*Toivola, M., P. Lintunen & L. M. Heikkola (toim.) 2024. Puheen tutkimuksen uusia suuntia – Aineistona vapaasti tuotettu puhe. New directions in speech research – Freely produced speech as data. AFinLA-teema / n:o 17, 191–215.*

**Elina Lehtilä, Pauliina Peltonen & Pekka Lintunen**
University of Turku

# Listener perceptions of L1 speech in L2 and L3 fluency assessment contexts: a qualitative approach

## Highlights

- Listeners' target language fluency assessments are influenced by both temporal and non-temporal L1 features.
- There are individual differences across listeners in how they attend to L1 speech when assessing L2/L3 fluency.
- Half of the listeners felt that listening to L1 speech before assessing L2/L3 fluency facilitated the rating process.

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

## Abstract

Studies on listeners' perceptions of speech fluency have typically focused on numeric L2 fluency assessments and rarely addressed individual differences in L1 speech fluency qualitatively. The objective of the present study was to examine how listeners attend to speakers' L1 speech when assessing the speakers' L2 and L3 fluency. In the study, 37 listeners were presented with freely produced L1 Finnish, L2 English (Experiment I), and L3 Swedish (Experiment II) speech by Finnish-speaking and Finnish–Swedish bilingual speakers (N=30). The listeners first listened to the speakers' L1 Finnish speech before assessing their target language fluency and then commented on how hearing the L1 speech samples affected their ratings. A qualitative analysis focusing on the listeners' comments revealed that listeners attend to both temporal and non-fluency related features in L1 speech when assessing target language fluency, but there were clear individual differences across listeners' intuitive approaches to L1 speech.

**Keywords:** fluency, speech perception, individual differences
**Asiasanat:** sujuvuus, puheen havaitseminen, yksilölliset erot

# 1 Introduction

Listener perceptions of a speaker's fluency play an essential role in all communicative settings. Examining subjective listener-based assessments of learners' speech fluency in different linguistic contexts and understanding the factors affecting listeners' perceptions of fluency is, therefore, crucial for the learning and teaching of second language (L2) spoken production skills. While a number of studies have investigated L2 perceived fluency from purely quantitative (e.g., Derwing et al. 2004; Bosker et al. 2013) and mixed-methods perspectives (e.g., Magne et al. 2019; Peltonen 2022), examining listeners' qualitative, freely produced comments across different target language assessment contexts can uncover important aspects of the factors underlying listeners' perceptions of fluency.

Over the last decade, connections between language learners' speech fluency in their first (L1) and target languages (L2, L3…) have received increased attention (e.g., De Jong et al. 2015; Duran-Karaoz & Tavakoli 2020; Peltonen & Lintunen 2022; Gao & Sun 2024), and the accumulating results from these studies indicate that several L2 fluency features are affected by L1 fluency features. However, to our knowledge, examining how a speaker's L1 fluency affects their target language fluency has not been studied from the listener's perspective, even though it has been suggested that individual differences in L1 speech fluency should be recognized in L2 fluency assessment contexts (De Jong 2018; Gao & Sun 2024). With a detailed qualitative analysis, the present study examines how listeners take L1 speech into account when assessing the speakers' fluency in the L2 and L3 and how helpful they perceive L1 speech to be in a fluency assessment context. Thus, the overall aim of the present

study is to gather empirical evidence on the connections between L1 and L2 fluency from the listener's perspective.

## 2 Approaches to listener perceptions of L2 fluency

In second language acquisition (SLA) research, fluency is traditionally defined as one of the three key components of L2 performance within the complexity, accuracy, and fluency framework (CAF; see, e.g., Housen et al. 2012), and it is typically regarded as a complex phenomenon between speaker and listener involving both speech production and perception (Lennon 2000). With reference to Lennon's (1990, 2000) distinction between broad and narrow senses of fluency, we approach fluency from the narrow sense, where speech fluency refers to the speed and efficiency of delivery, as opposed to global oral proficiency of the speaker (see also Lintunen et al. 2020). Using the term *perceived fluency* to refer to speech fluency from the listener's perspective, we subscribe to the following definition of the concept developed by Segalowitz (2010: 48): "the inferences listeners make about a speaker's cognitive fluency based on their perception of utterance fluency". Thus, although this study focuses on perceived fluency, we consider listeners' perceptions of fluency to be inherently linked to the actual temporal features of speech (i.e., *utterance fluency*) and the efficiency of the speaker's processes underlying speech production (i.e., *cognitive fluency*; Segalowitz 2010). The features of speech that are thought to reflect cognitive fluency are typically grouped into three main fluency dimensions: *speed* (speech tempo), *breakdown* (pausing), and *repair fluency* (corrections and hesitations; Tavakoli & Skehan 2005; Skehan 2009).

As stated in the introduction, L2 perceived fluency is often studied from a quantitative perspective, where numeric fluency assessments based on spoken productions are examined in terms of objective utterance fluency measures (for a recent meta-analysis, see Suzuki et al. 2021). Results from these studies indicate that listeners' perceived fluency ratings are strongly connected with measures related to speed (e.g., speech rate and articulation rate; Préfontaine et al. 2016; Kallio et al. 2023) and breakdown fluency (e.g., frequency, duration and distribution of silent pauses; Kahng 2018), and to a lesser extent with repair fluency measures (e.g., frequency of corrections; Cucchiarini et al. 2002; Bosker et al. 2013). While perceived fluency has rarely been studied from a purely qualitative perspective (for an exception, see Préfontaine & Kormos 2016), a number of studies have approached L2 perceived fluency using mixed-methods designs, where quantitative analyses are typically complemented with qualitative analyses based on post-assessment interviews (Freed 1995; Magne et al. 2019; Suzuki & Kormos 2020) or raters' written comments (Kormos & Dénes 2004; Rossiter 2009; Peltonen 2022). These qualitative analyses have demonstrated that L2 fluency ratings tend to be affected by non-fluency

related factors, even when the listeners have been instructed to base their assessments on temporal and hesitation phenomena (Rossiter 2009; Magne et al. 2019). The factors that have been found to influence listeners' perceptions of L2 fluency include other linguistic aspects of language production such as pronunciation, vocabulary, and grammar, but also non-linguistic speaker attributes such as perceived confidence (Freed 1995; Rossiter 2009) and nativelikeness (Préfontaine & Kormos 2016; Magne et al. 2019).

