LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Fluency in Dialogue: Turn-Taking Behavior Shapes Perceived Fluency in Native and Nonnative Speech

Marjolein van Os,[a] Nivja H. de Jong,[b]
and Hans Rutger Bosker[c]

[a]Language Science and Technology, Saarland University, Saarbrücken, Germany, [b]ICLON Graduate School of Teaching/Leiden University Centre for Linguistics (LUCL), Leiden University, Leiden, The Netherlands, and [c]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Abstract: Fluency is an important part of research on second language learning, but most research on language proficiency typically has not included oral fluency as part of interactions even though natural communication usually occurs in conversations. The present study considered aspects of turn-taking behavior as part of the construct of fluency and investigated whether these aspects differentially influence perceived fluency ratings of native and nonnative speech. Results from two experiments using acoustically manipulated speech showed that, in native speech, too "eager" answers (interrupting a question with a fast answer) and too "reluctant" answers (answering slowly after a long turn gap) negatively affected fluency ratings. However, in nonnative speech, only too "reluctant" answers led to lower fluency ratings. Thus, we demonstrated that acoustic properties of dialogue are perceived as part of fluency. By adding to the current understanding of dialogue fluency, these lab-based findings carry implications for language teaching and assessment.

We would like to thank Marie Stadtbäumer for her help in running Experiment 2 and our native and nonnative speakers for the voice recordings.

All data from the present study, together with an R analysis script, are available for download (under a CC BY-NC-ND 4.0 license) from: https://osf.io/b465y/

Correspondence concerning this article should be addressed to Hans Rutger Bosker, Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH, Nijmegen, The Netherlands. E-mail: HansRutger.Bosker@mpi.nl

The handling editor for this article was Judit Kormos.

DOI: 10.1111/lang.12416

Keywords fluency; turn-taking; nonnative fluency; dialogue; dialogic fluency; second language; speech; conversation

Introduction

Billions of people are currently learning to communicate in a second language (L2). It is therefore no surprise that researchers have expended considerable effort into understanding L2 learning. A growing consensus is that the primary goal of L2 education is to enable learners to become competent language users. For oral abilities, this implies that learners need to develop the ability to achieve communicative goals in monologues as well as in interaction. Because interaction is reciprocal, that is, interlocutors take turns in speaking and listening, L2 learners need to be able to produce and understand messages in real time, to adjust messages to what they perceive the speech partner's understanding to be, and to monitor and manage the interactional encounter itself (Bygate, 1987). Part of the management of interaction revolves around appropriate turn-taking behavior in conversation (Young, 2011). L2 learning therefore includes learning how to take turns in conversation. Likewise, when L2 oral abilities are assessed, turn-taking abilities should be part of the evaluation. The current study focused on how turn-taking behavior is evaluated when a speaker's fluency is assessed.

Some definitions of the term fluency tend to focus on the general linguistic abilities of L2 speakers, who, when fluent, sound nativelike (Lennon, 1990). This seems to imply that all native speakers of a language speak fluently and with the same degree of fluency. However, listening to several native speakers of one's own language reveals that this is not the case. Some speakers may indeed speak fluently, with few pauses, restarts, or corrections, but other speakers constantly stumble over their words, or speakers differ in how well they can say the appropriate thing in varying contexts, how creatively they can use language, or how coherent their utterances are (Fillmore, 1979; Goldman-Eisler, 1968; Lennon, 1990; Riggenbach, 1991). There can be additional variation within individual speakers as a function of their emotional state, speech register, and audience (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001).

It is not clear how to define fluency in both native and nonnative speech objectively because many different variables play a role, as does the purpose of the definition (Chambers, 1997; Götz, 2013). We have mentioned disfluencies such as pauses, restarts, and corrections, but other variables possibly play a role as well. Fluency has been used as a typical component in oral proficiency and has been shown to make up an important part of overall oral assessment (Iwashita, Brown, McNamara, & O'Hagan, 2008). The highest level of the Cambridge English Qualifications expects candidates to have "many features, including pausing and hesitating, [that] are 'native-like'" (Ffrench, 2003, p. 15). As we have mentioned, native speakers vary in their fluency, which makes it hard to assess whether nonnative speakers have reached nativelike levels of fluency.

In everyday conversation, speakers tend not to hold endless monologues, but they are rather engaged in dialogues with others. Because of the co-constructed nature of conversation, the perceived fluency of speakers in dialogue will not depend solely on the quality of their speech but also on the interaction with their interlocutors (McCarthy, 2010; Peltonen, 2017). In other words, when language learners are striving to develop their interactional competence, they will need to learn how to co-construct interaction that includes topic management, interactive listening, breakdown repair, nonverbal or visual behaviors, and turn management (Galaczi & Taylor, 2018, p. 226).

The present study aimed to contribute to the field of fluency research by investigating whether one such fundamental characteristic of dialogic spoken communication—turn-taking behavior—influences fluency perception of native and nonnative speech. We specifically tested the effect of gaps (a pause of 0, 300, 600, or 900 milliseconds) and overlaps (simultaneous speech for 300 or 600 milliseconds) between question and answer turns in dialogues on the perceived fluency of the answers produced by native (Experiment 1) and nonnative speakers of Dutch (Experiment 2). Thus, we aimed to contribute to the understanding of the construct of fluency and assess how aspects of dialogue fluency such as turn-taking affect fluency perception when perception may be partially based on interactive components.

In the present study, we have used the term *native speaker* to refer to speakers who have grown up with a particular language from birth. We have used the term *nonnative speaker* to refer to speakers who began learning a language at a later age in life and who are still in the process of learning that language. We acknowledge that this is a simplified view of language competence, one where nativelike speaker competence is a goal of the language learner, although this is not L2 speakers' goal in all circumstances (see Ortega, 2019).

Background Literature

Oral Fluency in Native and Nonnative Speech

Definitions of Fluency

Several studies have given definitions of fluency in L2 speech. In an influential paper, Lennon (1990) distinguished two senses of fluency: fluency in the broad sense and fluency in the narrow sense. Fluency in the broad sense can be seen as another term for oral proficiency in general. It encompasses speech that is grammatically correct, that uses a large vocabulary, and that is pronounced in a nativelike way. This is the definition most often used in everyday life when someone states, for example, "He is fluent in Italian" (Chambers, 1997). In contrast, fluency in the narrow sense constitutes one specific component of oral proficiency that deals with the flow and smoothness of speech and focuses on speech production at a rate similar to the rate of native speakers of the language, without pauses, hesitations, fillers, and corrections.

Beside this distinction between fluency in the broad and narrow sense, Segalowitz (2010) proposed another definition of fluency from a cognitive perspective. He distinguished three senses of fluency: perceived fluency, utterance fluency, and cognitive fluency. Perceived fluency refers to the impression listeners have of a speaker's cognitive fluency based on the speech signal. Lennon (1990) also mentioned this, stating that "fluency is an impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently" (p. 391). Utterance fluency refers to the features of an utterance that can be measured acoustically such as speech rate, pauses, and repairs. Utterance fluency is subdivided into measures of breakdown fluency (e.g., number, duration, and distribution of pauses), speed fluency (e.g., mean length of syllables), and repair fluency (e.g., number of corrections and repetitions; Tavakoli & Skehan, 2005; see Tavakoli & Uchihara, 2020, for a recent study examining these aspects of utterance fluency). Finally, cognitive fluency refers to a speaker's ability to efficiently coordinate the cognitive processes, such as lexical retrieval and phonetic encoding, needed for speech production. In the present study, we were concerned with fluency in the narrow sense and, in particular, with the relationship between utterance fluency and perceived fluency.

Acoustic Correlates of Nonnative Fluency

Three studies in particular have investigated the relation between utterance fluency (objective acoustic measurements) and perceived fluency (subjective ratings by listeners) in monologues and attempted to identify the acoustic correlates of oral fluency.

Cucchiarini, Strik, and Boves (2002) investigated the objective properties of speech that can be used to predict the perceived fluency of nonnative Dutch speakers. Trained judges rated both read speech and spontaneous speech for perceived fluency, and their ratings correlated with different quantitative measures of fluency. In spontaneous speech, the perceived fluency ratings related in particular to frequency and distribution of pauses, but speed of articulation (measured in number of phonemes per second) did not show such a relationship. In read speech, both speed of articulation and frequency of pauses were related to fluency ratings.

Rossiter (2009) compared fluency ratings of nonnative speech made by different groups of raters: trained native speakers, untrained native speakers, and proficient nonnative speakers. All three groups' fluency ratings were related to unfilled or nonlexical filled pauses, self-repetitions, and speech rate (measured in syllables per second).

