




ORIGINAL ARTICLE

Task-dependent consequences of disfluency in perception of native and non-native speech

Zachary Houghton^{1,2} , Misaki Kato¹, Melissa Baese-Berk^{1,4}  and Charlotte Vaughn^{1,3} 

¹Department of Linguistics, University of Oregon, Eugene, OR, USA, ²Department of Linguistics, University of California, Davis, Davis, USA, ³University of Maryland, College Park, MD, USA and ⁴Department of Linguistics, University of Chicago, Chicago, MD, USA

Corresponding author: Zachary Houghton; Email: zhoughton@ucdavis.edu

(Received 29 December 2022; revised 25 September 2023; accepted 6 November 2023; first published online 13 December 2023)

Abstract

Silent pauses are a natural part of speech production and have consequences for speech perception. However, studies have shown mixed results regarding whether listeners process pauses in native and non-native speech similarly or differently. A possible explanation for these mixed results is that perceptual consequences of pauses differ depending on the type of processing that listeners engage in: a focus on the content/meaning of the speech versus style/form of the speech. Thus, the present study examines the effect of silent pauses of listeners' perception of native and non-native speech in two different tasks: the perceived credibility and the perceived fluency of the speech. Specifically, we ask whether characteristics of silent pauses influence listeners' perception differently for native versus non-native speech, and whether this pattern differs when listeners are rating the credibility versus the fluency of the speech. We find that while native speakers are rated as more fluent than non-native speakers, there is no evidence that native speakers are rated as more credible. Our findings suggest that the way a non-native accent and disfluency together impact speech perception differs depending on the type of processing that listeners are engaged in when listening to the speech.

Keywords: credibility; fluency; ordinal regression; psycholinguistics; speech perception

Introduction

In speech communication, listeners experience considerable variation in the acoustic characteristics of speech, including speech that is produced with an unfamiliar accent and speech that contains disfluencies. One source of acoustic variation in speech that can impact speech perception is the speakers' native language background. Numerous previous studies have demonstrated that a non-native accent influences the way native listeners process speech (e.g., Bosker et al., 2014a; Bosker & Reinisch, 2015; Hanulíková et al., 2012; Munro & Derwing, 1998, 2001, 2006; Pinget et al., 2014). For example, for native listeners, non-native speech

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

is often less intelligible (e.g., Bent & Bradlow, 2003; Ferguson et al., 2010; Munro & Derwing, 1995a) and is processed more slowly (e.g., Floccia et al., 2009; Munro & Derwing, 1995b) than native speech. Furthermore, the information conveyed in non-native speech can be perceived to be less credible or less reliable than the same information conveyed in native speech (Foucart et al., 2019; Lev-Ari & Keysar, 2010; Livingston et al., 2017; cf. De Meo et al., 2011; Souza & Markman, 2013; Stocker, 2017).

In addition to speakers' native language background, disfluencies, which are present in both native and non-native speech (Goldman-Eisler, 1968; Lennon, 1990), also impact the perception of speech (e.g., Bosker et al., 2013; Bosker et al., 2014a; Cucchiariini et al., 2002; Kormos & Dénes, 2004; MacGregor et al., 2010; Rossiter, 2009). Particularly, mid-utterance disruptions, such as silent or filled pauses, have consequences for listeners' perception. For example, listeners are quicker to respond to a target word in an utterance when the target word is preceded by a silent pause or filled pause compared to when it is not, suggesting that certain forms of pausing facilitate listeners' processing (Brennan & Schober, 2001; Corley & Hartsuiker, 2011; Fox Tree, 2001). Other studies also show that pauses can impact listeners' holistic perception of the speech or the speaker, in terms of how eloquent the speech sounds, or how anxious or honest the speaker is (Christenfeld, 1995; Fox Tree, 2002).

While studies have shown that both a non-native accent and disfluencies impact speech perception (e.g., Bosker et al., 2014b), mixed patterns have been found in how these factors together impact native listeners' processing of native and non-native speech. Specifically, on one hand, manipulations of the characteristics of silent pauses, in terms of their number, duration (Bosker et al. 2014b), location, and distribution (Kahng, 2018), showed similar effects on perceived fluency of native speech and non-native speech. That is for both native and non-native speech, Dutch utterances with more pauses or longer pauses were rated as less fluent than utterances with less pauses or shorter pauses by native speakers (Bosker et al. 2014b). Further, for both native and non-native speech, English sentences that contain pauses placed within a clause were rated as less fluent by native speakers than sentences that contain pauses placed between clauses (Kahng, 2018). While these studies have shown that pauses impact perception of native and non-native speech in a similar manner, other work suggests that pauses are perceived differently across the two types of speech. For example, when listening to native speech, the filler word "um" triggered prediction of low-frequency referents for native listeners, but this effect was not found during perception of non-native speech (Bosker et al., 2014a), suggesting that disfluency impacted native listeners' perception differently depending on whether the speech was native or non-native speech.

One way to understand these mixed patterns may be to take listeners' approach to speech processing into consideration. For example, the above studies have all examined the effect of pauses in native and non-native speech in a single perception task (e.g., fluency rating task: Kahng, 2018; visual world paradigm: Bosker et al., 2014a; comprehension task: Hanulíková et al., 2012). That is, the previous studies did not use the same set of stimuli across different perception tasks. However, it is possible that perceptual consequences of certain acoustic characteristics (e.g., pausing patterns, features of a non-native accent) generally differ depending on how listeners are encouraged to approach the listening activity (e.g., they may

differ depending on the task). Literature on speech perception more broadly suggests that listening is an active process, in which task-specific demands impact the way listeners process auditory information (e.g., Fritz *et al.*, 2007; Heald & Nusbaum, 2014). As such, it seems likely that different tasks could encourage listeners to process the same auditory stimuli in different manners depending on the goal of the task. For example, listeners' behavioral and neural activation patterns differ for the same auditory stimuli when they are encouraged to pay attention to different aspects of the auditory stimuli (e.g., Brechmann & Scheich, 2005; Hugdahl *et al.*, 2003; Pallier *et al.*, 1993). Similarly, for the perception of disfluency, Christenfeld (1995) demonstrated that listeners were more likely to detect filled pauses in spontaneous speech when they were instructed to focus on the style of the speech (e.g., does the speaker use the language well?) as compared to the content of the speech (e.g., does the argument make logical sense?); such differences in the instructions also impacted subjective perception (e.g., speech with filled pauses was perceived to be more eloquent when listeners were focused on the content of the speech than the style of the speech). Thus, as mentioned earlier, since many of the previous studies have examined the influence of disfluency on perception in the context of a single task, it is possible that the mixed results are due to different task demands.

