 2005).

There are several limitations of this study that might limit its generalizability. With respect to the rating stimuli, the chosen nonunderstanding episodes focused narrowly on nonverbal behaviors associated with only one type of repair initiation, which was a clarification request (e.g., *what*, *hmm*, *sorry*). Therefore, it is presently unclear what visual cues are associated with other ways of initiating repair or whether those cues would be equally salient to raters. Similarly, the videos showed episodes in which holds co-occurred with verbal repair initiation. By silencing the videos, we ensured that raters oriented to the nonverbal contributions only while they were rating listener comprehension. Further research is needed to determine whether raters are equally adept at recognizing visual only repair (i.e., without any verbal contribution). Although infrequent, visual only repair has been shown to occur (Dingemanse, 2015; Levinson, 2015; Manrique, 2016; Seo & Koshik, 2010), but visual only repair was not tested here. By asking them to evaluate listener comprehension while watching silent videos, the raters may have oriented to visual cues more than they would have if they also had access to the speaker and listener voices. Although they were never told about holds or asked to evaluate holds

specifically, the instructions made it clear that they had to estimate listener comprehension based on the available visual cues. Consequently, future research should explore whether sensitivity to holds is enhanced or diminished when both visual and verbal information is available.

With respect to the raters recruited for this study, although they were all members of the same university community, there was likely individual variation in their ability to detect and interpret visual cues of nonunderstanding. Raters with a variety of L1 backgrounds were purposely recruited for both experiments to reflect the same linguistically diverse population of university students in the video episodes. Because there is generally little difference in how L1 and L2 raters evaluate global dimensions of L2 performance, such as comprehensibility, fluency, and accentedness (e.g., Derwing & Munro, 2013), comparisons of L1 versus L2 rater judgments were beyond the scope of the current experiments but could be explored systematically in future work. Although the CELFI corpus includes interaction between English L2 speakers only, the raters had no exposure to their speech as the rating stimuli were silent videos. In other words, the raters had no access to any verbal content that might provide information about the listeners' English proficiency. Nevertheless, as L2 proficiency might play a role in how readily interlocutors detect and use visual cues of nonunderstanding in real-time interaction, this variable should be considered as important in future research.

Future work investigating rater perceptions of visual cues, including holds, might also focus on raters' personality, social, and cognitive skills as possible individual differences that influence ratings. Similarly, raters' sensitivity to visual cues might have been impacted by individual differences in their lipreading ability, as it has been shown to vary by speakers' age (Feld & Sommers, 2009) and to determine their susceptibility to such audiovisual speech illusions as the McGurk effect (Strand et al., 2014). Although all visual materials in this study

were shown as silent videos, it is possible that at least some were better than others at lipreading, which would allow them to interpret mouth shapes as an explicit request for repair (e.g., *huh, what*) thereby influencing their ratings. Despite these limitations, the findings of this study point to an encouraging conclusion that listener holds are a clearly detectable visual signal of nonunderstanding. As an important next step to explore the applicability of these findings to L2 learning and teaching, our current research is exploring whether metacognitive training to raise university students' awareness about holds positively contributes to their interactional competence.

Notes

1 The naturalness ratings were slightly higher for the release videos (M = 77.51, SD = 14.17) than the onset videos (M = 66.93, SD = 16.53) where a rating of 100 meant *extremely natural*. The ease of understanding ratings were highly correlated with the listener comprehension ratings (r = .73 for onset videos and r = .79 for release videos), which contributed to the decision to analyze the comprehension ratings only.

2 The tendency for the final position to be salient was confirmed by additional testing of a manipulated turn sequence in which the hold onset appeared last.

3 The naturalness ratings were similar for the nonunderstanding episodes in the natural (M = 69.08, SD = 12.90) and manipulated turn videos (M = 68.81, SD = 13.04) but slightly higher for the understanding episodes (M = 74.05, SD = 11.63). As in Experiment 1, the ease of understanding ratings were highly correlated with the listener comprehension ratings (r = .80), so only comprehension ratings were analyzed.