Bosker, Pinget, Quené, Sanders, and De Jong (2013) investigated the contribution of pauses, speech rate, and repairs (utterance indices that were measured acoustically) to perceived fluency. They used six objective measures of fluency calculated on spoken time excluding silences. These measures were the mean length of syllables, the number of silent pauses per second spoken time, the number of filled pauses per second spoken time, the mean length of silent pauses, the number of repetitions per second spoken time, and the number of corrections per second spoken time. Results showed that the number and mean length of silent pauses and mean length of syllables were most strongly related to fluency perception of nonnative speech. The authors concluded that breakdown fluency and speed fluency play a larger role than repair fluency in perception of nonnative fluency. These findings were in line with other studies showing that nonnative speakers who are perceived as more fluent use fewer pauses (both filled and silent) and have a faster speech rate (Chambers, 1997; Lennon, 1990; Segalowitz, 2010).

Native Speakers' Fluency

Native speakers are typically deemed fluent by default (Davies, 2003; Riggenbach, 1991), but individual variation in utterance fluency (i.e., number of disfluencies) can be found also between native speakers of a language, for example, based on the age or the gender of speakers (Bortfeld et al., 2001). Although relatively few studies have investigated fluency in native speakers, some have compared native and nonnative speakers on comparable tasks (see Götz, 2013, for an overview for English).

Bosker, Quené, Sanders, and De Jong (2014) investigated perceived fluency in both native and nonnative Dutch speech and particularly focused on the question of whether listeners evaluate similarly fluency characteristics of native speech and of nonnative speech. They conducted two experiments with acoustic manipulations that allowed them to draw conclusions about the causal effects of particular fluency characteristics. In the first experiment, they manipulated silent pauses, but, in the second experiment, they manipulated speech rate. Results showed that for both native and nonnative speech, increasing the number of silent pauses and lengthening the duration of silent pauses had a negative effect on fluency ratings. There was no difference in the size of this effect between natives and nonnatives. For speech rate, they found a decrease in fluency ratings when they slowed down native speech to the nonnative level and found an increase in fluency ratings when they sped up nonnative speech to the native level. The relative decrease and increase were notably of similar magnitude. This suggested that a single silent pause or a particular speech rate is weighed similarly in native and nonnative fluency perception. Bosker et al. (2014) concluded that the relationship between utterance fluency and perceived fluency is similar for native and nonnative speakers.

Kahng (2014, 2018) also used speech manipulations to investigate how silent pause distributions in both native and nonnative speech affect raters' fluency judgments. Results showed a similar effect of pause distribution for both native and nonnative speech. Speech without pauses was rated as more fluent than speech with pauses either between or within clauses, and speech with pauses between clauses was judged to be more fluent than speech with pauses within clauses. These results, combined with those of Bosker et al. (2014) suggested that those aspects of speech affecting fluency ratings of nonnative speakers—speech rate and use of pauses and hesitations—affect the perception of native speakers' fluency in a similar way.

Fluency in Dialogues

The definitions of fluency mentioned earlier all focus on the speech of language users in monologues. However, in everyday life, speech is rarely produced in monologue settings but rather in collaboration with other speakers. As McCarthy (2010) noted, "fluency also involves the ability to create flow and smoothness across turn-boundaries and can be seen as an interactive phenomenon in discourse" (p. 1). He proposed the term *confluence* to refer to the joint process of two speakers who cooperate to create a fluent interaction. Here he focused particularly on turn-openings and turn-closings because these show how fluency is constructed interactively between speakers. Speakers in a dialogue share the responsibility of creating and maintaining a flowing conversation and of filling silences. Peltonen (2017) introduced the term *dialogue fluency* to refer to the contributions of individual speakers to the collaborative aspects of fluency in a dialogue, measured as the number of turn pauses, the mean length of turn pauses, the number of repetitions of what the other speaker said, and the number of collaborative completions. This differs from

McCarthy's confluence in that these can be measured objectively and take into account more than turn-taking and subjective impressions of the flow of conversation. Sato (2014) discussed *interactional fluency*, focusing on the perceived fluency of language learners engaged in paired decision-making tasks. He found that fluent speakers show natural patterns of back-channeling and turn-taking, but disfluent speakers are hesitant in starting their turn.

The term interactional fluency used by Sato (2014) followed from the term *interactional competence* (Galaczi & Taylor, 2018; Young, 2011). As noted above, research into fluency has so far capitalized on understanding the construct of fluency in monologue speaking competence. However, as yet, there has been little research on the construct of fluency in dialogues, which is a component of language learners' interactional competence. Many different elements make up interactional competence (Salaberry & Kunitz, 2019), a competence that interests not only fluency researchers but also those in the field of conversation analysis in L2 acquisition. The many facets of interactional competence make it difficult to define the concept clearly and completely. In many cases, it is only qualitatively described in raters' instructions, which point out interactional skills like taking turns, keeping the floor, and engaging in conversation in a clearly participatory manner, without being explicit about specific behaviors such as the timing of turn-taking. This complicates the objective assessment of learner's interactional competence. Another challenge with regard to speaking assessment in pairs is the question of how scores should be assigned to individuals based on a jointly constructed interaction (May, 2009). Sato (2014) has suggested an interactional oral fluency rating scale that was empirically based and succeeded in differentiating fluent from disfluent speakers. However, this scale has not been validated for groups of learners other than the group of Japanese university-level learners of English who participated in Sato's study. Better knowledge of how dialogue settings affect perceived fluency would help create more precise and focused definitions for raters and, in that way, lead to more valid and reliable assessments of language proficiency.

One of the first studies to investigate the effect of turn-taking on perceived fluency was Riggenbach (1991). She analyzed the speech of six nonnative speakers of English, three of whom had been rated as nonfluent by trained raters and three of whom had been rated as fluent. She examined hesitation phenomena including filled and unfilled pauses, repair phenomena such as restarts, rate of speech, and speed of speech, and various interactive phenomena present in dialogues, for example, backchannels, questions, and turn change types such as overlaps and gaps. Results of this analysis showed that slower speech rate and more use of unfilled pauses contributed to lower fluency ratings.

However, because the sample size of the study was very small, it would be hard to generalize these results to other nonnative speakers.

A more systematic approach was used by Galaczi (2014), who investigated how language learners manage interaction in paired speaking assignments. For this, she used the framework of conversation analysis in L2 acquisition. Some parts of her findings can be interpreted as reflecting dialogue fluency. Quantitative and qualitative analyses of turn-taking management showed that speakers who were more proficient were better able to create confluence than were less proficient speakers. Speakers with a lower proficiency level had a weak alignment with longer pauses between turns, but speakers with a higher proficiency level showed rapid speaker changes and typically managed their turns in a no-gap–no-overlap manner that native speakers also prefer.

Research focusing mainly on utterance fluency has shown that speakers are more fluent in dialogue tasks compared to monologue tasks (Michel, Kuiken, & Vedder, 2007; Sato, 2014; Tavakoli, 2016; Witton-Davies, 2014) and that high-proficiency speakers in dialogue settings are more fluent than speakers with lower proficiency (Peltonen, 2017).

In L2 testing practice, several formats of speaking assessment are used. For standardization purposes, these often make use of monologue tasks. For example, the official L2 Dutch exams contain a speaking assessment where the candidate sits in front of a computer and records several speaking assignments (College voor toetsen en examens, 2017). A similar procedure is used in the Test of English as a Foreign Language, where candidates record their spoken responses to several questions that are then scored by raters. Although it attempts to simulate some aspects of dialogue, the test takes place on a computer with candidates responding to recorded questions, and thus the candidates do not interact with other interlocutors. Other speaking assessments, such as the official Cambridge English qualifications, make use of dialogue settings. For these assessments, two candidates are tested at the same time, and they speak with each other and with a test leader. Candidates are assessed on, among other things, interactive communication, which refers to a candidate's sensitivity to turn-taking (Ffrench, 2003), that is, interactional competence. The International English Language Testing System assesses speaking ability based on an interactive setting between learner and examiner, although the scoring criteria do not include any interactional competences or aspects of turn-taking (International English Language Testing System, n.d.). The Common European Framework of Reference (Council of Europe, 2001) does mention turn-taking as part of discourse competence and focuses on speakers' taking the floor and initiating and maintaining conversations. Results from the present study can help in understanding the construct of fluency in dialogues as part of overall interactional competence.

Managing Dialogues

In contrast to monologues, a dialogue includes by definition more than one interlocutor and requires coordinated processing. This leads to added time pressure for speakers to plan utterances in comparison to the requirements of monologues (Garrod, 1999). Both interlocutors have to ensure that their utterances are understood in the way that they intend them to be and that they follow the intended meaning of their interlocutor, as well as ensuring that they perform these communicative actions at appropriate time points. Interlocutors facilitate these processes by constructing a common ground as well as by perspective taking (Traxler, 2012).

Interlocutors in a dialogue constantly change roles from being speakers to being listeners. Changing roles is done by turn-taking. Speakers follow a universal tendency to avoid overlapping turns and at the same time try to minimize pauses between turns (Sacks, Schegloff, & Jefferson, 1978; Stivers et al., 2009). Turn gaps, overlaps, and delays are all terms used to refer to the same concept—the time between the turn of the first speaker and the turn of the second speaker's replying to the first. We use the term *delay* to refer to this phenomenon in general; we use *gap* to refer to a pause between the two turns, thus a positive temporal relation between the two turns; and we use *overlap* to refer to two speakers' speaking at the same time (a negative temporal relation). Example 1 illustrates a gap where there is a pause of 600 milliseconds between two speakers' turns; Example 2 illustrates two speakers who overlap while speaking.