Furthermore, studies examining native listeners' perception of native and non-native speech also suggest that the extent to which listeners process native versus non-native speech similarly depends on what part of the speech listeners are encouraged to pay attention to. For example, as discussed above, listeners processed disfluencies in native and non-native speech in a similar manner when they were evaluating the form or style of the speech (e.g., in a fluency rating task: Bosker *et al.*, 2014b; Kahng, 2018). However, listeners processed native and non-native speech differently when they were required to process the content of the speech, such as rating credibility of the information (Lev-Ari & Keysar, 2010) and answering comprehension questions (Hanulíková *et al.*, 2012; Lev-Ari & Keysar, 2012). Given these studies, it is possible that listeners process certain acoustic characteristics—those that give rise to a non-native accent and disfluencies (in both native and non-native speech)—differently depending on how they are encouraged to approach the listening activity. However, it is unknown to what extent the listening focus of the task impacts the way listeners evaluate native and non-native speech with variable disfluency patterns.

Current study

The aim of the current study is to examine how the perception of silent pauses is affected by differing task demands in native and non-native speech. In order to do so, we ask how various characteristics of pauses impact perception of native and non-native speech in different tasks. Specifically, we examine the perceptual consequences of silent pauses, particularly, the effect of the presence of a pause, and location of a pause (e.g., within-clause location vs. between-clause location), as these characteristics have been shown to impact perception of native and non-native speech (Bosker *et al.* 2014b, Kahng, 2018). We examine the effects of task demands using two tasks: in a task where listeners are encouraged to focus on the content of the speech by evaluating the credibility of the trivia statements (e.g., *Polar bears can swim more than 60 miles without a rest*) and in a task where listeners are encouraged

to focus on the style/property of the speech by evaluating the fluency of the same trivia statements. Crucially, we use the same materials across tasks, allowing us to draw direct comparisons. It is possible that silent pauses might affect the perception of native and non-native speech differently depending on the task. For example, silent pauses might affect fluency ratings in a similar manner (e.g., speech with more pauses may be rated as less fluent than speech with fewer pauses for both native and non-native speakers), but might affect credibility ratings in a different manner (e.g., speech with pauses may be rated as less credible than speech without pauses for native speakers, but this might not be the case for non-native speakers).

In addition to the relationship between task demands and disfluencies, we ask whether there is a relationship between fluency and credibility ratings. For example, it is possible that speech that is rated as more fluent might also be rated as more credible, and we examine whether there is a correlation between fluency and credibility ratings for either native or non-native speakers. If there is a difference in the correlation between fluency and credibility ratings for native versus non-native speech (i.e., if the correlation between fluency and credibility ratings for native is different from those of non-native speech), then that would suggest that the way we process speech depends on whether the speech is native or non-native.

Method

Materials

Materials were 24 English trivia statements from Lev-Ari and Keysar (2010). Of the 24 statements, 12 were true (e.g., *Polar bears can swim more than 60 miles without a rest*) and 12 were false statements (e.g., *The koala is the only known animal that never gets sick*). The true and false statements were of comparable length; the average number of words were 9.25 for true and 8.58 for false statements.

Two native Japanese learners of English (20 and 23 years old) and two native English speakers (20 and 21 years old) recorded these statements. All speakers identified themselves as female and reported no history of speech or hearing impairment. The native Japanese speakers were international students in an intensive English program, who were studying English before entering an American university as matriculated students. They reported their TOEFL ITP scores to be 493 and 498, which are identified to be B1 level in the Common European Framework of Reference for Languages (CEFR)¹. All speakers were recorded in a sound-attenuated booth using a Blue Yeti USB microphone. The statements were written on a sheet of paper, and the reading was self-paced. Recording was done on a single channel at a sampling rate of 44,100 Hz (16 bit) using the Praat speech analysis software package (Boersma & Weenink, 2001). Speakers were given time to familiarize themselves with the list of statements in order to prevent disfluencies and mispronunciation when reading. The speakers were not told which statements were true or false. After the recording, speakers completed a language background questionnaire.

In order to examine whether the presence of a pause and the location of a pause influence listeners' perception, for each of the 24 statements produced by each speaker, we created three pause conditions: no pause, between-clause pause, and within-clause pause items (following Kahng, 2018). To create no pause items, all the

silent pauses in the speech samples longer than 100 ms (Idemaru *et al.*, 2019; Trofimovich & Baker, 2006) were cut from the sound files using Praat. These “no pause” items were used to create between-clause and within-clause pause items. Between-clause and within-clause pause items were created by adding a pause of 600 ms either in a between-clause or within-clause location (Kahng, 2018). Following Foster *et al.* (2000), a clause was defined to be a unit that consists minimally of a finite or non-finite verb and at least one other clause element, such as subject, object, complement, or adverbial. For between-clause items, a pause was added either between a subject and verb of the sentence or before an adverbial phrase (e.g., *Some crocodiles [Pause] may eat other crocodiles; Polar bears can swim more than 60 miles [Pause] without a rest*). For within-clause pause items, a pause was added within a noun phrase (e.g., *Some crocodiles may eat other [Pause] crocodiles; Polar [Pause] bears can swim more than 60 miles without a rest*). This resulted in 288 unique items (24 statements \times 4 speakers \times 3 pause conditions). The stimuli were root mean square (RMS) normalized to an approximately equal amplitude level across stimuli.

Participants

Participants were 277 native English listeners (123 females, 154 males; mean age = 37.2 years), recruited using Amazon Mechanical Turk. Of the 277 participants, 148 participants completed the credibility-rating task and 129 participants completed the fluency-rating task². None of the listeners reported a history of speech or hearing impairment. All participants resided in the United States and self-reported to be native speakers of American English. None of the participants reported experience with Japanese on the language background questionnaire.

