LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

The Perception of Fluency in Native and Nonnative Speech

Hans Rutger Bosker,[a] Hugo Quené,[b] Ted Sanders,[b] and Nivja H. de Jong[b]

[a]Max Planck Institute for Psycholinguistics and [b]Utrecht University

Where native speakers supposedly are fluent by default, nonnative speakers often have to strive hard to achieve a nativelike fluency level. However, disfluencies (such as pauses, fillers, repairs, etc.) occur in both native and nonnative speech and it is as yet unclear how fluency raters weigh the fluency characteristics of native and nonnative speech. Two rating experiments compared the way raters assess the fluency of native and nonnative speech. The fluency characteristics were controlled by using phonetic manipulations in pause (Experiment 1) and speed characteristics (Experiment 2). The results show that the ratings of manipulated native and nonnative speech were affected in a similar fashion. This suggests that there is no difference in the way listeners weigh the fluency characteristics of native and nonnative speakers.

Keywords fluency; disfluencies; hesitations; phonetic manipulations; nonnative speech

Introduction

Fluency in spoken language has been termed an automatic procedural skill (Schmidt, 1992) that encompasses the notion of "rapid, smooth, accurate, lucid,

Many thanks to the researchers within the "What Is Speaking Proficiency" project who kindly made their speech materials available. The WISP project was sponsored by the Netherlands Organisation for Scientific Research, grant number 254–70–030: Margarita Steinel, Arjen Florijn, Rob Schoonen, and Jan Hulstijn, Amsterdam Center for Language and Communication, Faculty of Humanities, University of Amsterdam. We are also grateful to Theo Veenker from the Utrecht institute of Linguistics OTS (UiL OTS) lab for technical support, to Iris Mulders for help with participant recruitment, to Heidi Klockmann for her help in collecting the data of Experiment 1, and to Cem Keskin and Erica Bouma for annotating the speech. Finally, we would also like to express our thanks to three anonymous reviewers for their comments and suggestions. This work was carried out at the UiL OTS, Utrecht University, The Netherlands, and was supported by Pearson Language Testing by means of a grant awarded to Nivja H. de Jong ("Oral fluency: production and perception").

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

DOI: 10.1111/lang.12067

and efficient translation of thought or communicative intention into language" (Lennon, 2000, p. 20). Lennon (1990) distinguished between fluency in the broad sense, that is, global speaking proficiency, and fluency in the narrow sense, that is, the "impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently" (Lennon, 1990, p. 391). Segalowitz (2010) distinguished between three facets of fluency, namely cognitive fluency—"the efficiency of operation of the underlying processes responsible for the production of utterances"; utterance fluency—"the features of utterances that reflect the speakers cognitive fluency," which can be acoustically measured; and perceived fluency—"the inferences listeners make about speakers' cognitive fluency based on their perceptions of their utterance fluency" (p. 165). In this study, we are concerned with the relationship between utterance fluency and perceived fluency.

Despite the fact that the aforementioned definitions of fluency may apply to both native and nonnative speech, fluency assessment has thus far mostly (if not exclusively) been aimed at nonnative speakers. Native speakers are supposedly perceived as fluent by default even though they, too, produce disfluencies such as *uhm*'s, silent pauses and repetitions. In fact, it is estimated that 6 in every 100 words is affected by disfluency (Fox Tree, 1995) and various factors have been found to influence native disfluency production, including speaker gender, speaker age, conversational topic, planning difficulty, and so on (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001). Therefore, the current study compares the way native and nonnative fluency characteristics are weighed by listeners.

Native and Nonnative Fluency

The production of nonnative disfluencies has been widely studied. Producing fluent speech is an important component of speaking proficiency for nonnative speakers as defined in the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). The descriptors in the global scale (p. 24) state that speakers at level B2 can communicate "with a degree of fluency"; at level C1, speakers can express themselves "fluently," and at level C2, "very fluently." In the language testing practice, human raters frequently assess nonnative speakers' fluency levels (e.g., Iwashita, Brown, McNamara, & O'Hagan, 2008). Many studies have investigated the acoustic fluency characteristics of nonnative speakers. The literature ranges from child second language (L2) learners (Trofimovich & Baker, 2007) to very advanced L2 speakers (Riazantseva, 2001). Nonnative speech is reported to contain more

disfluencies than native speech (e.g., Cucchiarini, Strik, & Boves, 2000) and nonnative speakers are seen to become more fluent as their proficiency in the nonnative language advances (e.g., Freed, 2000; Towell, Hawkins, & Bazergui, 1996). De Jong, Groenhout, Schoonen, and Hulstijn (2013) have argued that the fluency characteristics of one's L2 speech are strongly related to those in the talker's L1 (cf. Segalowitz, 2010). Both a person's individual traits and the speaker's nonnative proficiency level define the speaker's L2 cognitive fluency, with consequences for the fluency characteristics of the speech signal (utterance fluency). The utterance fluency of a speaker (i.e., the number of silent pauses per minute, the number of filled pauses, repetitions, corrections, and so on) affects, in turn, the fluency impression that listeners have of a particular speaker (perceived fluency).

There have been numerous studies investigating the subjective fluency level of nonnative speakers (e.g., Cucchiarini et al., 2000; Cucchiarini, Strik, & Boves, 2002; Derwing, Rossiter, Munro, & Thomson, 2004; Freed, 2000; Ginther, Dimova, & Yang, 2010; Kormos & Dénes, 2004; Mora, 2006; Rossiter, 2009; Wennerstrom, 2000). All these studies involved relating measures of perceived fluency (listener ratings, typically involving 7- or 9-point Likert scales) to utterance fluency (temporal speech measures) in order to assess the relative contributions of different speech characteristics to fluency perception. These studies indicate that temporal measures alone can account for a large amount of variance in perceived fluency ratings. Rossiter (2009) reported a correlation of $r = .84$ between subjective fluency ratings and pruned number of syllables per second (the total number of syllables minus disfluencies). She also compared ratings from untrained and expert fluency raters and did not find a statistically significant difference between the two groups. Derwing et al. (2004) used novice raters to obtain perceived fluency judgments. These raters listened to speech materials of 20 beginner Mandarin-speaking learners of English. Derwing et al. (2004) found that pausing and pruned syllables per second together accounted for 69% of the variance of their fluency ratings. Kormos and Dénes (2004) related acoustic measurements from nonnative Hungarian speakers to fluency ratings by native and nonnative teachers. They reported a correlation of $r = .87$ between the measure of speech rate and subjective fluency ratings. Cucchiarini et al. (2002) had teachers rate spontaneous speech materials obtained from nonnative speakers of Dutch. They found a correlation of $r = .65$ between the mean length of runs (mean number of phonemes between silent pauses) and the perceived fluency of spontaneous speech.

These studies suggest that temporal factors are major contributors to fluency judgments. However, many researchers have raised the question whether

nontemporal factors, such as grammatical accuracy, vocabulary use, or foreign accent, should also be considered as influencing fluency judgments (Freed, 1995; Lennon, 1990). Rossiter (2009) noted in her study that subjective ratings of fluency were influenced by nontemporal factors as well (on the basis of qualitative analyses of rater comments). The most important factor in this respect was learners' pronunciation of the nonnative language. More recently, a quantitative study by Pinget, Bosker, Quené, and De Jong (2014) has tackled the relationship between perceived fluency and perceived accent. Their results suggested that raters can keep the concept of fluency well apart from perceived foreign accent. Fluency ratings and accent ratings of the same speech samples were found to correlate only weakly ($r = -.25$) and, moreover, acoustic measures of accent did not add any explanatory power to a statistical model of perceived fluency. This suggests that, although the contribution of nontemporal factors to perceived fluency should not be ignored, these nontemporal factors likely play only a minor role.

Taking all the evidence together, studies targeting nonnative fluency perception converge on the view that acoustic measures of fluency can account for fluency ratings to a large extent. However, as noted, the emphasis of the aforementioned studies is on the level of fluency of *nonnative* speakers. Studies exploring the relationship between utterance fluency and perceived fluency of native speakers are rare. Native speakers are supposedly perceived as fluent by default (Davies, 2003; Riggenbach, 1991). Nevertheless, individual differences between native speakers in the production of disfluencies have been reported (Bortfeld et al., 2001). The psychological literature has primarily studied disfluency as a window into different stages of speech planning (e.g., Goldman-Eisler, 1958a, 1958b; Levelt, 1989; Maclay & Osgood, 1959). The study of speech pathology and speech therapy has primarily focused on the factors that influence (atypical) disfluency production (Christenfeld, 1996; Panico, Healey, Brouwer, & Susca, 2005; Susca & Healey, 2001). However, it is unclear how these disfluencies in native speech are perceived by the listener.

From the field of social psychology we know that listeners constantly make inferences about speakers based on the (nonlinguistic) content of speech, engaging in what is called person or speaker perception (Krauss & Pardo, 2006). Listener attributions may range from social status (Brown, Strong, & Rencher, 1975) and emotion (Scherer, 2003) to metacognitive states (Brennan & Williams, 1995) and even to physical properties of a speaker (Krauss, Freyberg, & Morsella, 2002). Nevertheless, it is as yet unknown how the fluency characteristics of native speech contribute to the perception of a native speaker's fluency level. The few studies that have included native speech

in their fluency research report that natives are consistently rated higher than nonnatives (Cucchiarini et al., 2000) and that they also produce fewer disfluencies than nonnatives do (Cucchiarini et al., 2000). Ginther et al. (2010) report higher overall oral proficiency for native speakers as measured by an oral English proficiency test as compared to nonnative speakers. From these studies, we cannot gather how listeners weigh native and nonnative fluency characteristics. In order to gain more insight into the perception of fluency in native and nonnative speech, the current work addresses the following research question: Do listeners evaluate fluency characteristics in the same way in native and nonnative speech?