Example 1
Question:
Do you enjoy cooking, are you good at it?
Answer:
(600 ms gap) I enjoy cooking, but it shouldn't be too complicated. Then it usually goes well.

Example 2
Question (brackets indicate overlapped speech):
Do you enjoy cooking, are you good [at it?]
Answer:
[As long] as I can follow a recipe, I can cook quite well. Improvising is not my strong suit, then it usually goes wrong.

Using a worldwide, typologically diverse sample of 10 different languages, Stivers et al. (2009) investigated the timing of turn-taking in polar questions, for example, yes–no questions. Results showed that the temporal relation between the answer and the question followed a unimodal distribution, with the overall mode at 0 milliseconds (language-specific modes between 0 and 200 milliseconds; Dutch: 100 milliseconds). The mean time between the question and the answer was 208 milliseconds, with language-specific means ranging from 7 to 469 milliseconds. They concluded that the language-specific means fell within approximately 250 milliseconds on either side of the cross-linguistic mean, leading the researchers to conclude that speakers generally avoid overlapping talk and minimize the silence between conversational turns (no-gap–no-overlap strategy). However, other studies have emphasized the considerable variation that is present in turn timing due to, among other variables, characteristics of the social setting, complexity of the response, and the general context, as well as some individual differences (Heldner & Edlund, 2010; Meyer, Alday, Decuyper, & Knudsen, 2018; Roger & Nesshoever, 1987).

Studies have shown that deviating from this no-gap–no-overlap strategy has communicative significance. Longer gaps may signify comprehension problems (Beňuš, Gravano, & Hirschberg, 2011), speech planning difficulties (Bull & Aylett, 1998), or production of longer (Torreira, Bögels, & Levinson, 2015) or even disconfirmative responses and nonanswers (Stivers et al., 2009). Speakers who produce larger gaps between their turn and that of the other speaker are seen as less affiliative and more distancing because the speaker is judged as less willing to comply with a request or agree with the other speaker (Roberts, Margutti, & Takano, 2011). At the same time, speakers who produce overlapping turns (i.e., interrupting their interlocutor) are seen as less affiliative (Van Leeuwen, 2017), as less agreeable and more assertive (Robinson & Reis, 1989; ter Maat, Truong, & Heylen, 2010), as more dominant (Beňuš et al., 2011), and as less sociable (Robinson & Reis, 1989) in comparison to speakers who do not produce overlapping turns. Additionally, overlaps are linked to displays of power and control and interpreted as rude and disrespectful (Goldberg, 1990).

The Present Study

Studies have shown that speakers prefer no-gap–no-overlap between turns in a dialogue (Sacks et al., 1978; Stivers et al., 2009) and that nonnative speakers improve their dialogue management skills as they improve their L2 proficiency (Galaczi, 2014; Peltonen, 2017). However, no study beside that of Riggenbach (1991) has investigated how turn-taking behavior in dialogue contributes to the perception of fluency of native and nonnative speech.

The aim of the present study was to investigate in Experiment 1 how native speakers of Dutch are rated on fluency when their turn-taking behaviors differ. In the present study, in both experiments, we investigated the effect of various delay steps—overlaps (−600 and −300 milliseconds) or gaps (0, 300, 600, and 900 milliseconds) between the turns of question–answer sequences between native speakers—on fluency ratings. Additionally, we manipulated the speaking rate (fast vs. slow) of the answers. We included this manipulation in the experiment as an additional check to see whether our experimental design could capture fluency differences between conditions. If we had manipulated only the delay step and had found no difference between different steps, this finding might have been due to no effect of delay step on fluency, or it might have been due to the experimental design's not being sensitive enough to find it. We expected that results would show an effect of speaking rate, where fast answers would yield higher fluency scores than would slow recordings because studies have shown that faster speech is perceived as more fluent than slower speech (Bosker et al., 2014). Replicating these findings would show that the manipulation in the present experiment was valid and that the experiment was measuring fluency in a similar fashion to that of previous studies.

For the effect of overlap and gap, we expected that results would show lower fluency scores for larger gaps between questions and answers because longer pauses are associated with less fluent speech even for native speakers (Bosker et al., 2014; Götz, 2013). A longer pause signals that a speaker needs more time to formulate an answer, thus the speaker appears less fluent. Because interlocutors of a wide variety of languages prefer the no-overlap–no-gap strategy for turn-taking management (Stivers et al., 2009), we expected that the fluency ratings would be higher for a gap of 0 milliseconds than for larger gaps. We also expected high fluency ratings for the overlap conditions because, in these cases, a speaker apparently is able to quickly formulate and produce a response to the question that has been asked.

Experiment 2 resembled Experiment 1, but in it nonnative speakers of Dutch produced all the answers in the question–answer conversations. As we had for Experiment 1, we predicted an effect of speech rate, that is, raters would find fast speech to be more fluent than slow speech because the speech of high-proficient nonnatives generally has been found to be faster than that of nonnative speakers with lower proficiency (Chambers, 1997; Cucchiarini et al., 2002; Lennon, 1990; Rossiter, 2009; Segalowitz, 2010). Moreover, we predicted that, also for the nonnative speakers' recordings, larger gaps would receive lower fluency scores and larger overlaps higher fluency scores. Again, we expected that gaps of 0 milliseconds would be rated as most fluent because research has shown that proficient nonnative speakers want to minimize gaps just as do native speakers (Galaczi, 2014; Peltonen, 2017). Because the answers in the two experiments involved the same lexical content, they already implied relatively high proficiency for the L2 speakers. Therefore, we presumed that raters would assume that these speakers showed the same pattern of minimizing gaps as that preferred by native speakers.

The assessment of the effect of turn gaps and speech rate on the fluency perception of both native and nonnative speech (i.e., across both experiments) allowed us to test whether effects are similar or different in perception of native versus nonnative fluency. Earlier studies (Bosker et al., 2014; Kahng, 2014, 2018) have consistently shown that pausing behavior and speech rate are weighed similarly in native and nonnative fluency perception. In line with these observations, we predicted similar effects for turn gaps and for speech rate across native and nonnative fluency perception.

We describe the design and methods for each experiment separately before presenting the results and discussion for the two experiments together.

Experiment 1: Native Speakers

Method

Listeners

We recruited 49 listeners from the Max Planck Institute for Psycholinguistics participant pool. We obtained informed consent from the participants before the experiment and paid them for their participation. We lost the data of one listener due to technical problems. Thus, we analyzed the ratings of 48 listeners. All listeners were native speakers of Dutch with normal hearing. The mean age of the group was 22.33 years (range = 18–30, $SD = 2.47$); 14 listeners were male, 34 were female.

Materials

We constructed 80 question–answer pairs, consisting of 40 different questions and 80 different answers, two per question. The questions covered everyday topics like hobbies, holiday plans, and the weather. We structured the questions so that it was plausible that the answer could be given before the end of the question had been heard. The answers had a length ranging from 19 to 44 syllables, with a mean length of 33.40 syllables. An example of a question–answer pair can be found in Example 3. For an overview of all 80 question–answer pairs, see Appendix S1 in the Supporting Information online.

Example 3

Question:

"*Hoe reis je naar je werk, normaal gesproken*?"

How do you travel to work, normally speaking?

Answer 1:

"*Ik ga altijd met de auto. Ik woon vrij ver van mijn werk en het is voor mij niet handig om de trein te nemen*."

I always go by car. I live quite far away from work and it's not convenient for me to take the train.

Answer 2:

"*Ik neem eigenlijk altijd de fiets om op mijn werk te komen, behalve als het regent. Dan ga ik met de bus*."

I actually always take the bike to get to my work, except when it is raining. Then I go by bus.

We made recordings of 10 native Dutch speakers (two male, eight female) reading a script consisting of eight questions and eight answers each, recording each speaker in a separate session. We constructed the lists so that no speaker answered the questions that they themselves had read and so that, for one rater's stimuli, when Speaker B answered questions asked by Speaker A, Speaker A did not answer questions asked by Speaker B. No speaker read both answers to a question. We instructed all speakers to read the sentences in a natural way as if the sentences were part of a natural dialogue, and we told the speakers that it did not matter if they hesitated or corrected themselves. All speakers read their questions and answers twice.

We isolated all questions and answers from the recordings using Praat (Boersma & Weenink, 2013) by cutting them at zero-crossings, and we selected for each the better one (as evaluated subjectively by the first and last author, both native speakers of Dutch) from the two recorded versions, paying attention to clarity of speech (no slips of the tongue) and intonation. The recordings of the questions had a mean length of 2.51 seconds ($SD = 0.62$).

