How do utterance measures predict raters' perceptions of fluency in French as a second language?

Yvonne Préfontaine

University of Illinois at Urbana-Champaign

Judit Kormos

Daniel Ezra Johnson

Lancaster University

Preprint version to be published in Language Testing

Abstract

While the research literature on second language (L2) fluency is replete with descriptions of fluency and its influence with regard to English as an additional language, little is known about what fluency features influence judgments of fluency in L2 French. This study reports the results of an investigation that analyzed the relationship between utterance fluency measures and raters' perceptions of L2 fluency in French using mixed-effects modeling. Participants were 40 adult learners of French at varying levels of proficiency, studying in a university immersion context. Speech performances were collected on three different types of narrative tasks. Four utterance fluency measures were extracted from each performance. Eleven untrained judges rated the

speech performances and we examined which utterance fluency measures are the best predictors of the scores awarded by the raters. The mean length of runs and articulation rate proved to be the most influential factors in raters' judgments, while the frequency of pauses played a less important role. The length of pauses was positively related to fluency scores indicating a prominent cross-linguistic variation specific to French. The relative importance of the utterance measures in predicting fluency ratings, however, was found to vary across tasks.

Keywords: L2 fluency, psycholinguistics, language assessment, fluency judgments, speech production and perception

How do utterance measures predict raters' perceptions of fluency in French as a second language?

Fluency is an important construct in the assessment of language proficiency and forms part of a large number of rating scales in various high stakes exams (e.g. IELTS (International English Language Testing System) and in descriptors of levels of second language (L2) language competence (e.g. Common European Framework for Languages (CEFR, Council of Europe, 2001). Previous investigations have analyzed L2 fluency in terms of native speaker judgments and perceptions (Derwing, Rossiter, Munro, & Thomson, 2004; Iwashita, Brown, McNamara, & O'Hagan, 2008; Kormos & Dénes, 2004; Rossiter, 2009) primarily with learners of English as a second language (ESL). While such research has contributed significantly to our understanding of L2 fluency in L2 English, little is known about how fluency is perceived and evaluated in L2 French despite the fact that previous cross-linguistic research has uncovered important differences between fluency phenomena in French and English (Grosjean & Deschamps, 1975; Raupach, 1987)

Research in task-based learning has shown that fluency is greatly influenced by the speech task (for a recent meta-analysis see Jackson & Suethanapornkul, 2013; Préfontaine & Kormos, 2015). Additionally, fluency might also vary within individuals, for example depending on the L2 learner's momentary feelings of anxiety (MacIntyre & Gardner, 1994; Wood, 2010). Recent developments in mixed-effects modeling (e.g. Baayen, Davidson, & Bates, 2008; Barr, Levy, Scheepers, & Tily, 2013) allow for treating these potentially confounding variables as random effects and can provide us with a better understanding of the nature of the relationship between utterance fluency and listeners' perceptions of fluency (see also Bosker, Pinget, Quené, Sanders, & De Jong, (2013). The present investigation with learners of L2 French is novel in that it uses the fluency descriptors of the CEFR rating scale, which frequently serves as a basis of

making inferences about students' fluency in a number of national and international language proficiency exams (e.g., Cambridge English Language Assessment, Cambridge Michigan Language Assessments (CaMLA), Test de connaissance du français (TCF), Diplôme d'études en langue française (DELF). Additionally, our research mirrors exam contexts where raters make judgments about learners' fluency based on their complete task performance, and thus has higher ecological validity than many of the previous research projects that only used very short segments of students' utterances (Bosker et al., 2013; Derwing, Munro, Thomson, & Rossiter, 2009; Derwing et al., 2004; Freed, 2000; Rossiter, 2009). Our study addressed three research questions:

1) How do four utterance fluency variables (articulation rate, mean length of runs, pause frequency and average pause time) predict fluency ratings using the Common European Framework of Reference (CEFR) (Council of Europe, 2001) scale?
2) How is the perception of pauses predicted by these utterance fluency variables?
3) How is the perception of speed predicted by the utterance fluency variables?

L2 speech characteristics that influence fluency judgments

Research in second language (L2) fluency has been concerned with two major themes to account for fluent speech production and perceptions of fluency: first, temporal variables to measure utterance fluency; and second, factors that affect rater evaluation of perceived fluency. According to Segalowitz (2010, p. 48), *utterance fluency* designates the temporal variables of speech or the "oral features of utterances that reflect the operation of underlying cognitive processes", while *perceived fluency* refers to the "inferences listeners make about a speaker's cognitive fluency based on their perception of utterance fluency". For the purposes of this study,

these definitions will be employed because they most accurately describe L2 speech production and perception referring both to the automatic nature of spoken language on the part of the speaker producing speech, and the listener, perceiving it. Accordingly, this perspective is also in line with Lennon's (2000) definition, namely that fluency is the "rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention under the temporal constraints of on-line processing" (p. 26).

