Fluency in dialogue:

The effect of turn-taking behavior on perceived fluency in native and non-native speech

Marjolein van Os [a], Nivja H. de Jong [b], and Hans Rutger Bosker [1 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

[c] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

ACCEPTED FOR PUBLICATION IN:

LANGUAGE LEARNING

(March 2020)

[1] Corresponding author: Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH, Nijmegen, The Netherlands. E-mail address: HansRutger.Bosker@mpi.nl.

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 interaction, 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 non-native speech. Results from two experiments using acoustically manipulated speech showed that, in native speech, too 'eager' (interrupting a question with a fast answer) and too 'reluctant' answers (answering slowly after a long turn gap) negatively affected fluency ratings. However, in non-native speech, only too 'reluctant' answers led to lower fluency ratings. Thus, we demonstrate that acoustic properties of dialogue are perceived as part of fluency. By adding to our current understanding of dialogue fluency, these lab-based findings carry implications for language teaching and assessment.

Running head: fluency in dialogue

Keywords: fluency, turn-taking, non-native fluency, dialogue, dialogic fluency

1. Introduction

Billions of people are currently learning to communicate in a second language (L2). It is therefore no surprise that considerable research efforts have been put into understanding L2 learning. There is a growing consensus that the primary goal of language education is to enable learners to become competent language users. With regard to 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, with interlocutors taking turns, language learners need to be able to produce and understand messages in real time, adjust messages to what they perceive the speech partner's understanding to be, and 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). Language learning therefore includes learning how to take turns in conversation. Likewise, when L2 oral abilities are being assessed, turn-taking abilities should be part of the evaluation. The current paper is focused on how turn-taking behavior is evaluated when assessing a speaker's fluency.

Some definitions of the term 'fluency' tend to focus on general linguistic abilities of second language speakers, who, when fluent, sound native-like (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 already reveals that this is not the case. One speaker might indeed speak fluently, with few pauses, restarts, or corrections, while another speaker constantly stumbles over their words, or the speakers differ in how well they can say the appropriate thing in varying context, how creatively they can use language, or how coherent their utterances are (Fillmore, 1979; Goldman-Eisler, 1968; Lennon, 1990; Riggenbach, 1991). Additionally, there is variation within one speaker as a function of their emotional state, speech register, and audience (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001).

Objectively, it is not clear how to define fluency both in native and non-native speech, as many different factors play a role, as well as the purpose of the definition (Chambers, 1997; Götz, 2013). Above we mentioned disfluencies such as pauses, restarts, and corrections, but possibly other factors play a role as well. Typically, fluency is used as a component in oral proficiency, and has been shown to make up an important part of the 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, 2003:15). As mentioned above, there is an issue here, in that native speakers vary in fluency, which makes it hard to assess whether non-native speakers have reached "native-like" levels of fluency.

In everyday conversation, we do not tend to hold endless monologues, but we rather are 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 interlocutor(s) (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, which includes topic management, interactive listening, breakdown repair, non-verbal 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, namely turn-taking behavior, influences fluency perception in native and non-native speech. Specifically, we tested the effect of gaps (a pause of 0, 300, 600 or 900 ms) and overlaps (simultaneous speech for 300 or 600 ms) between question and answer turns in a dialogue on the perceived fluency of the answer as produced by native (Experiment 1) and non-native speakers of Dutch (Experiment 2). Thus, we aim to contribute to the understanding of the construct of fluency, and assess how aspects of dialogue fluency such as turn-taking affect fluency perception, which may be partially based on interactive components.

In the present study, the term "native speaker" is used to refer to speakers growing up with a particular language from birth. The term "non-native speaker" is used to refer to speakers learning the language at a later age in life, and who are still in the process of learning the language. We acknowledge that this is a simplified view of language competence, where the native speaker's competence is the goal of the language learner, which does not have to be the case in all circumstances (see Ortega, 2019).

The following literature review will cover definitions of oral fluency in native and non-native speech (Section 1.2), how fluency of second language learners is assessed in dialogic tasks (Section 1.3), and research on how speaking turns in dialogues are managed (Section 1.4). Section 1.5 gives an overview of the research aims and hypotheses of the current study.

1.2 Oral fluency in native and non-native speech

1.2.1 Definitions of fluency

Several studies have given definitions of fluency in second language speech. In an influential paper, Lennon (1990) distinguishes 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 native-like way. This is the definition most often used in everyday life when we say sentences like '*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 the speech, and focuses on producing speech at a speech rate similar to native speakers of the language, without pauses, hesitations, fillers, and corrections.

Besides this distinction of fluency in the broad and narrow sense, another definition of fluency has been proposed by Segalowitz (2010), who uses a cognitive perspective. He distinguishes three senses of fluency: perceived fluency, utterance fluency, and cognitive fluency. Perceived fluency refers to the impression listeners have with regards to the speaker's cognitive fluency, based on the speech signal. This has also been mentioned by Lennon (1990): "fluency is an impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently." (1990:391). Utterance fluency refers to the features of an utterance that can be measured acoustically, such as speech rate, pauses, and repairs. Specifically, 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). Finally, cognitive fluency refers to the speaker's ability to efficiently coordinate the cognitive processes needed for speech production. The present study is concerned with fluency in the narrow sense, and in particular the relationship between utterance fluency and perceived fluency.

1.2.2 Acoustic correlates of non-native fluency

Several studies have investigated the relation between utterance fluency (objective acoustic measurements) and perceived fluency (subjective ratings by listeners) in monologues, attempting to identify the acoustic correlates of oral fluency, of which we discuss three studies in particular.

For instance, Cucchiarini, Strik, and Boves (2002) investigated which objective properties of speech can be used to predict the perceived fluency of non-native Dutch speakers. Trained judges rated both read and spontaneous speech on perceived fluency, which was correlated with different quantitative measures of fluency. In spontaneous speech, perceived fluency ratings were related in particular to the frequency and distribution of pauses, while speed of articulation (measured in number of phonemes per second) did not show such a relation. In read speech, both speed of articulation and frequency of pauses were related to fluency ratings.