Apropriate Methodologies for Comparison of Native and Nonnative Fluency

One could propose to address this question through correlational analyses (cf. Bosker, Pinget, Quené, Sanders, & De Jong, 2013; Cucchiarini et al., 2002; Derwing et al., 2004; Kormos & Dénes, 2004; Rossiter, 2009), which would involve collecting subjective fluency judgments of native and nonnative speech, collecting objective acoustic measurements from native and nonnative speech, and then statistically testing the extent to which the acoustic measures can account for the subjective ratings. This correlational approach is, however, unsuitable for the comparison of the perception of native and nonnative speech, because native and nonnative speech differs in many respects. The hypothetical observation that silent pauses play a large role when rating nonnative fluency, compared to rating native fluency, could simply be accounted for by a difference in pause incidence in native and nonnative speech (rather than by a difference in relative weight of pausing). Therefore, a comparison between native and nonnative fluency perception is only viable when native and nonnative speech samples have been matched for the acoustic dimensions under study.

In order to circumvent this problem, we propose a different method for investigating the contribution of acoustic variables to fluency judgments. We propose to use experiments with acoustic manipulations of the speech signal so as to ascertain that observed effects in fluency judgments may be directly attributed to particular fluency characteristics (cf. Munro & Derwing, 1998, 2001, who used phonetic manipulations to study perceived accent). The advantage of this method is that it becomes possible to compare native and nonnative fluency perception. For instance, we may compare how the same modification of silent pauses in native and nonnative speech affects the perception of

fluency. If different fluency ratings are given to two speech samples differing in a single manipulated phonetic property, then this perceptual difference may be reliably attributed to the minimal acoustic difference between the samples. This experimental method has the additional advantage that separate contributions of multiple acoustic factors can be investigated. Thus, the effect of one acoustic property on fluency judgments can be singled out through the use of phonetic manipulations targeting the disfluencies in the speech while keeping all other possibly interacting factors constant. Even different properties of one and the same acoustic phenomenon can thus be studied, such as the number and the duration of silent pauses. It is difficult to disentangle the contributions of these properties of silent pauses to fluency ratings using correlational analyses. The current approach could thus shed light on differential effects of two pause properties by manipulating pause duration while keeping the number of pauses constant.

The Present Research

The present study reports on two experiments that aim to answer the research question above by studying two different fluency dimensions, namely pausing and speed characteristics of native and nonnative speech. Both experiments make use of phonetic manipulations in native and nonnative speech. In Experiment 1, the silent pauses present in native and nonnative speech were manipulated. In Experiment 2, the speed of native and nonnative speech was modified. In our analyses, the main objective was to determine whether or not our manipulations affect fluency ratings of native and nonnative speech in a similar fashion.

Two possible hypotheses can be proposed with respect to the distinction between native and nonnative fluency. The effects of phonetic manipulations could be similar across native and nonnative fluency perception, such that both are equally affected by phonetic manipulations. The literature on nonnative fluency perception has shown that fluency judgments depend on the disfluencies in the speech signal (e.g., Bosker et al., 2013; Cucchiarini et al., 2000, 2002; Derwing et al., 2004; Rossiter, 2009). But native speech also contains disfluencies, and manipulating these might have similar effects on fluency ratings as compared to nonnative disfluencies.

Alternatively, manipulating characteristics of fluency in the speech signal may also have differential effects on the perception of native and nonnative fluency. For example, because natives are proficient in their native language, they are generally perceived as fluent. Therefore, the addition of disfluency

characteristics may affect native speech to a lesser extent than nonnative speech. The same line of reasoning could also support the opposite prediction: Because natives are generally perceived as fluent, the added disfluencies may—in the perception of listeners—stand out more than nonnative disfluencies. Therefore, our manipulations could also affect native speech to a larger extent than nonnative speech.

The production literature (e.g., Davies, 2003; Skehan & Foster, 2007; Skehan, 2009; Tavakoli, 2011) seems to suggest that native and nonnative fluency characteristics may be weighed differentially by listeners. For instance, Skehan and Foster (2007) observed that native speakers have a different pause distribution compared to nonnative speakers. Differences in the position of pauses may lead to differential perception of pauses in native and nonnative speech. It has even been argued that disfluencies in native speech can help the listener. For instance, eye-tracking data indicate that hesitations may aid the listener in reference resolution. In a study by Arnold, Hudson Kam, and Tanenhaus (2007), listeners were presented with both a known and a novel visual object on a computer screen. They found that hesitations in the speech signal created an expectation for a novel target word as judged by increased fixations on the novel object. Although research on the role of disfluencies produced by nonnatives in listener comprehension of speech is, as yet, still lacking, native disfluencies may differ from nonnative disfluencies in their function in speech processing. Nonnative disfluencies, for instance, may arise from incomplete knowledge (grammar and/or vocabulary), or insufficient skills (automaticity) in the nonnative language and thus hinder native speech processing. This difference in the psycholinguistic source of disfluencies may lead to differences in how listeners judge native and nonnative fluency.

Experiment 1

In Experiment 1, both the duration and the number of silent pauses were independently manipulated. These phonetic manipulations were performed both in native and nonnative speech. Native and nonnative speech materials were matched on the manipulated dimension. In Experiment 1, this was achieved by matching the native and nonnative speech materials for the number of silent pauses. The phonetic manipulations involved three pause conditions: speech materials in which silent pauses had been removed, speech materials in which the duration of silent pauses had been altered to be relatively short, and speech materials in which their duration was relatively long. We expected that native speech would be rated as being more fluent than nonnative speech due to

differences between native and nonnative speech in fluency characteristics irrespective of the phonetic manipulations (e.g., filled pauses). We also predicted that fluency would be rated lower when there are more pauses (increasing the number) and/or longer pauses (increasing the duration). We did not make a clear prediction for a possible interaction between the manipulation effects and nativeness. On the one hand, it was possible that the phonetic manipulations would affect ratings of native and nonnative fluency in a similar fashion. On the other hand, it was also possible that the phonetic manipulations would have different effects on native fluency perception, as compared to nonnative fluency perception (cf. the two hypotheses introduced above).

Method

Participants

Participants were 73 paid members of the participant pool of the Utrecht institute of Linguistics OTS (UiL OTS) at Utrecht University. All were native Dutch speakers who reported to have normal hearing ($M_{age} = 20.56$, $SD_{age} = 3.00$; 15m/58f) and who participated with implicit informed consent in accordance with local and national guidelines. A postexperimental questionnaire inquired (among other issues) whether they had noticed anything particular about the experiment. Specifically, they were asked whether they thought the speech had been digitally edited and, if so, how. In total, 27 of the 73 participants responded that they thought the stimuli had been edited in some particular way. Individual responses ranged from comments about nonnative accents to different amounts of background noise or the censoring of personal details. All responses from participants which could reasonably be interpreted as relevant to pause manipulations were taken as evidence of awareness of the experimental manipulation ($n = 14$; 19%). Data from these participants were excluded from any further analyses. The postexperimental questionnaire also assessed participants' prior experience in teaching L2 Dutch or rating fluency. One participant indicated having taught L2 Dutch previously and was excluded for this reason. The final sample size for Experiment 1 was 58. Their mean age was 20.39 years ($SD = 3.15$; 11m/47f).

Stimulus Description

Speech recordings from native speakers and nonnative speakers of Dutch were obtained from the What Is Speaking Proficiency corpus (WISP) in Amsterdam, created and described by De Jong, Steinel, Florijn, Schoonen, and Hulstijn (2012). This corpus was selected because it contains recordings from a large range of native and nonnative speakers of Dutch. All speech in the

Table 1 Descriptions of the selected topics

	CEFR-level	Characteristics	Description
Topic 1	B1	Simple, formal, descriptive	The participant, who has witnessed a road accident some time ago, is in a courtroom, describing to the judge what had happened.
Topic 2	B1	Simple, formal, argumentative	The participant is present at a neighborhood meeting in which an official has just proposed to build a school playground, separated by a road from the school building. Participant gets up to speak, takes the floor, and argues against the planned location of the playground.
Topic 3	B2	Complex, formal, argumentative	The participant, who is the manager of a supermarket, addresses a neighborhood meeting and argues which one of three alternative plans for building a car park is to be preferred.

WISP corpus was collected with signed informed consent from the speakers in accordance with local and national guidelines. All speakers in this corpus had performed computeradministered monologic speaking tasks on eight different topics. These topics had been designed to cover the following three dimensions in a $2 \times 2 \times 2$ fashion: *complexity* (simple, complex), *formality* (informal, formal), and *discourse type* (descriptive, argumentative). For each task, instruction screens provided a picture of the communicative situation and one or several visual-verbal cues concerning the topic. Participants were informed about the audience they were expected to address in each task and were requested to role play as if they were actually speaking to these audiences. From the eight topics, three topics were selected that covered a range of characteristics and that elicited sufficiently long stretches of speech (approximately 2 minutes). In Table 1, descriptions are given of the different topics, together with the proficiency level according to CEFR (Hulstijn, Schoonen, De Jong, Steinel, & Florijn, 2012).