A number of previous investigations have examined the temporal variables which best predict fluency (Préfontaine, 2013a; Bosker et al., 2013; Derwing et al., 2009; Derwing et al., 2004; Freed, 2000; Freed, Segalowitz, & Dewey, 2004; Ginther, Dimova, & Yang, 2010; Iwashita et al., 2008; Kormos & Dénes, 2004; Riggenbach, 1991; Rossiter, 2009; Towell, Hawkins, & Bazergui, 1996). Of particular interest is the ESL research of Kormos and Dénes (2004) who analyzed ten temporal variables and fluency judgments from untrained raters. Correlation analyses found that speech rate (SR) (syllables per second, unpruned), mean length of runs (MLR) (0.25 sec pause cut-off), phonation-time ratio (PTR) and the number of stressed words produced per minute (pace) were the best predictors of fluency. While Lennon (1990) and Foster and Skehan (1999) reported that the frequency of filled pauses and unfilled pauses correlated with fluency, these speech phenomena did not impact perceptions of fluency in Kormos and Dénes' study. With the exception of two raters, average pause time (APT) did not affect fluency judgments or account for much variation in the listeners' perceptions.

Derwing et al. (2004) also found associations between temporal variables and L2 perceived fluency with ESL learners. They analyzed SR (syllables per second, pruned), MLR (0.40 sec pause cut-off), the silent pause frequency (PF) and fluency judgments of untrained raters. An important result was a significant correlation between perceived fluency ratings and

SR and PF, accounting for between 65% and 69% of the variance across tasks. Similar to Kormos and Dénes (2004), Derwing et al. (2004) also used untrained raters, with perceived fluency as the independent variable and temporal speech variables as the dependent variables, but obtained different results with regard to pause phenomena.

Rossiter (2009) reported further analyses examining utterance and perceived ESL fluency using native and non-native listeners with differing experience rating L2 speech. Temporal measures analyses consisted of SR (syllables per second, pruned) and MLR (.40 sec pause cut-off). Rossiter found that speakers with higher speech rate and lower number of pauses per second were consistently assigned higher fluency scores. Pausing, self-repetition, speech rate and fillers were reported as having a negative influence on perceived fluency, while pronunciation, grammar and vocabulary were observed as important non-temporal characteristics predicting fluency. Contrary to Kormos and Dénes (2004), pausing phenomena in Rossiter's study strongly influenced the three different groups of raters and accounted for almost half of the variations in their scores.

Cucchiarini, Strik and Boves (2000) found strong associations between perceived fluency in read speech and SR (syllables per second, unpruned), MLR (.20 sec pause cut-off), PTR, PF and APT in L2 Dutch. The authors reported that raters found pauses acceptable when accompanied by sufficiently long uninterrupted stretches of speech and suggest that PF is more relevant than pause length in judging L2 fluency. In their later investigation examining both read and spontaneous speech in L2 Dutch, Cucchiarini et al. (2002) observed that speech rate and PTR were important measures of fluency for beginners, whereas MLR was more indicative of fluency in intermediate learners.

In a recent study of L2 Dutch, Bosker et al., (2013) further investigated the link between utterance fluency measures and perceived fluency also providing a detailed analysis of multicollinearity. They examined rater sensitivity to breakdown fluency (number of silent pauses (.25 sec pause cut-off), number of filled pauses and mean length of pauses), speed fluency (spoken time /number of syllables), and repair fluency (number of repetitions and number of corrections). Experiment 1 showed that pause and speed measures were good predictors of fluency in the judgment of untrained raters, while repairs were not. Next, they investigated perceptual sensitivity using three new groups of untrained raters to evaluate learners' use of silent and filled pauses (Experiment 2), speed of delivery (Experiment 3) and repetitions and repairs (Experiment 4). The authors concluded that raters were sensitive to all aspects of breakdown, speed, and repair fluency when judging speech samples.

Fluency in L2 French has also been studied using temporal variables and from various different learning conditions, including study abroad, immersion contexts and formal classroom settings. In a seminal longitudinal study, Towell et al., (1996) set out to determine how the conversion of controlled to automatic processing took place and how it impacted L2 fluency in French measured at two different times. Fluency was operationalized as SR (syllables per second, unpruned), AR, and MLR (0.28 sec pause cut-off) and PTR. The results showed increases in SR, AR and MLR between Time 1 and Time 2, but no change in PTR. They thus concluded that the significant increase in SR was due to longer runs, and not to a decrease in pausing, and that MLR was the most important temporal variable contributing to fluency. Moreover, qualitative changes were also noticed in the performance of two participants at different fluency levels. Longer runs, faster SR, less pausing and more use of fixed expressions, characterized the speech of the more fluent participant.