We removed leading and ending silent intervals as well as leading and ending filled pauses from the answer recordings using Praat. The mean duration of the trimmed answer recordings was 6.35 seconds ($SD = 1.37$) and ranged from 3.38 to 10.36 seconds. They had a length ranging from 19 to 44 syllables, with a mean length of 33.49 syllables ($SD = 6.03$). The mean number of syllables per second in the original (i.e., unmanipulated) answer recordings was 5.16 syllables per second ($SD = 1.11$) and ranged from 2.03 to 7.13 syllables per second.

We scaled the intensity of all questions and answers to 65 dB. We linearly compressed the speech rate of the answers by a factor of 0.83 (using PSOLA in Praat) to speed the speech rate up to create the fast speech condition. We also expanded the speech rate of the original answers by a factor of 1/0.83, or 1.20, to slow the speech rate down to create the slow speech condition. Bosker et al. (2014), who compared native and nonnative speech, also used these factors and found that a large majority of listeners (85%) judged these speech rates to be natural. We did not manipulate the speech rate of the questions. We manipulated the delay between the question and the answer in Praat to achieve an overlap of 600 milliseconds, 300 milliseconds, 0 milliseconds (i.e., no overlap or gap), or to have a gap of 300, 600, or 900 milliseconds. We chose these steps to fall within the peak of the distribution of turn transitions for Dutch as found by Stivers et al. (2009) and to have a step size larger than 250 milliseconds (De Jong & Bosker, 2013). This resulted in 12 unique conditions (two speech rate conditions, each with six delay steps).

We excluded four questions and their eight corresponding answers because at least one of the answers sounded relatively unnatural as a reply to the question (as judged by the first author). This resulted in the 72 question–answer pairs that we used in the experiment.

Design

We arranged the experimental items in a Latin square design: Listeners heard each question–answer pair in only one of the 12 conditions but heard all conditions during the experiment. There were 12 different pseudo-randomized lists of the stimuli. Each list consisted of mini-blocks of 10 speakers, so that the answers in the trials in a given mini-block were all spoken by different speakers. We made sure that the same question was not presented a second time immediately after itself and that the same speaker did not produce the answers of two consecutive trials across mini-block boundaries. We also reversed all 12 lists, and this resulted in 24 different orders of experimental items. Each list contained the same number of fast items and slow items ($n = 36$ each), and each delay step occurred 12 times.

Procedure

We ran the experiment using Presentation (v16.5; Neurobehavioral Systems, Albany, CA, USA). We conducted the experiment in a sound-attenuating booth and presented the audio through headphones at a comfortable volume. Before the experimental task started, listeners signed the informed consent sheet, and the experiment leader gave a short introduction to the task. The experiment started with written instructions that we presented on the screen. We told the listeners that we were interested in how listeners perceive fluency in different speakers. We instructed the listeners to listen to the question–answer pairs and rate the fluency of the second speaker, that is, the speaker who gave the answer. We asked them to do this using a 9-point scale ranging from *not fluent at all* on the lower end of the scale and *very fluent* on the higher end of the scale (and no other anchor points; see Bosker et al., 2013, 2014). We instructed them to do this not on the basis of fluency in the broad sense (i.e., overall language proficiency, as in the statement, "He is fluent in Italian"), but rather to base their judgments on the use of filled and silent pauses, speech rate, and use of hesitations and corrections and thus to focus on fluency in the narrow sense (see Appendix S2 in the Supporting Information online). Instructions to raters to judge fluency in the narrow sense have been used by previous fluency perception studies (Bosker et al., 2013, 2014; Derwing, Rossiter, Munro, & Thomson, 2004; Pinget, Bosker, Quené, & De Jong, 2014; Rossiter, 2009) and are compatible with instructions given to raters of language tests.

We informed the listeners that all the speakers were native speakers, and we instructed them to use the entire scale from 1 to 9. Important for the experiment was the fact that the listeners could only make their judgments after the entire speech stimulus (i.e., question and answer) had played. They could only listen to each recording once. The listeners self-paced their reading of the instructions and then started with four practice trials. These consisted of the four most naturally sounding question–answer pairs from the excluded items. Neither the practice questions nor the practice answers appeared in the experimental trials. The listeners heard two of the four practice trials in the fast condition and two in the slow condition. None of these practice trials contained the extreme overlap of 600 milliseconds or the extreme gap of 900 milliseconds. All the listeners heard the same four practice trials in the same order. The listeners did not receive feedback on their practice ratings. If they had no questions, they continued with the experiment. Halfway through the experiment, the listeners took a short self-timed break. After all 72 trials, the listener filled out a short post-experimental questionnaire investigating whether they had noticed anything about the recordings and what they thought about the experiment. Finally, we debriefed the listeners. The experimental procedure took approximately 30 minutes.

In the post-experimental questionnaire, we asked the listeners open questions about what they thought was the goal of the experiment, about whether they had noticed anything in particular, and about whether or not they thought the recordings sounded natural. Qualitative inspection of the responses showed that about half of the listeners (54.2%) had noticed that speech rate had been manipulated. Most of the listeners commented that, although most of the recordings sounded natural, some sounded accelerated or slowed down. This suggested that there was variation across speakers and listeners, as in fact all recordings had been rate-manipulated. Only four of the listeners (8.3%) noted that the answer sometimes started before the question had been finished, but they noted that this did not lead to unnatural-sounding fragments.

Experiment 2: Nonnative Speakers

Experiment 2 was identical to Experiment 1 except that we used nonnative speakers for the answer recordings.

Method

Listeners

We recruited a new group of 48 listeners (38 female, 10 male; $M_{\text{age}} = 23.00$ years, range = 18–31, $SD = 3$) from the Max Planck Institute for Psycholinguistics participant database and paid the listeners for their participation. None of these listeners had participated in Experiment 1. All listeners were native speakers of Dutch with normal hearing and sight; none had dyslexia and/or speech problems.

Design and Materials

The design and materials of Experiment 2 were identical to those of Experiment 1 except that we recorded nonnative speakers of Dutch (eight female, two male) producing the answers for the question–answer stimuli.

We used the 40 questions from Experiment 1, again combining each with two different answers, resulting in 80 different question–answer pairs. We used the same question recordings from Experiment 1, that is, they were spoken by the same native Dutch speakers. For Experiment 2, 10 nonnative Dutch speakers recorded the answers to the questions (the same answers as used in Experiment 1). Their oral proficiency in Dutch varied from very proficient—level C2 in the Common European Framework of Reference (Council of Europe, 2001) on an official Dutch as L2 exam—to very low proficiency—hardly any knowledge of Dutch. (Unfortunately, full proficiency data are not available). We scripted the sentence stimuli (and therefore all nonnative speakers produced the same grammatically coherent lexical content) to facilitate comparison between native and nonnative stimuli (see Appendix S2 in the Supporting Information online). The first and last author (both native speakers of Dutch) assessed all the speakers to have a high level of accentedness. All experimental comparisons involved within-speaker effects and were therefore not confounded with the speakers' accents. Each speaker read out eight of the answers. The nonnative speakers received the same instructions as the native speakers had received in Experiment 1, that is, we asked them to read in a natural manner and not to worry about accent, disfluencies, hesitations, or corrections.

We trimmed the answer recordings and linearly compressed or expanded them using the same procedure as we had used in Experiment 1. The mean duration of the 72 trimmed answer recordings (again, eight were excluded and some of them used as practice trials) was 11.67 seconds ($SD = 1.90$), ranging from 6.45 to 15.87 seconds. They had a length ranging from 19 to 44 syllables, with a mean length of 33.52 syllables ($SD = 6.07$). The mean number of syllables per second in the original answer recordings was 2.93 syllables per second ($SD = 0.65$) and ranged from 1.76 to 4.89 syllables per second. Again, we did not manipulate the speech rate of the questions.

Finally, we scaled the intensity of the answer recordings to 65 dB, after which we concatenated them with the question recordings from Experiment 1 using the same six conditions as we had used in Experiment 1 (−600, −300, 0, 300, 600, 900 milliseconds overlap/delay between question and answer), which again resulted in 12 unique conditions.

Procedure

The procedure was identical to that of Experiment 1, except that we informed the listeners that the persons answering the various questions were nonnative speakers of Dutch and thus had a nonnative accent.

We asked the listeners to fill in the same questionnaire as we had used in Experiment 1. Qualitative inspection showed that 19 listeners (39.6%) noticed the overlap manipulation, that is, that the answer began before the question ended. Twelve (25%) listeners mentioned the speech rate manipulation. Although the majority of listeners commented that most of the recordings sounded natural, many (60.4%) reported that at least some answers sounded less natural, mostly due to obvious speed manipulations, as had been the case in Experiment 1.