In total, 10 native speakers and 10 nonnative speakers of Dutch were selected. In order to avoid homogeneity in L1 background, nonnative speakers from two L1 backgrounds were selected (5 English and 5 Turkish). Proficiency

in Dutch was assessed by means of a productive vocabulary knowledge test with 116 items, shown to be strongly related to the speakers' overall speaking proficiency (De Jong et al., 2012): $M_{L1} = 106$, $SD_{L1} = 5$; $M_{L2} = 69$, $SD_{L2} = 22$ (max = 116). Comparing these scores to Hulstijn et al. (2012), we find that our nonnative speakers scored approximately at B2 level indicating an intermediate proficiency in Dutch. Their mean length of residence was 7.33 years ($SD = 5.42$) and their mean age of acquisition was 24.9 years ($SD = 3.38$). Fragments of approximately 20 seconds were excerpted from roughly the middle of the original recordings. Thus, 60 speech fragments from 20 speakers talking about three topics were created. All fragments started at a phrase boundary, according to the Analysis of Speech Unit or AS-unit (Foster, Tonkyn, & Wigglesworth, 2000). Most of the fragments also ended at a phrase boundary (native: $n = 23$ out of 30; nonnative: $n = 22$ out of 30), but all fragments ended at a pause (>250 milliseconds).

We attempted to manipulate our native and nonnative speech materials in a similar fashion. Therefore, the native and nonnative speakers were matched on the number of silent pauses per 100 syllables ($M_{L1} = 6.1$, $SD_{L1} = 2.0$; $M_{L2} = 6.5$, $SD_{L2} = 2.2$).

The excerpted speech fragments served as the basis of our stimulus materials. Each speech fragment was manipulated resulting in three different experimental conditions using Praat (Boersma & Weenink, 2012). The three conditions differed in the manipulations targeting pauses with a duration of more than 250 milliseconds. De Jong and Bosker (2013) have demonstrated that a silent pause threshold of 250 milliseconds leads to acoustic measures that have the highest correlation with L2 proficiency (but see Hieke, Kowal, & O'Connell, 1983).

In the NoPauses condition, all pauses of >250 milliseconds were removed by changing the duration to <150 milliseconds. This was achieved by excising silence in between two extremes at positive going zero-crossings in the speech signal. The other two conditions were designed on the basis of the NoPauses condition. In the ShortPauses condition, pauses that originally had a duration of >250 milliseconds, were now altered to have a duration of 250–500 milliseconds. This was achieved by adding silence to the NoPauses condition (extracted silent intervals of that particular recording). In the LongPauses condition, pauses of >250 milliseconds were altered to have a duration of 750–1000 milliseconds. We decided on these two duration intervals because research shows that silent pauses of 250–1000 milliseconds are very common in native speech (Campione & Véronis, 2002) and in nonnative speech (De Jong & Bosker, 2013). Also, in this fashion, the ShortPauses condition would

Table 2 Examples of speech fragments on topic 1 from a native and nonnative speaker

Native speech fragment
uh ik zag een [40; 364; 804] vrouw op de fiets bij een *uh* stoplicht [54; 352; 910] door een groen stoplicht fietsen [*breath of 966 milliseconds*] en ik zag een rode auto voor het stoplicht staan [42; 366; 792] en *uh* op het moment dat zij [40; 374; 896] *uh* voor de auto langs bijna reed begon de rode auto te rijden ik denk dus dat hij door rood reed. *uh* I saw a [40; 364; 804] woman on the bike at a *uh* traffic light [54; 352; 910] pass a green traffic light [*breath of 966 milliseconds*] and I saw a red car standing in front of the traffic light [42; 366; 792] and *uh* at the very moment that she [40; 374; 896] *uh* almost cycled past in front of the car the red car started to drive so I think that his light was red.
Nonnative speech fragment
uh ik z ik heb gezien dat dat die vrouw was aan het [136; 467; 905] rijden [120; 466; 939] toen *uh* met een groene licht op de fiets en een auto kwam van die *uh* rechterkant *uh* was een rooie auto [*breath of 1001 milliseconds*] die man heeft *uh* tegen die vrouw [143; 481; 913] gereden [137; 482; 955] en *uh* [138; 474; 907] ja ik heb de wel een *uh* rode licht denk ik want die *uh* die van die vrouw was nog *uh* groen. *uh* I z I have seen that that woman was [136; 467; 905] driving [120; 466; 939] when *uh* with a green light on the bike and a car came from the *uh* right side *uh* was a red car [*breath of 1001 milliseconds*] that man has *uh* against the woman [143; 481; 913] driven [137; 482; 955] and *uh* [138; 474; 907] yeah I have the well a *uh* red light I think be- cause that *uh* that of that woman was still *uh* green.

Note. Silent pause durations (ms) of the three conditions are given as [NoPauses; ShortPauses; LongPauses]. Translations from Dutch to English are provided below each example.

be clearly distinct from the LongPauses condition, with no overlap between the ShortPauses interval of 250–500 milliseconds and the LongPauses interval of 750–1,000 milliseconds. Pauses close to the silent pause threshold (i.e., between 150 and 250 milliseconds) were decreased in duration to $<$150 milliseconds in each of the three conditions. If a speech fragment contained fewer than three pauses of $>$250 milliseconds, then some pauses of $<$250 milliseconds were also manipulated such that the number of manipulated pauses per item would add up to at least three. Table 2 provides examples of each of the three pause conditions. Note that our phonetic manipulations involved adjustment of silent pauses already present in the original recordings, such that no supplementary silent pauses were added to the speech.

Table 3 Pause characteristics of native and nonnative speech in the three conditions of Experiment 1 ($N = 60$ per column, M (SD))

		NoPauses	ShortPauses	LongPauses
Native	Number of pauses per 100 syllables	0 (0)	6.1 (2.0)	6.1 (2.0)
	Silent pause duration (ms)	0 (0)	383 (40)	867 (32)
Nonnative	Number of pauses per 100 syllables	0 (0)	6.5 (2.2)	6.5 (2.2)
	Silent pause duration (ms)	0 (0)	393 (32)	873 (29)

Note. Silent pause threshold 150 milliseconds.

In natural speech, the ratio between inspiration time and expiration time is about 10% inspiration time and 90% expiration time (Borden, Raphael, & Harris, 1994, pp. 64–65). Therefore, the silent pauses in the NoPauses condition could not all be excised without impairing the naturalness of our materials. For that reason one pause containing a breath located roughly in the middle of a speech fragment was exempted from manipulations in all conditions (not included in the data shown in Table 3).

Prior to running the rating experiment, all items were evaluated for naturalness in a blinded control procedure by the first author. If a particular manipulated silent pause was perceived as unnatural, its duration was slightly altered while maintaining the range of silent pause durations of each manipulation condition. After the first corrections, the evaluation procedure was repeated by the last author. Finally, the second author listened to all the items and again corrections were made. If specific manipulated pauses were still deemed to sound unnatural after all these corrections, this particular pause was exempted from manipulation in all conditions. Table 3 summarizes the differences between the three conditions of Experiment 1 for both native and nonnative speech. All resulting audio stimuli were scaled to an intensity of 70 decibels.

Procedure

The manipulated versions of the speech fragments (i.e., no original recordings) were presented to participants by making use of the FEP experiment software (Veenker, 2006). Each experimental session started with written instructions, presented on the screen, which instructed participants to judge the speech fragments for overall fluency. Raters were instructed not to rate the items in a broad interpretation of fluency (i.e., overall language proficiency, as in "he is fluent in French"). Instead, they were asked to base their judgments on the use of silent and filled pauses, the speed of delivery of the speech, and the use of hesitations and/or corrections (see the Appendix for instructions). Bosker et al.

(2013) have demonstrated that raters, given these instructions, are able to give fluency ratings that correlate strongly with pause and speed measures. Pinget et al. (2014) reported that fluency ratings of this type are relatively independent from such interfering factors as perceived accent. The participants rated the speech fragments using an Equal Appearing Interval Scale (Thurstone, 1928): It included nine stars with labeled extremes ("not fluent at all" on the left; "very fluent" on the right).

Following these instructions but prior to the actual rating experiment four practice items were presented so that participants could familiarize themselves with the task and the items. The participants were given the opportunity to ask questions if they thought they did not understand the task. No instructions other than the written instructions were supplied to the participants by the experimenters.

After the practice items, the experimental session started. Participants listened to the speech fragments over headphones at a comfortable volume in sound-attenuated booths. The experimental items were arranged in a Latin Square design: Participants heard each item in only one condition, with three groups of listeners for counterbalancing. Participants themselves were unaware of this partitioning. In line with the three listener groups, there were three different pseudo-randomized presentation lists of the stimuli and three reversed versions of these lists resulting in six different orders of items.

Each session lasted approximately 45 minutes, but participants were allowed to take a brief pause halfway through the experiment. As introduced previously, at the end of each session the participants filled out a short questionnaire which inquired about personal details, prior experiences with teaching L2 Dutch and/or rating fluency, and which factors they thought had influenced their judgments. We also inquired whether they had noticed anything particular about the speech stimuli (as explained above).