Later longitudinal research by Towell (2002) also examined the rate of fluency development using temporal variables and hesitation phenomena for French learners using a personal adventure and a story continuation task. The utterance fluency variables examined were SR (syllables per second, unpruned), PTR, MLR and ALP (average length of pauses). While all the participants increased their scores on all the temporal variables over time, except for ALP, he found differences between those who performed at a higher level, as tested by a pre-university examination and a cloze test at outset, and those who performed at a lower level. Although the fluency scores for low level performers increased the most, they did not compare to those of the high level performers whose fluency did develop considerably. From this research, a useful perspective on pausing, in French, emerged. He reported that lower level performers altered their speech by pausing less and at different junctures, especially pausing at syntactic boundaries rather than within them. This behaviour in turn increased both the PTR and MLR. He identified pausing modification as the reason why higher scores on temporal variables were achieved between the two levels of learners.

With the intention to investigate how L2 learning differed from one context to another, Freed et al., (2004) conducted longitudinal research with students studying French in study abroad, immersion and at home settings. The oral performances were examined by two sets of analyses: 1) general measures which consisted of total words spoken, duration of speaking time, number of words (length) and 2) oral fluency based on a composite of speech rate (words per minute), hesit-free (mean length of run with no silent pauses of .40 sec or more), filler-free (mean run length with no filled pause dysfluencies), fluent-run (number of words as the longest run containing no dsyfluencies), repeat-free (mean run length spoken without repetitions), and repair-free (mean run length spoken without grammatical repairs). They reported that the

immersion group showed significant gains in the total number of words spoken, length of longest turn, speech rate, speech fluidity (derived from a composite of six fluidity measures) and used fewer silent and filled pauses over the other two groups. For the study abroad group, only gains in oral fluency were observed when compared with the at home group. No gains were reported for the learners in the formal classroom at home setting.

Although several investigations have employed temporal variables to evaluate L2 French fluency (Freed, 2000; Freed et al., 2004; Raupach, 1987; Towell, 2002; Towell et al., 1996), only Préfontaine (2013a) included utterance fluency and perceived fluency judgments of both native speaker (NS) and L2 speaker raters using three different tasks varying in cognitive demand. This perceived fluency data was then correlated with the utterance measures of SR (syllables per second, unpruned), AR, MLR (.25 sec pause cut-off), PTR, PF and APT. The analyses revealed fairly homogenous characterizations of French fluency skill existed between both L2 speakers and NS raters. Moreover, participants' self-perceptions of fluency showed a clear link to MLR and APT.

Given the empirical research in L2 fluency English, Dutch and French has produced different key findings by using a variety of operationalization measures and speech tasks, a compilation of the studies reviewed in the literature review is shown in Appendix A. The appendix shows relevant information about the study including the participants, L2, proficiency level, learning context, speech task, utterance fluency measures and findings.

As the overview of the literature shows, studies investigating L2 fluency in French are scarce, and research conducted on raters' perceptions of fluency and their relationship to utterance fluency using objective measures is negligible. Investigating how ratings of fluency are related to listeners' perceptions in various languages is important as there might be considerable

variation across languages in the acoustic features of speech that contribute to judgments of fluency. This is all the more important because there are important cross-linguistic differences between temporal variables of speech across languages (see e.g. Campione & Véronis, 2002 ; De Jong, Groenhout, Schoonen, & Hulstijn, 2013; Grosjean, 1980; Grosjean & Deschamps, 1975; Riazantseva, 2001) compared utterance fluency variables in English and French native speakers. In their study, they found English L1 speakers pause more often than French speakers, but the pauses were briefer in English than in French. Grosjean explains it thus:

> ...the pause time ratio in the two languages is almost identical … but that this equal pause time is organized differently in the two languages: there are fewer but longer pauses in French whereas in English pauses are more numerous but shorter. (p. 307)

Method

The present study analyzed the interrelationship between raters' judgments and utterance fluency measures based on three oral narrative tasks undertaken with learners of L2 French.

Participants

To perform three speaking tasks, 40 L2 speakers with varying levels of French proficiency were recruited in beginning, intermediate and advanced level classes from a 5-week immersion program at a university in Québec, Canada. All 40 participants were volunteer undergraduate and graduate students and native speakers of Canadian, American and British English who were enrolled in a wide range of academic fields. The speech data was collected from 21 women and 19 men ranging in age from 18 to 69 (M = 26 years, $SD = 10.57$). Of the sample group, 10 Canadian participants had spent an average of nine years in a French immersion setting in an English-speaking environment in Canada, while the remainder reported an average of six years of French study in a regular classroom.

