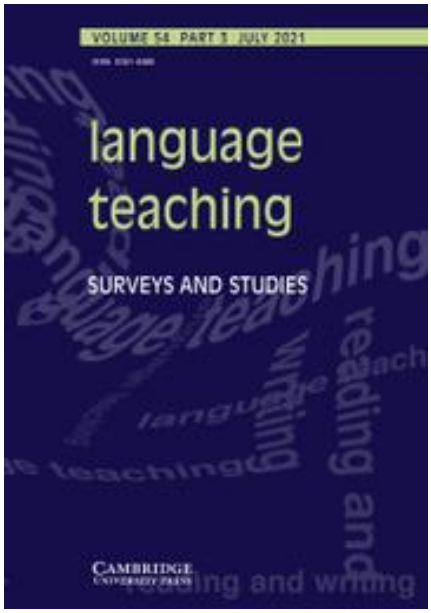

The visual signature of non-understanding: A systematic replication of McDonough, Trofimovich, Lu, and Abashidze (2019)

Kim McDonough, Rachael Lindberg, Pavel Trofimovich, and Oguzhan Tekin

McDonough, K., Lindberg, R., Trofimovich, P., & Tekin, O. (2021). The visual component of non-understanding: A systematic replication of McDonough, Trofimovich, Lu, & Abashidze (2019). *Language Teaching*. Published online 15 June 2021.
https://doi.org/10.1017/S0261444821000197

Abstract

This replication study seeks to extend the generalizability of an exploratory study (McDonough et al., 2019) that identified holds (i.e., temporary cessation of dynamic movement by the listener) as a reliable visual cue of non-understanding. Conversations between second language (L2) English speakers in the Corpus of English as a Lingua Franca Interaction (CELFI, McDonough & Trofimovich, 2019) with non-understanding episodes (e.g., *pardon?*, *what?*, *sorry?*) were sampled and compared with understanding episodes (i.e., follow-up questions). External raters ($N = 90$) assessed the listener's comprehension under three rating conditions: +face/+voice, –face/+voice, +face/–voice. The association between non-understanding and holds in McDonough et al. (2019) was confirmed. Although raters distinguished reliably between understanding and non-understanding episodes, they were not sensitive to facial expressions when judging listener comprehension. The initial and replication findings suggest that holds remain a promising visual signature of non-understanding which can be explored in future theoretically- and pedagogically-oriented contexts.

Keywords: Visual cues, non-understanding, clarification requests, holds

The visual component of non-understanding: A systematic replication of McDonough, Trofimovich, Lu, & Abashidze (2019)

A form of communication breakdown, non-understanding occurs when a listener fails to understand a speaker, after which the listener may choose to ignore it, employing the "let it pass" strategy (Firth, 1996), or to pursue verbal or visual means of remediation to achieve understanding (Bremer, 1996; Cogo & Pitzl, 2016). In English as a lingua franca (ELF) interactions, frequent verbal means of remediation include direct clarification questions (e.g., *what?*) and minimal incomprehension tokens with rising intonation (e.g., *hmm?*) (Mauranen, 2006; Pietikäinen, 2018). As for visual means of remediation, leans, head direction, head position, holds, facial expressions, and eye gaze have all been associated with non-understanding episodes during dyadic interaction (Floyd et al., 2016; Seo & Koshik, 2010). This research has shown that holds, which refer to the temporary cessation of dynamic movement (e.g., eyebrow raises, head tilts, eye gaze), frequently occur when a listener requests clarification. Once the utterance has been clarified, the listener then resumes dynamic movement. Other visual cues such as nodding, head shakes, and pointing have been used by second language (L2) instructors when providing students with corrective feedback (Davies, 2006; Faraco & Kida, 2008; Wang & Loewen, 2016). For instance, during dyadic interaction with L2 English speakers, a listener provided more head nods and blinks prior to recasting but longer eye gaze with more facial expression after non-corrective repetition (McDonough et al., 2020).

Although the specific visual cues associated with non-understanding and with other types of corrective feedback (e.g., recasts) have been identified, less research has investigated whether such cues are perceptible to external observers. For visual cues to serve communicative functions, it seems obvious that they should be detectable. If specific visual cues can reliably

signal listener non-understanding, speakers may be able to detect those cues and initiate self-repair to avoid a communication breakdown. Prior studies that elicited perceptions of visual cues have shown that the meaning of gestures may not be perceived accurately by L2 speakers (Kamiya, 2018; Mohan & Helmer, 1988). Carpenter et al. (2006) reported that external observers of recasts and non-corrective repetitions did not mention visual cues as being useful for differentiating between episode types. Similarly, McDonough et al. (2020) found that external raters either associated the same visual cue with both recasting and non-corrective repetition or claimed that a visual cue was unique to a conversational move when it actually occurred in both episode types. Even though they could not identify the visual cues associated with each episode type, those raters did attribute significantly higher ratings of corrective intent to recasts than to non-corrective repetitions. Taken together, it appears that external observers can differentiate between feedback and non-feedback episodes although they may not be able to describe the specific visual features associated with each one.

Turning specifically to the visual signature of non-understanding, in the exploratory study to be replicated here, McDonough et al. (2019) examined whether listeners used visual cues to signal their lack of understanding. The researchers sampled data from a larger study in which research assistants interacted with 74 English L2 university students to carry out two interactive tasks (interview task and TED talk discussion) during which the research assistant (henceforth listener) was asked to provide recasts to the L2 speakers whenever it seemed appropriate. Analyzing the interaction transcripts, the researchers identified 21 conversations in which the listener (a male graduate student from Quebec) both requested clarification (non-understanding episode) and asked a follow-up question (understanding episode) from the same L2 speaker. The two episodes from each speaker were matched in terms of the length of the speaker's initial

utterance and the listener's response. After identifying the episodes in the transcripts, video recordings of the conversations were examined for the occurrence of listener holds. Next, the videos were edited to show the listener's face during the speaker's initial utterance, following the logic that the listener would signal non-understanding while listening to the utterance that was not understood. Three versions of the video were created to manipulate access to the listener's face and the speaker's voice: clear face and clear voice (+face/+voice), blurred face and clear voice (–face/+voice), and clear face and distorted voice (+face/–voice). Videos were then coded to determine what visual cues the listener provided, which included head nods, blinks, and facial expressions. Raters ($N = 66$) from the same speech community as the L2 speakers were randomly assigned to a rating condition and asked to indicate the degree to which the listener had understood the L2 speaker using a 100-millimeter scale. The video analysis revealed that non-understanding episodes had a significantly higher number of holds and head nods than understanding episodes, while the rating data indicated that raters who had access to the listener's face (either with or without voice) gave lower comprehension ratings than raters who could not see the visual cues. In sum, there was an association between visual cues (specifically holds and head nods) and non-understanding, and raters were sensitive to those cues when assessing listener comprehension.