While a consistent finding in several utterance fluency studies using correlation and regression analyses has been that there are connections between learners' L1 and L2 fluency features (e.g., De Jong et al. 2015; Duran-Karaoz & Tavakoli 2020), perceived fluency studies have rarely addressed connections between L1 fluency and listeners' assessments of L2 fluency. Some utterance fluency studies also indicate that the connections between L1 and L2/L3 fluency measures are potentially influenced by cross-linguistic differences and learners' proficiency level in the target language (Huensch & Tracy-Ventura 2017; Peltonen 2018; Peltonen & Lintunen 2022). In one of the few studies examining L1 and L2 fluency links in an assessment context, Pinget et al. (2014) examined connections between L2 Dutch fluency ratings and both traditional L2 fluency measures and L1-adjusted fluency measures (residuals obtained from correlating L1 and L2 fluency measures) for L1 Turkish and English (N=30) speakers. The L1-adjusted measures were not found to be better predictors of perceived fluency than the traditional fluency measures, which is perhaps unsurprising since, as acknowledged by the authors, their listeners only had access to the speakers' L2 Dutch speech.

In a different approach, Derwing et al. (2009) examined the relationship between L1 and L2 fluency across 16 Mandarin- and 16 Russian/Ukrainian-speaking immigrants to Canada by collecting fluency ratings based on both the learners' L1 and L2 English productions. For both learner groups, positive correlations were found between the L1 and L2 fluency ratings at the first data collection time, but not at the later stages of data collection over the course of two years. Although these results point towards a complex relationship between L1 and L2 fluency, the authors recognize that their assessment data cannot be used for a direct comparison of L1 and L2 perceived fluency, as different listeners (native speakers of the learners' L1s vs. native speakers of English) were used to rate the L1 and L2 speech samples. As neither Pinget et al. (2014) nor Derwing et al. (2009) examined connections between L1 and L2 perceived fluency from the perspective of the same listeners, how listeners attend to L1–L2 connections when assessing L2 fluency remains an unexplored area in L2 fluency research.

To address this gap in research, the present study examined how listeners intuitively attend to speakers' L1 Finnish speech when assessing their target language (L2 English/L3 Swedish) fluency. Using freely produced L1 Finnish, L2 English, and L3 Swedish speech recorded from Finnish-speaking and Finnish–Swedish bilingual

speakers, the study included two experiments: in Experiment I, we examined all speakers' (N=30) L2 English perceived fluency, and, in Experiment II, the Finnish-speaking participants' (n=20) L3 Swedish perceived fluency. In both experiments, the listeners (N=37) were first presented with short excerpts from the speakers' L1 Finnish samples and then asked to assess their target language fluency.[1]

The data analyses focused on the listeners' open-ended comments on how hearing the Finnish speech samples affected their rating process. Furthermore, we examined the listeners' numeric evaluations of how helpful they felt the Finnish speech samples were in rating the target language speech samples. The study addressed the following research questions:

1) Which features and themes can be identified in the listeners' comments regarding the influence of L1 speech on L2 and L3 fluency assessment?
2) How do the themes and features identified from the comments relate to the listeners' helpfulness evaluations?

# 3 Methodology

## 3.1 Speech data

The speech data used in the present study were collected as part of a project Fluency across Multilingual Speakers (MultiFluency; funded by the Swedish Cultural Foundation in Finland) and selected based on our previous research on these speakers' utterance fluency in Finnish, Swedish, and English (see Peltonen & Lintunen 2022). The speakers were 20 Finnish-speaking and 10 Finnish–Swedish bilingual university students of language subjects from two Finnish universities, most of whom studied English as either their major (n=21) or minor (n=5) subject (see Table 1 for a summary of the speakers' background information). The target language speech samples used in Experiment I consisted of the L2 English speech samples from all 30 speakers, while the samples used in Experiment II consisted of the L3 Swedish speech samples produced by the 20 Finnish-speaking participants. All speakers volunteered to participate in the study, and informed consent was obtained from the speakers prior to participating.

To elicit spontaneous speech, all speakers were asked to perform cartoon-based narratives in Finnish, Swedish, and English. The speakers were presented with three different cartoon strips consisting of six pictures and given two minutes to plan their

1 The experiment design also included numeric fluency assessments from listeners who only heard the speakers' L2/L3 speech samples, making the total number of listeners in the two experiments 72. As the present study examines listeners' perceptions of L1 speech from a qualitative perspective, here we focus only on the listeners who had access to the speakers' L1 speech samples.

narrative before performing the task in each of the three languages. The speakers were not allowed to take notes during the planning time, but they could look at the cartoons when they were speaking. Two of the cartoon strips used in the study had been employed in previous research projects (e.g., Peltonen & Lintunen 2016), and the third cartoon strip was selected based on its comparability with the other two prompts in terms of key visual elements and narrative structure. In the first cartoon, a man sat down next to two people on a bench and lifted his hat in greeting, revealing a squirrel underneath. In the second cartoon, a man came across a group of children building a snowman, and the children asked the man to help them put a hat on the snowman's head. The third cartoon showed two children growing a tree, reflecting their growth into adults. The first task was always done with the same cartoon in either Finnish or Swedish depending on the speaker's L1, but the order of the additional two languages and the additional two cartoons was counterbalanced across speakers (for details, see Peltonen & Lintunen 2022). To ascertain the speakers' overall level of oral proficiency in the target languages, four trained raters evaluated the L2 English and L3 Swedish speech samples based on CEFR criteria for qualitative aspects of spoken language use (range, accuracy, fluency, and coherence; Council of Europe 2001: 28–29). All speakers were assessed to represent level C1 on average in spoken L2 English and the Finnish-speaking participants level A2 on average in spoken L3 Swedish.

TABLE 1. Speakers' background characteristics by language group.

| | L1(s) | | | |
|---|---|---|---|---|
| | Finnish (n=20) | | Finnish and Swedish (n=10) | |
| Background variable | *M* | *SD* | *M* | *SD* |
| Age (years) | 20.60 | 2.04 | 27.00 | 8.38 |
| Duration of English studies at school (years) | 10.18 | 1.09 | 10.33 | 2.40 |
| Duration of Swedish studies at school (years) | 6.63 | 1.56 | | |
| Self-rated English spoken production (1–5*) | 4.0 | 0.65 | 4.1 | 0.74 |
| Self-rated Swedish spoken production (1–5*) | 1.85 | 1.04 | | |

*Scale: 1 = weak, 2 = moderate, 3 = good, 4 = very good, 5 = excellent.