Rossiter (2009) compared fluency ratings of non-native speech by different groups of raters: trained native speakers, untrained native speakers, and proficient non-native speakers. All three groups' fluency ratings were related to unfilled or non-lexical 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, measured acoustically, on 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, 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 speech rate were most strongly related to fluency perception in non-native speech. The authors conclude that breakdown fluency and speed fluency play a larger role than repair fluency in non-native fluency perception. These findings are in line with other studies, showing that non-native 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).

1.2.3 Native fluency

Typically, native speakers are seen as 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 age or gender of the speaker (Bortfeld et al., 2011). While relatively few studies have investigated fluency in native speakers, some have compared native and non-native speakers in comparable tasks (see Götz, 2013, for an overview for English).

Bosker, Quené, Sanders, and De Jong (2014) investigated perceived fluency in both native and non-native Dutch speech, and particularly focused on the question whether fluency characteristics in native speech on the one hand and non-native speech on the other are evaluated similarly by listeners. They conducted two experiments with acoustic manipulations, which allowed them to draw conclusions on the causal effects of particular fluency characteristics. In the first experiment silent pauses were manipulated, while in the second experiment speech rate was manipulated. Results showed that for both native and non-native 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 non-natives. With regards to speech rate, a decrease in fluency ratings was found when slowing down native speech to the non-native level, and an increase in fluency ratings was found for non-native speech sped up to the native level. Notably, the relative decrease and increase were of similar magnitude. This suggests that a single silent pause or a particular speech rate is weighed similarly in native and non-native fluency perception. As such, Bosker et al. concluded that the relationship between utterance fluency and perceived fluency is similar for native and non-native speakers.

Kahng (2014; 2018) investigated how silent pause distributions in both native and non-native speech affect raters' fluency judgements, also using speech manipulations. Results showed a similar effect of pause distribution on both native and non-native speech, where speech without pauses was rated as more fluent than speech with pauses either between or within clauses, and where 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) suggest that those aspects of speech affecting fluency ratings of non-native speakers, namely speech rate, and use of pauses and hesitations, affect the perception of native speakers in a similar way.

1.3 Fluency in dialogues

Notably, the definitions of fluency mentioned earlier are all focusing on the speech of the language user *in monologue*. However, in everyday life, speech is rarely produced in monologue settings, but rather in collaboration with other speakers. As McCarthy noted, "fluency also involves the ability to create flow and smoothness across turn-boundaries and can be seen as an interactive phenomenon in discourse" (2010:1). He proposes the term *confluence* to refer to the joint process of two speakers who cooperate to create a fluent interaction. Here he focuses particularly on turn-

openings and turn-closings, as these show how fluency is constructed interactively between speakers. The speakers in a dialogue share the responsibility to create and maintain a flowing conversation and fill silences. The term *dialogue fluency* has been introduced by Peltonen (2017) to refer to the contributions of individual speakers to the collaborative aspects of fluency in a dialogue, measured in 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, finding that fluent speakers show natural patterns of back-channeling and turn-taking, while a disfluent speaker is hesitant to start their turn.

The term interactional fluency used by Sato (2014) follows from the term *interactional competence* (Young, 2011; Galaczi & Taylor, 2018). Research into fluency has so far capitalized on understanding the construct of fluency in monologue speaking competence (partly described in 1.2.2). However, as yet, there is little research on the construct of fluency in dialogue, which is a component of language learners' interactional competence. Interactional competence comprises many different elements (Salaberry & Kunitz, 2019), and is of interest not only to fluency researchers but, for instance, also to those in the field of conversation analysis in second language acquisition (CA-SLA). The many facets to interactional competence make it difficult to define the concept clearly and completely. In many cases, it is only qualitatively described in raters' instructions, pointing out interactional skills like "taking turns", "keeping the floor", and "engage 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 regards to speaking assessment in pairs is the question how scores should be assigned to individuals based on a jointly constructed interaction (May, 2009). One suggestion for an interactional oral fluency rating scale has been made by Sato (2014), which was empirically based and succeeded in differentiating fluent from disfluent speakers. However, the scale has not been validated for other groups of learners than the group of Japanese university-level learners of English in the study. Better knowledge on how dialogue settings affect perceived fluency will help create more precise and focused definitions for raters and in that way 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 non-native 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 and speed of speech, and various interactive phenomena present in dialogues, such as backchannels, questions, and turn change types such as overlaps and gaps. Results of this analysis showed that speech rate and use of unfilled pauses contribute to low fluency ratings. However, as the sample size of the study was very small, it is hard to generalize these results to other non-native 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 CA-SLA framework. Some parts of her findings can be seen to reflect dialogue fluency. Quantitative and qualitative analyses of turn-taking management showed that speakers who are more proficient are better able to create confluence than less proficient speakers. Speakers with a lower proficiency level had a weak alignment with longer pauses between turns, while speakers with a higher proficiency level show rapid speaker changes and typically manage their turns in a no-gap-no-overlap manner, as is preferred by native speakers.

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

In the language testing practice, several formats of speaking assessment are being used. For standardization purposes, these often make use of monologue tasks. For example, the official ‘Dutch as a second language’ 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 (TOEFL), where candidates record their spoken responses to several questions, which are then scored by raters. Although attempting to simulate some aspects of dialogue, the test takes place on a computer with candidates responding to recorded questions, thus not interacting with other interlocutors. Other speaking assessments make use of dialogue settings, such as the official Cambridge English qualifications. Here two candidates are tested at the same time, and they speak with each other and

with a test leader. Candidates are assessed on 'Interactive Communication', which refers to the candidate's sensitivity to turn-taking (Ffrench, 2003), or interactional competence. The International English Language Testing System (IELTS) assesses speaking ability based on an interactive setting between learner and examiner, although the scoring criteria do not include any interactional competences or criteria related to turn-taking (IELTS, n.d.). The Common European Framework of Reference (Council of Europe, 2001) does mention turn-taking as part of discourse competence, focusing on taking the floor and initiating and maintaining conversations. Results from the present study will help in understanding the construct of fluency in dialogues, as part of overall interactional competence.