Despite these findings, McDonough et al. (2019) cautioned that replication studies were needed due to the small sample size (only 21 conversations), the inclusion of only one listener who had been instructed to give recasts, and the possibility of cross-cultural variation in the use and interpretation of visual cues. Therefore, in response to their call for replication, this systematic replication study seeks to test the generalizability of the findings that (a) head nods and holds were associated with non-understanding episodes and (b) external raters detected the

visual signature of non-understanding, relative to the condition where visual information was not available. This study was conceived as a systematic (or approximate) replication in which the methods of the original study are duplicated as closely as possible, but some variables are altered with the goal of extending the initial findings and testing generalizability (Marsden et al., 2018; Porte & McManus, 2018). Therefore, the types of rating stimuli and procedure of the initial study (McDonough et al., 2019) are duplicated, but to improve the generalizability of the findings, several key variables have been adjusted: a larger, more diverse sample of L2 speakers, multiple L2 speakers as listeners instead of one research assistant, and additional communicative tasks. These changes were made to ensure that the initial findings extend to a broader sample of L2 interlocutors and that the visual cues identified in the initial study are not specific to one listener. In addition, this replication included a more refined classification of the specific type of movement that is held static during holds and elicited rater comments about their orientation to visual cues, which allowed for confirmation of the findings of previous studies that identified different types of held movements. Finally, the raters in the initial study came to a research laboratory to do the ratings, whereas the raters in the replication did an online survey due to the pandemic. This adjustment helped determine whether orientation to visual cues differs based on context (i.e., in a research laboratory vs. remote).

This replication study addresses three research questions, with the first two questions identical to the initial study. The first research question asked what visual cues are associated with non-understanding and the second research question asked whether raters can differentiate between non-understanding and understanding episodes based on access to a speaker's voice, listener's face, or both during the initial turn. Based on the findings of the initial study and those of prior non-understanding research (Floyd et al., 2016; Seo & Koshik, 2010), we predicted that

holds would occur in non-understanding episodes more often than would be expected by chance. We also predicted that non-understanding episodes would have more listener holds and more head nods in the initial turn than understanding episodes. Furthermore, in line with the initial finding that visual information provides unique, perceptible cues to listener non-understanding, we expected that the raters who could see the listener's face would show greater differentiation between understanding and non-understanding episodes than those with access to voice only. To obtain greater insight into raters' orientation to visual cues, which was not explored in the initial study, the third research question explored what visual cues the raters reported attending to while rating. We added this new question to shed further light on what specific nonverbal cues used by the listeners were salient to the raters. Because this question was not addressed in the initial study, we made no predictions about the types of visual cues raters would report.

Method

Corpus Overview

Whereas the speech episodes in the initial study came from an experiment in which research assistants interacted with L2 speakers and provided feedback for their non-targetlike forms (McDonough et al., 2018), this systematic replication study samples less controlled data from the Corpus of English as a Lingua Franca Interaction (CELFI; McDonough & Trofimovich, 2019). The corpus consists of conversations between L2 English speakers from Canadian English-medium universities who carried out three, 10-minute communicative tasks in pairs. Similar to the participants in the initial study, these students had met the minimum English proficiency required for admission to their universities (minimum TOEFL iBT score of 75 or equivalent) and were at the B2 to C1 levels in the Common European Framework of Reference. Students were randomly assigned to pairs ($N = 224$) to interact with someone from a different

language background. There was an equal distribution of dyads with same and different reported genders. Whereas the initial study had two interactive tasks, this corpus included three communicative tasks to elicit longer conversations: a discussion task about problems students encountered when moving to Quebec, a close-call narrative (i.e., sharing a personal story about something bad that almost happened to them, but turned out okay in the end), and an academic discussion task based on research studies about motivation, medical ethics, advertising, or nature versus nurture. The students' interaction while carrying out the three tasks was audio- and video-recorded, their eye gaze was tracked, and their skin conductance was measured. 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 targeting the final task, and a debriefing interview eliciting explanations for their task ratings. These data were collected as part of the corpus, but only transcripts of the audio-recordings and video extracts from the interactions were used for this replication study. Audio-recordings of the students' interaction were transcribed and verified by research assistants.

Data Sampling

All 224 transcripts were analyzed for instances of non-understanding, which was operationalized as a four-turn sequence consisting of (a) the speaker's initial utterance, (b) the listener's non-specific, open clarification request, such as *sorry*, *pardon*, *what*, or *huh*, (c) the speaker's repair, and (d) the listener's response. A total of 139 listeners in the corpus (139/448 or 31%) produced at least one clarification request (M = 1.66, SD = 1.32, *range* = 1–10). To ensure variability in listeners while maintaining consistency across episodes, we selected one clarification request from each listener in which the speaker's initial utterance contained at least

three words, there was minimal overlap between turns, and the listener's response turn indicated understanding. Application of these three inclusion criteria resulted in one non-understanding episode from 79 listeners, with these episodes used to determine whether holds were associated with non-understanding. These 79 listeners (40 women, 39 men) came from 30 different first language (L1) backgrounds, and they requested clarification from speakers (46 men, 33 women) of 29 different L1s. In comparison to the initial study, this replication has considerably more L2 speakers (21 vs. 79, respectively) and listeners (1 vs. 79, respectively).