For the assessment tasks, all speech samples were first resampled to a sampling frequency of 44100 Hz and scaled to an average intensity of 70 dB (see, e.g., Pinget et al. 2014). For each L1 Finnish narrative, approximately 20-second excerpts were extracted from the beginning of the original recording, excluding any initial hesitations and ending at a natural break in speech. Based on the piloting of the two experiments, the excerpts were deemed sufficient to get an impression of the speaker's L1 speaking style. However, to increase the ecological validity of the assessment, the entire target language recordings were presented to the listeners (Préfontaine et al. 2016; Suzuki & Kormos 2020). The mean duration of the target language recordings was 64.60 s in Experiment I (*SD* = 29.34) and 56.75 s in Experiment II (*SD* = 19.02).

## 3.2 Listeners and assessment task

Skilled non-native speakers of the target language (see, e.g., Rossiter 2009; Magne et al. 2019) were recruited as listeners in the present study. The listeners in Experiment I (n=15; 11 female, 3 male, 1 preferred not to say, mean age 27.07 years) were advanced Finnish-speaking English majors at a Finnish university. The average duration of the listeners' English studies at university was 4.17 years (*SD* = 0.92), and 12 listeners had either completed or were participating in teacher training at the time of the study. The listeners in Experiment II were advanced Finnish-speaking university students of Swedish (n=22; all female, mean age 23.86 years) from two Finnish universities. The listeners had studied Swedish at university for 3.86 years on average (*SD* = 0.79). Sixteen listeners studied Swedish as their major subject and six as their minor subject, and all listeners had either completed or were participating in teacher training. The listeners in both experiments reported high familiarity with Finnish-accented target language speech on a scale from 1 to 6 (Experiment I: *M* = 4.93, *SD* = 1.22; Experiment II: *M* = 4.91, *SD* = 1.06). While most listeners had no prior experience in language assessment, seven listeners in Experiment I and nine listeners in Experiment II reported having some, relatively limited previous experience in assessing learners' spoken or written language skills. All listeners reported normal hearing and completed the assessment task as part of an MA-level course (n=33) or as volunteers (n=4). All listeners participated in the study voluntarily and gave informed consent to participate in the study before beginning the assessment task.

The assessment tasks were implemented online using REDCap, a secure data collection software hosted at the University of Turku (Harris et al. 2009; Harris et al. 2019). All listeners completed the assessment task unsupervised on their own computers and received all the instructions in written format (see Appendix). The listeners were first presented with a narrow definition of fluency, focusing on the temporal aspects of speech, and then instructed to rate each speaker's target lan-

guage speech rate, pausing, corrections, and overall fluency (based on their ratings of speech rate, pausing, and corrections) using 9-point scales (see, e.g., Rossiter 2009; Bosker et al. 2013; Kahng 2018; Magne et al. 2019; Suzuki & Kormos 2020). The inter-rater reliability was high for all rated aspects in both experiments, and the results regarding these numeric assessments are reported in Lehtilä et al. (in review). To mitigate the potential influence of familiarity bias, the participants were also introduced to the cartoons used in the speaking tasks before they began the assessment task (see, e.g., Derwing et al. 2004; Rossiter 2009; Kahng 2018; Suzuki & Kormos 2020). Following two practice trials, the target language speech samples were presented to the listeners in one of three pseudo-randomized orders to minimize potential order effects. Regarding the L1 Finnish speech samples, the listeners received the following instructions: "Before each [target language] sample, you will hear a short speech sample in the learners' first language (Finnish). Listen to this speech sample once before listening to the [target language] sample to get an impression of the learner's individual speaking style, but do not rate the Finnish sample." Thus, the listeners were not instructed to assess the speakers' L1 speech or to compare first and target language fluency features, as the objective was to examine how listeners intuitively attend to L1 speech when assessing target language fluency.

After completing the assessment task, the listeners were asked to evaluate how helpful the Finnish speech samples were in rating the L2/L3 speech samples using a scale with labeled endpoints (1 = not helpful at all and 6 = very helpful). After assigning the score for the overall helpfulness, the listeners were asked to comment freely on how they themselves felt hearing the Finnish speech samples had affected their rating process. These comments and the numeric helpfulness evaluations form the qualitative and quantitative data of the present study.

## 3.3 Analysis

To answer the first research question, the listeners' comments on how they felt hearing the speakers' L1 Finnish speech samples affected their L2/L3 fluency assessments were analyzed qualitatively using qualitative content analysis (Dörnyei 2007). Due to limited prior research in the area, the comments were coded in an inductive approach, where themes and patterns are identified from the data without the imposition of preconceived labels or categories (Miles & Huberman 1994; Dörnyei 2007). As each listener provided one comment at the end of the assessment task, altogether 37 comments were analyzed. The coding frame was developed and evaluated following O'Connor and Joffe's (2020: 20) recommendations for intercoder reliability assessment in qualitative research. Each comment was treated as a single data unit, and the initial coding frame was first developed and applied to the data by the first author, after which the second author independently coded approximately 50% of the data using the initial

coding frame. Reliability statistics (Krippendorff's alpha coefficient; Krippendorff 2004) were then calculated in R 4.0.4 (R Core Team 2021) using the irr R package (v0.84.1; Gamer et al. 2019) and interpreted individually for each code. The four codes that fell short of the acceptability threshold ($\alpha \geq .800$; Krippendorff 2004) were discussed and revised, and the first and second author independently coded the remaining data in accordance with the revised coding frame. After the revision, all codes except for one reached sufficient intercoder reliability. The code with an insufficient reliability value was excluded from the finalized coding frame.

The finalized coding frame based on the listeners' comments included five main categories: (1) Degree of effect, (2) Effect type, (3) Fluency features, (4) Speaker characteristics, and (5) Listener characteristics. Each main category had two to four subcategories, which are discussed in detail in the following section. To answer the second research question, cross-tabulations between the coded subcategories and the listeners' helpfulness evaluations were prepared to examine whether any patterns would be detected between the listeners' comments and their evaluations regarding the helpfulness of the L1 Finnish speech samples in their L2/L3 rating process.

# 4 Results

## 4.1 Listeners' perceptions of L1 speech in L2 and L3 fluency assessment

The subcategories for the five main categories of the finalized coding frame are presented in Tables 2–6, including subcategory descriptions and the percentages of comments in each subcategory in Experiments I (L2 English) and II (L3 Swedish). The first main category 'Degree of effect' reflected the magnitude of the effect that the L1 speech samples had on the L2/L3 fluency assessments, as estimated by the listeners themselves. As illustrated in Table 2, the most common degree of effect in Experiment I was mild (40.0%), whereas the most common degree of effect in Experiment II was moderate (45.5%). In addition, there were more listeners in Experiment I whose comments were classified under the subcategory 'No effect' (20.0%) compared to Experiment II (9.1%). Very few listeners in both experiments stated or implied that listening to the L1 speech samples had had a significant effect on their L2/L3 fluency assessments (Experiment I: 6.7%; Experiment II: 9.1%; note that the percentages in Experiment II do not add up to 100% due to rounding).