Additionally, 11 French native speakers (8 women, 3 men), instructors of L2 French in several different immersion programs in Québec, were selected to rate the L2 speech performances according to quantitative and qualitative speech features. We deliberately chose French language instructors as they, rather than random native speakers, are most often involved in assessing learner speech in pedagogical and testing contexts and are likely to provide more consistent and accurate judgments of fluency. The rationale behind our choice was that teachers are the ones who most frequently assess their students' fluency either in the form of continuous assessment to inform further pedagogical intervention or as summative assessment at the end of a language course or module. No training was provided to avoid that the authors' interpretations of fluency influence the raters and to reflect the situation in classrooms in this context where teachers rarely receive training in fluency assessment before having to evaluate their learners' performance.

Instruments

Speaking tasks. Hypothesized to vary in impact on utterance and perceived fluency, three narrative speech tasks were operationalized according to their different task conditions, level of difficulty in processing performance and demand on stage of speech production. In Task 1, an unrelated picture narration, participants told a story based on six random pictures. Unlike the other tasks, this one entailed a more creative performance as no storyline or context was provided. In Task 2, a story retell, L2 speakers were asked to retell a story about a horseback riding accident from a short text in English. Participants were informed that the goal of the task was not to test their translation skill, but rather to retell the story as if they had read about it or experienced it. Although an uncommon task in second language research studies, the actual real life event of reading a story in one's L1 and relaying it in an L2, is not. In Task 3, an 11-frame

cartoon strip, participants narrated a story according to the sequence of events presented in the pictures. In this task, they were expected to connect the pictures in consecutive order to create a main storyline (for a detailed description of the particular cognitive characteristics of these tasks please see Préfontaine & Kormos, 2015).

L2 Fluency Assessment Grid. To gauge the general assessment of a participant's fluency in French, raters indicated their perceptions on the grid after listening to each of the three speech performances (See Appendix B). The grid consisted of six quantitative and qualitative can-do statements from the Council of Europe's (2001)Common European Framework of Reference (CEFR, see Table 3, pp. 28-29). Ranging from the lowest level (A1) to the highest (C2), the raters selected one descriptor to represent their assessment of oral performance in French, in other words the CEFR scale ranging from A1 to C2 was converted into a six-point numerical scale.

Fluency Perception Semantic Scale. To investigate more specific perceptions of fluency in French, 11 raters completed the Fluency Perception Semantic Scale (see Appendix C). After listening to each task performance, raters indicated their perceptions of fluency on the scale by marking a continuum ranging between two opposing extremes. The scale, which was specifically developed for the purposes of this study, included two items pertaining to qualitative and quantitative speech features: pauses and speed. These characteristics and descriptors were selected based on two important components of fluency identified in previous studies: *break-down fluency,* which is related to pausing behavior, and *speed fluency*, which expresses the speed with which speech is delivered (for a recent discussion see Bosker et al., (2013).

Procedures

Perceived Fluency. First, the 40 participants responded to the three narrative speaking tasks, for each of which they were allotted three minutes of planning time. The L2 speakers were told they could speak for as long as they wished to complete the task, which generated speech productions between three and four minutes. The tasks were administered to the participants in a counter-balanced design to control for task order effects. Second, 11 raters listened to each speech performance and indicated their overall impressions of fluency in French using the grid and scale described in the preceding section. Contrary to previous research in which fluency judgments were based on short excerpts of speech between 20-30 seconds, in this study the raters were instructed to listen to the entire oral performance as is the norm in real-life speech perception between interlocutors and in testing contexts. The raters listened to the speech samples per task, with a few days/weeks interval between. Third, the participants' 3 speech samples were analyzed according to four utterance fluency temporal variables by use of Praat (see below).

Utterance Fluency. Utterance fluency was first analyzed by calculating AR, MLR, PF and APT using Praat (Boersma & Weenink, 2010) and a software script (De Jong & Wempe, 2009), which was modified by the authors of this study to automatically extract the aforementioned temporal measures. As the recordings were of high quality and contained no background noise, no filters had to be applied. The output of the automated analysis was checked for accuracy and unexpected outlying values. The four utterance fluency variables were chosen because previous research showed they are salient predictors of fluency in French. AR was selected as a measure of speed fluency, while PF and APT were intended to assess breakdown fluency. The mean length of run variable combines both speed and breakdown features (see Bosker et al., (2013). As it is a combined measure, it has been shown to be one of the strongest

predictors of fluency in previous studies in both French and English, and therefore we felt it important to include it for the sake of comparability with previous research. These variables were operationalized as follows:

1. Articulation rate (AR): The total number of number of syllables divided by the total phonation time (excluding pauses) expressed in seconds. Following Riggenbach (1991), the articulation rate was unpruned with all partial words and asides counted. Praat was configured to detect pauses of 0.25 seconds and above.[i]
2. Mean length of runs (MLR): The total number of syllables divided by the number of utterances between pauses of 0.25 seconds and above.
3. Pause frequency (PF): The total number of pauses divided by the total duration in seconds of the speech sample. Only pauses of 0.25 seconds and above were used in the calculations.
4. Average pause time (APT): The total duration of all pauses (of 0.25 seconds and above) divided by the number of pauses in a given speech sample.