As an additional test of the association between holds and non-understanding, these 79 transcripts were analyzed for the occurrence of an understanding episodes from the same listener. To select these matching episodes, the 79 transcripts were analyzed for the occurrence of four-turn understanding sequences consisting of (a) the speaker's initial utterance, (b) the listener's follow-up question, (c) the speaker's response, and (d) the listener's continuation move. Unlike non-understanding episodes, the listener's turn in (b) requested additional information about the topic, as opposed to clarification of the speaker's initial utterance. Examples of each episode type are provided in Table 1. So that the non-understanding and understanding episodes from each listener were comparable, the speaker's initial utterance and the follow-up question were approximately the same length (1–3 word difference) as the first two turns in the non-understanding episode. This resulted in 35 matched sets of episodes involving listeners (19 men, 16 women) from 16 different L1 backgrounds who elicited clarification or more information from speakers (19 men, 16 women) with 19 different L1s. Whereas the initial study compared visual cues in understanding and non-understanding in 21 matched episodes with only one listener, the sample here is larger (35 matched episodes) and has greater listener diversity (35 unique listeners).

Table 1. *Sample Non-Understanding and Understanding Episodes*

Turn	Non-understanding	Understanding
1	S: Are you planning to do a PhD?	S: Yeah and the Opus card is also a problem.
2	L: Hmm?	L: Why?
3	S: Wanna do your PhD?	S: Because you have to wait till like… I think September for you to be able to get it.
4	L: No no.	L: Oh the Opus for students.

Note. S = Speaker, L = Listener

To validate our classification of the episodes as understanding and non-understanding, we compared the ratings of the speakers' comprehensibility, defined as the ease or difficulty in understanding speech (Derwing & Munro, 2015), on a continuous 0–100 scale provided by the raters as part of the testing procedure (see below). The speakers were rated as being more comprehensible in the understanding episodes (M = 54.34, SD = 23.10) than they were in the non-understanding episodes (M = 49.06, SD = 21.74), $t(89) = 8.04$, $p = .001$, $d = 0.24$. In sum, whereas the original sample had only one listener (a French–English bilingual) who interacted with 21 speakers from six L1s, the current sample included a much wider variety of both listeners and speakers. By including a larger listener sample, it was possible to explore whether the visual cues provided by the original listener in response to understanding and non-understanding episodes would be replicated with listeners from more diverse language and cultural backgrounds.

Rating Stimuli and Materials

The initial turn from the 35 matched sets of understanding and non-understanding episodes (described above) were extracted and used as stimuli for external raters. These very

short video clips (M = 2.96 seconds, SD = 1.71) showed the listener's upper body (face, arms, and torso) while listening to their interlocutor's initial utterance of the four-turn sequence. The audios of all videos were normalized using MP3Gain Express 2.4.0 to have the same volume (90 dB). The video and audio tracks in each clip were then manipulated to vary the access to visual and verbal information, resulting in three different conditions: (a) *+face/+voice* (full video with original audio), (b) *–face/+voice* (listener's face blurred with original audio; see Figure 1), and (c) *+face/–voice* (full video with speaker's voice distorted). The listener's face was blurred in the *–face* condition using the pixelated censor feature in a video editing application (VideoPad). To make the speaker's utterance unintelligible in the *–voice* condition, Audacity was used to distort the voice by first lowering the pitch to between –8 and –10 semitones, then by using the distortion effect set to rectifier distortion and 35/100 on the scale of distortion amount.

Figure 1. Screenshot of the *–face/+voice* condition.

The video clips of the 70 target episodes were presented to raters in an online interface through LimeSurvey (https://www.limesurvey.org) and were preceded by two practice videos with listeners not used in the testing session. Videos appeared in a unique random order for each rater. Each video was displayed on a separate survey page and played automatically as soon as the rater advanced to the next page. However, as the video clips were short and could only be watched once, a three-second countdown was added to the beginning of all videos to give the raters time to readjust and be prepared for the clip to play.

Located below each video were two continuous 100-point slider scales (with the initial slider position at 50) which the raters used to evaluate the speaker's comprehensibility and the listener's comprehension (i.e., how much they thought the listener in the video understood the speaker). The scale endpoints were labeled with a negative anchor point on the left and a positive anchor point on the right. Just as in the initial study, the endpoints for comprehensibility were *hard for me to understand* and *easy for me to understand*, and for listener comprehension they were *this student understood 0%* and *this student understood 100%.* In contrast to the initial study, the raters did not watch the video a second time to assess intelligibility, which captured whether the speakers' intended message was actually understood by the raters (Derwing & Munro, 2015), in a word transcription task because this measure largely overlapped with the rating of comprehensibility. The online survey also included a background questionnaire and two open-ended debrief questions where the raters were asked to explain what informed their ratings when they judged that the listener understood little or nothing, and when they judged that the listener understood most or everything. Overall, compared to the initial study, this replication included more target episodes to rate (70 compared to 42) and the ratings were done online rather than on paper, allowing for more accurate and precise measurements of their ratings.

Raters

The 90 raters (58 females, 32 male) were sampled from the same population of English speakers as the initial study (university-level multilinguals in Montreal). As before, the raters also represented the same speech community as the listeners in the videos (i.e., potential classmates) to provide raters with familiar and relatable interactions. They ranged in age from 18 to 55 (M = 24.23 years, SD = 5.67), which is representative of the original sample (M = 22.9 years, SD = 5.3, *range* = 18–56). Their length of residence in Canada varied greatly from 9

months to 47 years (M = 10.45 years, SD = 10.29). In addition to including more raters in this replication study (90 instead of 66), it was also a more diverse sample, including graduate students (22) in addition to undergraduate students (68), whereas the initial study only included undergraduates, and represented a wider variety of L1s (22 compared to 15). While 65% of the raters in the initial study spoke English or French as their L1, only 36% in this study reported these language backgrounds (21 English, 11 French), while the other most common L1s were Mandarin (9), Spanish (8), Arabic (5), Farsi (4), and Hindi (4). The L2 English raters had been studying English for an average of 13.43 years (SD = 5.83). All raters reported that they listened to English 83.03% of their day on average (SD = 17.41%) and self-rated their English listening skills on average at 91.72 (SD = 12.50), where 100 is *very fluent*.