Procedure

The experiment was conducted online using a Qualtrics (www.qualtrics.com) link provided to participants via Amazon Mechanical Turk. Each participant completed either the fluency- or credibility-rating task³. Before participants began the trials, they were given clarification information and a series of tasks based on the procedure used in Lev-Ari and Keysar (2010). Specifically, in order to help participants understand that speakers' background (e.g., native language status) is irrelevant to the credibility of the statements, they were told that the speakers were not expressing their own knowledge but only reading aloud statements that were provided (following Lev-Ari & Keysar, 2010). They were also asked to read aloud three trivia statements (different from those in the test trials), so that they would better understand that the speakers (that the participants would listen to later in the listening task) were just reciting statements provided by the experimenter. They were then informed of the truth value of each statement that they read aloud (True or False), so that they would understand that the speakers had learned the truth value only after reciting the sentence. Then, they were told that they would listen to statements recorded by past participants.

In the credibility-rating task, participants listened to each statement and were asked to evaluate the truthfulness of the statement, using a 6-point Likert scale (1: Definitely

false, 6: Definitely true). Next to the scale there were two boxes labeled “I know the answer” and “I didn’t understand what was said” (following Hanzlíková & Skarnitzl, 2017); these responses were later excluded from the analysis in order to ensure that participants’ prior knowledge or the intelligibility of the speech did not affect the credibility ratings (e.g., to exclude the possibility that participants rated a statement to be “definitely false” because they did not understand what the speaker said).

In the fluency-rating task, participants listened to each statement and were asked to evaluate how “easily and smoothly” the speech is delivered, using a 6-point Likert scale (1: Extremely disfluent, 6: Extremely fluent). They were also asked to not evaluate the speech based on overall language proficiency of the speaker.

In both credibility- and fluency-rating tasks, participants could listen to the sentence only once, but could take as much time as needed to answer. Prior to the test trials, they completed two practice trials with speakers and sentences that were different from the 24 test trials. In the test trials, each participant listened to all 24 statements. They listened to 8 no pause items, 8 between-clause pause items, and 8 within-clause pause items. In each of the three pause conditions (i.e., 8 items), 4 items (2 true and 2 false statements) were produced by native speakers, and the other 4 items (2 true and 2 false statements) were produced by non-native speakers. Each listener heard all the speakers (i.e., two native and two non-native speakers). The presentation order of the 24 statements (pause conditions and speakers) was randomized for each listener. The combinations of the statement, speaker, and pause conditions were counterbalanced across listeners. The experiment took 10–12 minutes to complete.

Analysis

There was a total of 6,143 data points analyzed: 3,047 data points⁴ for the credibility-rating task and 3,096 data points for the fluency-rating task. In order to analyze whether the effect of speakers’ native status (native vs. non-native English speakers) and the effect of pause conditions (no pause, between-clause pause, within-clause pause) on listeners’ ratings differed depending on the task (credibility- vs. fluency-rating), we implemented a Bayesian logistic ordinal regression model⁵ with uninformative priors, using R package *brms* (Bürkner, 2017), with the rating as the dependent variable. The fixed effects were Task (credibility vs. fluency), Statement type (false vs. true), Speaker group (native vs. non-native), Pause (no pause, between-clause pause, within-clause pause), Speaking Rate⁶, and the interaction among these factors. Sum coding was used for the model. The intercepts represent the threshold or cutoff value to move from one rating to the subsequent rating (i.e., the first intercept corresponds to the threshold to go from a rating of 1 to a rating of 2, the second intercept corresponds to the threshold to go from a rating of 2 to a rating of 3, and so on and so forth. See Barreda and Silbert (2023) for a more in-depth explanation of thresholds in ordinal regression models). For example, if a threshold to move from a rating of 1 to a rating of 2 was -0.3 , then the model would predict a rating of 1 for any latent variable value below that number. The coefficient value represents the estimated increase of rating in log-odds for a 1-unit change in the variable: a positive coefficient value corresponds to a positive effect of the variable on rating. For binary level factors (in our case, Task), in order to get the

estimate of the other task (fluency rating task), we simply reverse the sign on the estimate for Task1. That is, if the coefficient value is positive, then the coefficient value for the fluency rating is negative, and vice versa if the coefficient value is positive. For variables with 3 levels (In our case, the variables with three levels are the Pause variable and all the interactions with Pause), we recover the third level by taking the negative sum of the estimates for the other 2 levels. In other words, Pause_3 (within-clause pause) = $-(\text{Pause}_1 + \text{Pause}_2)$, which correspond to no pauses and between-clause pauses respectively.

The random effects structure included random intercepts for speaker, listener, and item. The random effects structure also included by-speaker random slopes for Task, Pause, and Statement type, by-item slopes for Speaker group, Pause, and Task, and by-listener slopes for Pause, Statement type, and Speaker group.

In this paper, we report 95% credible intervals for the effects of interest as well as the effects that were found to be credible. Bayesian analyses do not force us to interpret the results in terms of significance, but we can interpret an effect to be meaningful if the 95% credible interval does not contain zero. Additionally, in some cases, we also report the percentage of posterior samples greater than zero. This is useful in cases where the credible interval crosses zero because it allows us to determine specifically how confident we can be in the direction of the effect.

Results

We present the results in two sections. First, we present the results of a Bayesian cumulative ordinal mixed-effects regression model with a logit link, examining the effect of task and pause characteristics on perception of native and non-native speech. Next, we present the results of a correlation analysis, examining the relationship between fluency and credibility ratings and whether it differs for perception of native and non-native speech.

The effect of task and pause characteristics on perception of native and non-native speech

Figure 1 shows the credibility ratings and fluency ratings for the statements produced by native and non-native speakers in the three pause conditions (no pause, between-clause pause, within-clause pause). In this figure, the false and true statements are collapsed as the rating patterns were similar across the two statement types. The figure suggests that while silent pauses affected fluency ratings, they did not affect credibility ratings. Additionally, sentences with pauses within clauses were rated as less fluent than sentences with pauses between clauses, and both were rated as less fluent than sentences with no pauses. For the credibility rating task, patterns appear similar for responses to native and non-native speakers, while for the fluency rating task, native speakers appear to be rated as more fluent than non-native speakers.