1.4 Managing dialogues

In contrast to monologues, in a dialogue there is by definition more than one interlocutor, which requires coordinated processing and leads to added time pressure to plan utterances compared to monologues (Garrod, 1999). Both interlocutors have to make sure that their utterances are understood the way they intend them to be, and that they follow the intended meaning from their interlocutor, and that they perform these communicative actions at appropriate time points. These processes are facilitated by construction of a common ground, as well as by perspective taking (Traxler, 2012).

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

Example 1. Example of gap in dialogue A (indicated by silent interval in seconds), example of overlap in dialogue B (overlapping speech indicated by square brackets). Q = question; A = answer.

A. *Q:* Do you enjoy cooking, are you good at it?

A: (0.6 s) I enjoy cooking, but it shouldn't be too complicated. Then it usually goes well.

B. *Q:* Do you enjoy cooking, are you good [at it?]

A: [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 consisting of ten different languages, Stivers and colleagues (2009) investigated the timing of turn-taking in polar questions, e.g. 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 ms (language-specific modes between 0 and +200 ms; Dutch: +100 ms). The mean time between the question and the answer was +208 ms, with language-specific means ranging from +7 to +469 ms. They concluded that the language-specific means fell within approximately 250 ms on either side of the cross-linguistic mean, leading the authors to conclude that talkers 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 factors, characteristics of the social setting, complexity of the response and the general context, as well as some individual differences (Roger & Nesshoever, 1987; Heldner & Edlund, 2010; Meyer, Alday, Decuyper, & Knudsen, 2018).

Studies have shown that deviating from this 'no-gap-no-overlap' strategy has communicative significance. Longer gaps may signify problems in comprehension (Beňuš, Gravano, & Hirschberg, 2011), speech planning difficulties (Bull & Aylette, 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, since 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 one's interlocutor) are seen as less affiliative (Van Leeuwen, 2017), as less agreeable and as more assertive (Maat, Truong, & Heylen, 2010; Robinson & Reis, 1989), as more dominant (Beňuš et al., 2011), and as less sociable (Robinson & Reis, 1989) compared to speakers who do not produce overlapping turns. Additionally, overlaps are linked to displays of power and control, and seen as rude and disrespectful (Goldberg, 1990).

1.5 Research aims and hypotheses

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 non-native speakers improve their dialogue management skills as they improve their second language proficiency (Galaczi, 2014, Peltonen, 2017). However, no study, besides Riggenbach (1991), has investigated how turn-taking behavior in dialogue contributes to the perception of fluency in native and non-native 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 behavior differs. In the present study, we investigated the effect of various delay steps, namely overlaps (-600 ms and -300 ms) or gaps (0 ms, +300 ms, +600 ms, and +900 ms) between the turns of a question-answer sequences between native speakers on fluency ratings. Additionally, the speaking rate (fast vs. slow) of the answer was manipulated. This manipulation was added to the experiment as an additional check to see whether our experimental design could capture fluency differences between conditions. If we only manipulated delay step, and found no difference between different steps, this might have been due to there being no effect of delay step on fluency, or it might be due to the experimental design 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 slow recordings, as studies have shown that faster speech is seen 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 as previous studies.

With regards to the effect of overlap and gap, it was expected that results would show lower fluency scores for larger gaps between the question and answer, as longer pauses are associated

with less fluent speech, also for native speakers (Bosker et al., 2014; Götz, 2013). The longer pause signals that the speaker needs more time to formulate the answer, thus being less fluent. As interlocutors of a wide variety of languages prefer the no-overlap-no-gap strategy for turn-taking management (Stivers et al., 2009), it was expected that the fluency ratings would be higher for a gap of 0 ms than for larger gaps. High fluency ratings were also expected for the overlap conditions, as in these cases the speaker apparently is able to quickly formulate and produce the response to the question that was asked.

Experiment 2 resembled Experiment 1, but this time all the answers in the question-answer conversations were produced by non-native speakers of Dutch. Similar as for Experiment 1, an effect of speech rate was predicted, where fast speech would be rated as more fluent than slow speech. Speech of high-proficient non-natives generally is faster than that of non-native speakers with a lower proficiency (Chambers, 1997; Cucchiarini, et al., 2002; Lennon, 1990, Rossiter, 2009; Segalowitz, 2010), and it was expected that this would be reflected in the ratings given in the experiment. Moreover, we predicted that also for the non-native speakers' recordings, larger gaps would be rated with lower fluency scores and larger overlaps with higher fluency scores. Again, gaps of 0 ms would be rated as most fluent, as research shows that proficient non-native speakers want to minimize gaps just like native speakers (Galaczi, 2014, Peltonen, 2017). As the answers in the two experiments involved the same lexical content, they already suggest a relatively high proficiency. Therefore, we presumed that raters would assume these speakers to show the same pattern of minimizing gaps as preferred by native speakers.

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

2. Method

2.1 Experiment 1 – native speakers

Listeners

Forty-nine listeners were recruited from the Max Planck Institute for Psycholinguistics participant pool. Informed consent was obtained before the experiment and all were paid for their participation. Data of one listener was lost due to technical problems. Thus, the ratings of forty-eight listeners were analyzed. All listeners were native speakers of Dutch with normal hearing. The mean age of the group was 22.3 years (*SD* = 2.47), 14 listeners were male, 34 were female.

Materials

Eighty question-answer pairs were constructed, consisting of forty different questions and eighty different answers, two per question. The questions had every-day topics like hobbies, holiday plans, and the weather. The questions were structured such that it was plausible that the answer could be given before the end of the question was heard. The answers had a length ranging from 19 to 44 syllables, with a mean of 33.4 syllables. An example of a question-answer pair can be found in Example 2. For an overview of all eighty question-answer pairs, see the Supporting Information.