Procedure

Whereas the rating procedure took place in a research laboratory in the initial study, ratings were conducted remotely through an online survey for the replication study due to the pandemic. Raters were randomly assigned to one of the three video conditions (30 per condition) while keeping the ratio of L1 to L2 English speakers identical across the three groups (7 to 23, respectively). After completing the consent form (2 minutes), they were given instructions for the rating procedure, and the rating criteria were explained (2 minutes). Next, they watched and rated the two practice videos using the slider scales (1 minute). They then provided their ratings for the first 35 target videos (15 minutes). To give the raters a break from the video-rating task, they were prompted to fill in the background questionnaire (5 minutes) before rating the second set of 35 videos (15 minutes). After having completed all the videos, the raters responded in a text box to the two open-ended debrief questions regarding how they judged the listeners' understanding and non-understanding (5 minutes). To conclude the 45-minute session, the raters

gave their email to receive their compensation by *interac* e-transfer ($20 CAD) and then submitted their survey responses. To ensure that the survey was completed appropriately and no videos were skipped, information on the amount of time the raters spent on each page was collected and verified. In comparison to the initial study, the testing procedure was slightly shorter as the raters were not asked to transcribe the speaker's utterance to assess intelligibility.

Data Analysis

First, the second author analyzed the four-turn videos of the 79 non-understanding episodes to determine whether or not a hold occurred (a binary variable). In the event of a hold, the dynamic movements that were held were classified as head pokes (i.e., head extends forward), head turns (i.e., head turns slightly to one side to bring the ear closer to the speaker), head tilts (i.e., head tilts to the side), upward head tilts (i.e., head tilts slightly back), downward head tilts, forward leans (i.e., upper body leans closer to the speaker), open mouth, raised/scrunched eyebrows, smiling, and eye gaze (i.e., eye movements become fixed on the speaker). Next, the four-turn videos of the 35 understanding episodes were analyzed using the same categories. This coding was checked by the fourth author and any disagreements were resolved through discussion. A research assistant subsequently coded 25% of the episodes independently, and interrater reliability yielded a Cohen's κ value of .91 for the occurrence of holds (identical to the initial study), which is close to the median κ value (.92) for applied linguistics research (Plonsky & Derrick, 2016). In contrast with the initial study which only analyzed holds during the listener's second turn, this study also included holds that began in the first turn while listening to the speaker's initial utterance although only three occurred in this dataset.

The visual cues provided by the listener during the first turn of the 35 understanding and 35 non-understanding episodes were coded into five categories. Following the initial study, the first two categories were (a) head nods and (b) blinks. The initial study's category of facial expressions was refined into (c) instances of smiling, laughing, and lip movement (e.g., curling, rounding) and (d) eyebrow movement (e.g., raises, furrows). The final category (e) included other infrequent movements such as head shakes, head tilts, body shifts, eye contact, and glancing away. The second author recorded the frequency counts for each category. Once the fourth author verified the coding and disagreements were resolved, the raw frequency counts per category were summed separately for the understanding and non-understanding episodes. Interrater reliability was assessed by a research assistant who independently coded the visual cues for 25% of the videos. The two-way mixed average-measures intraclass correlation coefficients were as follows (values from the initial study in parentheses): nods = .99 (.96), blinks = .99 (.99), smiling = .97 (.92 combined with eyebrow category), eyebrow movement = .94 (.92 combined with smiling), and other movements = .87 (n/a). The raters' assessments of speaker comprehensibility and listener comprehension (out of 100) were exported to a spreadsheet. All ratings were checked for internal consistency using two-way mixed average-measures intraclass correlation coefficients, and the values (initial study values in parentheses) were .95 for speaker comprehensibility (.99) and .96 for listener comprehension (.98). With the exception of infrequent movements, all intraclass correlation coefficients were higher than the median value (.94) for applied linguistics research (Plonsky & Derrick, 2016).

The qualitative coding of the raters' responses to the two open-ended debrief questions regarding what informed their ratings followed a bottom-up approach and resulted in the following categories: (a) facial expressions (e.g., showing a confused facial expression), (b)

smiling and laughing, (c) eyes/eyebrows (e.g., avoiding eye contact, raising eyebrows), (d) body language (e.g., leaning forward), (e) head movement (e.g., nodding, shaking head), (f) posture (e.g., slouching), and (g) hand movement (e.g., touching the back of their head). These nonverbal behaviors were coded by the second author separately depending on if they were mentioned for informing understanding or non-understanding. A subset of the responses (30%) was independently coded by the fourth author, and interrater reliability assessed using two-way consistency average-measure intraclass correlations was above .93 for each category, which approached the median value (.94) for applied linguistics research (Plonsky & Derrick, 2016).

To address the first research question, a one-sample chi-square test was used to test the prediction that holds in non-understanding episodes occur more frequently than chance. The question was further addressed by comparing the occurrence of holds in non-understanding and understanding episodes using a 2 × 2 chi-square test. If holds are associated with non-understanding, they should appear more frequently in non-understanding episodes. To compare the occurrence of other visual cues in understanding and non-understanding episodes, Wilcoxon signed-ranks tests were used, which are non-parametric paired-samples *t* tests. For the second research question about the effect of access to visual information when rating, first, the ratings for understanding and non-understanding episodes were compared for each rater group using paired-samples *t* tests to determine whether the raters could differentiate between episode types. Next, the difference between the understanding and non-understanding ratings was compared across the three rater groups to explore whether access to visual information facilitated the ability to differentiate between episode type. The third research question was addressed by summarizing the raters' comments about their orientation to visual cues when rating. Alpha was set at .05 for all statistical tests but adjusted for multiple comparisons as needed.

Results

Non-Understanding and Visual Cues

The first research question asked which visual cues are associated with non-understanding. We first considered the occurrence of holds in the videos of the four-turn non-understanding episodes. Of the 79 non-understanding episodes in the sample, 63 (80%) contained holds while 16 (20%) did not have a hold. A one-sample chi-square test against equal probability (i.e., 39.5 with holds and 39.5 without holds) was significant, $\chi^2(1, 79) = 27.96$, $p < .001$. In other words, holds occurred during non-understanding episodes more frequently than expected by chance. To gain greater insight into the non-understanding holds, the dynamic movement that was held in all 63 non-understanding episodes were categorized (see Table 2). Some episodes contained more than one held gesture, such as if a listener held both eye gaze and a head poke. Eye gaze was the most frequently held gesture, occurring in 94% of the holds, followed by open mouth (40%), head poke (40%), and forward lean (37%).