The results of the regression model revealed a main effect of Speaker group (native vs. non-native, $\beta = 0.826$, CI-2.5% = 0.04, CI-97.5% = 1.54) and an interaction effect between Speaker group and task ($\beta = -1.19$, CI-2.5% = -1.87 , CI-97.5% = -0.43). Note that the credible intervals for both of these effects do not cross zero, suggesting we can be confident in the direction of the effect. This indicates that while

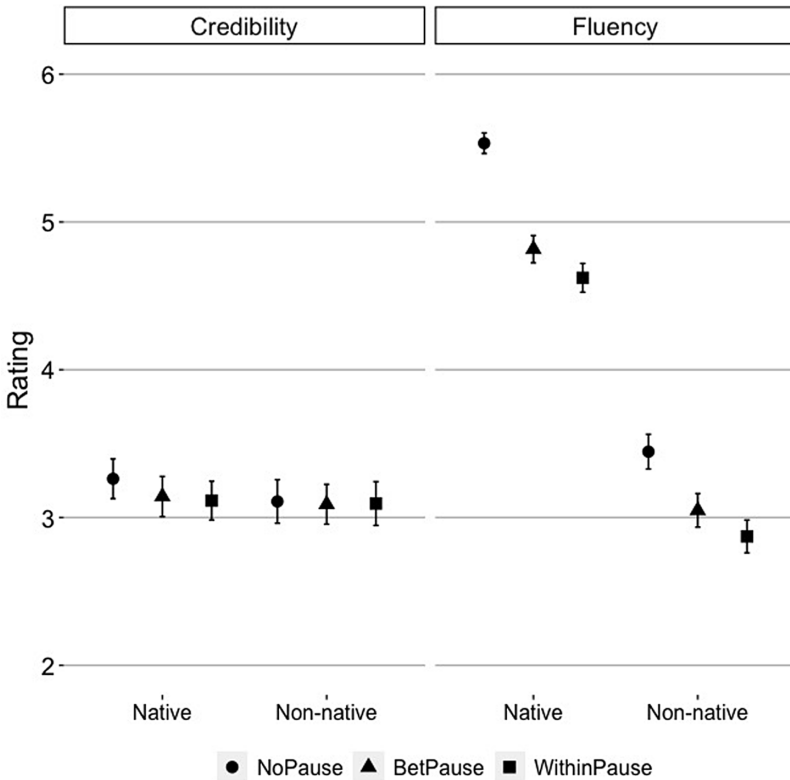


Figure 1. Credibility (1: Definitely false, 6: Definitely true) and fluency (1: Extremely disfluent, 6: Extremely fluent) ratings for statements produced by native and non-native speakers, presented in three pause conditions (no pause, between-clause pause, and within-clause pause). The error bars represent 95% confidence intervals.

the ratings generally were higher for the statements spoken by native speakers than those spoken by non-native speakers, this effect was larger in the fluency task than in the credibility task. To follow-up on this interaction effect, we examined the effect of speaker group in both tasks. We found a meaningful effect of speaker group for fluency ratings ($\beta = 2.02$, CI-2.5% = 0.91, CI-97.5% = 3.03), but not for credibility ratings ($\beta = -0.36$, CI-2.5% = -1.39, CI-97.5% = 0.68). This suggests that while listeners perceived statements spoken by native speakers to be more fluent than those spoken by non-native speakers, they did not perceive native speakers' statements to be more credible than non-native speakers' statements.

The credibility intervals for the main effect of no-pause, between-pause, and within-pause all contained zero, though it is worth noting that the within-clause pause condition's credible interval is mostly negative ($\beta = -0.43$, CI-2.5% = -1.10, CI-97.5% = 0.23), suggesting that sentences with pauses within clauses receive slightly lower ratings than the other two pause conditions. We also examined the effects of pauses to determine whether there was a difference in the ratings between the no pause condition and the within-clause pause condition. We found a small positive effect for this, but note that the credible interval crosses zero

($\beta = 0.65$, CI-2.5% = -0.35 , CI-97.5% = 1.65). While the credible interval crosses zero, we examine the posterior samples to determine what percentage of the sampled coefficients were greater than or less than zero and found that 90% of the posterior samples were greater than zero.⁷ This suggests that in general, regardless of task (credibility or fluency), sentences without pauses received a higher rating (i.e., perceived to be more credible and more fluent) than sentences with pauses within a clause. We also examined whether the simple effects for each pause condition varied by task (i.e., we examined the difference between the effect of each pause condition in each task). We found that for the no pause condition in the fluency task, there was a small effect ($\beta = 0.68$, CI-2.5% = -0.21 , CI-97.5% = 1.57). While the credible interval crosses zero, over 93% of the posterior samples were greater than zero. Compared to this, the effect of no pause in the credibility task was negligible ($\beta = -0.25$, CI-2.5% = -1.17 , CI-97.5% = 0.68), and the credible interval is almost centered around zero. This suggests that speech with no pauses is rated higher in the fluency task, but not in the credibility task. Further, we also found that for the within-clause condition in the fluency task, there was a decrease in perceived fluency ($\beta = -0.77$, CI-2.5% = -1.71 , CI-97.5% = 0.21). While the credible interval for this effect also crosses zero, over 94% of the posterior samples were less than zero. Contrastively, the effect of within-clause pauses in the credibility rating task is also rather negligible ($\beta = -0.10$, CI-2.5% = -1.15 , CI-97.5% = 0.94). This suggests that the effect of pauses within-clauses was greater in the fluency rating task than the credibility rating task. Finally, to confirm if there was a meaningful difference between the no pause and within-clause pause condition in the fluency task, we examined the difference between the two estimates and found a small difference ($\beta = 1.45$, CI-2.5% = -0.24 , CI-97.5% = 3.10). While the credible interval crosses zero, over 95.5% of the posterior samples were greater than zero. On the other hand, the difference between these two pause conditions in the credibility task was rather negligible ($\beta = -0.14$, CI-2.5% = -1.91 , CI-97.5% = 1.63). In other words, while sentences with no pauses are rated as more fluent than sentences with pauses within-clauses, sentences with no pauses are not rated as more credible than sentences with pauses within-clauses. Thus, the effect of pauses is different across tasks. The full model results are included in the appendix section, and the code that we used for the ordinal regression model as well as the simple effects calculations are all included in the supplemental materials.