Results for Experiments 1 and 2

Table 1 presents the descriptive statistics for all speech rate and gap and overlap conditions in both experiments. The results showed that many listeners in Experiment 1 used the entire scale of 1 to 9 to make their fluency judgments as we had instructed them to do (22 used the whole scale; 14 used the scale

Table 1 Descriptive statistics for fluency rating scores in both rate conditions and in all overlap/gap conditions

Speech rate	−600 ms *M*(*SD*) 95% CI	−300 ms *M*(*SD*) 95% CI	0 ms *M*(*SD*) 95% CI	300 ms *M*(*SD*) 95% CI	600 ms *M*(*SD*) 95% CI	900 ms *M*(*SD*) 95% CI	Overall *M*(*SD*) 95% CI
Experiment 1: Native speech ($N = 48$)							
Fast	6.51 (1.97)	6.67 (1.81)	6.93 (1.67)	6.93 (1.62)	6.84 (1.70)	6.86 (1.68)	6.79 (1.75)
	[6.28, 6.74]	[6.46, 6.88]	[6.73, 7.12]	[6.75, 7.12]	[6.64, 7.04]	[6.66, 7.05]	[6.71, 6.87]
Slow	5.38 (1.98)	5.42 (1.94)	5.45 (1.96)	5.28 (1.99)	5.23 (1.84)	5.27 (1.86)	5.34 (1.93)
	[5.15, 5.60]	[5.19, 5.65]	[5.23, 5.68]	[5.05, 5.51]	[5.02, 5.44]	[5.05, 5.48]	[5.25, 5.43]
Overall	5.94 (2.05)	6.04 (1.98)	6.19 (1.96)	6.11 (1.99)	6.03 (1.95)	6.06 (1.94)	6.06 (1.98)
	[5.77, 6.11]	[5.88, 6.21]	[6.03, 6.35]	[5.94, 6.27]	[5.88, 6.19]	[5.90, 6.22]	[6.00, 6.13]
Experiment 2: Nonnative speech ($N = 48$)							
Fast	6.88 (1.78)	6.83 (1.85)	6.77 (1.91)	6.89 (1.84)	6.82 (1.81)	6.84 (1.79)	6.84 (1.83)
	[6.67, 7.09]	[6.62, 7.04]	[6.55, 6.99]	[6.67, 7.10]	[6.61, 7.03]	[6.63, 7.05]	[6.75, 6.92]
Slow	5.24 (1.86)	5.28 (1.82)	5.31 (1.86)	5.24 (1.89)	5.03 (1.83)	5.08 (1.89)	5.20 (1.86)
	[5.03, 5.46]	[5.07, 5.49]	[5.10, 5.53]	[5.02, 5.45]	[4.82, 5.25]	[4.86, 5.30]	[5.11, 5.29]
Overall	6.06 (2.00)	6.06 (2.00)	6.04 (2.02)	6.06 (2.04)	5.93 (2.03)	5.96 (2.04)	6.01 (2.02)
	[5.90, 6.22]	[5.89, 6.22]	[5.88, 6.21]	[5.89, 6.23]	[5.76, 6.09]	[5.79, 6.13]	[5.95, 6.08]

Note. Negative column values indicate overlap; positive column values indicate gaps. Scale = 1 (*not fluent at all*) to 9 (*very fluent*). ms = millisecond.

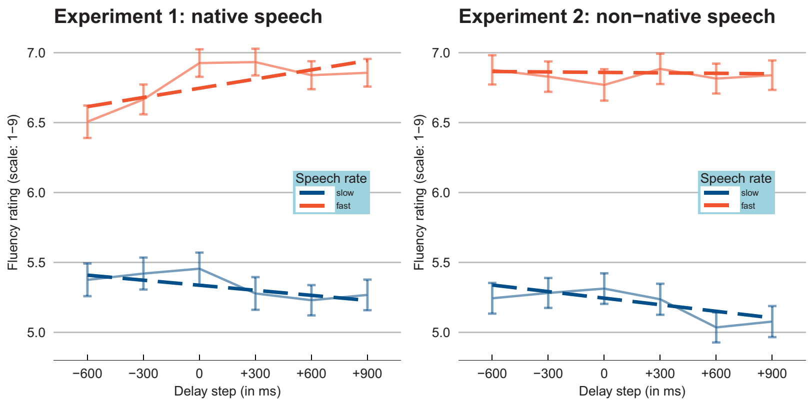

Figure 1 Mean fluency ratings for each condition (speech rate and delay step), for the native speech in Experiment 1 (left panel) and the nonnative speech in Experiment 2 (right panel). Error bars show standard error; dashed lines show model fit. [Color figure can be viewed at wileyonlinelibrary.com]

from 2 to 9). The remaining listeners had a bias for the higher end of the scale. The overall mean fluency judgment score was 6.06 ($SD = 1.98$). The left panel of Figure 1 illustrates the mean fluency judgment for each delay step and both speech rates in Experiment 1. The results shown in the left panel suggested that listeners rated the fast recordings as more fluent than the slow recordings and that they rated the recordings with 0 milliseconds overlap between the question and answer (no gap and no overlap) as the most fluent of all delay steps.

In Experiment 2, most listeners also made use of the entire 9-point scale (20 used the whole scale; nine used the scale from 2 to 9). The remaining listeners had a bias for the higher end of the scale. The overall mean fluency judgment score given to the nonnative answers was 6.02 ($SD = 2.02$). The right panel of Figure 1 illustrates the mean fluency ratings for each of the 12 conditions in Experiment 2. The results shown in the right panel suggested that the listeners rated fast recordings as more fluent than slow recordings. In the slow condition, the no-gap–no-overlap conditions received the highest fluency ratings.

The Shapiro–Wilk test showed that the fluency ratings were not normally distributed ($p = .02$). Because our data observations were nested within listeners, speakers, and items, we opted to use linear mixed models for the analyses. We entered the listeners' fluency ratings (scores 1 to 9) from both Experiment 1 and 2 into an omnibus linear mixed model (Baayen, Davidson, & Bates, 2008) implemented in the lme4 library (Bates, Maechler, Bolker, & Walker, 2015) in

R (R Core Team, 2012). We started with a series of models, beginning with simpler models with only a subset of predictors and gradually increased the model complexity by adding interactions and more complex random effects structures.

The first simpler model included the predictors speech rate, delay step, and experiment (deviation coded; Experiment 1 coded as -0.5 and Experiment 2 coded as 0.5), but no interactions. Given our experimental design, we entered listeners, items, and speakers into the model as random effects. We operationalized speech rate as a categorical variable (fast vs. slow; with the fast condition mapped onto the intercept), not as a continuous variable, in line with our acoustic manipulation that was a dichotomous variable. Of course, there was considerable variation between items in their original speech rate, but items also differed on many other dimensions, such as the occurrence of hesitations, corrections, mispronunciations, lexical content, and so on. To avoid confounding the interpretation of any potential effect of speech rate, we performed our speech rate manipulation within items. Therefore, any potential effect of speech rate could be directly attributed to the speech rate manipulation without being confounded with other variables. Moreover, we coded the predictor delay step as a continuous variable (scaled around the mean), not as a categorical variable distinguishing all gaps from all overlaps, in line with our continuous acoustic manipulations. Nevertheless, we checked whether this analysis decision made any difference to our interpretations, and found that models with the predictor delay entered as a categorical variable (excluding the 0 milliseconds delay step because it is neither a gap nor an overlap) led to a qualitatively similar interpretation of the results reported below.

A second, more extended model was identical to the first, except that it additionally included all possible interactions between the three fixed effects. This second model fit the data better than the first model, according to a likelihood ratio test, $\chi^2(4) = 28.45, p < .001$. Finally, a third full model extended the second model by additionally including by-listener random slopes for speech rate and delay, by-item random slopes for speech rate, delay, and experiment, and by-speaker random slopes for speech rate and delay. We used this maximal random effects structure (i.e., all possible random slopes) because this best fit the data based on a likelihood ratio test, $\chi^2(19) = 1{,}386.90, p < .001$, in line with suggestions by Barr, Levy, Scheepers, and Tily (2013). We assessed statistical significance at the .05 significance level by checking whether effects had absolute t values exceeding 2, since for large numbers of degrees of freedom (>100) the t distribution approximates the normal distribution (Baayen, 2008, p. 270).

This full, omnibus model (marginal $R^2 = .15$; conditional $R^2 = .64$; see Table 2) revealed a significant effect of speech rate: Fluency judgments were lower for slow fragments than for fast fragments of the manipulated answers across both experiments (recall that only the speech rate of the answers and not of the questions was manipulated). Additionally, the model revealed a significant effect of (scaled) delay step: Fluency judgments were slightly higher for fragments with a larger delay step than for fragments with a smaller delay step. This effect of delay step should be interpreted only with respect to the fast condition because this condition was mapped onto the intercept. In fact, the model also revealed a significant interaction between speech rate and delay step. However, there was a marginal three-way interaction for speech rate, delay step, and experiment. Therefore, we carried out separate analyses for both experiments.

Each model (one for each experiment) included the two predictors speech rate—a categorical variable with the fast condition mapped onto the intercept—and delay step—a continuous variable scaled around the mean—together with their interaction. We entered listeners, items, and speakers into the model as random effects. We also included by-listener, by-item, and by-speaker random slopes for speech rate and delay step.