Table 2. *Static Gestures During Non-Understanding Holds (out of 63 Holds)*

Gesture held	*k* holds	Percent
Eye gaze	59	94
Forward lean	23	37
Head poke	25	40
Head tilt	4	6
Head turn	7	11
Open mouth	25	40
Raised/scrunched eyebrows	2	3
Smile	13	21

Upward head tilt	6	10

Note. Column totals exceed $k = 63$ (100%) because a hold could contain one or more statically held gestures.

To further confirm the association between holds and non-understanding, the 35 matched sets of understanding and non-understanding episodes were compared. Whereas 32/35 (91%) of the non-understanding episodes contained a hold, only 9/35 (26%) of the understanding episodes had one. A chi-square test with a continuity correction indicated that the relationship between the occurrence of holds and episode type was significant, $\chi^2(1, 70) = 28.50$, $p < .001$, Cramer's $V = .67$. This confirms the findings of the initial study, which also found a significant relationship between holds and episode type and reported a slightly higher effect size in the analysis of holds produced by a single listener (Cramer's $V = .87$), compared to the current analysis of held gestures produced by 35 different listeners. The static gestures that occurred in the understanding episodes were classified to explore similarities in held gestures across episode types. As shown in Table 3, head poke, forward lean, and raised/scrunched eyebrows were more frequent in non-understanding episodes, whereas downward head tilt, head turn, and head tilt occurred more frequently in understanding episodes. Eye gaze was frequently held in both understanding and non-understanding episodes.

Table 3. *Comparison of Static Gestures in Holds by Episode Type*

	Understanding (9 holds)		Non-understanding (32 holds)	
Gesture held	*k* holds	Percent	*k* holds	Percent
Downward head tilt	3	33	0	0
Eye gaze	9	100	30	94

Forward lean	0	0	14	44
Head poke	1	11	14	44
Head tilt	2	22	3	9
Head turn	3	33	5	16
Open mouth	2	22	10	31
Raised/scrunched eyebrows	1	11	12	38
Smile	1	11	6	19
Upward head tilt	1	11	2	6

Note. Column totals exceed $k = 9$ (100%) for understanding episodes and $k = 32$ (100%) for non-understanding episodes because a hold could contain one or more statically held gestures.

Finally, in addition to examining the occurrence of holds in the complete four-turn episodes, we also compared the occurrence of other visual cues in the first turn of the understanding and non-understanding episodes. Following the same logic as in the initial study, we tested whether the listener's face during the speaker's initial turn provides visual cues that signal non-understanding prior to the spoken clarification request. As shown in Table 4, descriptively head nods were more frequent in understanding episodes, whereas instances of smiling, laughter, and lip movement (curling, rounding) were more prevalent in non-understanding episodes. The number of blinks, eyebrow movements (raises, frowns), and other movements (a combined category including infrequently occurring head shakes, head tilts, body shifts, eye contact, and glancing away) were similar between episode types.

Table 4. *Frequency of Listener Visual Cues (Sum Across All Episodes) by Episode Type*

	Understanding	Non-understanding

Visual cue	Sum	*M*	*Mdn*	*SD*	95% CI		Sum	*M*	*Mdn*	*SD*	*95%CI*	
Head nods	32	0.91	0.00	1.27	0.48	1.35	8	0.23	0.00	0.56	0.04	0.42
Blinks	46	1.31	1.00	1.45	0.82	1.81	45	1.29	1.00	1.41	0.80	1.77
Smiling, laughing, & lip movement	8	0.23	0.00	0.49	0.06	0.40	20	0.57	1.00	0.56	0.38	0.76
Eyebrow movement	6	0.17	0.00	0.45	0.02	0.33	8	0.23	0.00	0.43	0.08	0.37
Other movements	40	1.14	1.00	0.49	0.97	1.31	49	1.40	1.00	0.60	1.19	1.61

Wilcoxon-signed ranks tests (appropriate when data are not normally distributed) using an adjusted alpha level to account for multiple comparisons (.05/5 = .01) revealed a statistically significant difference for smiling, laughing, and lip movements [$Z = 3.00$, $p = .003$, $d = .31$] and a trend for head nods [$Z = 2.49$, $p = .013$, $d = .69$]. There were no statistically significant differences for blinks [$Z = .24$, $p = .809$, $d = .01$], eyebrow movements [$Z = .63$, $p = .527$, $d = .14$], or other movements [$Z = 1.65$, $p = .098$, $d = .48$). Whereas the initial study reported a statistically significant difference for head nods only, which was a trend here, separating the coding of mouth and eyebrow movements in the replication study revealed a difference in smiling, laughing, and lip movements.

To summarize the findings for the first research question, both the occurrence of holds in the larger set of non-understanding episodes and the smaller comparison of non-understanding and understanding episodes confirmed the association between non-understanding and holds. Whereas static gestures involving head poke, forward lean, and raised/scrunched eyebrows were predominantly associated with non-understanding holds, eye gaze was held in both understanding and non-understanding episodes. In addition to the occurrence of holds in the four-turn episodes, smiling, laughing, and lip movements were used by the listener significantly more frequently during the initial turn of non-understanding episodes than understanding episodes.

Non-Understanding and Rater Assessments

The second research question asked whether external raters are able to distinguish understanding and non-understanding episodes under rating conditions that manipulated access to speaker voice and listener face during the initial turn. When asked to assess the degree to which the listener had understood the speaker, the raters gave higher comprehension scores to understanding episodes than to non-understanding episodes in all three rating conditions (see Table 5). Paired-samples *t* tests indicated that the comprehension scores were significantly higher for understanding episodes in all three rating groups: +face/+voice, $t(1, 29) = 4.26$, $p < .001$, $d = 0.31$; +face/–voice, $t(1, 29) = 5.05$, $p < .001$, $d = 0.26$; and –face/+voice, $t(1, 29) = 3.20$, $p = .003$, $d = 0.28$. In other words, the raters evaluated the listener's comprehension lower in non-understanding episodes regardless of which type of information they received—voice, face, or both.