The above results demonstrate that native listeners' perception of native and non-native speech was impacted by the task (credibility-rating vs. fluency-rating) as well as by the characteristics of a pause in speech (no pause, between-clause pause, within-clause pause). Specifically, the effect of task was clear, such that speakers' native language status (native vs. non-native) influenced fluency ratings but not credibility ratings. Further, there is some support that pause characteristics influenced ratings differently depending on the task (though the credible interval for this did cross zero). That is, the presence of a pause (no pause vs. pause) impacted fluency ratings more readily than credibility ratings, for perception of both native and non-native speech. These results suggest that particular acoustic characteristics of speech, originating from speakers' native language background and pause characteristics, had different perceptual consequences depending on how listeners were encouraged to evaluate the speech in different tasks.

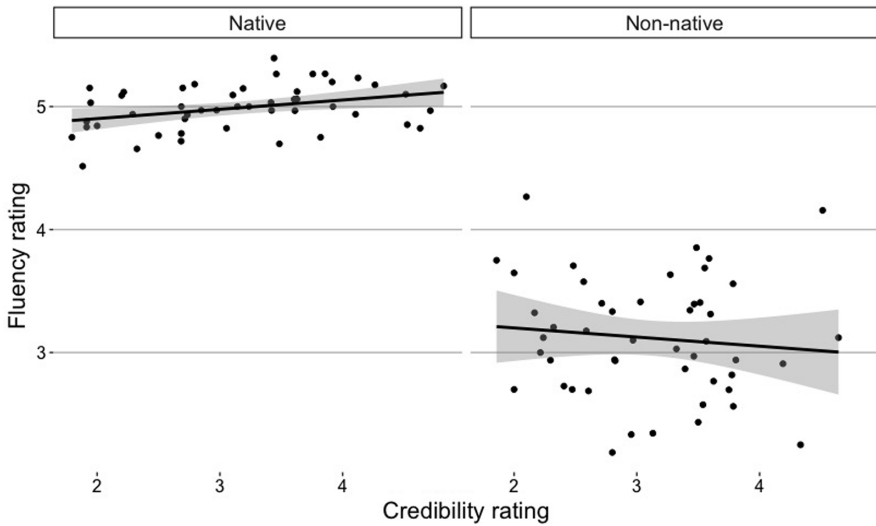


Figure 2. Scatterplots of correlation between credibility ratings (x-axis) and fluency ratings (y-axis) for native speakers' speech (left panel) and non-native speakers' speech (right panel). Pause conditions (no pause, between-clause pause, within-clause pause) are collapsed. The solid lines are best-fitting linear regression lines.

The relationship between credibility- and fluency-ratings for native and non-native speech

In order to better understand the nature of task demands when evaluating native and non-native speech, we examined whether fluency and credibility ratings correlated with one another. Figure 2 shows the relationship between the credibility ratings (x-axis) and fluency ratings (y-axis) for native (left panel) and non-native speech (right panel). Each data point represents a mean rating for an item for a speaker; here, pause conditions (i.e., no pause, between-clause pause, within-clause pause) are collapsed as the observed patterns were similar across all conditions. The solid lines represent the best-fitting linear regression lines. A Pearson correlation test was conducted, and the correlation between the two types of ratings was significant for native speakers' speech ($r = .35, t(46) = 2.51, p = .016$), indicating that items that were perceived to be more credible were also perceived to be more fluent. However, the correlation was not significant for non-native speakers' speech ($r = .11, t(46) = -.73, p = .47$). These results suggest that there was a relationship between how native listeners evaluated perceived credibility of the statements (meaning-focused evaluation) and fluency of the same statements (form-focused evaluation) produced by native speakers, but not for those produced by non-native speakers.

Discussion

The present study examined how characteristics of silent pauses (i.e., presence vs. absence of a pause and pause location) in native and non-native speech impact native listeners' subjective ratings of these types of speech in different tasks, namely, a credibility rating task and a fluency rating task. The current results suggest that

native listeners' processing of native and non-native speech with silent pauses is impacted by task demands. The findings have implications for our understanding of how disfluencies in speech have different perceptual consequences depending on how listeners approach the listening task.

The effects of native language status and pauses on perception in different tasks

The current results demonstrated that the effects of speakers' native language status and pauses on listeners' perception were different across tasks (fluency vs. credibility ratings). Specifically, while silent pauses affected the fluency ratings of both native and non-native speech, we found no meaningful effect of this for the credibility ratings. Since different tasks were designed to draw attention to different parts of the speech, it is likely that the fluency task may have highlighted the form of the speech, while the credibility task may have highlighted the meaning of the speech. Indeed, this would help to explain why the location of pauses impacted fluency ratings of native and non-native speech, but not credibility ratings. We address this possibility in more depth later on in this section.

The present study replicated findings from previous studies that native speech is perceived to be more fluent than non-native speech by native listeners (e.g., Bosker *et al.*, 2014b; Kahng, 2018). However, it contrasted with previous findings that the same statements produced by non-native speakers sound less credible than those produced by native speakers (Lev-Ari & Keysar, 2010). That is, the present study is one among several studies that fails to find a relationship between speakers' non-native status and their reduced perceived credibility as compared to native speakers' (e.g., De Meo *et al.*, 2011; Souza & Markman, 2013; Stocker, 2017). One difference between the present study and Lev-Ari and Keysar (2010) which may have influenced our contradictory results is the speaker population: the speakers in our study were different from the speakers in their study, which may have influenced the results. For example, since we did not explicitly control for overall accentedness and intelligibility of the speakers (though Lev-Ari and Keysar (2010) also do not explicitly control for these either), it is possible that the speakers in their study were more accented than speakers in our study. Additionally, the speakers in Lev-Ari and Keysar (2010) came from different native language backgrounds, which may have further contributed to potentially increased processing costs for the listeners in their study than ours, where the non-native speakers were of the same native language background. Despite this difference, however, there are many studies that have failed to replicate the effect of the non-native speaker status on perceived credibility (e.g., De Meo *et al.*, 2011; Souza & Markman, 2013; Stocker, 2017), suggesting that the relationship between a speaker's non-native speaker status and their perceived credibility may not be as straightforward as it seems.