The statistical analyses included computing descriptive statistics, calculating correlations among utterance fluency variables and performing mixed-effects modeling.

Results

First, we calculated the descriptive statistics for the 11 native speaker ratings of L2 perceived fluency (see Table 1).

Table 1

Descriptive Statistics for CEFR Descriptors and Ratings of L2 Fluency Perception Variables (n=40)

Perception Variable	Task 1 *M* (*SD*)	Task 2 *M* (*SD*)	Task 3 *M* (*SD*)
CEFR Rating	3.24 (1.40)	3.31 (1.38)	3.88 (1.30)
Pauses	3.28 (1.25)	3.23 (1.25)	3.68 (1.23)
Speed	3.66 (1.22)	3.68 (1.70)	4.06 (1.19)

Second, we computed descriptive statistics of the utterance fluency measures elicited by the three speech elicitation tasks (See Table 2). Shapiro-Wilk normality tests showed the utterance fluency measures and the rating data were normally distributed. Based on these results, we decided that parametric statistical procedures would be used in further analyses.

Table 2

Descriptive Statistics for L2 Utterance Fluency Measures (n=40)

Utterance Fluency Measure	Task 1	Task 2	Task 3
	M (*SD*)	*M* (*SD*)	M (SD)
AR	7.69 (2.79)	7.48 (3.05)	8.59 (3.43)
MLR	5.94 (1.38)	5.96 (1.56)	6.33 (1.54)
PF	0.45 (0.05)	0.44 (0.07)	0.44 (0.06)
APT	1.40 (0.32)	1.42 (0.34)	1.49 (0.33)

(AR = articulation rate, MLR = mean length of runs, PF = pause frequency, APT = average pause time)

Third, intraclass correlation coefficients were computed to measure the degree of rater consistency and absolute agreement: ICC(3, 1) and ICC(2, 1) in the terminology of Shrout and Fleiss (1979), and ICC(C, 1) and ICC(A, 1) in the classification of McGraw and Wong (1996). The consistency values, which ignore systematic differences between raters, are somewhat higher than the absolute agreement values. The much higher values for Cronbach's alpha – ICC(3, k) or ICC(C, k) – measure the consistency of raters' total scores (for all participants), and are included for comparison purposes (See Table 3).

The ICC values in Table 3, calculated individually for each perception variable and task, suggest reasonable agreement and consistency. Task 2 (the story retell) was rated most reliably, though as we saw below, Task 3 (the cartoon narration) was judged most fluent, and Task 1 (the unrelated picture narration) was judged least fluent, overall.

Table 3

Intraclass Correlation Coefficients (Agreement / Consistency / Cronbach's alpha) Across Tasks

Perception Variable	Task 1	Task 2	Task 3
CEFR Rating	0.49 / 0.53 / 0.93	0.58 / 0.62 / 0.95	0.51 / 0.55 / 0.93
Pauses	0.38 / 0.48 / 0.91	0.48 / 0.59 / 0.94	0.46 / 0.56 / 0.93
Speed	0.41 / 0.48 / 0.91	0.52 / 0.59 / 0.94	0.42 / 0.52 / 0.92

Finally, correlational analyses were conducted, which revealed a high degree of intercorrelation between the measures of utterance fluency (see Table 4). PF, which was included as a measure of breakdown fluency, showed only a weak relationship with AR in Task 2 and 3, and a moderately strong correlation in Task 1. All the other utterance fluency variables were strongly intercorrelated with APT and MLR demonstrating correlations as high as .90. The magnitude of these intercorrelations is not unexpected (see e.g., Derwing et al. (2009) and French Segalowitz & Guay (under review). AR, APT and MLR can all be seen as variables measuring speed fluency and can be assumed to constitute the same underlying construct of speed fluency (Bosker et al., 2013). Pause frequency is a measure of breakdown fluency and as such it draws upon a different aspect of fluency, an assumption which was also ascertained by factor analysis. A principal component analysis of the fluency variables confirmed this hypothesis as it showed that AR, APT and MLR formed one factor (Eigenvalue= 4.64) and PF another one (Eigenvalue= 1.07). It is surprising to find, however, that the correlation between APT and PF is negative indicating that speakers who pause less frequently make longer pauses.