TABLE 2. Subcategories for the main category 'Degree of effect' and the percentages of all comments in Experiments I (L2 English) and II (L3 Swedish) for each subcategory.

| Subcategory | Description | Percentage of all comments | |
|---|---|---|---|
| | | Experiment I (L2 English) | Experiment II (L3 Swedish) |
| No effect | Listener states/implies that the L1 samples did not have any effect on their ratings | 20.0% | 9.1% |
| Mild | Listener states/implies that the L1 samples had or might have had a mild, potentially unspecified effect on their ratings | 40.0% | 36.4% |
| Moderate | Listener states/implies that the L1 samples had or might have had a moderate effect on their ratings and identifies at least one feature they attended to in the L1 samples | 33.3% | 45.5% |
| Large | Listener states/implies that they actively used the L1 sample (e.g., for all the speakers) in the rating process at least for one specified feature | 6.7% | 9.1% |

The second main category, 'Effect type', reflected the ways in which the listeners reported the L1 speech samples had affected their rating process (note that, from Table 3 onwards, the percentages do not add up to 100% as one comment could be coded in multiple categories). As indicated in Table 3, the most common effect type was 'Facilitative' in both experiments (Experiment I: 53.3%; Experiment II: 50.0%), i.e., the listeners stated or implied that hearing the L1 speech samples made the rating process easier in some way (e.g., helped to prepare them for the L2/L3 samples). Approximately a quarter of the participants in both experiments reported consciously comparing the L1 sample to the L2/L3 sample for at least one specified feature (Experiment I: 26.7%; Experiment II: 27.3%). In addition, a few listeners in both experiments indicated that the L1 samples were not helpful in assessing L2/L3 fluency (Experiment I: 13.3%; Experiment II: 9.1%).

TABLE 3. Subcategories for the main category 'Effect type' and the percentages of all comments in Experiments I (L2 English) and II (L3 Swedish) for each subcategory.

| Subcategory | Description | Percentage of all comments | |
|---|---|---|---|
| | | Experiment I (L2 English) | Experiment II (L3 Swedish) |
| Facilitative | Listener states/implies that the L1 samples made the rating process easier in some way | 53.3% | 50.0% |
| Comparative | Listener states/implies that they consciously compared the L1 samples to the L2/L3 samples | 26.7% | 27.3% |
| Non-facilitative | Listener states/implies that the L1 samples were not helpful in the rating process for some or all of the rated variables | 13.3% | 9.1% |

The third main category covered the three fluency features the listeners were instructed to base their L2/L3 fluency assessments on (speech rate, pausing, and corrections). As can be seen from Table 4, the most commonly mentioned fluency feature in both experiments was speech tempo (Experiment I: 60.0%; Experiment II: 45.5%), whereas pauses were less frequently mentioned (Experiment I: 20.0%; Experiment II: 13.6%). Very few listeners in Experiment II (9.1%) and none of the listeners in Experiment I mentioned attending to hesitations or corrections in the speakers' L1 speech.

TABLE 4. Subcategories for the main category 'Fluency features' and the percentages of all comments in Experiments I (L2 English) and II (L3 Swedish) for each subcategory.

| Subcategory | Description | Percentage of all comments | |
|---|---|---|---|
| | | Experiment I (L2 English) | Experiment II (L3 Swedish) |
| Speech tempo | Listener mentions speech rate as one of the features they attended to in the L1 speech samples | 60.0% | 45.5% |
| Pauses | Listener mentions pauses as one of the features they attended to in the L1 speech samples | 20.0% | 13.6% |
| Hesitations and corrections | Listener mentions corrections or hesitations as one of the features they attended to in the L1 speech samples | 0.0% | 9.1% |

The two remaining main categories consisted of two subcategories each, covering all other speaker (Table 5) and listener (Table 6) characteristics that were identified from the comments. In addition to the three fluency features, listeners in both experiments reported attending to other, non-fluency related features (e.g., vocabulary choices or narrative strategies) in the L1 samples (Experiment I: 26.7%; Experiment II: 22.7%). Some listeners also mentioned typological or proficiency-related differences between the two languages spoken by the participants (Experiment I: 6.7%; Experiment II: 9.1%). While the frequencies of comments in the subcategories for 'Speaker characteristics' were very similar in both experiments, both subcategories for 'Listener characteristics' were more prominent in Experiment I compared to Experiment II. As summarized in Table 6, compared to Experiment II, more listeners in Experiment I reported attempting to consciously focus only on the L2 samples in the rating process (26.7%) and mentioned their personal attitudes, expectations or behaviors regarding the rating procedure (33.3%).

TABLE 5. Subcategories for the main category 'Speaker characteristics' and the percentages of all comments in Experiments I (L2 English) and II (L3 Swedish) for each subcategory.

| | | Percentage of all comments | |
|---|---|---|---|
| Subcategory | Description | Experiment I (L2 English) | Experiment II (L3 Swedish) |
| Personal speaking style | Listener mentions other, non-fluency related features they attended to in the L1 speech samples | 26.7% | 22.7% |
| L1–L2/L3 difference | Listener mentions either typological differences or differences in speakers' proficiency levels between the L1 and L2/L3 | 6.7% | 9.1% |

TABLE 6. Subcategories for the main category 'Listener characteristics' and the percentages of all comments in Experiments I (L2 English) and II (L3 Swedish) for each subcategory.

| Subcategory | Description | Percentage of all comments | |
|---|---|---|---|
| | | Experiment I (L2 English) | Experiment II (L3 Swedish) |
| Deliberate disregard | Listener states/implies attempting to consciously focus only on the L2/L3 samples and/or ignore the L1 samples in the rating process | 26.7% | 9.1% |
| Personal responses | Listener mentions their personal attitudes, expectations or behaviors regarding the rating procedure | 33.3% | 18.2% |

## 4.2 Relationship between the listeners' comments and their helpfulness evaluations

The distribution of the listeners' evaluations regarding the helpfulness of the L1 Finnish speech samples when assessing L2/L3 fluency in Experiments I and II are illustrated in Figure 1. As shown in the figure, the majority of the responses in Experiment I fell in the middle of the scale (3–4; $M = 3.67$, $SD = 1.23$), whereas the most common response in Experiment II was 2 ($M = 3.09$, $SD = 1.31$). Furthermore, some listeners in Experiment II evaluated the L1 samples as 'not helpful at all' (1), and none of the listeners evaluated the L1 samples as 'very helpful' (6). Overall, the distributions suggest that the listeners in Experiment II felt that the L1 Finnish samples were less helpful in assessing target language fluency than the listeners in Experiment I.