The model for Experiment 1 (marginal $R^2 = .14$; conditional $R^2 = .57$; see Table 2) revealed a significant effect of speech rate: Fluency judgments were lower for slow fragments than for fast fragments. Additionally, the model revealed a significant effect of (scaled) delay step: Fluency judgments were slightly higher for fragments with a larger delay step. This effect of delay step should be interpreted only with respect to the fast condition because this condition was mapped onto the intercept. In fact, the model also revealed a significant interaction between speech rate and delay step. This interaction indicated that, where there was a positive effect of delay step in the fast condition, there was a negative effect of delay step for slow speech fragments. A mathematically equivalent model, this time mapping the slow condition onto the intercept, indeed showed a negative effect of delay step on fluency judgments, $b = -0.07$, $SE = 0.02$, $t = -2.84$, $p < .05$.

We constructed a similar linear mixed effects model for the nonnative fluency ratings obtained in Experiment 2 (marginal $R^2 = .16$; conditional $R^2 = .72$; see Table 2). We observed a significant main effect of speech rate, with fast answers receiving higher fluency ratings than did slow answers. Because the fast speech rate was mapped onto the intercept of the predictor speech rate, the absence of a main effect of delay step suggested that the slope of the red (or light gray in black and white print versions) line in Figure 1 was flat. The model

Table 2 Model outcomes for fixed effects and their interactions in the omnibus model (combined data from Experiments 1 and 2) and in the models for the individual experiments

	Model: Omnibus			Model: Experiment 1			Model: Experiment 2		
Parameter	*b*	*SE*	*t*	*b*	*SE*	*t*	*b*	*SE*	*t*
Intercept	6.80	0.22	30.44	6.75	0.23	29.38	6.86	0.39	17.82
Speech rate	−1.51	0.16	−9.66	−1.41	0.03	−5.21	−1.62	0.16	−10.42
Delay step	0.03	0.01	2.23	0.07	0.02	2.84	0.00	0.02	−0.19
Experiment	0.11	0.45	0.25						
Speech Rate × Delay Step	−0.07	0.02	−4.32	−0.10	0.03	−3.94	−0.04	0.02	−2.06
Speech Rate × Experiment	−0.21	0.31	−0.66						
Delay Step × Experiment	−0.07	0.03	−2.50						
Speech Rate × Delay Step × Experiment	0.06	0.03	1.73						

Note. Absolute *t* values greater than 2 are significant at .05.

also showed a significant interaction between speech rate and delay step. This indicated that the effect of delay step had a significantly more negative slope in the slow condition relative to the slope of delay step in the fast condition. That is, longer delays between questions and answers received lower fluency ratings in the slow condition (compared to the scores for the fast condition). A mathematically equivalent model, mapping the slow (rather than fast) speech rate onto the intercept of the predictor speech rate supported this interpretation. This model showed a significant negative effect of delay step, $b = -0.05$, $SE = 0.02$, $t = -2.80$, $p < .05$, which should be interpreted only with regard to the slow condition.

Discussion of Experiments 1 and 2

Summary of the Present Study

The present study investigated what variables contribute to perception of native and nonnative fluency in dialogue and what the specific effect of gaps and overlaps between question and answer turns is in dialogue. The effect of speech rate on fluency ratings was also examined in order to test the validity of the protocol.

Speech Rate

We had expected to find an effect of speech rate in both experiments, with listeners rating fast recordings as more fluent than slow recordings, thus replicating the results found by Bosker et al. (2014). We confirmed this hypothesis because the present study indeed showed that fast fragments were rated as more fluent than slow fragments, both for native and for nonnative speakers. These findings indicate that the experiment measured fluency in a fashion similar to that of previous fluency studies. Overall, the difference between fast and slow fragments was 1.45 on the 9-point scale in Experiment 1 (with native speakers) and 1.65 in Experiment 2 (with nonnative speakers). This suggests that speech rate is an important variable in determining fluency, both for native speech and for nonnative speech.

Furthermore, we did not find an interaction effect between speech rate and experiment, which suggests that speech rate is weighed similarly in perception of both native and nonnative fluency. This has also been found by Kahng (2014) and Bosker et al. (2014). This would mean that the perception of nonnative speakers' fluency is based on criteria similar to those used in fluency perception of native speakers. However, the fact that speech rate is weighed similarly in fluency evaluation tasks does not mean that the speech rate of native and nonnative speakers is processed similarly in online speech perception (e.g., Bosker & Reinisch, 2015; Kaufeld, Ravenschlag, Meyer, Martin, & Bosker, 2020; Maslowski, Meyer, & Bosker, 2019), nor that it is achieved by means of similar underlying cognitive regimes in speech production (Rodd et al., 2020).

Delay Step

For delay step, we had predicted that the lowest fluency ratings would be found for the largest gaps between questions and answers because this would signal that the speaker had had trouble formulating the answer. We had expected to find the highest fluency ratings for conditions with overlap because the speaker was quickly ready to start speaking. Because proficient nonnative speakers have previously been found to show nativelike turn-taking behavior (Galaczi, 2014), we had expected to find the lowest fluency ratings for the largest gaps between question and answer and the highest fluency ratings for conditions with an overlap.

In the slow speech rate condition, we did indeed find that overlap was rated as more fluent than conditions with a gap both in Experiment 1 (native speakers) and Experiment 2 (nonnative speakers). However, for the fast condition in Experiment 1, we found the opposite pattern: Fluency scores were higher in conditions with a gap than in conditions with overlap. In Experiment 2, the fast condition did not show a difference in fluency ratings as a function of delay condition.

The interaction effect between speech rate and delay step that we found in both experiments shows that the various delay steps, either gap or overlap, have a differential effect on fluency ratings for either fast or slow speech. That is, raters judged the turn gaps relative to the speech rate of the answer. Although for fast speech, overlaps in turns led to lower fluency ratings than gaps, the opposite was true in slow speech. In slow speech, larger gaps led to lower fluency ratings than did overlaps. This is a striking finding because the speech production literature has not found a strong relationship between the timing of turns and the speech rate of the next speaker (as reported by Roberts, Torreira, & Levinson, 2015).

We interpret this interaction to indicate that the listeners combined cues from both speech rate and delay step to make their judgments about fluency. Because the listeners made their judgments only after hearing the entire speech fragment, they could take into account all acoustic cues available to them. It seems that the listeners specifically “punished” the two extremes: fast speech that was also initiated too early (“too eager”) and slow speech that was initiated very late (“too reluctant”). Fast speech in itself is relatively fluent, as robustly indicated by the large size of the speech rate effect, and larger gaps did not negatively affect this fluency enhancing cue. However, overlap does have this effect because speakers then may sound too eager, or too fast, and not synchronized with the interlocutor (see the interactive alignment account; Pickering & Garrod, 2004), which makes the turn seem less fluent. Additionally, in the fast speech condition there was relatively more overlapping speech than in the slow condition because speech was faster but the same absolute time of overlap was used. This might also have affected fluency ratings in the fast speech overlap condition. Also, literature on pragmatics has demonstrated that speakers who produce overlapping turns (i.e., interrupting their interlocutor) are seen as less affiliative (Goldberg, 1990; Robinson & Reis, 1989; ter Maat et al., 2010). Hence, it may be argued that pragmatic perceptions influence fluency ratings even in the absence of instructions to take such dimensions into account.

A possible explanation for the finding that, for nonnative speakers, the fast condition showed no effect of delay step could be that the threshold for being perceived as too fast had not yet been reached. The nonnative recordings had a slower overall speech rate than did the native recordings, so the rate manipulation resulted in relatively slower fast fragments for nonnative speakers than for the native speakers. In fact, a post hoc analysis predicting the fluency ratings on the basis of a numerical (continuous) speech rate predictor (i.e., replacing the categorical predictor speech rate with the numerical predictor number of syllables per second; *z*-scored) did not establish a three-way interaction between number of syllables per second, delay step, and experiment, $t = 0.41$, $p =$ n.s. This indicates that the lack of an effect of delay step in the fast condition in Experiment 2 was (at least partly) due to the fact that the fast speech condition in Experiment 2 was slower than the fast condition in Experiment 1. Because of this difference between the native and nonnative speech fragments, the latter were fast and rated as fluent also in conditions of overlap, which was not the case in Experiment 1, where fast speech in overlap conditions was rated as less fluent than in gap conditions. One reason for this might be that although the overlaps took the same length of time in the two experiments, they contained less speech in Experiment 2 compared to Experiment 1, thus the overlaps were perhaps perceived less negatively (e.g., as less intrusive).

Slow speech in itself is less fluent, but, in the case of overlap, it is rated more fluent because this shows that speakers do not have trouble constructing their utterances. Larger gaps, on the other hand, affect fluency ratings in a negative way for slow speech because these show that slow speakers also need more time to construct their responses, leading to lower fluency ratings. For nonnative speakers, the threshold of being too slow, based on the slow speech rate and long pause, was reached, which contrasted to the fast condition, because the nonnative speech was already relatively slow compared to the native speech, and subsequently slowed down even more in the rate manipulation.

Thus, listeners seemed to take into account both the effects of speech rate and the effects of gaps or overlap on fluency and combined these two inferences to make their overall fluency judgment for speakers. It seems that extremes are rated as less fluent: Fast speech in itself is fluent, but, with overlap, it is perceived as too fast and becomes less fluent; slow speech is fluent in itself (albeit slow), but, with gaps, it is perceived as being too slow and rated as less fluent.