Table 4

Intercorrelations between Utterance Fluency Measures

Utterance fluency measure	AR	MLR	PF	APT
Task 1				
AR	1	.917**	−.504**	.881**
MLR		1	−.742**	.927**
PF			1	−.806**
APT				1
Task 2				
AR	1	.816**	−.233	.829**
MLR		1	−.681**	.933**
PF			1	−.686**
APT				1
Task 3				
AR	1	.876**	−.378*	.822**
MLR		1	−.696**	.909**
PF			1	−.778**
APT				1

* Indicates $p<0.05$.
** Indicates $p<0.01$.

In a multiple regression analysis, strong intercorrelation of independent variables – known as multicollinearity – can lead to regression coefficients and partial R-squared values that are unstable and even arbitrary. In our study, we compared simple regression models, each of which contained only one of the four independent variables measuring utterance fluency. This

allows us to see which of the four had the greatest (and least) effect on each of the three perceived fluency variables. We can thus determine which utterance variable is perceived as the best (and worst) reflection of different types of perceived fluency.

It needs to be noted, however, that because the four utterance fluency measures are highly intercorrelated, any determination of their relative importance is difficult. Another approach would be to evaluate the effect of each variable while controlling for the other three (the unique variance explained). However, a variable could score low according to this method while still being better than the others on its own. For this reason, we have opted to use the simpler type of comparison mentioned above.

We fit models to the combined data from the three speech production tasks, including an interaction between task and the utterance fluency measure. These models yielded substantial results for the effects of utterance fluency, which were fairly consistent across tasks. In these models, the dependent variables (fluency perception ratings) were either CEFR rating, pauses rating, or speed rating. These were scored from 1 to 6, and were treated as linear. The independent variables (utterance fluency measures) were AR, APT, MLR, and PF. These were numeric variables, on different scales, and therefore they were standardized into z-scores.

A two-step procedure was used, following Bosker et al., (2013). In the first step, a mixed-effects linear regression model was fit in R (R Core Team, 2014), using the lme4 package (Bates, Maechler, & Bolker, 2013). This step used random effects to model a) the severity of each rater overall (a random intercept), and that of each rater on each task (a random slope). We then subtracted the random effect estimates (or BLUPs) from the fluency perception ratings. The resulting numbers can be seen as having been corrected for the differing severity of the raters, including regular between-rater differences across tasks.

The second step was an ordinary linear regression, with one of the three corrected perceived fluency ratings (CEFR, pauses, speed) as the dependent variable and one of the four utterance fluency measures (AR, APT, MLR, or PF) as the independent variable. Each of these twelve models estimates the effect of one independent variable on one dependent variable. The effect of each independent variable was assessed in two ways. First, the regression coefficients were compared directly: these represent the estimated increase in the perceived fluency rating for a one-standard-deviation increase in the utterance fluency measure (averaged across tasks). Second, we compared increases in R-squared, which represent the proportion of total variance accounted for by the independent variable (and its interaction with task). As seen in Table 5, these methods produced similar results regarding the relative importance of the four utterance fluency measures. Across the three fluency perception ratings, the effects emerged in the following order: MLR > APT > AR > PF.

Table 5

Effects of Utterance Fluency Measures on CEFR, Pause and Speed Ratings

	CEFR ratings		Pause ratings		Speed ratings	
	Coefficient*	R^2 increase**	Coefficient*	R^2 increase**	Coefficient*	R^2 increase**
AR	0.656	.225	0.546	.216	0.550	.229
MLR	0.770	.324	0.623	.293	0.608	.289
PF	−0.561	.175	−0.470	.166	−0.441	.154
APT	0.682	.259	0.563	.241	0.558	.247

*All coefficients were significantly different from 0, $p < .001$.
** R2 of model with utterance fluency measure * Task – R2 of model with only Task.

PF had the smallest effect on the dependent variables. The coefficient was always negative, between -0.441 and -0.561; more frequent pauses were associated with lower perceived fluency. MLR was the most important predictor, with coefficients between 0.608 and 0.770. Longer runs were associated with higher perceived fluency.

AR and APT were intermediate in their effects, showing less predictive power than MLR but more than PF. Interestingly, the coefficients for APT, which ranged from 0.558 to 0.682, were positive. This meant that speakers with longer average pause times were judged to be more fluent in French.

Turning to a comparison of the three dependent variables, these results show that CEFR rating is most strongly associated with the utterance fluency measures. Pause rating and speed

rating are less strongly associated with the acoustic variables, and are similar to each other in this respect.

Among the most notable task-related differences, some apply overall, and some are interactions with the utterance fluency variables. The models confirm what Table 2 already suggested: Task 1 (the unrelated picture narration) and Task 2 (the story retell) received lower perceived fluency ratings, while Task 3 (the cartoon narration) received higher ratings. Perhaps relatedly, the effects of utterance fluency on perceived fluency tended to go in the opposite direction: these effects were usually larger for Tasks 1 and 2 and considerably smaller for Task 3. However, the effect of pause frequency was different; it was larger (that is, more negative) for Task 3 and smaller (less negative) for Tasks 1 and 2 (see Appendix D for details).