Example 2. Example of two question-answer pairs, with two answers (A1 and A2) to the question Q. English translation in *italics.*

Q: Hoe reis je naar je werk, normaal gesproken?

Q: How do you travel to work, normally speaking?

A1: 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.

A1: I always go by car. I live quite far away from work and it's not convenient for me to take the train.

A2: Ik neem eigenlijk altijd de fiets om op mijn werk te komen, behalve als het regent. Dan ga ik met de bus.

A2: I usually always take the bike to get to my work, except when it is raining. Then I go by bus.

Recordings were made of ten native Dutch speakers (two were male, eight were female) reading a script consisting of eight questions and eight answers each. Lists were made so that no speaker answered the questions they read themselves, and so that when Speaker B answered questions by

Speaker A, Speaker A did *not* answer questions by Speaker B. A speaker would not read both answers to a question. All speakers were instructed to read the sentences in a natural way, as part of a natural dialogue, and were told that it did not matter if they hesitated or corrected themselves. All speakers read their questions and answers twice.

All questions and answers were isolated from the recordings using Praat (Boersma & Weenink, 2013) by cutting them at zero-crossings, and for each the best (as evaluated subjectively by the first and last author, both native speakers of Dutch) was selected 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).

Leading and ending silent intervals, as well as leading and ending filled pauses, were removed from the answer recordings using Praat. The mean duration of the trimmed answer recordings was 6.35 seconds (*SD* = 1.37), ranging from 3.38 to 10.36 seconds. They had a length ranging from 19 to 44 syllables, with a mean 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.

The intensity of all questions and answers was scaled to 65 dB. A 2 (speech rate) x 6 (delay) design was used in the experiment. The speech rate of the answer was linearly compressed by a factor of 0.833 (using PSOLA in Praat), so that it was sped up, and expanded by a factor of 1/0.833 = 1.2, so that it was slowed down. These factors were also used by Bosker et al. (2014), who compared native and non-native speech, and found that a large majority of listeners (85%) judged these speech rates to be natural. The speech rate of the questions was not manipulated. The delay between the question and the answer was manipulated in Praat to have an overlap of 600 ms or 300 ms, be 0 ms, or have a gap of 300 ms, 600 ms or 900 ms. These steps were chosen to fall in 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 ms (cf. De Jong & Bosker, 2013). This resulted in twelve unique conditions (2 Speech Rate conditions x 6 Delay steps).

Four questions and the eight corresponding answers were excluded as at least one of the answers sounded relatively unnatural as a reply to the question (as judged by the first author). This resulted in 72 question-answer pairs used in the experiment.

Design

The experimental items were arranged in a Latin Square design: Listeners heard each question-answer pair in only one of the twelve conditions, but heard all conditions during the experiment. There were twelve different pseudo-randomized lists of the stimuli. Each list consisted of mini-blocks of ten speakers, so that the answers in the trials in a given mini-block were all spoken by different speakers. Here it was made sure that the same question would not be presented twice in a row, and that the answers of two consecutive trials across mini-block boundaries were not produced by the same speaker. All twelve lists were reversed, resulting in twenty-four 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 twelve times.

Procedure

The experiment was run using Presentation software. The experiment was conducted in a sound-attenuating booth and the audio was presented over 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, presented on the screen. Listeners were told that we were interested in how listeners perceived fluency in different speakers. They were instructed to listen to question-answer pairs and rate the fluency of the second speaker, that is, the talker who gave the answer. They should do this using a nine-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 (cf. Bosker et al. 2013; 2014). They were instructed to do this *not* based on fluency in the broad sense (i.e., overall language proficiency, as in: "he is fluent in Italian"), but rather base their judgements on the use of filled and silent pauses, speech rate, and use of hesitations and corrections, thus focusing on fluency in the narrow sense. Instructions 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 et al., 2014; Rossiter, 2009) and are compatible with instructions given to raters of language tests.

Listeners were warned that all speakers were native speakers, but were instructed to use the entire scale from 1 to 9. Importantly, listeners could only make their judgements after the entire speech stimulus had played. They could only listen to each recording once. The listeners read the instructions self-paced, and then started with four practice trials. These consisted of the four most naturally sounding question-answer pairs from the excluded items. As such, neither the practice

questions nor answers were repeated in the experimental trials. Two of the four practice trials were presented in the fast condition, and two in the slow. None of these practice trials contained the extreme overlap of 600 ms or extreme gap of 900 ms. All listeners heard the same four practice trials, in the same order. Listeners did not receive feedback on their practice ratings. If there were no further questions, they continued with the experiment. Halfway through the experiment there was a short self-timed break. After all 72 trials, the listener filled out a short post-experimental questionnaire investigating whether they noticed anything about the recordings and what they thought about the experiment. Finally, listeners were debriefed. The experimental procedure took approximately thirty minutes.

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

2.2 Experiment 2 – non-native speakers

Experiment 2 was identical to Experiment 1 except for the fact that now non-native speech materials were used in the answer recordings.

Listeners

A new group of 48 listeners (38 f/10 m; mean age in years = 23; *SD* = 3) was recruited from the Max Planck Institute for Psycholinguistics participant database and listeners were paid 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 in Experiment 1, except that non-native speakers of Dutch (8 f/2 m) were recorded producing the answers for the question-answer stimuli.

Recall that there were forty questions, each combined with two different answers, resulting in eighty different question-answer pairs. Question recordings remained the same as in Experiment 1; that is, they were spoken by the same native Dutch speakers. Answers were re-recorded by ten non-native Dutch speakers. Their oral proficiency in Dutch varied from very proficient (achieving level C2 in the CEFR framework on an official Dutch as second language exam) to very low proficiency (hardly knowing any Dutch at all). However, considering that our sentence stimuli were scripted (and therefore all non-native speakers produced the same grammatically coherent lexical content), we were mostly concerned with the accentedness of the non-native speakers. All speakers demonstrated a high level of accentedness, as assessed by the first and last author (both native speakers of Dutch). Note also that all experimental comparisons involved within-talker effects and are therefore not confounded with the speakers' accents. Each speaker read out eight of the answers. The non-native speakers received the same instructions as the native speakers in Experiment 1: to read in a natural manner and not to worry about accent, disfluencies, hesitations, or corrections.