Implications, Limitations, and Future Directions

Results from the present study showed that turn-taking behavior in a dialogue affected fluency ratings for the speaker. Overlap led to higher fluency ratings for slow speech produced by both native and nonnative speakers but to lower fluency ratings for fast speech produced by native speakers. In the fast speech rate condition, nonnative speakers showed no effect of overlap or delay on fluency ratings. Gaps between question and answer turns led to lower fluency ratings for slow speech produced by nonnative and native speakers but to higher fluency ratings for native speakers' fast speech, possibly because the gaps perceptually slow down the speech, preventing it from being perceived as too fast. We observed these modulating effects of turn-taking behavior even in the absence of explicit instructions to the listeners to rate this interactional characteristic of the dialogues. Hence, this suggests that turn-taking behavior may be weighed to some extent automatically when listeners rate fluency, revealing a novel contributor to fluency perception.

Of course, we acknowledge that our design—despite its use of question–answer sequences—still does not represent everyday spoken interaction. Moreover, although our use of acoustic manipulations of turn gaps and speech rate provided the required experimental control, it also resulted in artificial speech (i.e., not original recordings). Therefore, we are careful about directly generalizing our findings to everyday communication. However, as an initial attempt to provide implications for our understanding of L2 learning, the present study suggests that the perceived fluency of L2 learners with slower speech rates in dialogue settings may be improved by learners' increasing their speech rate and by the early initiation of turns. In fact, early initiation of turns can mitigate, to some degree, the negative effect of a slow speech rate. For L2 learners with faster speech rates, larger gaps between a question and learners' answers would not have quite such a negative effect on perceived fluency, when compared with learners with slower speech rates.

Previous studies have found that native and nonnative speakers' pausing behavior and speech rate are judged similarly for fluency (Bosker et al., 2014; Kahng, 2014, 2018). The present study corroborates the finding that speech rate is weighed in a similar fashion for native and nonnative speech. However, this study is the first to show that perception of native and nonnative fluency also involves differential weighting of the same acoustic characteristic, namely, turn-taking behavior. Early initiation of a speaker's turn can be perceived as an indication of lower fluency for fast native speakers but as an indication of higher fluency for slow nonnative speakers. This raises the question of L2 proficiency. In the present study, all nonnative speakers demonstrated accentedness in their Dutch utterances. Future work may investigate whether the differential weighting of turn-taking behavior in native versus nonnative speech is modulated by proficiency: Is the turn-taking behavior of highly proficient nonnative speakers, with a low level of accentedness, weighed similarly to that of native speech when judging fluency?

The small yet robust effect of turn gaps that we observed in the present study suggests that the term dialogue fluency can be adopted (Peltonen, 2017) as an additional category of utterance fluency measures in interactional settings, in addition to breakdown fluency, speed fluency, and repair fluency (Segalowitz, 2010). Dialogue fluency refers to the acoustic properties of the speech of a given speaker that reflect confluence (McCarthy, 2010). Turn-taking behavior, that is the timing of one interlocuter's speech relative to the speech of the other interlocutor, can be used as a basis for operationalizing dialogue fluency. The present study should be seen as first evidence for effects of dialogue fluency measures on perceived fluency; future studies may investigate whether other properties of dialogues are also used as a cue for fluency perception. Examples are syntactic and lexical alignment (Pickering & Garrod, 2004), collaborative completions (Peltonen, 2017), and the degree of phonetic convergence (Van Leeuwen, 2017).

The present results are also of interest for L2 testing practice. This study shows that the current practices in L2 testing, where assessment is primarily based on monologues, do not entirely reflect spoken interaction in dialogue. In language testing, fluency is often assessed in monologues and based on a narrow definition, but, in reality, speech occurs in dialogues where speakers are at least partially judged on turn-taking behavior—particularly, the current study has shown, for learners with low speech rates. Thus, although some language tests assess a L2 learner to be fluent on the basis of monologue tasks, this speaker might, for example, leave long gaps before answering questions in a dialogue and thus be rated as less fluent in everyday communication. Although the present study only showed a small effect of turn-taking gaps, it suggests nonetheless that L2 testing practice might implement changes in fluency assessments to better capture fluency in everyday communication.

All delay steps that we used in the present experiment are reported to occur in natural Dutch conversation, though the extremes (−600 and 900 milliseconds) occur only rarely (Stivers et al., 2009). As such, the stimuli in the present experiment reflected natural behavior in dialogues. Stivers et al. (2009) showed that disconfirmative answers and nonanswers are prefaced by longer gaps than confirmation responses. Although most questions in this experiment were open questions, some were closed and paired with nonpreferred answers. For these items, longer gaps might sound more natural and therefore more fluent. This might have led to variation between question–answer pairs that we did not control in the present study. Additionally, in the present study, we did not control for between-speaker variation for speech rate and fluency (as our design manipulated these variables within-speaker). Future studies could implement a more controlled design, albeit at the expense of the naturalness of the stimulus materials. The design could also be expanded with other delay steps, for instance, by using a more densely sampled delay continuum.

Conclusion

The present study investigated the effect of speech rate (fast or slow) and delay between question and answer (various gaps and overlaps) in a dialogue setting on fluency ratings for native and nonnative speakers' answers to questions. The listeners heard short dialogues and were instructed to rate, on a 9-point scale, the fluency of the speaker who provided the answer. Results showed that fast speech was rated as more fluent than slow speech and, additionally, showed an interaction effect between speech rate and delay step. In fast speech, overlap with an interlocutor was rated as less fluent than gaps (for native speakers), but, in slow speech, overlap was rated as more fluent than gaps (for both native and nonnative speakers). The findings from the present study help in understanding the construct of fluency and how the construct of fluency should be expanded to dialogue settings, suggesting that, with respect to interactional competence, turn-taking behavior is part of perceived dialogue fluency and interacts with a speaker's nativelikeness and speed fluency. In turn, this better understanding of the construct of fluency carries practical implications: If fluency is not just the sum of an individual's temporal speech characteristics such as speech rate and pausing but also includes the smooth turn-taking behavior of all speakers in a conversation, this knowledge should be incorporated in L2 teaching and assessment practices.

Open Research Badges

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data are available at https://osf.io/b465y/.

Final revised version accepted 3 March 2020

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beňuš, Š., Gravano, A., & Hirschberg, J. (2011). Pragmatic aspects of temporal accommodation in turn-taking. *Journal of Pragmatics*, *43*, 3001–3027. https://doi.org/10.1016/j.pragma.2011.05.011

Boersma, P., & Weenink, D. (2013). *Praat: Doing phonetics by computer* (Version 5.3.56) [Computer software]. Retrieved from http://www.praat.org/

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, *44*, 123–147. https://doi.org/10.1177/00238309010440020101

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*, 159–175. https://doi.org/10.1177/0265532212455394

Bosker, H. R., Quené, H., Sanders, T., & De Jong, N. H. (2014). The perception of fluency in native and nonnative speech. *Language Learning*, *64*, 579–614. https://doi.org/10.1177/0265532212455394

Bosker, H. R., & Reinisch, E. (2015). Normalization for speech rate in native and nonnative speech. In M. Wolters, J. Livingstone, B. Beattie, R. Smith, M. MacMahon, J. Stuart-Smith, & J. Scobbie. (Eds.), *Proceedings of the 18th international congress of phonetic sciences* (pp. 1–5). *ICPhS 2015*. London, UK: International Phonetic Association. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0324.pdf

Bull, M. C., & Aylett, M. P. (1998). An analysis of the timing of turn-taking in a corpus of goal-orientated dialogue. In R. Mannel & J. Robert-ribes (Eds.), *Proceedings of the 5th International Conference on Spoken Language Processing, ICSLP 1998* (pp. 1175–1178). Canberra, Australia: Australian Speech Science & Technology Association.

Bygate, M. (1987). *Speaking*. Oxford, UK: Oxford University Press.

Chambers, F. (1997). What do we mean by fluency? *System*, *25*(4), 535–544. https://doi.org/10.1016/S0346-251X(97)00046-8

College voor toetsen en examens (2017). *Staatsexamens Nederlands als tweede taal: Examenjaar 2018* [Brochure]. Amsterdam, Netherlands: Steunpunt Staatsexamens Nt2.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, *111*, 2862–2873. https://doi.org/10.1121/1.1471894

Davies, A. (2003). *The native speaker: Myth and reality*. Clevedon, UK: Multilingual Matters.

De Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In R. Eklund (Ed.), *Proceedings of the 6th workshop on disfluency in spontaneous speech* (pp. 17–20). *DiSS* Stockholm, Sweden: KATH Royal Institute of Technology.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*, 655–679. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Ffrench, A. (2003). The development of a set of assessment criteria for speaking tests. *Research Notes*, *13*, 8–16.

Fillmore, C. J. (1979). On fluency. In C. Filmore, d. Kempler, & W. S-Y. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 85–101). New York, NY: Academic Press.

Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, *35*, 553–574. https://doi.org/10.1093/applin/amt017

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, *15*, 1–18. https://doi.org/10.1080/15434303.2018.1453816

Garrod, S. (1999). The challenge of dialogue for theories of language processing. In S. Garrod & M. J. Pickering (Eds.), *Language processing* (pp. 389–415). Hove, UK: Psychology Press.

Goldberg, J. A. (1990). Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power- and rapport-oriented acts. *Journal of Pragmatics*, *14*, 883–903. https://doi.org/10.1016/0378-2166(90)90045-F

Goldman-Eisler, F. (1968). *Psycholinguistics experiments in spontaneous speech*. New York, NY: Academic Press.

Götz, S. (2013). *Fluency in native and nonnative English speech*. Amsterdam, Netherlands: John Benjamins.

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*, 555–568. https://doi.org/10.1016/j.wocn.2010.08.002

International English Language Testing System. (n.d.). *SPEAKING: Band descriptors* (public version). Retrieved from https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx?la=en on 23-07-2019.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*, 24–49. https://doi.org/10.1093/applin/amm017

Kahng, J. (2014). *Exploring the production and perception of second language fluency: Utterance, cognitive, and perceived fluency* (Unpublished doctoral dissertation). East Lansing: Michigan State University. Retrieved from https://d.lib.msu.edu/etd/3005/datastream/OBJ/download/Exploring_the_production_and_perception_of_second_language_fluency___utterance__cognitive__and_perceived_fluency.pdf

Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39*, 569–591. https://doi.org/10.1017/S0142716417000534

Kaufeld, G., Ravenschlag, A., Meyer, A. S., Martin, A. E., & Bosker, H. R. (2020). Knowledge-based and signal-based cues are weighted flexibly during spoken language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*, 549–562. https://doi.org/10.1037/xlm0000744

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*, 387–417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2019). How the tracking of habitual rate influences speech perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*, 128–138. https://doi.org/10.1037/xlm0000579

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, *26*, 397–421. https://doi.org/10.1177/0265532209104668

McCarthy, M. (2010). Spoken fluency revisited. *English Profile Journal*, *1*(1), 1–15. https://doi.org/10.1017/S2041536210000012

Meyer, A. S., Alday, P. M., Decuyper, C., & Knudsen, B. (2018). Working together: Contributions of corpus analyses and experimental psycholinguistics to understanding conversation. *Frontiers in Psychology*, *9*, 1–13. https://doi.org/10.3389/fpsyg.2018.00525

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *IRAL-International Review of Applied Linguistics in Language Teaching*, *45*, 241–259. https://doi.org/10.1515/iral.2007.011

Ortega, L. (2019). SLA and the study of equitable multilingualism. *The Modern Language Journal*, *103*, 23–38. https://doi.org/10.1111/modl.12525

Peltonen, P. (2017). Temporal fluency and problem-solving in interaction: An exploratory study of fluency resources in L2 dialogue. *System*, *70*, 1–13. https://doi.org/10.1016/j.system.2017.08.009

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–190. https://doi.org/10.1017/S0140525X04000056

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*, 349–365, https://doi.org/10.1177/0265532214526177

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*, 423–441. https://doi.org/10.1080/01638539109544795

R Core Team. (2012). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org

Roberts, F., Margutti, P., & Takano, S. (2011). Judgments concerning the valence of inter-turn silence across speakers of American English, Italian, and Japanese. *Discourse Processes*, *48*, 331–354. https://doi.org/10.1080/0163853X.2011.558002

Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: A corpus study. *Frontiers in Psychology*, *6*, 1–16. https://doi.org/10.3389/fpsyg.2015.00509

Robinson, L. F., & Reis, H. T. (1989). The effects of interruption, gender, and status on interpersonal perceptions. *Journal of Nonverbal Behavior*, *13*, 141–153. https://doi.org/10.1007/BF00987046

Rodd, J., Bosker, H. R., Ernestus, M., Alday, P. M., Meyer, A. S., & ten Bosch, L. (2020). Control of speaking rate is achieved by switching between qualitatively distinct cognitive "gaits": Evidence from simulation. *Psychological Review*, *127*, 281–304. https://doi.org/10.1037/rev0000172

Roger, D., & Nesshoever, W. (1987). Individual differences in dyadic conversational strategies: A further study. *British Journal of Social Psychology*, *26*, 247–255. https://doi.org/10.1111/j.2044-8309.1987.tb00786.x

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review*, *65*, 395–412. https://doi.org/10.3138/cmlr.65.3.395

Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In J. Schenkein (Ed.), *Studies in the* *organization of conversational interaction* (pp. 7–55). New York, NY: Academic Press.

Salaberry, M. R., & Kunitz, S. (Eds.). (2019). *Teaching and testing L2 interactional competence: Bridging theory and practice*. New York, NY: Routledge.

Sato, M. (2014). Exploring the construct of interactional oral fluency: Second language acquisition and language testing approaches. *System*, *45*, 79–91. https://doi.org/10.1016/j.system.2014.05.004

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York, NY: Routledge.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., … Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*, 10587–10592. https://doi.org/10.1073/pnas.0903616106

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching*, *54*, 133–150. https://doi.org/10.1515/iral-2016-9994

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). Amsterdam, Netherlands: John Benjamins.

Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning*, *70*, https://doi.org/10.1111/lang.12384

ter Maat, M., Truong, K. P., & Heylen, D. (2010). How turn-taking strategies influence users' impressions of an agent. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Lecture Notes in Computer Science: Vol. 6356. Intelligent Virtual Agents. IVA 2010* (pp. 441–453). Berlin, Germany: Springer. https://doi.org/10.1007/978-3-642-15892-6_48

Torreira, F., Bögels, S., & Levinson, S. C. (2015). Breathing for answering: The time course of response planning in conversation. *Frontiers in Psychology*, *6*, 1–11. https://doi.org/10.3389/fpsyg.2015.00284

Traxler, M. J. (2012). *Introduction to psycholinguistics: Understanding language science*. Chichester, UK: John Wiley.

Van Leeuwen, A. R. (2017). *Right on time: Synchronization, overlap, and affiliation in conversation* (Unpublished doctoral dissertation). Utrecht, Netherlands: Universiteit Utrecht. Retrieved from https://www.lotpublications.nl/Documents/478_fulltext.pdf

Witton-Davies, G. (2014). *The study of fluency and its development in monologue and dialogue* (Unpublished doctoral dissertation). Lancaster, UK: Lancaster University.

Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 426–443). New York, NY: Routledge.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Overview of all Questions and Answers in the Experiment.
Appendix S2. Literal Rating Instructions (in Dutch) to Participants.

Appendix: Accessible Summary (also publicly available at https://oasis-database.org)

Taking Your Turn in Dialogues: Speaking Too Soon or Too Late Can Change How Fluency Is Perceived

What This Research Was About and Why It Is Important

Being fluent is important for people learning a new language. Most research on spoken fluency, however, has not investigated dialogues, which are a very natural way of communicating. In this study, we investigated whether one particular characteristic of dialogue influences how fluent speakers are perceived to be, namely, their turn-taking behavior. We recorded short dialogues and artificially varied the time between a question and an answer in these dialogues, so that the second speaker sometimes overlapped with the first, or left a long pause. We also changed the speech rate of the second speaker to be either fast or slow. We then asked other people to rate the fluency of the second speaker, who was either a native or a nonnative speaker. We found that the fluency ratings depended on the pause/overlap between the two speakers.

What the Researchers Did

We recorded short question–answer dialogues from native speakers of Dutch (Experiment 1) and nonnative speakers of Dutch (Experiment 2).

- We artificially created different versions of the recordings: sometimes the second speaker overlapped with the first, and sometimes there was a long pause. We also made the second speaker have a fast or slow speech rate.
- We asked another group of people to rate how fluent they found the person giving the answer. They gave their ratings on a scale from 1 to 9.

What the Researchers Found

Fast talkers were always perceived as more fluent than slow talkers.

- In native speech, too "eager" answers (interrupting a question with a fast answer) and too "reluctant" answers (answering slowly after a long turn gap) negatively affected fluency ratings.
- In nonnative speech, only too "reluctant" answers led to lower fluency ratings.

Things to Consider

We never instructed our participants to pay attention to the time between the question and answer, but they still used it to make their judgments.

- Of course, it is important to note that we did not use completely natural speech; we manipulated the speed and pauses. However, we think that the results will be similar for natural speech.
- Language tests often use monologues to assess oral fluency, with no interaction between speakers. Our results suggest that it might be more natural to use dialogues as well or even instead.

Materials and data: All data from the present study, together with an R analysis script, are available for download (under a CC BY-NC-ND 4.0 license) from https://osf.io/b465y/.

How to cite this summary: Van Os, M., de Jong, N. H., & Bosker, H. R. (2020). Taking your turn in dialogues: Speaking too soon or too late can change how fluency is perceived. *OASIS Summary* of Van Os, De Jong, & Bosker (2020) in *Language Learning*.