More than the other three utterance fluency variables, articulation rate varied noticeably in importance depending on the task. For Task 1, AR was the most important independent variable, slightly ahead of MLR. For Task 2, AR came out in third place (as it did overall, see Table 5). For Task 3, AR was the least important variable; its coefficients (and R-squared values) were smaller in magnitude than those of PF, which was clearly the least important variable for the other tasks.

Discussion and conclusions

Our research questions enquired into how utterance fluency measures predict CEFR ratings (RQ1) and evaluations of pausing behavior (RQ2) and speed (RQ3). The modeling of the data has shown that raters' judgments of fluency in terms of the CEFR scale, speed and pausing are influenced in a rather similar way by the utterance fluency variables. The MLR is always one of the most important variables, while depending on the task, AR and APT are too. While this

finding is not unexpected given the prominence of descriptors related to speed, flow and efficiency of encoding in the CEFR scale, it was surprising that PF featured as the weakest predictor of ratings of pausing behavior in two out of the three tasks.

Our analyses have brought somewhat different outcomes from previous studies with regard to the role of PF in perceptions of fluency. The frequency of pauses was a significant predictor of fluency judgments, but its contribution to variance in scores was relatively smaller in comparison with other variables. Previous research has shown that the location of pauses and their distribution within and at clause boundaries might play a more important role in rater perceptions than their frequency (for a summary see Ejzenberg, 2000; Götz, 2013; Pawley & Syder, 2000; Riggenbach, 1991; Wennerstrom, 2001) and indeed the qualitative comments provided by three of the raters lend support to our assumptions (Préfontaine, 2013b; Préfontaine & Kormos, forthcoming). It also needs to be noted that earlier research on pause perception in studies where transcribers were asked to indicate the location of pauses in the text revealed that listeners were inaccurate in identifying pauses (Arlington, Brenninkmeyer, Arn, Grundhauser, & O'Connell, 1992). This can provide an additional explanation for the finding that the frequency of pauses might not be as reliable an indicator of perceptions of fluent performance in L2 French.

It is also useful to consider the results of our study concerning PF together with the findings on the role of APT. Interestingly, our models suggest that the longer the unfilled pauses, the more favorable raters' perceptions of participants' fluency are. This finding seems unique in the field of L2 fluency, as previous research has found that mean length of pauses was either not related to raters' perceptions in the case of L2 learners of Dutch (Cucchiarini et al., 2002) and English (Kormos & Dénes, 2004) or that its relationship to fluency ratings was negative (Bosker et al., 2013). The intercorrelations of PF and APT, together with earlier cross-linguistic research

on pausing in L1 French, can provide a partial explanation for these results. In all three tasks, PF demonstrated a strong negative relationship with APT. Grosjean and Deschamps' (1975) study yielded similar results as they found that L1 speakers of French tended to pause less frequently but for longer than L1 speakers of English. It can then be assumed that on the one hand, fluent L2 users of French approximate this pausing profile, and on the other hand, the raters awarded high scores to those learners whose pausing behavior mirrors that of the L1 French. It can also be assumed that longer silent pauses are used by speakers for content planning, whereas shorter pauses might be indicative of encoding breakdowns (see e.g.Götz, 2013). The qualitative data on justifications of fluency scores (Préfontaine & Kormos, forthcoming) suggests that raters could differentiate pauses used for message conceptualization from those which are indicative of linguistic encoding problems. The raters' sensitivity to the purpose of pauses might explain why longer silent pauses, if they occurred relatively infrequently, were associated with positive fluency perceptions.

The results of our study seem to confirm the importance of the MLR in perceptions of fluency found earlier in the case of learners of English as an L2 (e.g. Kormos & Dénes, 2004). In previous studies of L2 French, MLR was also a significant factor in influencing raters' perceptions (e.g. Towell, 2002). In addition, the MLR was also found to improve as a result of a study abroad program in Towell et al.'s (1996) study, and this change was assumed to be indicative of the development of automaticity in the participants' speech encoding mechanisms. The fact that in our research MLR was consistently one of the most important predictors of L2 fluency judgments also highlights that the automaticity and encoding efficiency is one of the most significant factors in how raters award scores on a fluency scale. Our findings indicate that the length of unbroken speech produced by L2 French speakers can be reliably perceived by

raters and forms an important basis for the evaluation of encoding speed and the fluency component of the CEFR scale.

Our analyses reveal that AR as a measure of the number of syllables per second excluding pause time is an additional important variable in predicting perceptions of fluency. According to Towell et al., "any increase in the AR can be taken as an indication of proceduralization within the articulator" (1996, p. 92), and therefore this variable also seems to function as a similar indicator of the efficiency of encoding processes as the MLR. De Jong, Steinel, Florijn, Schoonen and Hulstijn (2012) argue that AR is one of the best measures of speed fluency, yet in our study it was the strongest predictor of speed ratings only in Task 1, which required the participants to generate the content of their stories. This suggests that the raters might show different sensitivity to this utterance fluency measure in the different types of task.