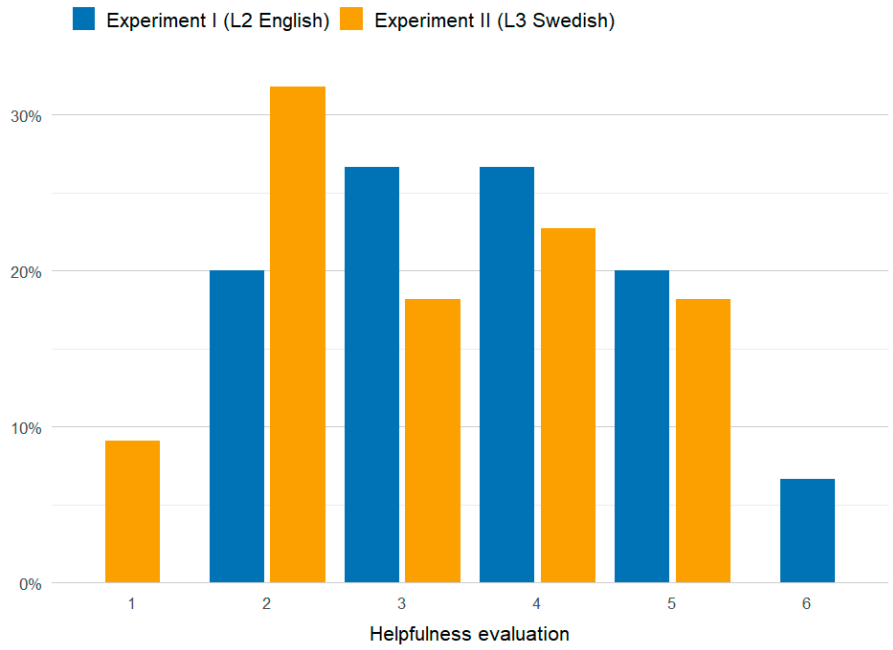


FIGURE 1. Percentage of responses regarding the perceived feeling of helpfulness of the L1 Finnish speech samples when assessing fluency (scale 1–6; 1 = not helpful at all, 6 = very helpful) in Experiments I (L2 English) and II (L3 Swedish).

The cross-tabulations between the coded subcategories and the listeners' helpfulness evaluations are summarized in Tables 7 (Experiment I) and 8 (Experiment II).

TABLE 7. Cross-tabulations of the coded subcategories according to the helpfulness evaluations in Experiment I (L2 English).

| | Helpfulness evaluation | | | | |
|---|---|---|---|---|---|
| Subcategory | 2 | 3 | 4 | 5 | 6 |
| Degree of effect | | | | | |
| No effect | 66.7% | 33.3% | – | – | – |
| Mild | 16.7% | 33.3% | 33.3% | 16.7% | – |
| Moderate | – | 20.0% | 40.0% | 40.0% | – |
| Large | – | – | – | – | 100.0% |

| | Helpfulness evaluation | | | | |
|---|---|---|---|---|---|
| Subcategory | 2 | 3 | 4 | 5 | 6 |
| Effect type | | | | | |
| Facilitative | 25.0% | 25.0% | 12.5% | 25.0% | 12.5% |
| Comparative | – | 25.0% | 50.0% | 25.0% | – |
| Non-facilitative | 50.0% | 50.0% | – | – | – |
| Fluency features | | | | | |
| Speech tempo | 11.1% | 22.2% | 44.4% | 11.1% | 11.1% |
| Pauses | – | – | 33.3% | 33.3% | 33.3% |
| Speaker characteristics | | | | | |
| Personal speaking style | – | – | 25.0% | 50.0% | 25.0% |
| L1–L2 difference | 100.0% | – | – | – | – |
| Listener characteristics | | | | | |
| Deliberate disregard | 75.0% | 25.0% | – | – | – |
| Personal responses | 20.0% | 20.0% | 60.0% | – | – |

TABLE 8. Cross-tabulations of the coded subcategories according to the helpfulness evaluations in Experiment II (L3 Swedish).

| | Helpfulness evaluation | | | | |
|---|---|---|---|---|---|
| Subcategory | 1 | 2 | 3 | 4 | 5 |
| Degree of effect | | | | | |
| No effect | – | 100.0% | – | – | – |
| Mild | 25.0% | 50.0% | 25.0% | – | – |
| Moderate | – | 10.0% | 20.0% | 50.0% | 20.0% |
| Large | – | – | – | – | 100.0% |
| Effect type | | | | | |
| Facilitative | – | 18.2% | 27.3% | 27.3% | 27.3% |
| Comparative | – | – | 16.7% | 33.3% | 50.0% |

| | Helpfulness evaluation | | | | |
|---|---|---|---|---|---|
| Subcategory | 1 | 2 | 3 | 4 | 5 |
| Non-facilitative | 50.0% | 50.0% | – | – | – |
| Fluency features | | | | | |
| Speech tempo | – | 20.0% | 30.0% | 20.0% | 30.0% |
| Pauses | – | – | 100.0% | – | – |
| Hesitations and corrections | – | 50.0% | – | – | 50.0% |
| Speaker characteristics | | | | | |
| Personal speaking style | – | – | – | 80.0% | 20.0% |
| L1–L3 difference | 50.0% | 50.0% | – | – | – |
| Listener characteristics | | | | | |
| Deliberate disregard | 50.0% | 50.0% | – | – | – |
| Personal responses | – | 50.0% | 25.0% | 25.0% | – |

Regarding the degree of effect, tables 7 and 8 show that in both experiments the larger degrees of effect were associated with higher helpfulness evaluations. In particular, the subcategories of 'No effect' and 'Large' were respectively associated with lower and higher helpfulness evaluations, as demonstrated in comments 1 and 2 (the examples come directly from the responses without modifications or omissions; English translations by the first author are provided below each Swedish comment). With respect to effect type, comments 1 and 2 also illustrate that there was no clear relationship between the helpfulness evaluations and the 'Facilitative' effect type in either experiment. However, in both experiments the majority of the listeners who reported actively comparing the two speech samples gave the L1 samples helpfulness evaluations of 4 or 5 in the 6-point scale (see comment 2). Unsurprisingly, the listeners whose comments were coded as 'Non-facilitative' also evaluated the L1 samples to be less helpful in both experiments, as demonstrated by comment 3.