The answer recordings were trimmed and linearly compressed or expanded using the same procedure as in Experiment 1. The mean duration of the 72 trimmed answer recordings (eight were excluded and partially 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 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, the speech rate of the question was not manipulated.

Finally, the answer recordings' intensity was scaled to 65dB, after which they were concatenated with the question recordings from Experiment 1 using the same six delay conditions as in Experiment 1 (-600, -300, 0, 300, 600, 900 ms overlap between question and answer), again resulting in twelve unique conditions.

Procedure

The procedure was identical to Experiment 1, except that listeners were instructed that the persons answering the various questions were non-native speakers of Dutch and thus had a non-native accent.

Listeners were asked to fill in the same questionnaire as 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. While the majority of listeners commented that most of the recordings sounded natural, many (60.4%) could report at least some answers that sounded less natural, mostly due to obvious speed manipulations, like in Experiment 1.

3. Results

Behavioral results show that in Experiment 1, many listeners used (almost) the entire scale from 1 to 9 to make their fluency judgements, as instructed (n = 22 used the whole scale, n = 14 used the scale from 2 to 9). Remaining listeners had a bias for the higher end of the scale. The overall mean fluency judgement score was 6.06 (SD = 1.98). The left panel of Figure 1 illustrates the mean fluency judgement for each delay step and both speech rates in Experiment 1. This panel suggests that listeners rated the fast recordings as more fluent than the slow recordings, and the recordings with 0 ms overlap between the question and answer (no gap and no overlap) were rated as most fluent of all delay steps.

In Experiment 2, most listeners also made use of the entire 9-point scale (n = 20 used the whole scale, n = 9 used the scale from 2 to 9). Remaining listeners had a bias for the higher end of the scale. The overall mean fluency judgement score given to the non-native answers was 6.02 (SD = 2.02). The right panel of Figure 1 illustrates the mean fluency ratings for each of the twelve conditions in Experiment 2. They suggest that fast recordings were rated as more fluent than slow recordings. In the slow condition, the no-gap-no-overlap conditions received the highest ratings.

Table 1. Means and standard deviation (between brackets) of the fluency scores in both rate conditions and all gap/overlap conditions (negative values indicate overlap, positive values are gaps), ranging on a scale from 1 (not fluent at all) to 9 (very fluent).

		-600 ms	-300 ms	0	+300 ms	+600 ms	+900 ms	Overall
Exp. 1: native speech	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)
	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)
	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)
Exp. 2: non-native speech	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)
	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)
	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)

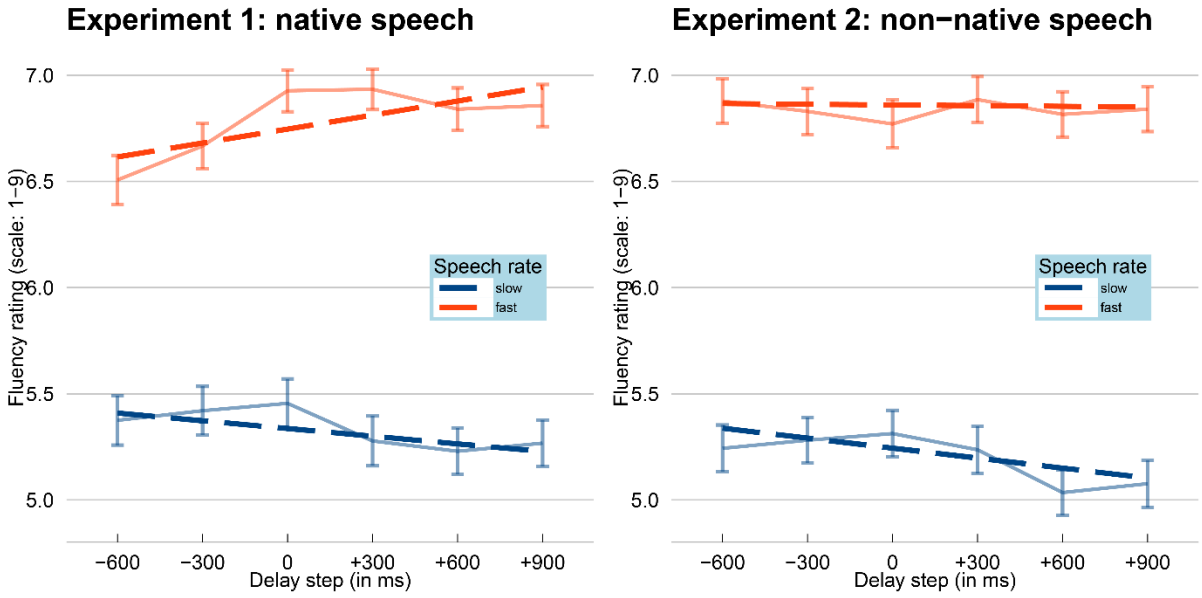

Figure 1. *Mean fluency ratings for each condition (Speech Rate and Delay Step), for the native answers in Experiment 1 and the non-native answers in Experiment 2. Error bars show standard error, dashed lines show model fit.*