References

Chenail, R. (2010). Getting specific about qualitative research generalizability. *Journal of Ethnographic and Qualitative Research, 5*(1), 1-11.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning, 63*(2), 163–185. https://doi.org/10.1111/lang.12000

Dingemanse, M. (2015). Other-initiated repair in Siwu. *Open Linguistics, 1*(1), 232–255. https://doi.org/10.1515/opli-2015-0001

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26*(3), 423–443. https://doi.org/10.1177/0265532209104669

Edwards, D. (2004). Proof procedure. In M. Lewis-Beck, A. Bryman, & Liao, T. (Eds.), *The Sage encyclopedia of social science research methods* (pp. 875-876). Sage. http://dx.doi.org/10.4135/9781412950589.n763

Enfield, N. J., & Levinson, S. C. (Eds.). (2006). *Roots of human sociality: Culture, cognition, and human interaction.* Berg. https://doi.org/10.4324/9781003135517

Enfield, N. J., Dingemanse, M., Baranova, J., Blythe, J., Brown, P., Dirksmeyer, T., & . . . Torreira, F. (2013). Huh? What? A first survey in 21 languages. In M. Hayashi, G. Raymond, & J. Sidnell (Eds.), *Conversational repair and human understanding* (pp. 343–380). Cambridge University Press. https://doi.org/10.26530/oapen_630828

Feld, J. E., & Sommers, M. S. (2009). Lipreading, processing speed, and working memory in younger and older adults. *Journal of Speech, Language, and Hearing Research, 52*(6), 1555–1565. https://doi.org/10.1044/1092-4388(2009/08-0137)

Firth, A. (1996). The discursive accomplishments of normality: On 'lingua franca' English and conversation analysis. *Journal of Pragmatics, 26*(2), 237–259. https://doi.org/10.1016/0378-2166(96)00014-8

Floyd, S., Manrique, E., Rossi, G., & Francisco, T. (2016). Timing of visual bodily behavior in repair sequences: Evidence from three languages. *Discourse Processes, 53*(3), 175–204. https://doi.org/10.1080/0163853X.2014.992680

Galaczi, E. D., & Taylor, L. (2018). Interactional competence: Conceptualizations, operationalizations, and outstanding questions. *Language Assessment Quarterly*, *15*(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816

Groeber, S., & Pochon-Berger, E. (2014). Turns and turn-taking in sign language interaction: A study of turn-final holds. *Journal of Pragmatics, 65*, 121—136. http://dx.doi.org/10.1016/j.pragma.2013.08.012

Heritage, J. (2011). Conversation analysis: Practices and methods. In D. Silverman (Ed.), *Qualitative research: Theory, method and practice, 3rd edition* (pp. 208-230). Sage.

Jenkins, S., & Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment decisions. *The Modern Language Review*, *87*(1), 90-107. https://doi.org/10.1111/1540-4781.00180

Kelley, M. R., Neath, I. & Surprenant, A. M. (2013). Three more semantic serial position functions and a SIMPLE explanation. *Memory & Cognition*, *41*(4), 600–610. https://doi.org/10.3758/s13421-012-0286-1

Kelley, M. R., Neath, I., & Surprenant, A. M. (2015). Serial position functions in general knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1715–1727. https://doi.org/10.1037/xlm0000141

Kendrick, K. (2015). Other-initiated repair in English. *Open Linguistics, 1*(1), 164–190. https://doi.org/10.2478/opli-2014-0009

Levinson, S. (2015). Other-initiated repair in Yélî Dnye: Seeing eye-to-eye in the language of Rossel Island. *Open Linguistics, 1*(1), 386–410. https://doi.org/10.1515/opli-2015-0009

Li, X. (2014). Leaning and recipient intervening questions in Mandarin conversation. *Journal of Pragmatics, 67*, 34-60. http://dx.doi.org/10.1016/j.pragma.2014.03.011

Manrique, E. (2016). Other-initiated repair in Argentine Sign Language. *Open Linguistics, 2*(1), 1–34. https://doi.org/10.1515/opli-2016-0001

Matsumoto, Y. (2018). Functions of laughter in English-as-a-lingua-franca classroom interactions: A multimodal ensemble of verbal and nonverbal interactional resources at miscommunication moments. *Journal of English as a Lingua Franca, 7*(2), 229–260. https://doi.org/10.1515/jelf-2018-0013

May, L. (2011) Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*(2), 127-145. https://doi.org/10.1080/15434303.2011.565845