Results

Cronbach's alpha coefficients, as measures of inter-rater agreement, were calculated using the ratings within the three participant groups ($\alpha_1 = .95$; $\alpha_2 = .96$; $\alpha_3 = .95$). Linear Mixed Models (Baayen, Davidson, & Bates, 2008; Lachaud & Renaud, 2011; Quené & Van den Bergh, 2004, 2008) as implemented in the lme4 library (Bates, Maechler, & Bolker, 2012) in R (R Development Core Team, 2012) were used to analyze the data. Our analyses consisted of two phases. In the first phase, a correction procedure was carried out. A model was built with random effects for individual differences between speakers (Speaker), individual differences between raters (Rater), and individual differences in

Table 4 Estimated parameters of mixed-effects modeling on Experiment 1 (standard errors in parentheses)

	estimates	*t* values	significance ($df = 6$)
fixed effects			
Intercept, γ0(00)	5.58 (.15)	37.75	$p < .001$ ***
Nativeness, γA(00)	2.33 (.24)	9.84	$p < .001$ ***
Number contrast, γB(00)	−.79 (.06)	−13.06	$p < .001$ ***
Duration contrast, γC(00)	−.55 (.05)	−10.45	$p < .001$ ***
Nativeness × Number contrast, γD(00)	−.18 (.12)	−1.45	$p = .197$
Nativeness × Duration contrast, γE(00)	−.17 (.11)	−1.62	$p = .156$
Topic 2, γF(00)	.21 (.05)	3.96	$p = .007$ **
Topic 3, γG(00)	.42 (.05)	7.96	$p < .001$ ***
Nativeness × Topic 2, γH(00)	−.25 (.11)	−2.4	$p = .053$
Nativeness × Topic 3, γI(00)	−.70 (.11)	−6.63	$p < .001$ ***
random effects			
Speaker intercept, $\sigma^2_{u\_0(j0)}$	.25		
Rater intercept, $\sigma^2_{v\_0(0k)}$	.46		
Order, $\sigma^2_{w\_Order0(0k)}$	< .01		
Residual, $\sigma^2_{e\_i(jk)}$	1.59		

Note. $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

order effects, varying within raters (Order). Simple models, containing one or two of these predictors, were compared to more complex models that contained one additional predictor. In order to allow such comparisons of models in our analysis, coefficients of models were estimated using the full maximum likelihood criterion (Hox, 2010; Pinheiro & Bates, 2000). Likelihood ratio tests (Pinheiro & Bates, 2000) showed that the most complex model proved to fit the data of Experiment 1 better than any simpler model. This model showed effects of Speaker $\sigma^2_{u\_0(j0)}$, Rater $\sigma^2_{v\_0(0k)}$, and Order, varying within raters, $\sigma^2_{w\_Order0(0k)}$ and contained a residual component $\sigma^2_{e\_i(jk)}$. Extending this model with a fixed effect Order, testing for general learning or fatigue effects, did not improve it ($\chi^2(1) < 1$). Furthermore, we also tested a supplementary model with a maximal random part including random slopes (cf. Barr, Levy, Scheepers, & Tily, 2013; also Barr, 2013). Because this did not lead to a different interpretation of results, we only report the model with a simple random part.

The second phase of our analyses involved the addition of fixed effects to the model. These fixed effects tested for effects of our particular interest, resulting in the model given in Table 4. A fixed effect of Nativeness γA was included to test

for differences between native and nonnative speakers. In the contrasts matrix, native speech was coded with .5 and nonnative speech with -.5. Two Condition contrasts were tested. The first contrast γB compared the NoPauses condition (contrast coding -.5) against the ShortPauses and LongPauses conditions (each receiving the contrast coding of .25), thus testing for an effect of the number of silent pauses. The second contrast γC compared the ShortPauses condition (-.5) against the LongPauses condition (.5), thus testing for an effect of the duration of silent pauses. Matching our first research question, interactions between the two Condition contrasts and the factor Nativeness were also included (γD and γE), thus testing whether the effect of the number or the duration of silent pauses differed across native and nonnative speakers. Finally, fixed effects of the topics tested for differences between the three speaker topics (denoted as γF and γG). Adding additional interactions between fixed effects did not improve the model, as neither interactions between topics and the two Condition contrasts ($\chi^2(4) = 7.07$, $p = .13$) nor three-way interactions between topics, Nativeness, and the two Condition contrasts ($\chi^2(8) = 14.60$, $p = .07$) significantly improved the predictive power of the model. No effect of the L1 background of our nonnative speakers (Turkish vs. English) was observed and, therefore, this factor was excluded from the analysis. The additional interaction between Nativeness and Topic (γH and γI) did improve the model and was therefore included. Results of this model are listed in Table 4. Degrees of freedom (*df*) required for statistical significance testing of *t* values was $df = J - m - 1$ (Hox, 2010), where J is the most conservative number of second-level units ($J = 20$ speakers) and m is the total number of explanatory variables in the model ($m = 13$) resulting in $df = 6$. In Figure 1, the mean fluency ratings are represented graphically.

The significant effect of Nativeness showed that native speakers were rated as more fluent than nonnative speakers. Also, both condition contrasts were found to be statistically significant. Specifically, the condition NoPauses was rated as more fluent than the conditions LongPauses and ShortPauses taken together (the number contrast γB), and the condition ShortPauses was rated as more fluent than the LongPauses condition (the duration contrast γC). The effects of the manipulations on fluency ratings did not differ between native and nonnative speakers, that is, no interaction between either of the two condition contrasts and Nativeness was found. However, effects of the different topics were found in nonnative speech: The significant interaction between Topic 3 and Nativeness showed that only nonnative speech fragments on topic 3 were rated to be more fluent as compared to topic 1.

It is possible to estimate how much of the variability of the fluency ratings the model accounts for by calculating the proportional reduction in unexplained

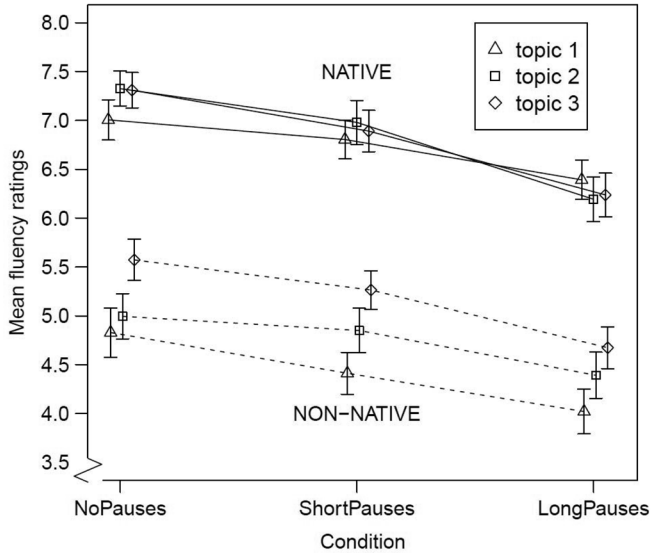

Figure 1 Mean fluency ratings in Experiment 1 (error bars enclose 1.96 × SE, 95% CIs). Plot points were jittered along the *x*-axis to avoid overlap of error bars.

variance (Snijders & Bosker, 1999, pp. 99–103). The proportion of explained variance was estimated by comparing the random variance of the full model (in Table 4) to the simple model without fixed effects. The proportional reduction in unexplained variance of the full model relative to simple model was .34. We also investigated what proportion of the predicted error was accounted for by our manipulation conditions (the Number and the Duration contrasts). For this, we compared the full model with a simpler model without the Number and Duration contrasts as predictors. The proportional reduction in unexplained variance was then found to be .06. This means that our manipulations accounted for 6% of the predicted error.

In Experiment 1, one interaction involving the factor Nativeness was found, namely the interaction between Topic and Nativeness. Our models showed that nonnatives were rated as more fluent when talking about topic 2 and 3 than when talking about topic 1 (cf. Table 1), but this effect was absent in native speech. There may have been acoustic differences between the topics in nonnative speech. For instance, compared to natives, nonnatives could have produced more filled pauses when talking about topic 1 relative to topics 2 and 3. This was assessed in posttest 1 in which the acoustic differences between topics in native and nonnative speech were investigated using Linear Mixed Models.

Based on transcriptions of the speech stimuli, acoustic speech measures were calculated for the stimuli in all three manipulation conditions. The speech measures that were investigated were: (a) the number of silent pauses per second spoken time, (b) the number of filled pauses per second spoken time, (c) the *log* of the mean silent pause duration, (d) the *log* of the mean syllable length, (e) the number of repetitions per second spoken time, and (f) the number of corrections per second spoken time. We tested models that predicted these acoustic speech measures using fixed effects of Topic, Nativeness, and Condition and their interaction (and the random effect Speaker). Indeed one interaction between Topic, Nativeness, and Condition was found: Nonnatives produced significantly fewer silent pauses when talking about topic 3 relative to topic 1 (in the two conditions in which silent pauses were present, namely, ShortPauses and LongPauses). Thus, discussing a more difficult topic pushed the nonnative speakers in our sample to speak more fluently. The decrease in the production of silent pauses may explain, at least in part, the higher ratings on nonnative speech from topic 3.

Another possible account for why nonnatives were rated to be more fluent when talking about topics 2 and 3 may possibly be found in the vocabulary range found in the different topics. Hulstijn et al. (2012) established that successfully produced speech on topic 3 would demonstrate a higher CEFR language proficiency level (B2) than speech on topic 1 or 2 (B1). Adopting this classification of the speaking tasks, raters may have considered the possibly more elaborate vocabulary of the topic when judging fluency. This was investigated in posttest 2, which analyzed the frequency of occurrence of the words produced by native and nonnative speakers. To test whether more complex speaker topics lead to more complex language among nonnative speakers, vocabulary differences between topics were investigated in posttest 2 using Linear Mixed Models. The frequency of occurrence of each token in our speech materials was obtained from SUBTLEX-NL, a database of Dutch word frequencies based on 44 million words from film and television subtitles (Keuleers, Brysbaert, & New, 2010). We tested models that predicted the *log* frequency of each token using Topic and Nativeness and their interaction as fixed effects and Speaker as random effect. One interaction between Topic and Nativeness was found: Nonnatives produced more low-frequency words in fragments from topic 3 relative to topic 1, whereas this did not apply to natives. Thus, discussing a more difficult topic pushed nonnatives to use more low-frequency words. Listeners may have been influenced by lexical sophistication in their assessment of the complexity of the different topics, which may have caused the higher ratings of nonnative speech from topic 3.