Table 5. *Rater Assessments of Listener Comprehension by Episode Type and Rating Condition*

	Understanding				Non-understanding				Difference
Condition	*M*	*SD*	*Mdn*	*95% CI*	*M*	*SD*	*Mdn*	*95% CI*	M_{diff} (*SD*)
+face/+voice	65.37	13.08	66.96	60.49, 70.26	61.32	12.99	60.81	56.47, 66.17	4.05 (5.21)
+face/–voice	53.97	19.16	59.22	46.81, 61.12	49.03	18.18	52.34	42.24, 55.81	4.94 (5.36)
–face/+voice	67.12	10.41	67.12	63.23, 71.10	64.14	11.18	63.81	59.97, 68.36	2.98 (5.09)

When considering the difference between their comprehension ratings for understanding and non-understanding episodes, the raters who had access to the listener's face showed greater divergence in their ratings (4.05 and 4.94), than the raters who only had access to the speaker's voice (2.98). However, a one-way ANOVA indicated that the difference in mean divergence scores was not statistically significant, $F(2, 87) = 1.07$, $p = .349$, partial $\eta^2 = .02$. In sum, the findings indicated that external observers could identify that the listener had greater comprehension difficulties during non-understanding episodes regardless of whether they had access to the speaker's voice, the listener's face, or both. This contrasts with the initial study, which found benefits for access to visual cues.

Non-Understanding and Rater Perceptions

Turning to the third research question about rater orientation to visual cues, which was not investigated in the initial study, the raters' responses to the two debriefing questions about how they recognized understanding and non-understanding episodes in the two +face conditions

(where visual cues were visible to the raters) were compiled and coded. The values in Table 6 show the number of raters who focused on each type of facial or body gesture when judging listener understanding or non-understanding.

Table 6. *Number of Raters (out of 60) Commenting on Specific Visual Cues by Episode Type*

Visual cue	Understanding	Non-understanding
Body language	7	10
Eyes/eyebrows	16	29
Facial expressions	35	47
Hand movement	8	3
Head movement	24	8
Laugh	14	1
Posture	2	9
Smile	19	3

For both episode types, the raters mentioned facial expressions as the cue to listener comprehension; however, the specific types of facial expressions differed. The raters who commented on facial expressions that signaled listener comprehension described these expressions as engaged (3), happy (3), interested (2), focused (2), relaxed (2), positive, receptive, confident, not hesitant, and attentive. In contrast, facial expressions signalling non-understanding included descriptors such as confused (23), expressionless and blank (11), hesitant (2), emotionless (2), embarrassed (2), bored (2), disinterested (2), surprised (2), tense, stressed, uncomfortable, serious, and quizzical. Although smiling, laughing, and head movement (e.g., nodding) were deemed more indicative of understanding, they were occasionally mentioned for

non-understanding as well, such as when the listener would *make a puzzled face by smiling*, *laugh nervously*, or *tilt their head*.

On the other hand, eyes or eyebrows and posture were more commonly associated with non-understanding, where the listener's eye gaze seemed *far away*, *sort of distant*, or they would *stare off into space*. Both *raised eyebrows* and *slumped over* posture were mentioned as signals of non-understanding. Whereas a *blank stare* or avoiding eye contact seemed to determine non-understanding (9), *direct* and *engaged* eye contact appeared to be mentioned in relation to understanding (8). Similarly, although hand movement was the visual cue the least informative of comprehension, different types of hand movements signified different levels of comprehension. For example, the use of hand gestures was more commonly described as indicating understanding, whereas hand movements indicating non-understanding involved nervous mannerisms such as *touching the back of their head*, *adjusting hair*, or *fidgeting with an object*.

Discussion

The goal of this replication was to revisit McDonough et al.'s (2019) initial study, examining visual cues associated with non-understanding in interactions involving L2 English speakers. Compared to the initial study, the current dataset included a larger corpus of target episodes (79 vs. 21), a more extensive set of matched examples of understanding versus non-understanding (35 vs. 21), a wider range of individual listeners whose visual cues were evaluated (35 vs. 1), and a larger sample of raters assessing the target episodes (90 vs. 60). A key finding of this study, which confirmed the initial result, was that holds signal non-understanding, with head pokes, forward leans, and raised or scrunched eyebrows emerging as the most frequent gestural configurations held static by the listener during holds. However, unlike the initial study

which found evidence for frequent head nods to be linked to non-understanding, this study revealed that listeners tended to use smiling, laughter, and lip movements (curling, rounding) more often when listening to the speaker's initial turn in non-understanding episodes. With respect to rater sensitivity to visual cues, the raters distinguished reliably between instances of understanding and non-understanding. However, unlike the initial study where the raters evaluated listener comprehension lower when they had access to the listener's face, this study revealed little evidence that the raters were particularly sensitive to facial expressions in judging listener comprehension.

Described as temporary cessation of all body movement, with the listener briefly holding their facial expression and body posture fixed until a problematic utterance is resolved (Floyd et al., 2016; Seo & Koshik, 2010), holds appear to be a reliable, unambiguous marker of non-understanding for the listener. Focusing on a single listener's reactions to non-understanding, McDonough et al. (2019) reported holds in 86% (18/21) of the sampled non-understanding episodes, compared to only 5% (1/21) of the matched understanding episodes. In this study, which targeted episodes from 35 listeners, the holds were attested in 91% (32/35) of non-understanding versus 26% (9/35) of understanding episodes, revealing a 3.5-fold increase in the likelihood of holds being associated with non-understanding. While McDonough et al.'s initial finding could be explained by idiosyncratic behaviors of one listener, in this dataset, the association of holds with non-understanding was robust, as holds were detected across 35 listeners representing 16 different L1 backgrounds, minimizing the chances that holds reflected speaker- or culture-specific reactions.