Another possible explanation for our results is that credibility may no longer be influenced by a non-native accent once an intelligibility threshold is met. In the current credibility task, in addition to the rating scale, listeners were also given the option of choosing "I didn't understand what was said," meaning that the stimuli that were given numeric credibility ratings were likely intelligible enough for listeners to evaluate the credibility of the meaning of the statements (though note that this option was also included in Lev-Ari & Keysar, 2010). As shown in previous studies (e.g., Munro & Derwing, 1995a), more strongly non-native accented speech is not

necessarily less intelligible to listeners. Thus, it is possible that acoustic characteristics of the speech that do not impede listeners' understanding of the meaning of the speech (e.g., mild non-native accent, disfluencies) do not significantly impact perceived credibility. Additionally, we do not find evidence that the effect of pauses was different across the speakers' native language status. This is interesting because previous research has demonstrated that filled pauses (e.g., "um") affect the perception of native speech differently than non-native speech (Bosker et al., 2014a). Specifically, as mentioned earlier, Bosker et al. (2014a) demonstrated that listeners predict a low frequency word to follow "um" when listening to native speech, but not when listening to non-native speech. One possible explanation for the different results is that perhaps the processing consequences of filled pauses (e.g., "um") is different from silent pauses. Indeed, previous research has provided evidence that filled and silent pauses have different functions (e.g., Rose, 2019; Swerts, 1998). An interesting avenue for future research then might be to examine the differences between the perception of silent and filled pauses across different tasks.

Lastly, as discussed throughout this paper, differences in task demands also influenced the way pause characteristics impacted the perception of native and non-native speech. Specifically, while the location of pauses (i.e., within vs. between-clauses vs. no pauses) impacted the fluency ratings of native and non-native speech, this was not true for credibility ratings. That is, the location of the pauses had little effect on the perceived credibility of the statements. In other words, when listeners were focusing on the form of the speech (i.e., in the fluency task), the disfluencies had a large effect on the ratings, but when listeners were focusing on the content of the speech (i.e., the credibility task), the disfluencies had almost no meaningful effect on the ratings. One possible explanation for this result is that the differences in the task demands draw the listener's attention to different parts of the speech; fluency rating tasks encourage participants to pay attention to the form of the speech, while credibility rating results encourage participants to pay attention to the meaning of the speech. Thus, we suggest that the differences in the effects of pauses on fluency ratings versus credibility ratings is due to the differences in the task demands. The current results provide support for the claim that the perception of speech is influenced by the task that people are carrying out (e.g., Christenfeld, 1995), and further extends such results to perception of native and non-native speech. We explore the relationship between different task demands further in the section below.

Relationship between fluency and credibility ratings

The present study also examined the relationship between fluency and credibility ratings, and whether this relationship differed depending on the native versus non-native status of the speakers. The results indicate that while there was a relationship between fluency and credibility for native speech (i.e., native speech that was rated as more fluent was also rated as more credible), there was no such relationship for the perception of non-native speech.

One possible explanation for this result is that disfluencies in native speech may be more surprising than disfluencies in non-native speech. Listeners' experience with speech may drive their expectations of upcoming speech (e.g., Hanulíková et al., 2012). Thus, speech that violates expectations (i.e., the expectation that native speech should

be relatively free from disfluencies) may draw attention to the form of the speech even when the task itself does not. This might explain why less fluent speech is rated as less credible for native speech; the results above indicate that silent pauses in native speech have a larger effect on fluency ratings than silent pauses in non-native speech, suggesting that disfluencies in native speech are more noticeable. If disfluencies are more noticeable in native speech than non-native speech, it is possible that this is the reason that disfluent native speech is rated as less credible than fluent native speech, while disfluent non-native speech is not rated as less credible than fluent non-native speech. Further, it is possible that the reason that disfluent non-native speech is not rated as less credible is simply because listeners are not as attentive to the silent pauses (since disfluencies in non-native speech are not as surprising as those in native speech). In other words, the present results may indicate that even when the task (i.e., the credibility task) does not draw attention to the form of the speech, speech characteristics that are surprising enough may call attention to its form. Future studies could examine whether the number of disfluencies impacts credibility ratings in native speech. If the number of disfluencies directly affects credibility ratings in the perception of native speech, this would provide further evidence that speech with more salient disfluencies is rated as less credible.

Conclusion

Together, these results revealed that pauses and speakers' native language status impact listeners' perception when they are encouraged to pay attention to the form of the utterance (via fluency rating task), but much less so when they were encouraged to listen for the meaning of the utterance (via credibility rating task). These results suggest that listeners may perceive disfluencies in speech differently depending on whether those variations are relevant to the listening task at hand, adding support to the body of research demonstrating that listening is a goal-oriented, attentionally guided behavior (e.g., Christenfeld, 1995; Fritz *et al.*, 2007; Heald & Nusbaum, 2014; Hugdahl *et al.*, 2003). Additionally, our results suggest that listeners' expectations of upcoming speech may further play a role in the effects of attention on processing: listeners confronted with surprising patterns of speech (i.e., speech that is not in line with their expectations) may focus more on the form of that speech, even when performing a task that does not draw attention to the form of the speech. However, it is an open question as to what extent form-oriented evaluation and meaning-oriented evaluation are related to one another.

In summary, the present study contributes to the current body of literature by demonstrating task effects on the processing of silent pauses in native and non-native speech. Specifically, we demonstrate that the perception of non-native and native speech differs depending on whether the listener is attending to the form or the meaning of the speech, which is directly affected by the task they are performing. We also demonstrate the need for future studies exploring the relationship among different types of holistic perception.

Acknowledgments. The authors would like to thank Dr. Santiago Barreda and Dr. Phillippe Rast for their insightful discussion and advice on the statistical analyses and the reporting of the results in this paper.