Discussion

In summary, Experiment 1 was designed to provide an answer to the question of how listeners weigh the fluency characteristics of native and nonnative speech. Therefore, Experiment 1 targeted the effect of the number of silent pauses and the effect of the duration of silent pauses on both native and nonnative fluency perception. Native and nonnative speech was manipulated such that there were three experimental conditions: NoPauses (<150 milliseconds), ShortPauses (250–500 milliseconds), and LongPauses (750–1000 milliseconds). Participants who reported to have noticed pause manipulations in the speech stimuli were excluded from the analyses ($n = 14$). Adding these participants to the analyses did not lead to a different interpretation of results.

The high Cronbach's alpha coefficients demonstrated that the raters strongly agreed among each other. The main effect of Nativeness showed that, overall, native speakers were perceived to be more fluent than nonnatives (a difference of 2.33 on our 9-point scale). The native and nonnative speakers had been matched on the number of silent pauses, but they still differed in other aspects that have been shown to contribute to fluency perception (Bosker et al., 2013; Cucchiarini et al., 2002; Ginther et al., 2010; Rossiter, 2009). For example, nonnatives produced more filled pauses (*uh*) per second spoken time, more repetitions per second spoken time, and had longer syllable durations than natives. Any of these temporal but also non-temporal factors (e.g., vocabulary, grammar, and so on) may have contributed to the fact that, overall, nonnative speech was rated to be less fluent than native speech.

Furthermore, it has been observed that pauses in native speech occur in different positions in the sentence as compared to those in nonnative speech (e.g., Skehan & Foster, 2007). Our native and nonnative speech materials had been matched for silent pauses, but pause distribution was not taken into account. If pauses in our native materials occurred in different positions than those in our nonnative materials, it may be expected that there would be differential effects of our manipulation conditions across native and nonnative speech. However, inspection of our stimuli showed that our speech fragments of approximately 20 seconds were too short to provide the listener with a firm idea of pause distribution. In fact, the median number of pauses in between AS-units (Foster et al., 2000) per speech fragment was 1.5 for native speakers and 1 for nonnative speakers.

It was also established that increasing the number of silent pauses while keeping all other possibly interacting factors constant led to a decrease in fluency ratings. More specifically, the addition of one pause every 15 syllables

(approximately; see Table 3) led to an average decrease in fluency ratings of .79 on the 9-point scale. Also, increasing the duration of silent pauses resulted in a decrease in fluency judgments: Lengthening pauses by roughly 480 milliseconds (see Table 3) led to an average decrease in fluency ratings of .55 on our 9-point scale. These effects, together with a proportional reduction in unexplained variance of only 6% afforded by the full model with our manipulations of Number and Duration over a simpler model without such manipulations, may seem at first blush to be relatively small contributions of silent pauses to fluency judgments. However, one should note that silent pauses are not the only contributors to perceived fluency ratings. The observed variance in perceived fluency may be explained by a range of factors, such as silent pauses but also filled pauses, speaking rate, corrections, repetitions, and so on. As such, our results are in line with previous research (Bosker et al., 2013; Cucchiarini et al., 2002; Ginther et al., 2010), showing that both the number and the duration of silent pauses have significant effects on fluency ratings. The approach of the current study (manipulating speech in one factor while keeping all else constant) has allowed us (a) to attribute the observed effects to controlled manipulated variables and (b) to distinguish between the contributions of the two properties of silent pauses (i.e., number vs. duration).

With respect to the two hypotheses posed in Experiment 1, our statistical model did not show any difference in the effects of our manipulations across native and nonnative speech. There was no indication that the manipulations affected native speech any differently from nonnative speech. Natives were rated more fluent than nonnatives, and adding silent pauses or lengthening their duration led to lower fluency ratings overall, with no indications for differential effects across native and nonnative speech.

The acoustic differences between topics in nonnative speech and the vocabulary of the nonnative speech from topic 3 may have influenced raters in Experiment 1 to rate nonnatives to be more fluent when talking about topic 2 and 3 relative to topic 1. Still other factors that we did not control for and we have not investigated further can be argued to have influenced the raters (e.g., grammatical accuracy). All these differences between native and nonnative speech may have been partially responsible for the difference between native and nonnative speech in fluency perception. However, these differences between natives and nonnatives were independent from our experimental manipulations. We found no indications for differential effects of our pause manipulations on the perception of fluency in native versus nonnative speech.

Experiment 2

In addition to the speaker's pausing behavior, the speed of speech has been shown to play an important role in fluency perception (e.g., Bosker et al., 2013; Cucchiarini et al., 2002). Experiment 2 extends the insights from Experiment 1 by studying the effect of the speed of speech on fluency ratings of native and nonnative speech. The original native and nonnative speech materials from Experiment 1 (i.e., not the manipulated versions) were reused and manipulated in terms of the speed with which speakers were speaking.

Previously, Munro and Derwing (1998, 2001) also applied speed manipulations to native and nonnative speech. Munro and Derwing (1998), in their Experiment 2, adjusted the speaking rate of native English speech to the mean speaking rate of L2 English speakers and vice versa. Their dependent variable was the rated appropriateness of the speed. They found that some speeded nonnative speech was found to be more appropriate by their raters than unmodified nonnative speech. Munro and Derwing (2001) made use of speed manipulations to study different dependent variables, namely perceived accentedness and comprehensibility. In that study, only nonnative speech materials were analyzed. Results indicated that the speaking rate could account for 15% of the variance in accentedness ratings. The phonetic manipulations in both studies by Munro and Derwing involved speech compression-expansion applied to the entire speech signal including silences. This entails that not only the articulation rate (the number of syllables divided by speaking time, excluding pauses) but also the duration of the pauses was altered (i.e., manipulations of speech rate; the number of syllables divided by total time, including pauses).

In the present Experiment 2, the dependent variable is perceived fluency. Because in the materials of Experiment 1 the articulation rate of the native speakers was not matched to the articulation rate of the nonnative speakers (see the discussion of Experiment 1 above), we used a cross-wise experimental design to match the two groups (cf. Munro & Derwing, 1998). The speed of nonnative speech was sped up to the mean value of the native speakers, and the native speech was slowed down to the mean value of nonnative speakers. This procedure made comparisons across native and nonnative speakers possible. The increase in speed in nonnative speech was expected to lead to an increase in fluency ratings and the decrease in speed in native speech to a decrease in perceived fluency. The magnitude of these two effects may either be similar or different from each other (e.g., speed manipulations affecting nonnative fluency perception more than native fluency perception or vice versa).

An important distinction between Munro and Derwing's (1998, 2001) studies and the current work is that not only the speech rate (including pauses) but

also the articulation rate (excluding pauses) was manipulated. Thus, the contribution of silent pauses to fluency perception (Experiment 1) was clearly separated from the contribution of the speed of the speech (Experiment 2). Experiment 2 thus consisted of three conditions: the original speech, speech with its speech intervals manipulated (i.e., articulation rate manipulations), and speech with both its speech intervals and its silent intervals manipulated simultaneously (i.e., speech rate manipulations). The effect of manipulations in speech rate is expected to be larger than that of manipulations in articulation rate because pause duration has already been shown to contribute to perceived fluency in Experiment 1.

Method

Participants

Seventy-three members from the same participant pool took part in the experiment with implicit informed consent. All were native Dutch speakers with normal hearing (M_{age} = 21.22, SD_{age} = 4.30, 7m/66f). None had previous experience in teaching L2 Dutch or rating fluency. The postexperimental questionnaire inquired (among other issues) whether they had noticed anything particular about the experiment. Of all participants, 19 responded that they thought the stimuli had been edited in some way. Again, individual responses ranged from comments about nonnative accents to different amplitudes. All responses from participants that could reasonably be interpreted as relevant to the pause and also the speed manipulations were taken as evidence of awareness of the experimental manipulation (n = 11; 15%). Data from these participants were excluded from the analyses. Data from an additional four participants were lost due to technical reasons. One participant had already taken part in Experiment 1 and, for that reason, was excluded from further analyses. The final data set included the remaining 57 participants (M_{age} = 21.44, SD_{age} = 3.18, 6m/51f).

Stimulus Description

The original recordings from the native and nonnative speakers from Experiment 1 (i.e., not the pause-manipulated speech fragments) served as the basis for the materials of Experiment 2. As explained above, nonnative speech was increased in speed to match the mean speaking rate of the native speakers and native speech was slowed down to match the mean speaking rate of the nonnative speakers, thus making comparisons across native and nonnative speakers possible. Two types of speed manipulations were performed in Experiment 2, relating to two different measures of the speed of speech. Based on manual

transcriptions of the speech stimuli, both the speech rate and the articulation rate of every speech fragment were calculated. Speech rate is calculated as the number of produced syllables per second of the total time (i.e., including silences). In contrast, articulation rate is calculated per second of spoken time (i.e., excluding silences). In line with this distinction, two types of speed alterations were part of Experiment 2: a manipulation of the spoken time and a manipulation of the total time.

Together with the original recording this resulted in three conditions: Original, Articulation Rate Manipulations (ARM), and Speech Rate Manipulations (SRM). In the ARM condition, native speakers were slowed down to the mean value of the nonnatives (ratio = 1.21) and the speed of nonnative speech was increased to the mean value of the natives (ratio = .83). This manipulation was performed only on the speech intervals in between pauses of >250 milliseconds using Pitch-Synchronous OverLap-and-Add, a method for manipulating the pitch and duration of speech (Moulines & Charpentier, 1990) as implemented in Praat (Boersma & Weenink, 2012). The settings used for the manipulation were the following: minimum frequency = 75 Hz, maximum frequency females = 420 Hz, maximum frequency males = 220 Hz. In this manner, items in the ARM condition differed from the original speech only in the speed of articulation. The duration of silent pauses was identical in both conditions. Table 5 provides examples exemplifying the three manipulation conditions.