(1) They didn't really affect the rating since they were not supposed to be rated. They only provided a little soundcheck before hearing the actual sample that was supposed to be rated.

(Listener 21, Experiment I; 'No effect', 'Facilitative', 'Deliberate disregard'; helpfulness = 2)

(2) Jag anser att de finska talproven var väldigt nyttiga eftersom jag visste en natural talhastighet för alla läsare och kunde jämföra det till de svenska talproven.

I think the Finnish speech samples were extremely helpful because I knew a natural speech rate for all the readers and could compare it to the Swedish speech samples.

(Listener 52, Experiment II; 'Large', 'Facilitative', 'Comparative', 'Speech tempo'; helpfulness = 5)

(3) Genom finska talproven fick man bekanta sig med rösten av personen men talproven i finska och svenska var ändå så jätteolika så de hjälpte inte jättemycket med bedömning.

Through the Finnish speech samples one got to familiarize oneself with the speaker's voice but the Finnish and Swedish speech samples were still so very different that they didn't help that much with the assessment.

(Listener 46, Experiment II; 'Mild', 'Non-facilitative', 'L1–L3 difference'; helpfulness = 1)

In terms of the three fluency features, there was no clear association between 'Speech tempo' and helpfulness evaluations, as indicated by the relatively even distribution of the helpfulness evaluations for the subcategory in both experiments. In comparison, attending to pauses seemed to be associated with somewhat higher helpfulness evaluations in Experiment I, but not in Experiment II, as illustrated in comment 4.

(4) Hearing the Finnish samples were handy because they gave perspective to certain behaviours that might not have been discernable from the English samples alone. For example, longer pauses might not be due to disfluency, but rather a linguistic strategy for the speaker. This, however, poses a problem for assessing fluency: Should L2 fluency be assessed based on the speaker's L1 fluency, or a 3rd party standard (native speakers)?

(Listener 20, Experiment I; 'Moderate', 'Facilitative', 'Pauses', 'Personal speaking style'; helpfulness = 5)

In both experiments, however, the listeners who reported attending to features related to personal speaking styles evaluated the L1 samples to be more helpful (see comments 4 and 5), whereas the listeners who mentioned differences between the L1 and L2/L3 or reported attempting to ignore the L1 samples evaluated the L1 samples to be less helpful (see comment 6).

(5) It gave me an idea of the persons style of speaking and also often gave a "hint" whether or not the person was going to simply describe the comic or come up with a story with character names and even lines.

(Listener 17, Experiment I; 'Moderate', 'Facilitative' 'Personal speaking style'; helpfulness = 5)

(6) Jag tycker att de finska talproven påverkade nästan inte alls mina bedömningar. Jag försökte att inte ta inte så mycket hänsyn till de finska talproven, för att jag burde ju

> inte bedöma de. Men det är ju möjligt att de påverkade bedömningen, men på ett omedvetet sätt.
>
> I feel that the Finnish speech samples hardly affected my ratings. I tried not to pay too much attention to the Finnish speech samples, since I wasn't supposed to assess them. However, it's possible that they did affect the assessment, but in an unconscious way.
>
> (Listener 43, Experiment II; 'Mild', 'Deliberate disregard'; helpfulness = 1)

## 5 Discussion and conclusion

The purpose of this study was to examine how listeners intuitively attend to freely produced L1 speech when assessing L2 and L3 fluency and how helpful they perceive information about L1 speech to be in terms of fluency assessment. Based on an analysis of open-ended comments and numeric helpfulness evaluations provided by 37 listeners, the present study provided information on the thus far understudied effects of L1 speaking style on listeners' assessments of target language fluency. Furthermore, by using speech samples from the same speakers to investigate listener perceptions of L1 Finnish speech in both L2 English (Experiment I) and L3 Swedish (Experiment II) fluency assessment contexts, the study design enabled us to investigate these effects across multiple languages with differing levels of language proficiency.

The listeners' comments were categorized into five main categories and 14 subcategories that were identified from the data based on the qualitative content analysis. The five main categories were (1) Degree of effect, (2) Effect type, (3) Fluency features, (4) Speaker characteristics and (5) Listener characteristics. With respect to the perceived degree of effect, we found that most listeners did not feel that the L1 speech samples had had a significant effect on their target language fluency assessments, with the majority of the comments falling under the subcategories of 'Mild' and 'Moderate' degree of effect in both experiments. However, the cross-tabulation analyses revealed that the perceived degree of effect was connected with how helpful the listeners felt the L1 samples were in the rating process, with higher helpfulness evaluations coinciding with larger degrees of effect in both experiments.

The different effect types revealed that approximately half of the listeners in both experiments felt that hearing the L1 speech samples somehow facilitated their rating process. However, the cross-tabulation analyses revealed no connection between the facilitative effect type and how helpful the listeners evaluated the L1 samples to be in the rating process. A possible explanation for this could be that, in the listeners' comments from both experiments, there was variation within the subcategory of facilitative effect. While some listeners felt that the L1 speech samples helped to assess target language fluency characteristics, the facilitative effect was

often more generally attributed to the opportunity to familiarize oneself with the speaker before having to assess their target language fluency. Furthermore, how helpful this familiarizing effect was perceived to be varied across listeners, as indicated by the variation found in their helpfulness evaluations. Based on our findings, it seems that giving listeners access to the speakers' L1 speech helps them to adjust and prepare for the target language speech samples, thus potentially releasing cognitive resources for language assessment, but the extent to which listeners find this helpful for L2/L3 fluency assessment specifically depends on the listener's individual characteristics.

A comparison of the three fluency features showed that speech tempo was the feature most often mentioned by the listeners in both experiments, indicating that the rate of speech is the fluency feature most consciously attended to in L1 speech. Interestingly, none of the listeners in Experiment I and very few listeners in Experiment II mentioned hesitations or corrections. Although this finding might seem surprising considering the prevalence of hesitations and self-corrections in language examinations and frameworks (Huhta et al. 2020), it echoes the results of Magne et al. (2019), who also focused on advanced L2 listeners and found that very few commented on repair fluency as influencing their fluency assessments, despite the significant correlation between their numeric ratings and the frequency of self-corrections. Thus, it is possible that the listeners in the present study attended to repair fluency features on a more subconscious level, but it is also possible that their relatively limited experience in language testing and assessment led them to focus more on speed and breakdown fluency features than repairs. Future studies could apply the assessment procedure introduced in the present study to listeners with more language testing experience (such as experienced language teachers) to corroborate the results.