Notes

- 1 This comparison between TOEFL ITP and CEFR is based on ETS website: https://www.ets.org/toefl_itp/research/performance-descriptors.
- 2 We aimed to have at least 10 participants in each counter-balancing group (discussed below), though we did not eliminate the data when we obtained a few more than 10 participants in a group. This resulted in the number of participants described here.
- 3 Note that in order to avoid participants' experience with one task biasing their choices in the other task, the task was a between-subject manipulation (i.e., participants only completed either the fluency or the credibility task, not both).
- 4 For the credibility-rating task, there were 3552 data points (148 listeners \times 24 items). In the analysis, the responses, "I didn't understand what was said" and "I know the answer" were removed (following Hanzlíková & Skarnitzl, 2017). This resulted in: 3,552 data points - 505 data points (i.e., 14.2% of the credibility data points) = 3,047 data points (from 148 listeners) analyzed for the credibility-rating task.
- 5 See Liddell and Kruschke (2018) and Wu and Leung (2017) for detailed discussion on analyzing ordinal data.
- 6 Speaking rate was included to avoid erroneously attributing effects of speaking rate to effects of speaker group since these may be correlated.
- 7 In Bayesian mixed-effects models, credible intervals are created by sampling possible values for the coefficients and then finding the 2.5 and 97.5 quantiles of those values. One advantage of this process is that we can directly access the samples to determine the percentage of samples greater than zero. This is useful because it allows us to determine if there's some evidence of an effect in cases when the credible interval crosses zero.

References

- Barreda, S., & Silbert, N. (2023). *Bayesian multilevel models for repeated measures data: A conceptual and practical introduction* in R. Taylor & Francis.
- Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, *114*(3), 1600–1610.
- Boersma, P., & Weenink, D. (2001). *Praat speech processing software*. Institute of Phonetics Sciences of the University of Amsterdam. <http://www.praat.org>.
- Bosker, H. R., Pinget, A. F., Quené, H., Sanders, T., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159–175.
- Bosker, H. R., Quené, H., Sanders, T., & de Jong, N. H. (2014a). Native "um"s elicit prediction of low-frequency referents, but non-native "um"s do not. *Journal of Memory and Language*, *75*, 104–116.
- Bosker, H. R., Quené, H., Sanders, T., & de Jong, N. H. (2014b). The perception of fluency in native and nonnative speech. *Language Learning*, *64*, 579–614.
- Bosker, H. R., & Reinisch, E. (2015). Normalization for Speechrate in Native and Nonnative Speech. In *Proceedings of the 18th international congress of phonetic sciences (ICPhS 2015)*, 1–5.
- Brechmann, A., & Scheich, H. (2005). Hemispheric shifts of sound representation in auditory cortex with conceptual listening. *Cerebral Cortex*, *15*(5), 578–587.
- Brennan, S. E., & Schober, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, *44*(2), 274–296.
- Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1). <https://doi.org/10.18637/jss.v080.i01>.
- Christenfeld, N. (1995). Does it hurt to say um? *Journal of Nonverbal Behavior*, *19*(3), 171–186.
- Corley, M., & Hartsuiker, R. J. (2011). Why um helps auditory word recognition: The temporal delay hypothesis. *PLoS One*, *6*(5), e19792.
- Cucchiari, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America*, *107*(2), 989–999.
- Cucchiari, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, *111*(6), 2862–2873.
- De Meo, A., Vitale, M., Pettorino, M., & Martin, P. (2011). Acoustic-perceptual credibility correlates of news reading by native and chinese speakers of Italian. In *ICPhS* (pp. 1366–1369).

- Ferguson, S. H., Jongman, A., Sereno, J. A., & Keum, K. (2010). Intelligibility of foreign-accented speech for older adults with and without hearing loss. *Journal of the American Academy of Audiology*, *21*(3), 153–162.
- Floccia, C., Butler, J., Goslin, J., & Ellis, L. (2009). Regional and foreign accent processing in English: Can listeners adapt? *Journal of Psycholinguistic Research*, *38*(4), 379–412.
- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375.
- Foucart, A., Santamaría-García, H., & Hartsuiker, R. J. (2019). Short exposure to a foreign accent impacts subsequent cognitive processes. *Neuropsychologia*, *129*, 1–9.
- Fox Tree, J. E. (2002). Interpreting pauses and ums at turn exchanges. *Discourse Processes*, *34*(1), 37–55.
- Fox Tree J.E. (2001). Listeners' uses of *um* and *uh* in speech comprehension. *Memory and Cognition*, *29*, 320–326.
- Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention—focusing the searchlight on sound. *Current Opinion in Neurobiology*, *17*(4), 437–455.
- Goldman-Eisler, F. (1968). Psycholinguistics: Experiments in spontaneous speech.
- Hanulíková, A., van Alphen, P. M., Goch, M. M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience*, *24*(4), 878–887. https://doi.org/10.1162/jocn_a_00103
- Hanzlíková, D., & Skarnitzl, R. (2017). Credibility of native and non-native speakers of English revisited: Do non-native listeners feel the same? *Research in Language*, *15*(3), 285–298.
- Heald, S., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience*, *8*, 35.
- Hugdahl, K., Thomsen, T., Erslund, L., Rimol, L. M., & Niemi, J. (2003). The effects of attention on speech perception: an fMRI study. *Brain and Language*, *85*(1), 37–48.
- Idemaru, K., Wei, P., & Gubbins, L. (2019). Acoustic sources of accent in second language Japanese speech. *Language and Speech*, *62*(2), 333–357.
- Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39*(3), 569–591.
- Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145–164.
- Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 387–417.
- Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of experimental social psychology*, *46*(6), 1093–1096.
- Lev-Ari, S., & Keysar, B. (2012). Less-detailed representation of non-native language: Why non-native speakers' stories seem more vague. *Discourse Processes*, *49*(7), 523–538.
- Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong?. *Journal of Experimental Social Psychology*, *79*, 328–348.
- Livingston, B. A., Schilpzand, P., & Erez, A. (2017). Not what you expected to hear: Accented messages and their effect on choice. *Journal of Management*, *43*(3), 804–833.
- MacGregor, L. J., Corley, M., & Donaldson, D. I. (2010). Listening to the sound of silence: Disfluent silent pauses in speech have consequences for listeners. *Neuropsychologia*, *48*(14), 3982–3992.
- Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97.
- Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, *38*(3), 289–306.
- Munro, M. J., & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning*, *48*(2), 159–182.
- Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech the role of speaking rate. *Studies in Second Language Acquisition*, *23*(4), 451–468.
- Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, *34*(4), 520–531.
- Pallier, C., Sebastiángalles, N., Felguera, T., Christophe, A., & Mehler, J. (1993). Attentional allocation within the syllabic structure of spoken words. *Journal of Memory and Language*, *32*(3), 373–389.
- Pinget, A. F., Bosker, H. R., Quené, H., & De Jong, N. H. (2014). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing*, *31*(3), 349–365.
- Rose, R. L. (2019). The structural signaling effect of silent and filled pauses. In *The 9th workshop on disfluency in spontaneous speech* (p. 19).

- Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review*, 65(3), 395–412.
- Souza, A. L., & Markman, A. B. (2013). Foreign accent does not influence cognitive judgments. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35, No. 35).
- Stocker, L. (2017). The impact of foreign accent on credibility: An analysis of cognitive statement ratings in a Swiss context. *Journal of Psycholinguistic Research*, 46(3), 617–628.
- Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, 30(4), 485–496.
- Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, 28(1), 1–30.
- Wu, H., & Leung, S. O. (2017). Can Likert scales be treated as interval scales?—A Simulation study. *Journal of Social Service Research*, 43(4), 527–532.

Appendix

Full model result

	Estimate	Est. Error	Q2.5	Q97.5
Intercept [1]	−2.481	0.527	−3.51	−1.444
Intercept [2]	−1.017	0.526	−2.046	0.016
Intercept [3]	0.375	0.525	−0.649	1.409
Intercept [4]	1.93	0.526	0.904	2.965
Intercept [5]	3.79	0.528	2.762	4.829
Pause 1 (No Pauses)	0.218	0.314	−0.402	0.835
Pause 2 (Between clause pauses)	0.216	0.293	−0.361	0.785
Speaker Background (Native)	0.826	0.382	0.038	1.543
Statement Type (False)	−0.086	0.418	−0.901	0.74
Task (Credibility)	−0.254	0.48	−1.195	0.686
Speaking Rate (Syllables per second)	0.179	0.124	−0.06	0.424
Pause 1: Speaker Background	0.229	0.323	−0.414	0.859
Pause 2: Speaker Background	0.005	0.315	−0.607	0.631
Pause1: Statement Type	0.011	0.322	−0.624	0.638
Pause 2: Statement Type	−0.593	0.318	−1.212	0.028
Speaker Background: Statement Type	0.228	0.305	−0.373	0.83
Pause 1: Task	−0.464	0.339	−1.122	0.2
Pause 2: Task	0.132	0.306	−0.466	0.734
Speaker Background: Task	−1.19	0.367	−1.868	−0.427
Statement Type: Task	−0.027	0.403	−0.817	0.772
Pause 1: Speaker Rate	0.086	0.075	−0.061	0.234
Pause 2: Speaker Rate	−0.099	0.071	−0.238	0.04
Speaker Background: Speaking Rate	−0.01	0.078	−0.163	0.145
Statement Type: Speaking Rate	−0.014	0.102	−0.215	0.186

(Continued)

Full model result (*Continued*)

	Estimate	Est. Error	Q2.5	Q97.5
Task: Speaking Rate	-0.168	0.111	-0.388	0.048
Pause 1: Speaker Background: Statement Type	0.037	0.329	-0.603	0.69
Pause 2: Speaker Background: Statement Type	-0.093	0.339	-0.743	0.582
Pause 1: Speaker Background: Task	0.099	0.313	-0.515	0.713
Pause 2: Speaker Background: Task	-0.201	0.317	-0.817	0.422
Pause 1: Statement Type: Task	0.108	0.347	-0.579	0.787
Pause 2: Statement Type: Task	0.057	0.326	-0.598	0.689
Speaker Background: Statement Type: Task	-0.361	0.276	-0.893	0.188
Pause 1: Speaker Background: Speaking Rate	0.003	0.078	-0.15	0.155
Pause 2: Speaker Background: Speaking Rate	-0.028	0.077	-0.179	0.123
Pause 1: Statement Type: Speaking Rate	0.017	0.077	-0.136	0.168
Pause 2: Statement Type: Speaking Rate	0.148	0.075	0.002	0.295
Speaker Background: Statement Type: Speaking Rate	-0.082	0.073	-0.226	0.061
Pause 1: Task: Speaking Rate	0.015	0.082	-0.147	0.174
Pause 2: Task: Speaking Rate	-0.003	0.075	-0.15	0.144
Speaker Background: Task: Speaking Rate	0.117	0.07	-0.019	0.259
Statement Type: Task: Speaking Rate	-0.057	0.099	-0.253	0.136
Pause 1: Speaker Background: Statement Type: Task	0.283	0.321	-0.347	0.913
Pause 2: Speaker Background: Statement Type: Task	-0.134	0.336	-0.8	0.523
Pause 1: Speaker Background: Statement Type: Speaking Rate	0.002	0.079	-0.155	0.156
Pause 2: Speaker Background: Statement Type: Speaking Rate	0.004	0.08	-0.155	0.161
Pause 1: Speaker Background: Task: Speaking Rate	-0.081	0.075	-0.228	0.066
Pause 2: Speaker Background: Task: Speaking Rate	0.076	0.078	-0.076	0.227
Pause 1: Statement Type: Task: Speaking Rate	-0.006	0.085	-0.172	0.161
Pause 2: Statement Type: Task: Speaking Rate	-0.015	0.079	-0.168	0.142
Speaker Background: Statement Type: Task: Speaking Rate	0.109	0.066	-0.021	0.237
Pause 1: Speaker Background: Statement Type: Task: Speaking Rate	-0.059	0.077	-0.211	0.094
Pause 2: Speaker Background: Statement Type: Task: Speaking Rate	0.029	0.081	-0.128	0.19

Cite this article: Houghton, Z., Kato, M., Baese-Berk, M., & Vaughn, C. (2024). Task-dependent consequences of disfluency in perception of native and non-native speech. *Applied Psycholinguistics* 45, 64–80. <https://doi.org/10.1017/S0142716423000486>