Prior to the standard manipulation, native speech fragments that had an exceptionally slow articulation rate (such that, after manipulation, they would fall below the slowest speaking rate of the nonnatives) were either changed to nonoutlier value ($n = 3$), or they were slowed down with a smaller ratio (i.e., a ratio of 1.166; $n = 1$) such that it matched the syllable duration of the slowest nonnative speech fragment. A similar procedure was adopted for exceptionally fast nonnative speech fragments: Prior to the standard manipulation, they were either changed to nonoutlier value ($n = 2$) or their speed was increased with smaller ratios (.88 and .90; $n = 2$), such that they matched the syllable duration of the fastest native speech fragment. Similar to the method of Experiment 1, all manipulated items were evaluated for their naturalness by the first author and corrected accordingly. Subsequently, this procedure was repeated by the last and, finally, also by the second author. For instance, four very fast nonnative sentences within the speech fragments and seven very slow native sentences were exempted from manipulation.

In the SRM condition, the same modifications in native and nonnative speech were made as in the ARM condition but this time the manipulation

Table 5 Examples of speech fragments on topic 1 from a native and nonnative speaker

Native speech fragment
uh ik zag een {**1150; 1387; 1387**} [562; 562; 655] vrouw op de fiets bij een *uh* stoplicht {**3382; 4080; 4080**} [341; 341; 397] door een groen stoplicht fietsen {**1772; 2138; 2138**} [*breath of 966 milliseconds*] en ik zag een rode auto voor het stoplicht staan {**3105; 3746; 3746**} [609; 609; 710] en *uh* op het moment dat zij {**1986; 2397; 2397**} [349; 349; 407] *uh* voor de auto langs bijna reed begon de rode auto te rijden ik denk dus dat hij door rood reed {**7622; 9085; 9085**}
Nonnative speech fragment
uh ik z ik heb gezien dat dat die vrouw was aan het {**2535; 2102; 2102**} [433; 433; 359] rijden {**520; 431; 431**} [373; 373; 308] toen *uh* met een groene licht op de fiets en een auto kwam van die *uh* rechterkant *uh* was een rooie auto {**6905; 5723; 5723**} [*breath of 1001 milliseconds*] die man heeft *uh* tegen die vrouw {**2028; 1681; 1681**} [545; 545; 452] gereden {**883; 732; 732**} [835; 835; 692] en *uh* {**792; 657; 657**} [1209; 1209; 1002] ja ik heb de wel een *uh* rode licht denk ik want die *uh* die van die vrouw was nog *uh* groen {**5648; 4682; 4682**}

Note. Durations of speech intervals (ms) are given in bold as {**Original; ARM; SRM**} and subsequently silent pause durations as [Original; ARM; SRM]. Translations can be found in Table 2.

was performed on the entire speech fragment including the silent pauses. Thus, items in the SRM condition differed from the ARM condition only in the duration of silent pauses. The speed of articulation was identical in the ARM and SRM condition. Table 6 summarizes the differences between conditions of Experiment 2 for both native and nonnative speech. This table illustrates that the values for the two manipulation conditions of native speech were matched to the original values of nonnative speech (and vice versa). All resulting audio stimuli were scaled to an intensity of 70 decibels.

Procedure

The instructions, scales, pseudo-randomization and postexperimental questionnaire in Experiment 2 were the same as those used in Experiment 1 (for instructions, see the Appendix).

Results of Experiment 2

Cronbach's alpha coefficients were calculated on the ratings within the three participant groups ($\alpha_1 = .93$; $\alpha_2 = .93$; $\alpha_3 = .92$). Similar to the analyses in Experiment 1, the ratings were analyzed using Linear Mixed Models. Again, random effects of Speaker, Rater, and Order, varying within raters, were

Table 6 Speed characteristics of native and nonnative speech in the three conditions of Experiment 2

		Original	ARM	SRM
Native	Number of syllables per second spoken time (articulation rate)	4.87 (.53) [3.86–5.72]	4.04 (.44) [3.20–4.74]	4.04 (.44) [3.20–4.74]
	Number of syllables per second total time (speech rate)	3.94 (.51) [3.26–5.13]	3.37 (.41) [2.77–4.33]	3.26 (.42) [2.70–4.26]
Nonnative	Number of syllables per second spoken time (articulation rate)	3.88 (.39) [3.20–4.79]	4.68 (.47) [3.86–5.78]	4.68 (.47) [3.86–5.78]
	Number of syllables per second total time (speech rate)	3.26 (.42) [2.41–4.37]	3.82 (.53) [2.77–5.17]	3.94 (.51) [2.91–5.27]

Note. Mean, (Standard Deviations), [Minimum-Maximum]. $N = 60$ per column. Silent pause threshold 250 milliseconds.

included in the model. We also tested a supplementary model with a maximal random part including random slopes (cf. Barr et al., 2013; also Barr, 2013). Because this did not lead to a different interpretation of results, we only report the model with a simple random part. Subsequently, fixed effects were added to the model, resulting in the model given in Table 7.

Similar to the model of Experiment 1, a fixed effect of Nativeness (γA) compared ratings of native items with ratings of nonnative items. Again, native speech was coded with .5 and nonnative speech with -.5. A fixed effect of ARM (γB) tested for differences between original versions and ARM versions. In the contrast matrix, the original speech received the coding -.5 and the manipulated speech the code .5. Also an interaction with Nativeness was included (γC). Recall that the articulation rate was manipulated in two directions: The articulation rate in nonnative speech was increased whereas it was slowed down in native speech. If the speed manipulations would affect native speech to a similar extent as nonnative speech, then it is expected that slowed down native speech would lead to a decrease in fluency ratings, and that nonnative speech that has been increased in speed would lead to an increase in fluency ratings. In a statistical analysis the decrease in native fluency and the increase in nonnative fluency are, then, expected to cancel each other out. Therefore, we do not expect to find a main ARM effect (γB) but rather an interaction with Nativeness (γC). However, if the speed manipulations affect native speech differently from nonnative speech, this would have to show in a main effect of ARM (γB). The same holds for the SRM condition; therefore, a fixed main effect of SRM and an interaction with Nativeness (γD and γE) were also included.

Table 7 Estimated parameters of mixed-effects modeling on Experiment 2 (standard errors in parentheses)

	estimates	*t* values	significance ($df = 5$)
fixed effects			
Intercept, γ0(00)	5.45 (.17)	32.31	$p < .001^{***}$
Nativeness, γA(00)	1.57 (.29)	5.32	$p = .003^{**}$
ARM, γB(00)	−.09 (.06)	−1.38	$p = .226$
ARM × Nativeness, γC(00)	−.64 (.13)	−4.84	$p = .005^{**}$
SRM, γD(00)	−.14 (.06)	−2.11	$p = .089$
SRM × Nativeness, γE(00)	−1.11 (.13)	−8.41	$p < .001^{***}$
Topic 2, γF(00)	.24 (.06)	4.22	$p = .008^{**}$
Topic 3, γG(00)	.33 (.06)	5.82	$p = .002^{**}$
Nativeness × Topic 2, γH(00)	−.38 (.11)	−3.38	$p = .019^{*}$
Nativeness × Topic 3, γI(00)	−.73 (.11)	−6.48	$p = .001^{**}$
Order, γJ(00)	−.01 (.00)	−2.50	$p = .054$
random effects			
Speaker intercept, $\sigma^2_{u\_0(j0)}$	.40		
Rater intercept, $\sigma^2_{v\_0(0k)}$	.39		
Order, $\sigma^2_{w\_Order0(0k)}$	< .01		
Residual, $\sigma^2_{e\_i(jk)}$	1.78		

Note. $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

In addition, a fixed effect of Topic (γF and γG) was included to investigate main topic effects, along with interactions between Topic and Nativeness (γH and γI). A fixed effect of Order (γJ), testing for overall learning or fatigue effects, improved the explanatory power of the model and was therefore included in the model. No effect of the L1 background of our nonnative speakers (Turkish vs. English) was observed and, therefore, this factor was excluded from the analysis.

The estimates from our statistical model are listed in Table 7. Degrees of freedom (*df*) required for testing of statistical significance of *t* values were computed as follows: $df = J - m - 1$ (Hox, 2010), where J is the most conservative number of second-level units ($J = 20$ speakers) and m is the total number of explanatory variables in the model ($m = 14$) resulting in $df = 5$. Figure 2 illustrates mean fluency ratings from this experiment.

A significant effect of Nativeness showed that, overall, native speakers were rated as more fluent than nonnative speakers. With respect to the ARM condition, no main effect of ARM was found but only an interaction with Nativeness. This interaction reflected the different directions of the ARM manipulations.