In addition to replicating the initial finding for holds, this study also provided further insight into the various visual configurations held static, clarifying whether various hold types

are unique to non-understanding. The configurations most frequently associated with holds were head pokes, forward leans, and raised/scrunched eyebrows, whereas in the few cases where holds occurred when communication was not compromised, holds were characterized by downward head tilts, head turns, and sideways head tilts (see Table 3). The occurrence of head pokes, forward leans, and eyebrow shapes as part of holds corresponds nearly perfectly to one gestural signature of non-understanding reported by Seo and Koshik (2010) in a corpus of 23 hours of conversations between native English-speaking tutors and L2 speakers, where a hold involved "a head poke and upper body movement forward toward the speaker of the trouble source, sometimes with eyebrows scrunched" (p. 2221). Similarly, Floyd et al. (2006), who analyzed 120 non-understanding sequences by speakers from three ethnolinguistic groups, also listed various head positions (including up–down movements), eyebrow shapes, and upper body leans as the most frequent configurations held static to indicate non-understanding. Therefore, a tentative conclusion emerging from this dataset is that holds are not only robust markers of non-understanding but that their specific types, particularly involving head pokes, forward leans, and raised/scrunched eyebrows, may be more frequent cues to non-understanding than others.

Whereas holds appeared important as visual markers of non-understanding, there was little consensus between the initial study and this replication as to whether visual cues provided by the listener during the speaker's initial turn are associated with communication breakdowns. In McDonough et al. (2019), head nods were significantly associated with non-understanding, and blinks showed a trend in the same direction. In this dataset, however, there was only one reliable difference, with instances of smiling, laughter, and lip movement occurring more often in non-understanding episodes, and a trend for head nods to occur more frequently in understanding episodes. Given that the single listener in McDonough et al. produced a total of 43

head nods and 61 blinks across 21 non-understanding episodes, compared to a total of 8 head nods and 45 blinks attested for 35 different listeners here, it appears that the frequent nodding and blinking may have been the visual behaviors specific to that listener. In fact, as shown by a non-significant trend in this study, it might be more intuitive to interpret head nods as a sign of understanding, because speakers use nodding to track interlocutor comprehension (Aoki, 2011) and to provide interlocutors with supportive back-channels (Bavelas et al., 2002; Knapp et al., 2013). In turn, the tendency for smiling and laughter to co-occur more frequently with non-understanding aligns with prior work showing that these behaviors provide subtle cues to non-understanding (Matsumoto, 2018), where the listener uses them to indicate that non-understanding occurred but to mitigate its impact for the speaker (Pitzl, 2010) and promote positive interaction dynamics and cohesion (Canagarajah, 2013). Despite these interesting trends, however, only holds rather than any other visual cue emerged in this dataset as an unambiguous sign of non-understanding.

From the point of view of external observers, non-understanding episodes were clearly distinguishable from listener-matched understanding sequences, which is consistent with the initial study's findings. However, the raters who had access to the listener's face (i.e., in +face conditions) did not appear to rate listener understanding significantly lower compared to those whose access to the listener's face was obstructed by blurring, implying that the listener's facial expressions were inconsequential to the raters' judgments. Although the raters in the –face condition could not observe the listeners' facial expressions, they nevertheless had full visual access to their body movements such as head nods, head tilts, hand movements or body posture. This could mean that such body movements may be a more important visual cue of non-understanding than facial expressions. It is then little surprise that the ratings of listener

comprehension between the +face/+voice (M = 61.32) and –face/+voice (M = 64.14) conditions differed so little, ostensibly because certain visual cues which involved body movement were visible in both conditions. Similarly, a decrease in perceived listener comprehension in the +face/–voice condition (M = 49.03) might be attributable to the salience of body gestures, in addition to facial expressions, without the distraction of the speaker's utterance for the raters to consider.

Apart from the salience of body movements (body language, posture, head and hand movements) as visual cues of non-understanding, another reason for the raters' less extensive use of facial information to judge listener comprehension, relative to the initial findings, can be explained by the frequency and variability in the raters' visual experience (in +face conditions). In McDonough et al. (2019), the raters were exposed to the same listener's facial expressions 42 times. In that situation, the raters experienced a low type frequency input, where a single listener provided the raters with high token frequencies of his (person-specific) facial expressions such as head nods and blinks. This type of input is especially useful for the so-called fast mapping of novel information onto meaning (Goldberg et al., 2007; McDonough & Nekrasova-Becker, 2014), which in this case corresponds to raters associating a facial expression with non-understanding. In this study, however, the raters saw 35 different listeners' reactions, once each in an understanding and a non-understanding episode. This visual input involved high type frequency, where the raters saw multiple listeners' faces, with predictably lower token frequencies of individual facial cues, which corresponds to a learning condition known to promote further development and extension of a novel pattern rather than its initial detection (Gómez, 2002; Matthews & Bannard, 2010). Faced with high type frequency making it harder for the raters to associate individual facial expressions with non-understanding, the raters were

likely hard-pressed to detect individual facial expressions and use them as unambiguous signals of listener comprehension.

Finally, this study extended the initial study by asking the raters to report the visual cues that they considered in their rating, as a way of clarifying which specific nonverbal cues used by the listener were perceptible to the raters. Although L2 speakers do not always interpret the meaning of gestures (Kamiya, 2018; Mohan & Helmer, 1988) and do not rely on visual information in distinguishing recasts from non-corrective repetitions (Carpenter et al., 2006), the raters in this study seemed to orient to somewhat different visual cues in understanding and non-understanding episodes. Eyebrow shapes and body posture such as leaning forward, in particular, appeared to be linked to non-understanding, suggesting that the raters not only detected certain gestural configurations but also explicitly associated them with low comprehension. Although facial expressions were mentioned in relation to both understanding and non-understanding, the raters distinguished between the expressions that patterned with understanding (e.g., engaged, happy, interested, relaxed) and those that signaled non-understanding (e.g., confused, expressionless and blank, hesitant, emotionless).

Nevertheless, apart from eyebrow shapes, body posture, and several facial expressions, the raters appeared unaware of various other potential visual cues to non-understanding. For example, as shown in previous work and in this study, head movements (up–down, right–left) are clear configurations of holds (Floyd et al., 2016; Seo & Koshik, 2010), and laughing and smiling are subtle markers of communication breakdowns (Matsumoto, 2018; Pitzl, 2010), while sustained eye gaze is typical of both understanding and non-understanding episodes (Floyd et al., 2016). However, the raters predominantly associated head movements as well as instances of laughing and smiling with understanding and reported eye gaze (along with eyebrow shapes)

more frequently as signs of non-understanding. What emerges from these data, by way of summary, is that the external observers in this study (multilingual speakers residing in a multicultural, urban context) were only partially aware of various visual cues signalling non-understanding and that they might generally benefit from explicit instruction or awareness-raising activities targeting specific cues to non-understanding.