In addition, a similar proportion of listeners in both experiments reported attending to other, non-fluency related features in the L1 speech samples, such as narrative strategies or vocabulary choices (see, e.g., comment 5). These results are in line with previous mixed-methods and qualitative perceived fluency studies, where listeners' fluency assessments have been found to be influenced by factors other than temporal and hesitation phenomena (e.g., Freed 1995; Rossiter 2009; Préfontaine & Kormos 2016; Magne et al. 2019). Furthermore, cross-tabulation analyses showed that the listeners who reported attending to these features also assigned higher helpfulness scores for the L1 samples, which might point towards potential connections between these non-fluency related first and target language features. As, to our knowledge, connections between L1 and L2/L3 speech have only been studied with regards to fluency measures, a possible future research direction could be to examine the relationship between first and target language speech across other domains of language production (e.g., grammatical and lexical complexity). In addition, as our study focused on perceived fluency, future studies could investigate

whether information about L1 speaking style has an effect on the assessment of other target language oral skills, such as complexity or accuracy. However, acknowledging the limitations based on the results from the two small-scale experiments and the potential effects of the populations under investigation, our findings should be confirmed with large sample sizes and different speaker and listener populations.

We also found clear differences across individual listeners in how they intuitively attended to L1 speech when assessing target language fluency: approximately a fifth of the listeners in both experiments reported actively comparing first and target language fluency features, whereas some listeners reported attempting to consciously ignore L1 fluency features and focus only on target language fluency. As the listeners in both experiments were advanced Finnish-speaking university students of the target language and the majority of the listeners had participated in teacher training, the results suggest that even listeners with similar language and educational backgrounds may employ varying approaches to how they attend to L1 speech when assessing target language fluency. Since these differences across listeners represent a potential source of rater bias especially in assessment settings where the raters are familiar with the speakers' L1 speaking styles (e.g., in classroom-based speaking tests), the findings imply that there is a need to establish a common approach as to whether and how individual differences in L1 fluency should be taken into account in these assessment contexts. Furthermore, there is a need to increase pre-service language teachers' awareness of how individual differences may affect intuitive judgements of target language fluency (see also Peltonen 2023). Regarding the implications for large-scale language testing (e.g., in high-stakes standardized language tests), if and how potential L1 speaking style effects should be taken into account is a much more complex issue due to practical considerations and potential threats to test validity. As our findings provide the first insights into how listeners attend to L1 speaking styles when assessing fluency, more research on this topic is needed to ensure fair and equal language assessment practices across different contexts (see also De Jong 2018; Gao & Sun 2024).

In conclusion, the present study has been among the first to study the connections between first and target language speech fluency from a perceived fluency perspective. The research questions addressed the themes identified from the listeners' open-ended comments and the relationship between these themes and the listeners' helpfulness evaluations. The results for the first research question revealed a wide range of approaches to L1 speech in both L2 and L3 fluency assessment contexts. In addition, while the feature most often mentioned in both experiments was L1 speech tempo, the listeners in both experiments also commented on features of L1 speech that were not related to fluency. The results for the second research question also showed that these listeners perceived information about L1 speech to be more helpful than the listeners who commented only on L1 fluency features. Overall, the study has extended previous perceived fluency research by qualitatively

examining listeners' perceptions of L1 speech when assessing fluency in different target languages and, thus, provided important insights regarding the relationship between L1 and L2/L3 perceived fluency. The findings also have implications beyond fluency research, as they indicate that the potential effect of L1 speech fluency on target language fluency assessment needs to be addressed explicitly in both teacher training and language assessment.

## Literature

Bosker, H. R., A-F. Pinget, H. Quené, T. Sanders & N. H. de Jong 2013. What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30* (2), 159–175. https://doi.org/10.1177/0265532212455394

Council of Europe 2001. *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Cucchiarini, C., H. Strik & L. Boves 2002. Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America, 111* (6), 2862–2873. https://doi.org/10.1121/1.1471894

De Jong, N. H. 2018. Fluency in second language testing: insights from different disciplines. *Language Assessment Quarterly, 15* (3), 237–254. https://doi.org/10.1080/15434303.2018.1477780

De Jong, N. H., R. Groenhout, R. Schoonen & J. H. Hulstijn 2015. Second language fluency: speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36* (2), 223–243. https://doi.org/10.1017/S0142716413000210

Derwing, T. M., M. J. Munro, R. I. Thomson & M. J. Rossiter 2009. The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition, 31* (4), 533–557. https://doi.org/10.1017/S0272263109990015

Derwing, T. M., M. J. Rossiter, M. J. Munro & R. I. Thomson 2004. Second language fluency: judgments on different tasks. *Language Learning, 54* (4), 655–679. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Duran-Karaoz, Z. & P. Tavakoli 2020. Predicting L2 fluency from L1 fluency behaviour: the case of L1 Turkish and L2 English speakers. *Studies in Second Language Acquisition, 42* (4), 671–695. https://doi.org/10.1017/S0272263119000755

Dörnyei, Z. 2007. *Research methods in applied linguistics: quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.

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*. Amsterdam: John Benjamins, 123–148. https://doi.org/10.1075/sibil.9.09fre

Gamer, M., J. Lemon & I. F. P. Singh 2019. irr: Various coefficients of interrater reliability and agreement (R package version 0.84.1). *The Comprehensive R Archive Network*. https://CRAN.R-project.org/package=irr

Gao, J. & P. P. Sun 2024. How does learners' L2 utterance fluency relate to their L1? A meta-analysis. *International Journal of Applied Linguistics, 34* (1), 276–291. https://doi.org/10.1111/ijal.12493

Harris, P. A., R. Taylor, B. L. Minor, V. Elliott, M. Fernandez, L. O'Neal, L. McLeod, G. Delacqua, F. Delacqua, J. Kirby, S. N. Duda & REDCap Consortium 2019. The REDCap consortium: building an international community of software platform partners. *Journal of Biomedical Informatics, 95*, Article 103208. https://doi.org/10.1016/j.jbi.2019.103208

Harris, P. A., R. Taylor, R. Thielke, J. Payne, N. Gonzalez & J. G. Conde 2009. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42* (2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Housen, A., F. Kuiken & I. Vedder (eds) 2012. *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA.* Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.32