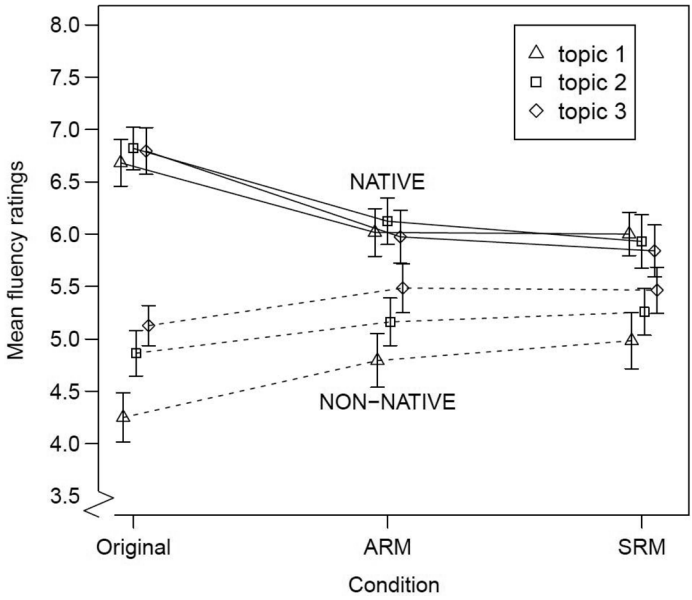

Figure 2 Mean fluency ratings in Experiment 2 (error bars enclose 1.96 × SE, 95% CIs). Plot points were jittered along the *x*-axis to avoid overlap of error bars.

Slowed down native speech was rated as less fluent than the original native speech, and nonnative speech that had received an increased speed was rated as more fluent than the original nonnative speech. The decrease in fluency perception in native speech was found to be similar to the increase in perceived fluency in nonnative speech, as evidenced by the absence of a main effect of ARM. A similar picture is observed for the SRM condition: No effect of this condition was found, but the interaction with Nativeness was statistically significant. The effect of the SRM manipulation was, as expected, larger than the effect of the ARM manipulation (i.e., the effect of SRM × Nativeness was larger than the effect of ARM × Nativeness). In addition, main effects of Topic were found, as well as interactions with Nativeness. Specifically, in nonnative speech, the more difficult topics (2–3) were rated higher in fluency than the easy topic (1). Finally, a very small, statistically marginal overall order effect was also found. The proportion of explained variance was estimated through a comparison of the random variance of the full model, given in Table 7, and the simple model without any fixed effects: .16. The proportional reduction in unexplained variance that was due to the manipulation conditions (i.e., the

ARM and the SRM predictors) was estimated by comparing the full model to a simpler model without ARM and SRM as predictors. The proportional reduction in unexplained variance was then found to be .04. This means that our manipulations accounted for 4% of the predicted error.

Discussion

In summary, Experiment 2 was designed to provide an answer to the question of how listeners weigh the fluency characteristics of native and nonnative speech. Therefore, Experiment 2 focused on the effect of the speed of the speech on both native and nonnative fluency perception. Native and nonnative speech was manipulated such that there were three conditions: original recordings, recordings that had been manipulated in their ARM, and recordings that had been manipulated in their speech rate. In these last two manipulated conditions, the direction of the manipulation differed for native and nonnative speech: nonnative speech was increased to match the native speech whereas native speech was slowed down to match the nonnative speech. Again, those participants who reported to have noticed the manipulations in the speech stimuli were excluded from the analyses ($n = 11$). Adding these participants to the analyses did not lead to a different interpretation of results.

Statistical analyses demonstrated that, overall, natives were perceived to be more fluent than nonnatives (a difference of 1.57 on our 9-point scale). This effect replicates the Nativeness effect found in Experiment 1. It was expected that the increase in speed in nonnative speech would lead to an increase in fluency ratings and that the decrease in speed in native speech would lead to a decrease in perceived fluency. The statistical analyses corroborated this expectation. Crucially, the relative increase and decrease in fluency ratings were of similar magnitude. Natives were rated higher than nonnatives overall, with no indication that manipulation in the speed of speech affected natives and nonnatives differently. Similar to Experiment 1, an interaction between Topic and Nativeness was found: nonnatives were rated to be more fluent when talking about topics 2 and 3 relative to topic 1 (cf. Table 1). Because the same speech materials were used for Experiments 1 and 2, vocabulary differences and acoustic differences between the speech of natives and nonnatives may explain this interaction in the same way as for Experiment 1.

The manipulations of speech intervals in between silent pauses (ARM condition) may not only have affected the perception of these speech intervals but also the perception of the duration of the (unedited) silent pauses. Slowing down speech may cause the duration of pauses to be perceived as subjectively shorter. The expected negative effect of slowing down speech on perceived fluency

could then be countered by a positive effect of shorter pauses. Although we cannot rule out such a countereffect in Experiment 2, it certainly was not strong enough to neutralize the primary effect of our speed manipulations. However, we did observe a stronger effect of the SRM as compared to the ARM, because the former included pauses. In fact, the SRM manipulations can be viewed as a combination of Experiment 1 (silent pauses) and the ARM manipulation within Experiment 2 (speed): the faster the articulation rate and the shorter the pauses, the higher the fluency ratings, both in native and nonnative speech.

General Discussion

The current study carries several implications. First of all, it has demonstrated that fluency characteristics present in the speech signal affect the perception of fluency in both native and nonnative speech: the more disfluency in the utterance, the lower the fluency ratings. This observation extends our current knowledge of the concept of fluency. Previous work has shown that such temporal factors as acoustic measures of the speech signal could explain variation in fluency ratings to a large degree (e.g., Bosker et al., 2013; Cucchiarini et al., 2000). Nontemporal factors such as perceived foreign accent have been shown to play a much smaller role (e.g., Pinget et al., 2014). The finding that the perception of fluency depends on the produced fluency characteristics of speech is relevant, because it confirms that variation in fluency judgments between different speakers can be accounted for by quantitative differences.

Furthermore, our study has demonstrated that the relationship between utterance fluency and perceived fluency is similar across native and nonnative speech. Manipulations of four phonetic factors (number of silent pauses, their duration, articulation rate, and speech rate) showed similar effects on perceived fluency for native and nonnative speakers. This is a striking result considering that native and nonnative speech differs in many respects (e.g., prosody, grammar, lexis, pronunciation, and so on). The main effect of the Nativeness factor in both our experiments testifies to this clear distinction: Our listeners easily discriminated native and nonnative speakers. Nevertheless, our experiments demonstrate that it is possible, through careful phonetic manipulation, to measure how specific acoustic properties contribute to fluency judgments of native and nonnative speech, while keeping some other possibly interacting factors constant. Thus, we observe that silent pause manipulations (Experiment 1) and speed manipulations (Experiment 2) affected subjective fluency ratings of native and nonnative speech to a similar degree.

Our study has demonstrated that, at least within the context of the experimental manipulations reported here, there is no difference in the way listeners weigh the fluency characteristics of native and nonnative speech. One should note, however, that we provided our fluency raters with particular instructions to judge the pausing, speed, and repair behavior of the native and nonnative speakers (see the Appendix). Our instructions were formulated in such a way that raters assessed fluency in its narrow sense (Lennon, 1990), as one of the components of speaking proficiency. The alternative to this approach would be to have raters assess fluency without any instructions on what comprises fluency. This alternative approach is expected to result in ratings of fluency in the broad sense (Lennon, 1990), as a synonym of overall speaking proficiency.

There were several reasons why the experiments reported above used ratings of fluency in the narrow sense. First of all, this approach is consistent with previous studies of fluency perception that have also used specific fluency instructions (cf. Bosker et al., 2013; Derwing et al., 2004; Rossiter, 2009). These studies made use of narrow definitions of fluency in instructions given to listeners (compared to broad or undefined instructions), precisely because the authors wished to collect reliable ratings on how listeners interpret fluency in its narrow sense as one aspect of spoken language. If, in contrast, the interpretation of the concept of fluency were left up to the listener, considerable variability in the subjective ratings would be the expected result. The findings from previous literature we reviewed in this article indicate that instructing listeners to specifically assess fluency in the narrow sense results in subjective ratings that can be accounted for to a large extent by the temporal characteristics of the speech signal.

Another reason for instructing raters to assess the narrow sense of fluency is that this approach is compatible with language testing practice. Most standardized language tests use speaking rubrics with explicit mention of different aspects of fluency, such as speed of delivery and hesitations. Therefore, the raters of these tests are provided with explicit instructions about how to assess oral fluency. Our conclusions about the similarity of native and nonnative fluency perception, based on subjective ratings of the narrow sense of fluency, are therefore directly applicable to the language testing practice where similar methods are used.

Although the narrowly defined fluency definition adopted in this study is fully compatible with existing research and assessment literature, it may still be argued that, by instructing raters to evaluate the pause, speed, and repair behavior of speakers, they were discouraged to take into account other factors that may influence fluency assessment with respect to potential differences between

native and nonnative speech. Thus, the criticism remains that our finding of no difference could be attributed to the specific nature of the instructions given to listeners in making their fluency judgments.

However, our results do not suggest that our specific instructions guided listeners to ignore the distinction between native and nonnative speech. In fact, we observed a consistent main effect of the Nativeness factor in both our experiments, testifying to listeners' ability to perceive a reliable difference in their rating of fluency in native and nonnative speech. If the specificity of our instructions had precluded listeners from taking the distinction between native and nonnative fluency into account, we would not have expected to find these main effects in our two experiments. Therefore, we conclude that the specificity of our instructions cannot fully explain why our listeners weighed the fluency characteristics of native and nonnative speech in a similar fashion.

Our justifications for collecting ratings targeting the narrow sense of fluency do not imply that an alternative approach to fluency perception (i.e., collecting ratings of fluency defined in its broad sense) should not be pursued. In fact, there have been several studies looking into the factors that contribute to perceived oral proficiency. For instance, Kang, Rubin, and Pickering (2010) reported that a combined set of suprasegmental features of nonnative speech (e.g., measures of speech rate, pausing, and intonation) accounted for 50% of the variance in overall proficiency ratings. Ginther et al. (2010) found moderate to strong correlations between overall oral proficiency scores and speech rate, speech time ratio, mean length of run, and the number and length of silent pauses. Taken together, these studies suggest that ratings of fluency in its broad sense are also to a great extent determined by temporal characteristics of nonnative speech. However, it remains to be shown whether native and nonnative fluency characteristics are also weighed in a similar fashion when it comes to perceived fluency in its broad sense. As yet, the relationship between the perception of fluency in its broad and narrow sense is underinvestigated, and so are potential differences between native and nonnative fluency. Our present findings can thus be viewed as an initial attempt to fill these particular gaps in our understanding of fluency perception.