The fluency ratings were not normally distributed (Shapiro-Wilk test: $p < 0.05$). Since our data observations were nested within Listeners, Speakers, and Items, we opted for analysis by Linear Mixed Models (LMMs). Listeners' fluency ratings (scores 1 to 9) from both Experiment 1 and 2 were entered into an omnibus Linear Mixed Model (Baayen, Davidson, & Bates, 2008) as implemented in the lme4 library (Bates, Maechler, Bolker, & Walker, 2015) in R (R Development Core Team, 2012). We started with a series of models, beginning with simpler models with only a subset of predictors, gradually increasing 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, Listeners, Items, and Speakers were entered into the model as random effects. Note that 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. 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, etc. To avoid confounding the interpretation of any potential effect of Speech Rate, our speech rate manipulation was performed within items; therefore, any potential effect of Speech Rate can be directly attributed to the speech rate manipulation without being confounded with other factors. Moreover, the predictor Delay Step was coded 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. However, models with a categorical predictor Delay (excluding the 0 ms delay step, since 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 fitted the data better than the first model, as assessed by means of a Likelihood Ratio Test (LRT; $\chi^2(4) = 28.446$, $p < 0.001$). Finally, a third 'full' model extended the second model by additionally including by-listener random slopes for Rate and Delay, by-item random slopes for Rate, Delay, and Experiment, and by-speaker random slopes for Rate and Delay. This maximal random effects structure was used (i.e., all possible random slopes) since this best fitted the data, as assessed by

an LRT ($\chi^2(19) = 1386.9$, $p < 0.001$), in line with the suggestions by Barr et al. (2013). Statistical significance was assessed at the 0.05 significance level by checking whether effects had absolute *t*-values exceeding 2 (Baayen, 2008).

This 'full' omnibus model (marginal $R^2 = 0.151$; conditional $R^2 = 0.643$; see Table 2) revealed a significant effect of Speech Rate: Fluency judgements were lower for slow fragments than for fast fragments across both experiments. Recall here 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 judgements were slightly higher for fragments with a larger Delay Step than for fragments with a smaller Delay Step. Note, however, that this effect of Delay Step should be interpreted only with respect to the fast condition, since 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 between Speech Rate, Delay Step, and Experiment. Therefore, separate analyses were carried out for both experiments.

Table 2. Model outcomes for the omnibus model (on the combined data from Experiment 1 and 2) and for each individual experiment. SR = Speech Rate; Expt = Experiment.

	Omnibus model			***Model Expt. 1***			***Model Expt. 2***		
	β	***SE***	***t***	***β***	***SE***	***t***	***β***	***SE***	***t***
intercept	6.803	0.224	30.436	6.747	0.230	29.378	6.860	0.385	17.823
SR	-1.512	0.157	-9.656	-1.410	0.027	-5.214	-1.616	0.155	-10.418
Delay	0.031	0.014	2.226	0.066	0.023	2.844	-0.003	0.017	-0.194
Expt	0.113	0.446	0.253						
SR * Delay	-0.073	0.017	-4.317	-0.102	0.026	-3.935	-0.044	0.021	-2.056
SR * Expt	-0.205	0.308	-0.663						
Delay * Expt	-0.069	0.028	-2.501						
SR * Delay * Expt	0.058	0.034	1.725						

Each model included the predictors Speech Rate, which was a categorical variable with the fast condition mapped onto the intercept, and Delay Step, which was a continuous variable scaled around the mean, together with their interaction. Listeners, Items, and Speakers were entered into

the model as random effects. By-listener, by-item, and by-speaker random slopes for Speech Rate and Delay Step were also included.

The model for Experiment 1 (marginal R^2 = 0.137; conditional R^2 = 0.572; see Table 2) revealed a significant effect of Speech Rate: Fluency judgements were lower for slow fragments than for fast fragments. Additionally, the model revealed a significant effect of (scaled) Delay Step: Fluency judgements were slightly higher for fragments with a larger Delay Step. Note, however, that this effect of Delay Step should be interpreted only with respect to the fast condition, since this condition was mapped onto the intercept. In fact, the model also revealed a significant interaction of Speech Rate and Delay Step. This interaction indicates 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 (β = -0.066, SE = 0.023, t = -2.844).

A similar Linear Mixed Effects Model was constructed for the non-native fluency ratings obtained in Experiment 2 (marginal R^2 = 0.164; conditional R^2 = 0.723; see Table 2). A significant main effect of Speech Rate was observed, with fast answers scoring higher fluency ratings than slow answers. Since the 'fast speech rate' was mapped onto the intercept of the predictor Speech Rate, the absence of a main effect of Delay Step suggests that the slope of the red line in Figure 1 is flat. The model also showed a significant interaction between Speech Rate and Delay step. This indicates 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 scores in the slow condition (as compared to the fast condition). This interpretation was supported by a mathematically equivalent model, this time mapping the 'slow speech rate' onto the intercept of the predictor Speech Rate. This model showed a significant negative effect of Delay Step (β = -0.047, SE = 0.017, t = -2.801), which should be interpreted only with regards to the slow condition.

4. Discussion

4.1 Summary of the present study

The present study investigated what factors contribute to native and non-native fluency perception in dialogue, and specifically what the effect is of gaps and overlaps between question and answer turns in a dialogue between speakers of Dutch. In order to test this, two experiments were set up. Recordings were made of questions, spoken by ten different native speakers of Dutch, and answers to those questions were recorded by the same group of native Dutch speakers (Experiment 1), and a group of non-native Dutch speakers (Experiment 2). The questions and answers were concatenated with various delay steps, namely overlaps (-600 ms and -300 ms) or gaps (0 ms, +300 ms, +600 ms, and +900 ms) between the turns. Additionally, the speech rate in the answers was manipulated to be fast or slow, resulting in twelve unique conditions. Listeners listened to the fragments and rated the fluency of the answer on a nine-point scale, ranging from 'not fluent at all' to 'very fluent'. They were instructed to make their judgements based on the narrow definition of fluency and to pay attention to pauses, speech rate, and repairs and hesitations.

4.2 Speech rate

It was expected that an effect of speech rate would be found in both experiments, where fast recordings would be rated as more fluent than slow recordings, thus replicating the results found by Bosker et al. (2014). This hypothesis was confirmed, as the present study indeed showed that fast fragments were rated as more fluent than slow fragments, both for native and for non-native speakers. These findings indicate that the experiment was measuring fluency in a similar fashion as previous fluency studies. Overall, the difference between fast and slow fragments was 1.45 on the nine-point scale in Experiment 1, and 1.65 in Experiment 2. This suggests that speech rate is an important factor in determining fluency, both for native speech and for non-native speech.