Although this replication effort addressed several shortcomings of the initial study, several limitations might impact the generalizability of the present findings. For example, the non-understanding episodes targeted here included only one type of clarification request (e.g., *what*, *hmm*, *sorry*). It is possible that other ways of requesting information in response to a communication breakdown might elicit different visual signatures from the listener. Similarly, the approach to data sampling taken in the initial study and this replication relied on the analysis of transcripts to select the target episodes, so as not to bias the frequency and distribution of various visual cues to non-understanding. However, as shown in prior work (e.g., Seo & Koshik, 2010), sometimes the listener provides only a visual cue to non-understanding, without an accompanying clarification request. By virtue of our data sampling procedure, such instances were excluded from the target materials.

In future work, researchers might wish to focus on the temporal dynamics of non-understanding. For instance, the listener might hold a gestural configuration static for longer than a single turn, or the listener might demonstrate a hold that is not synchronized with a trigger (problematic) utterance (Floyd et al., 2016) or might show a hold that precedes a verbal appeal for repair (Seo & Koshik, 2010). Therefore, researchers might need to establish whether a hold is an unambiguous sign of non-understanding no matter when it occurs or whether it functions as a cue to non-understanding only when it is synchronized with the speaker's trigger utterance and

the listener's appeal for repair. Future studies can explore these issues by exposing raters to the onset and release of holds with a variety of static movements to determine whether they can differentiate between the initiation of non-understanding and the return to understanding. Finally, given the possibility that interlocutors may be largely unaware of the visual cues of non-understanding, future work should explore pedagogical ways of raising awareness of visual cues so that speakers can detect, anticipate, and avert communication breakdowns. In the interim, a cautious take-home message emerging from this replication work is that holds remain a reliable visual signature of non-understanding to be explored in future descriptive and experimental work.

References

Aoki, H. (2011). Some functions of speaker head nods. In C. Goodwin, C. LeBaron, & J. Streeck (Eds.), *Embodied interaction: Language and body in the material world* (pp. 93–105). Cambridge University Press.

Bavelas, J., Coates, L., & Johnson, T. (2002). Listener responses as collaborative process: The role of gaze. *Journal of Communication*, *52*(3), 566–580. https://doi.org/10.1111/j.1460-2466.2002.tb02562.x

Bremer, K. (1996). Causes of understanding problems. In K. Bremer, C. Roberts, M. Vasseur, M. Simonot, & P. Broeder (Eds.), *Achieving understanding: Discourse in intercultural encounters* (pp. 37–64). Longman.

Canagarajah, S. (2013). *Translingual practice: Global Englishes and cosmopolitan relations*. Routledge.

Carpenter, H., Jeon, K. S., MacGregor, D., & Mackey, A. (2006). Learners' interpretation of recasts. *Studies in Second Language Acquisition*, *28*(2), 209–236. https://doi.org/10.1017/S0272263106060104

Cogo, A., & Pitzl, M.-L. (2016). Pre-empting and signalling non-understanding in ELF. *ELT Journal*, *70*(3), 339–345. https://doi.org/10.1093/elt/ccw015

Davies, M. (2006). Paralinguistic focus on form. *TESOL Quarterly*, *40*(4), 841–855. https://doi.org/10.2307/40264316

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

Faraco, M., & Kida, T. (2008). Gesture and the negotiation of meaning in a second language classroom. In S. G. McCafferty & G. Stam (Eds.), *Gesture: Second language acquisition and classroom research* (pp. 280–297). Routledge.

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

Goldberg, A. E., Casenhiser, D., & White, T. R. (2007). Constructions as categories of language. *New Ideas in Psychology*, *25*(2), 70–86. https://doi.org/10.1016/j.newideapsych.2007.02.004

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436. https://doi.org/10.1111/1467-9280.00476

Kamiya, N. (2018). The effect of learner age on the interpretation of the nonverbal behaviors of teachers and other students in identifying questions in the L2 classroom. *Language Teaching Research*, *22*(1), 47–64. https://doi.org/10.1177/1362168816658303

Knapp, M. L., Hall, J. A., & Horgan, T. G. (2013). *Nonverbal communication in human interaction*. Boston, MA: Wadsworth.

Marsden, E., Morgan-Short, K., Thompson, S. & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, *68*(2), 321–391. https://doi.org/10.1111/lang.12286

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

Matthews, D., & Bannard, C. (2010). Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child-directed speech. *Cognitive Science*, *34*(3), 465–488. https://doi.org/10.1111/j.1551-6709.2009.01091.x

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

McDonough, K., & Nekrasova-Becker, T. (2014). Comparing the effect of skewed and balanced input on English as a foreign language learners' comprehension of the double-object dative construction. *Applied Psycholinguistics*, *35*(2), 419–442. https://doi.org/10.1017/S0142716412000446

McDonough, K., Trofimovich, P., Dao, P., & Abashidze, D. (2018). 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.

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

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

Mohan, B., & Helmer, S. (1988). Context and second language development: Preschoolers' comprehension of gestures. *Applied Linguistics*, *9*(3), 275–292. https://doi.org/10.1093/applin/9.3.275

Pietikäinen, K. (2018). Misunderstandings and ensuring understanding in private ELF talk. *Applied Linguistics, 39*(2), 188–212. https://doi.org/10.1093/applin/amw005

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

Plonsky, L., & Derrick, D. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, *100*(2), 538-553. https://doi.org/10.1111/modl.12335

Porte, G., & McManus, K. (2018). *Doing replication research in applied linguistics.* Routledge.

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

Wang, W., & Loewen, S. (2016). Nonverbal behavior and corrective feedback in nine ESL university-level classrooms. *Language Teaching Research*, *20*(4), 459–478. https://doi.org/10.1177/1362168815577239