The results of our study carry consequences for how we understand the concept of the native speaker. Disfluencies contribute to the perceived fluency level of native speakers in the same way as they affect nonnative fluency levels. From the literature on social psychology (Brown et al., 1975; Krauss & Pardo, 2006), we know that listeners assess the speech of others on an everyday basis. People make attributions about speakers' social status, background, and even physical properties (Krauss et al., 2002; Krauss & Pardo, 2006; Lindemann & Subtirelu,

2013). Our results show that individual differences between native speakers in their production of disfluencies carry consequences for listeners' perceptions of a native speaker's fluency level. Thus, the idea that native speakers are generally fluent by default can be called into question. Indeed, our results add to the ongoing debate on the notion of the native speaker. For instance, Hulstijn (2011) advocates that a closer look be given to the distinction between native and nonnative, suggesting that the distinction may be a gradient rather than a categorical one. Our study provides some support for this statement, in that our experiments show that variation in fluency production affects subjective fluency judgments. We found no reason to believe that listeners make a qualitative distinction between native and nonnative speakers in fluency assessment. This view also has implications for language testing practice. The fluency level of nonnative speakers is regularly assessed in language tests on the grounds of an idealized native-speaker norm. Our results show that there is variation in the perceived fluency of native speakers. As a consequence, we conclude that a single ideal native fluency standard does not exist.

Last but not least, the results of the current study should not be taken to indicate that native and nonnative fluency characteristics are perceptually equivalent. Despite our finding that native and nonnative fluency characteristics are weighed similarly by listeners, it is likely that the psycholinguistic origins of native and nonnative disfluency in production do differ. Nonnative disfluency, for instance, is likely to be caused by incomplete linguistic knowledge of or skills in the nonnative language, whereas this is unlikely for native disfluency. These different psycholinguistic origins of disfluency could lead to different functions of native and nonnative disfluencies in speech processing. For instance, it has been previously found that native disfluencies may help the listener in word recognition (Corley & Hartsuiker, 2011), in sentence integration (Corley, MacGregor, & Donaldson, 2007), and in reference resolution (Arnold et al., 2007). Whether or not nonnative disfluencies can have similar functions in speech comprehension remains an open empirical question. The current study, which has revealed no essential differences in the way listeners weigh the fluency characteristics of native and nonnative speech, can provide a baseline for future investigations into this and similar issues.

Final revised version accepted 21 January 2014

References

Arnold, J. E., Hudson Kam, C. L., & Tanenhaus, M. K. (2007). If you say -thee uh- you're describing something hard: The on-line attribution of disfluency during

reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 914–930.

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.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*.

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.

Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes*. Retrieved from http://CRAN.R-project.org/package=lme4(R package version 0.999375-39)

Boersma, P., & Weenink, D. (2012). *Praat: doing phonetics by computer* [computer program]. Retrieved from http://www.praat.org/(Version5.3.18).

Borden, G. J., Raphael, L. J., & Harris, K. S. (1994). *Speech science primer: Physiology, acoustics, and perception of speech* (3rd ed.). Baltimore, MD: Lippincott, Williams & Wilkins.

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.

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T. J. M., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*, 157–175.

Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, *34*, 383–398.

Brown, B. L., Strong, W. J., & Rencher, A. C. (1975). Acoustic determinants of perceptions of personality from speech. *International Journal of the Sociology of Language*, *1975*(6), 11–32.

Campione, E., & Véronis, J. (2002). A large-scale multilingual study of silent pause duration. In B. Bel & I. Marlien (Eds.), *Proceedings of the Speech Prosody 2002 Conference* (pp. 199–202). Aix-en-Provence, France: Laboratoire Parole et Langage.

Christenfeld, N. (1996). Effects of a metronome on the filled pauses of fluent speakers. *Journal of Speech, Language and Hearing Research*, *39*, 1232–1238.

Corley, M., & Hartsuiker, R. J. (2011). Why um helps auditory word recognition: The temporal delay hypothesis. *PLoS One*, *6*(5), e19792.

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, *er*, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*, 658–668.

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. (2000). Quantitative assessment of second language learners fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America*, *107*, 989–999.

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.

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 *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech (DISS)* (pp. 17–20).

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2013). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*. 10.1017/S0142716413000210

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*, 5–34.

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.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*, 354–375.

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, *34*, 709–738.

Freed, B. F. (1995). What makes us think that students who study abroad become fluent? In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 123–148). Amsterdam: John Benjamins.

Freed, B. F. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 243–265). Ann Arbor: The University of Michigan Press.

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*, 379–399.

Goldman-Eisler, F. (1958a). The predictability of words in context and the length of pauses in speech. *Language and Speech*, *1*, 226–231.

Goldman-Eisler, F. (1958b). Speech production and the predictability of words in context. *The Quarterly Journal of Experimental Psychology*, *10*, 96–106.

Hieke, A. E., Kowal, S., & O'Connell, D. C. (1983). The trouble with "articulatory" pauses. *Language and Speech*, *26*, 203–214.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*, 229–249.

Hulstijn, J. H., Schoonen, R., De Jong, N. H., Steinel, M. P., & Florijn, A. F. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for languages (CEFR). *Language Testing*, *29*, 202–220.

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

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *Modern Language Journal*, *94*, 554–566.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*, 643–650.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*, 145–164.

Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers physical attributes from their voices. *Journal of Experimental Social Psychology*, *38*, 618–625.

Krauss, R. M., & Pardo, J. S. (2006). Speaker perception and social behavior: Bridging social psychology and speech science. In P. Van Lange (Ed.), *Bridging social psychology: Benefits of transdisciplinary approaches* (pp. 273–278). Hillsdale, NJ: Erlbaum.

Lachaud, C. M., & Renaud, O. (2011). A tutorial for analyzing human reaction times: How to filter data, manage missing values, and choose a statistical model. *Applied Psycholinguistics*, *32*, 389–416.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*, 387–417.

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25–42). Ann Arbor: The University of Michigan Press.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Lindemann, S., & Subtirelu, N. (2013). Reliably biased: The role of listener expectation in the perception of second language speech. *Language Learning*, *63*, 567–594.

Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, *15*, 19–44.

Mora, J. C. (2006). Age effects on oral fluency development. In C. Muñoz (Ed.), *Age and the rate of foreign language learning* (pp. 65–88). Clevedon, UK: Multilingual Matters.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*, 453–467.

Munro, M. J., & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning*, *48*, 159–182.

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition*, *23*, 451–468.

Panico, J., Healey, E. C., Brouwer, K., & Susca, M. (2005). Listener perceptions of stuttering across two presentation modes: A quantitative and qualitative approach. *Journal of Fluency Disorders*, *30*, 65–85.

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.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer Verlag.

Quené, H., & Van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, *43*, 103–121.

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*, 413–425.

R Development Core Team. (2012). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from http://www.R-project.org/(ISBN3-900051-07-0)

Riazantseva, A. (2001). Second language proficiency and pausing. *Studies in Second Language Acquisition*, *23*, 497–526.

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

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and nonnative speakers of English. *Canadian Modern Language Review*, *65*, 395–412.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*, 227–256.

Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition*, *14*, 357–385.

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

Skehan, P. (2009). Modeling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*, 510–532.

Skehan, P., & Foster, P. (2007). Complexity, accuracy, fluency and lexis in task-based performance: A meta-analysis of the Ealing research. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy, and fluency in*

second language use, learning, and teaching (pp. 207–226). Brussels, Belgium: University of Brussels Press.
Snijders, T. A., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: SAGE.
Susca, M., & Healey, E. C. (2001). Perceptions of simulated stuttering and fluency. *Journal of Speech, Language and Hearing Research*, *44*, 61–72.
Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal*, *65*, 71–79.
Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529–554.
Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, *17*, 84–119.
Trofimovich, P., & Baker, W. (2007). Learning prosody and fluency characteristics of second language speech: The effect of experience on child learners' acquisition of five suprasegmentals. *Applied Psycholinguistics*, *28*, 251–276.
Veenker, T. J. G. (2006). *FEP: A tool for designing and running computerized experiments, Version 2.4.19* [computer software].
Wennerstrom, A. (2000). The role of intonation in second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 102–127). Ann Arbor: The University of Michigan Press.

Appendix

Literal instructions to participants in the two experiments (in Dutch; English translation given below):

> *"Jouw taak is om spraakfragmenten te beluisteren en te beoordelen op vloeiendheid. Baseer je oordeel telkens op: (1) het gebruik van pauzes: bijv. geen en/of zeer korte stille en gevulde pauzes, of juist zeer veel en/of zeer lange stille en gevulde pauzes; (2) de snelheid van spreken: bijv. zeer langzaam of zeer snel; (3) het gebruik van herhalingen en correcties: bijv. geen of juist zeer veel herhalingen en/of correcties."*
>
> "It is your task to rate the speech fragments on fluency. Base your judgments on: (1) the use of pauses: e.g., none and/or very short silent and filled pauses vs. very many and/or very long silent and filled pauses; (2) the speed of speaking: e.g., very slow vs. very fast; (3) the use of repetitions and corrections: e.g., none vs. very many."