Furthermore, we did not find an interaction effect between Speech Rate and Experiment, which suggests that speech rate is weighed similarly in native and non-native fluency perception. This has also been found by Kahng (2014) and Bosker et al. (2014). This would mean that the perception of non-native speakers' fluency is based on similar criteria as fluency perception of native speakers. Note, however, that the fact that speech rate is *weighed* similarly in fluency evaluation tasks does not mean that the speech rate of native and non-native talkers is *processed* similarly in online speech perception (cf. Bosker & Reinisch, 2015; Kaufeld et al., in press; Maslowski et al., 2019), nor that it is achieved by means of similar underlying cognitive regimes in speech production (Rodd et al., in press).

4.3 Delay step

With regards to delay step, we predicted that the lowest fluency ratings would be found for the largest gaps between the question and answer, as this would signal that the speaker had trouble formulating the answer. The highest fluency ratings were expected for conditions with overlap as here the speaker is quickly ready to start speaking. As proficient non-native speakers have previously been found to show native-like turn-taking behavior (Galaczi, 2014), we expected to find the lowest fluency ratings for the largest gaps between question and answer, and the highest fluency ratings for conditions with overlap.

In the slow speech rate condition, it was indeed found that overlap was rated as more fluent than conditions with a gap, both in Experiment 1 (native) and Experiment 2 (non-native). However, for the fast condition in Experiment 1, the opposite pattern was found: Here 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 based on delay condition.

The interaction effect that was found in both experiments between speech rate and delay step 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 judge the turn gaps relative to the speech rate of the answer. While for fast speech, overlaps in turns lead to lower fluency ratings than gaps, this is the opposite in slow speech. In slow speech, larger gaps lead to lower fluency ratings than overlaps. This is a striking finding, since the speech production literature has not found a strong relationship between turn timing and the speech rate of the next talker (Roberts, Torreira, & Levinson, 2015).

We interpret this interaction to indicate that listeners combine cues from both speech rate and delay step to make their judgement. As listeners made their judgements only after hearing the entire speech fragment, they could take all acoustic cues available to them into account. Specifically, it seems that listeners '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 effect size, and larger gaps do not negatively affect this fluency enhancing cue. However, overlap does, as the speaker then might sound too eager, or *too fast*, and not synchronized with the interlocutor (see the interactive alignment account; Pickering & Garrod, 2004), making the turn less fluent. Additionally, in the

fast speech condition there was relatively more overlapping speech than in the slow condition, as 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 one's interlocutor) are seen as less affiliative (Goldberg, 1990; Maat, Truong, & Heylen, 2010; Robinson & Reis, 1989). 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 non-native speakers, the fast condition showed no effect of Delay Step could be that here the threshold for being *too fast* was not yet reached. The non-native recordings had a slower overall speech rate than the native recordings, so the rate manipulation resulted in relatively slower fast fragments for non-native speakers than for the native speakers. In fact, a post-hoc analysis predicting the fluency ratings based on a numerical 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 ($t = 0.406$). 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 non-native 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.

Slow speech in itself is less fluent, but in case of overlap rated more fluent as this shows that the speaker does not have trouble constructing their utterance. Larger gaps, on the other hand, affect fluency ratings in a negative way for slow speech as these show that the slow speaker also needs more time to construct their response, leading to lower fluency ratings. Here, for non-native speakers, the threshold of being *too slow*, based on the slow speech rate and long pause, is reached, as opposed to in the fast condition, as the non-native speech was already relatively slow compared to native speech, and subsequently slowed down even more in the rate manipulation.

Thus, listeners seem to take into account both the effects of speech rate and the effects of gaps or overlap on fluency and combine these two inferences to make their overall fluency judgement for the speaker. Here it seems that the 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; similarly, slow speech is fluent in itself (albeit slow), but with gaps it is perceived as being *too slow* and rated as less fluent.

4.4 Implications

Results from the present study show that turn-taking behavior in a dialogue affects fluency ratings for the speaker. Overlap leads to higher fluency ratings in slow speech for both native and non-native speakers, but to lower fluency ratings in fast speech for native speakers. In this speech rate condition, non-native speakers showed no effect of overlap. Gaps between question and answer turns lead to lower fluency ratings in slow speech of non-native and native speakers, but to higher fluency ratings in native speakers' fast speech, possibly because the gaps perceptually slow down the speech, preventing it from being perceived as *too* fast. Critically, these modulating effects of turn-taking behavior were observed even *in the absence of explicit instructions* to rate this interactional characteristic of the dialogues. Hence, this suggests that turn-taking behavior is weighed automatically when rating fluency, revealing a novel contributor to fluency perception.

Of course, we acknowledge that our design – despite involving question-answer sequences – is still not representative of everyday spoken interaction. Moreover, while the use of acoustic manipulations of turn gaps and speech rate provided the required experimental control, it also resulted in artificial speech (i.e., no original recordings). Therefore, we should be careful to directly generalize our findings to everyday communication. However, considering the current study as an initial attempt to provide implications for language learning, the present results suggest that the perceived fluency of beginning learners in dialogue settings may improve by increasing speech rate and early initiation of turns. In fact, early initiation of turns can mitigate, to some degree, the negative effect of a slow speech rate. For more proficient learners, with faster speech rates, potential larger gaps between a question and the learner's answer will not have quite such a negative effect on perceived fluency.

Previous studies have found that speakers' pausing behavior and speech rate are judged similarly in native and non-native speech when it comes to fluency (Bosker et al., 2014; Kahng, 2014; 2018). The present study corroborates the finding that speech rate is weighed in a similar fashion in native and non-native speech. However, this study is the first to show that native and non-native fluency perception also involve *differential* weighting of the same acoustic

characteristic, namely, turn-taking behavior. Early initiation of one's turn can be perceived as an indication of lower fluency for fast native speakers, but an indication of higher fluency for slow non-native speakers. This raises the question of L2 proficiency. In the present study, all non-native speakers demonstrated accentedness in their Dutch utterances. Future work may investigate whether the differential weighting of turn-taking behavior in native vs. non-native speech is modulated by proficiency: is the turn-taking behavior of highly proficient non-native speakers, with a low level of accentedness, weighed similarly as native speech?