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

McDonough, K., Crowther, D., Kielstra, P., & Trofimovich, P. (2015). Exploring the potential role of eye gaze in eliciting English L2 speakers' responses to recasts. *Second Language Research, 31*(4), 563-575. https://doi.org/10.1177/0267658315589656

McDonough, K., Lindberg, R., Trofimovich, P., & Tekin, O. (2021). The visual signature of nonunderstanding: A systematic replication of McDonough, Trofimovich, Lu, & Abashidze (2019). *Language Teaching*, 1-15. Advance online publication. https://doi.org/10.1017/S0261444821000197

McDonough, K., Trofimovich, P., Dao, P., & Abashidze, D. (2020). Eye gaze and L2 speakers' responses to recasts: A systematic replication study of McDonough, Crowther, Kielstra, and Trofimovich (2015). *Language Teaching, 53*(1), 81-95. doi:10.1017/S0261444818000368

McDonough, K., Trofimovich, P., Lu, L., & Abashidze, D. (2019). The occurrence and perception of listener visual cues during nonunderstanding episodes. *Studies in Second Language Acquisition, 41*(5), 1151-1165. https://doi.org/10.1017/S0272263119000238

McDonough, K., Trofimovich, P., Lu, L., & Abashidze, D. (2020). Visual cues during interaction: Are recasts different from noncorrective repetition? *Second Language Research*, *36*(3), 359-370. https://doi.org/10.1177/0267658320914962

Nakatani, Y. (2005). The effects of awareness-raising training on oral communication strategy use. *The Modern Language Journal*, *89*(1), 76–91. https://doi.org/10.1111/j.0026-7902.2005.00266.x

Pitzl, M.-L. (2010). *English as a Lingua Franca in international business: Resolving miscommunication and reaching shared understanding*. VDM-Verlag Müller.

Plonsky, L. & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Roever, C., & Kasper, G. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing*, *35*(3), 331–355. https://doi.org/10.1177/0265532218758128

Schegloff, E. A. (1997). Practices and actions: Boundary cases of other-initiated repair. *Discourse Processes*, *23*(3), 499–545. https://doi.org/10.1080/01638539709545001

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis*. Cambridge University Press. https://doi.org/10.1017/CBO9780511791208

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(2), 361–382. https://doi.org/10.1353/lan.1977.0041

Schegloff, E. A., & Sacks, H. (1973). Opening up closings. *Semiotica, 8*(4), 289–327. https://doi.org/10.1515/semi.1973.8.4.289

Seo, M. S., & Koshik, I. (2010). A conversation analytic study of gestures that engender repair in ESL conversational tutoring. *Journal of Pragmatics*, *42*(8), 2219–2239. https://doi.org/10.1016/j.pragma.2010.01.021

Sidnell, J. (2014). Basic conversation analytic methods. In J. Sidnell & T. Stivers (Eds.). *Handbook of conversation analysis* (pp. 77-100). Blackwell. https://doi.org/10.1002/9781118325001.ch5

Statistics Canada (2017). *Montréal [Economic region], Quebec and Quebec [Province]* (table). *Census Profile*. 2016 Census. Statistics Canada Catalogue no. 98-316-X2016001. Ottawa. Released November 29, 2017. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E (accessed July 2, 2021).

Stivers, T. (2013). Sequence organization. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 191-209). Blackwell. https://doi.org/10.1002/9781118325001.ch10

Strand, J., Cooperman, A., Rowe, J., & Simenstad, A. (2014). Individual differences in susceptibility to the McGurk effect: Links with lipreading and detecting audiovisual incongruity. *Journal of Speech, Language, and Hearing Research*, *57*(6), 2322–2331. https://doi.org/10.1044/2014_JSLHR-H-14-0059

Toerin, M. (2014). Conversations and conversation analysis. In U. Flick (Ed.), *The SAGE handbook of qualitative data analysis* (pp. 327-340). Sage. https://dx.doi.org/10.4135/9781446282243

Wagner, J. (1996). Foreign language acquisition through interaction: A critical review of research on conversational adjustments. *Journal of Pragmatics*, *26*(2), 215–235. https://doi.org/10.1016/0378-2166(96)00013-6

Wenden, A. L. (1999). An introduction to metacognitive knowledge and beliefs in language learning. *System*, *27*(4), 435–441. https://doi.org/10.1016/S0346-251X(99)00043-3