Huensch, A. & N. Tracy-Ventura 2017. Understanding second language fluency behaviour: the effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. *Applied Psycholinguistics, 38* (4), 755–785. https://doi.org/10.1017/S0142716416000424

Huhta, A., H. Kallio, S. Ohranen & R. Ullakonoja 2020. Fluency in language assessment. In P. Lintunen, M. Mutta & P. Peltonen (eds) *Fluency in L2 learning and use.* Bristol: Multilingual Matters, 129–145. https://doi.org/10.21832/9781788926317-011

Kahng, J. 2018. The effect of pause location on perceived fluency. *Applied Psycholinguistics, 39* (3), 569–591. https://doi.org/10.1017/S0142716417000534

Kallio, H., M. Kautonen & M. Kuronen 2023. Prosody and fluency of Finland Swedish as a second language: investigating global parameters for automated speaking assessment. *Speech Communication, 148*, 66–80. https://doi.org/10.1016/j.specom.2023.02.003

Kormos, J. & M. Dénes 2004. Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32* (2), 145–164. https://doi.org/10.1016/j.system.2004.01.001

Krippendorff, K. 2004. Reliability in content analysis: some common misconceptions and recommendations. *Human Communication Research, 30* (3), 411–433. https://doi.org/10.1093/hcr/30.3.411

Lehtilä, E., P. Peltonen & P. Lintunen in review. How does access to L1 speech affect L2 and L3 perceived fluency? Evidence from Finnish-speaking and Finnish–Swedish bilingual speakers.

Lennon, P. 1990. Investigating fluency in EFL: a quantitative approach. *Language Learning, 40* (3), 387–417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

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

Lintunen, P., M. Mutta & P. Peltonen (eds) 2020. *Fluency in L2 learning and use.* Bristol: Multilingual Matters. https://doi.org/10.21832/9781788926317

Magne, V., S. Suzuki, Y. Suzukida, K. Saito, M. Ilkan & M. Tran 2019. Exploring the dynamic nature of second language listeners' perceived fluency: a mixed-methods approach. *TESOL Quarterly, 53* (4), 1139–1150. https://doi.org/10.1002/tesq.528

Miles, M. B. & A. M. Huberman 1994. *Qualitative data analysis: an expanded sourcebook* (2nd edition). Thousand Oaks: Sage.

O'Connor, C. & H. Joffe 2020. Intercoder reliability in qualitative research: debates and practical guidelines. *International Journal of Qualitative Methods, 19*, 1–13. https://doi.org/10.1177/1609406919899220

Peltonen, P. 2018. Exploring connections between first and second language fluency: a mixed methods approach. *The Modern Language Journal, 102* (4), 676–692. https://doi.org/10.1111/modl.12516

Peltonen, P. 2022. Connections between measured and assessed fluency in L2 peer interaction: a problem-solving perspective. *International Review of Applied Linguistics in Language Teaching, 60* (4), 983–1011. https://doi.org/10.1515/iral-2020-0030

Peltonen, P. 2023. Fluency revisited. *ELT Journal*. Advance online publication. https://doi.org/10.1093/elt/ccad047

Peltonen, P. & P. Lintunen 2016. Integrating quantitative and qualitative approaches in L2 fluency analysis: a study of Finnish-speaking and Swedish-speaking learners of English at two school levels. *European Journal of Applied Linguistics, 4* (2), 209–238. https://doi.org/10.1515/eujal-2014-0018

Peltonen, P. & P. Lintunen 2022. Multilingual speakers' L1, L2, and L3 fluency across languages: a study of Finnish, Swedish, and English. *Nordand, 17* (1), 48–63. https://doi.org/10.18261/nordand.17.1.4

Pinget, A.-F., H. R. Bosker, H. Quené & N. H. de Jong 2014. Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing, 31* (3), 349–365. https://doi.org/10.1177/0265532214526177

Préfontaine, Y. & J. Kormos 2016. A qualitative analysis of perceptions of fluency in second language French. *International Review of Applied Linguistics in Language Teaching, 54* (2), 151–169. https://doi.org/10.1515/iral-2016-9995

Préfontaine, Y., J. Kormos & D. E. Johnson 2016. How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing, 33* (1), 53–73. https://doi.org/10.1177/0265532215579530

R Core Team 2021. *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/

Rossiter, M. J. 2009. Perceptions of L2 fluency by native and non-native speakers of English. *The Canadian Modern Language Review, 65* (3), 395–412. https://doi.org/10.3138/cmlr.65.3.395

Segalowitz, N. 2010. *The cognitive bases of second language fluency*. New York: Routledge. https://doi.org/10.4324/9780203851357

Skehan, P. 2009. Modelling second language performance: integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30* (4), 510–532. https://doi.org/10.1093/applin/amp047

Suzuki, S. & J. Kormos 2020. Linguistic dimensions of comprehensibility and perceived fluency: an investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition, 42* (1), 143–167. https://doi.org/10.1017/S0272263119000421

Suzuki, S., J. Kormos & T. Uchihara 2021. The relationship between utterance and perceived fluency: a meta-analysis of correlational studies. *The Modern Language Journal, 105* (2), 435–463. https://doi.org/10.1111/modl.12706

Tavakoli, P. & P. Skehan 2005. Strategic planning, task structure, and performance testing. In R. Ellis (ed.) *Planning and task performance in a second language*. Philadelphia: John Benjamins, 239–273. https://doi.org/10.1075/lllt.11.15tav

# Appendix:

## Rating instructions in English (a Swedish version of the instructions was used in Experiment II).

You will hear 30 speech samples in English produced by learners of English. Your task is to listen to the speech samples and rate them in terms of their **fluency**. Before each English sample, you will hear a short speech sample in the **learners' first language (Finnish)**. Listen to this speech sample once before listening to the English speech sample to get an impression of the learner's **individual speaking style**, but do not rate the Finnish sample.

In this study, fluency refers to the **ease and effortlessness** of speech and is thus separate from other aspects of oral proficiency, such as grammatical complexity and accuracy. Please give ratings for each of these three **temporal aspects of speech** using a 9-point scale:

- speech rate **(1 = very slow, 9 = very fast)**
- pausing **(1 = very many and/or very long disruptive pauses, 9 = very few and/or very short pauses)**
- corrections **(1 = very many and/or very long corrections, 9 = very few and/or very short corrections)**

Based on your ratings of speech rate, pausing and corrections, rate the overall fluency of the speaker on a scale from 1 to 9 *(1 = extremely disfluent, 9 = extremely fluent)*.

You will first rate **two practice samples** in order to familiarize yourself with the rating procedure.