The small yet robust effect of turn gaps observed in the present study suggests that we can adopt the term *dialogue fluency* (Peltonen, 2017) as an additional category of utterance fluency measures in interactional settings, beyond breakdown fluency, speed fluency, and repair fluency (Segalowitz, 2010). Dialogue fluency refers to the *acoustic properties* of speech of a given talker that reflect confluence (McCarthy, 2010). Turn-taking behavior, that is the timing of one's speech relative to the speech of one's interlocutor, can be used as a basis for operationalizing dialogue fluency. The present study should be seen as the first evidence for effects of dialogue fluency measures on perceived fluency; future studies may investigate whether other properties of dialogue 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 to the language testing practice. This study shows that the current practices in language 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 according to the narrow definition, while in reality speech occurs in dialogues where speakers are at least partially judged on turn-taking behavior – particularly in low-proficiency learners with low speech rates. Thus, while some language tests assess a learner to be fluent based on 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. While the present study only showed a small effect of turn-taking gaps, it suggests nonetheless that the language testing practice may implement changes with regards to fluency assessments to better capture fluency in everyday communication.

All delay steps used in the present experiment are reported to occur in natural Dutch conversation, though the extremes (-600 ms and +900 ms) occur only rarely (Stivers et al., 2009).

As such, the stimuli in the present experiment reflect natural behavior in dialogues. Stivers and colleagues showed that disconfirmative answers and nonanswers are prefaced by longer gaps than confirmation responses. While most questions in this experiment were open questions, some were closed and paired with non-preferred 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 was not controlled in the present study. Additionally, in the present study between-speaker variation with regards to speech rate and fluency was not controlled. 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 using a more densely sampled delay continuum.

4. 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 non-native speakers of the answer. Listeners listened to short dialogues and were instructed to rate the fluency of the speaker giving their judgment on a nine-point scale. 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 was rated as less fluent than gaps, while in slow speech overlap was rated as more fluent than gaps. The findings from the present study help in understanding the construct of fluency, and how the construct of fluency should be expanded to dialogic settings, suggesting that with respect to interactional competence, turn-taking behavior is part of perceived *dialogue fluency*, in interaction with the talker's nativeness 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 smooth turn-taking behavior of all speakers in a conversation, this knowledge should be incorporated in the language teaching and assessment practice.

5. Data availability statement

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/

6. Acknowledgements

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

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*(4), 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*(3), 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*(12), 3001-3027. https://doi.org/10.1016/j.pragma.2011.05.011

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

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*(2), 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*(2), 159-175. https://doi.org/10.1177/0265532212455394

Bosker, H. R., Quené, H., Sanders, T., & Jong, N. H. (2014). The perception of fluency in native and nonnative speech. *Language Learning*, *64*(3), 579-614. https://doi.org/10.1177/0265532212455394

Bosker, H. R., & Reinisch, E. (2015). Normalization for speechrate in native and nonnative speech. In M. Wolters, J. Livingstone, B. Beattie, R. Smith, M. MacMahon, J. Stuart-Smith, et al. (Eds.), *Proceedings of the 18th International Congresses of Phonetic Sciences (ICPhS 2015).* London: International Phonetic Association.

Bull, M. C., & Aylett, M. P. (1998). An analysis of the timing of turn-taking in a corpus of goal-orientated dialogue. *Proceedings of the 5th International Conference on Spoken Language Processing, ICSLP* 1998, 1175-1178

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: 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*(6), 2862-2873. https://doi.org/10.1121/1.1471894

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

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 (DiSS)* (pp. 17-20).

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, 54 (4), 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 *Individual differences in language ability and language behavior* (pp. 85-101). New York, USA: Academic Press Inc.

Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests?. *Applied Linguistics, 35*(5), 553-574. https://doi.org/10.1093/applin/amt017

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, 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 Ltd.

Goldberg, J. A. (1990). Interrupting the discourse on interruptions – an analysis in terms of relationally neutral, power-oriented and rapport-oriented acts. *Journal of Pragmatics*, *14*(6), 883-903. https://doi.org/10.1016/0378-2166(90)90045-F

Goldman-Eisler, F. (1968). *Psycholinguistics Experiments in Spontaneous Speech*. New York, USA: Academic Press Inc.

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

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics, 38*(4), 555–568. https://doi.org/10.1016/j.wocn.2010.08.002

IELTS. (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*(1), 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* (Dissertation, 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*(3), 569-591. https://doi.org/10.1017/S0142716417000534

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

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 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* (1): 128–38. https://doi.org/10.1037/xlm0000579

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, *26*(3), 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*. 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*(3), 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*(2), 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*. https://doi.org/10.1177/0265532214526177

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*(4), 423-441. https://doi.org/10.1080/01638539109544795

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*(5), 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*. 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*(3), 141-153. https://doi.org/10.1007/BF00987046

Rodd, J., Bosker, H. R., Ernestus, M., Alday, P. M., Meyer, A. S., & ten Bosch, L. (in press). Control of speaking rate is achieved by switching between qualitatively distinct cognitive "gaits": Evidence from simulation. *Psychological Review*. 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*(3), 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*(3), 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, USA: Academic Press Inc.

Salaberry, M. R., & Kunitz, S. (2019) (Eds.). *Teaching and Testing L2 Interactional Competence: Bridging Theory and Practice*. New York, USA: 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, USA: Routledge.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J.P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 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*(2), 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, the Netherlands: John Benjamins.

Ter Maat, M., Truong, K. P., & Heylen, D. (2010). How turn-taking strategies influence users' impressions of an agent. In *International Conference on Intelligent Virtual Agents.* 441-453. Berlin, Germany: Springer.

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

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

Van Leeuwen, A. R. (2017). *Right on time. Synchronization, overlap, and affiliation in conversation* (Dissertation, 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 University, Lancaster, UK.

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. 426–43). New York, USA: Routledge.

5. 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.