The analyses also reveal subtle variations in the relative importance of the different utterance fluency variables across tasks. Speed and hesitation measures can be assumed to be sensitive to both the linguistic demands of tasks as well as the need to creatively generate the content of one's message (for more detail see Préfontaine & Kormos, 2015). The tasks used in our study varied with regard to whether the participants had to narrate a given story or conceptualize a new story using the picture cues. The intraclass correlations of the raters' judgments reveal that the lowest values in terms of the agreement and consistency of the CEFR, speed and pause ratings were obtained in Task 1, in which students had to create their own stories. This suggests that raters found it more difficult to evaluate fluency in the task where the content varied.

Our study suggests that it is important to consider what characterizes fluent performance in the norms of the target language community in designing rating scales, rater training and

automated methods of assessment. Our research also reveals that depending on the characteristics of the tasks, the relative importance of the utterance measures in predicting fluency ratings vary to some extent. This indicates that if automated measures are to be used in fluency assessment, either a combination of various utterance fluency variables should be applied in the statistical analyses or those measures should be selected that best reflect the fluency demands of the given task.

Notes

In this study, the cut-off point was set to 0.25 seconds or more, as consistently used by Goldman-Eisler (1968), Kormos & Dénes (2004) and Ginther et al. (2010) in ESL, Bosker et al., (2013) in L2 Dutch, and for studies of French speech production (Grosjean & Deschamps, 1972, 1973, 1975); Raupach, (1987). As explained in Towell et al. (1996) and Towell (2002), 0.28 seconds was used for purely practical reasons. In two subsequent French language learning studies (Freed, 2000; Freed et al., 2004), the calculation of speech rate was the mean number of words per minute without dysfluencies, rather than syllables per second. Given this methodological difference in the calculation of speech rate, the results of studies using cut-off points at 0.25 and 0.40 are therefore not comparable. Thus, using a 0.25 cut-off point in this research study is not an arbitrary decision but rather one based on empirical research, using the same speech rate measurement.

References

Arlington, J., Brenninkmeyer, S. M., Arn, D., Grundhauser, R., & O'Connell, D. (1992). A usual extreme case: Pause reports of informal spontaneous dialogue. *Bulletin of the Psychonomic Society, 30*, 161–163.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bates, D., Maechler, M., & Bolker, B. (2013). lme4.0: Linear mixed-effects models using S4 classes (Version R package 0.999999-4/r1876). Retrieved from http://R-Forge.R-project.org/projects/lme4/

Boersma, P., & Weenink, D. (2010). Praat: doing phonetics by computer (Version 5.0.25) [Computer software]. Retrieved http://www.praat.org/

Bosker, H., Pinget, A., Quené, H., Sanders, T., & De Jong, N. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing, 30*, 159–175.

Campione, E., & Véronis, J. (2002). *A large-scale multilingual study of silent pause duration*. Paper presented at the Speech Prosody Conference, Aix-en-Provence, France.

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

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners'' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America, 107*, 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.

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. (2013). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, 1–21.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics, 33*, 1-24.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*, 385–390.

Derwing, T., Munro, M., Thomson, R., & Rossiter, M. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition, 31*, 533–557.

Derwing, T., Rossiter, M., Munro, M., & Thomson, R. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54*, 655–679.

Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (Ed.), *Perspectives on Fluency* (pp. 287–313). Ann Arbor, MI: University of Michigan Press.

Foster, P., & Skehan, P. (1999). The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research, 3*, 215–247.

Freed, B. (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, MI: University of Michigan Press.

Freed, B., Segalowitz, N., & Dewey, D. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition, 26*, 275–301.

French, L., Segalowitz, N., & Guay, J. D. (under review). Short-term immersion and the development of adults' L2 utterance fluency.

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. (1968). *Psycholinguistic experiments in spontaneous speech*. London: Academic Press.

Grosjean, F. (1980). Comparative studies of temporal variables in spoken and sign languages: A short review. In H. W. Dechert & M. Raupach (Eds.), *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler* (pp. 307–312). The Hague: Mouton.

Grosjean, F., & Deschamps, A. (1972). Analyse des variables temporelles du français spontané. *Phonetica, 26*, 129–156.

Grosjean, F., & Deschamps, A. (1973). Analyse des variables temporelles du français spontané II. *Phonetica, 28*, 191–226.

Grosjean, F., & Deschamps, A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français: Vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica, 31*, 144–184.

Götz, S. (2013). *Fluency in native and nonnative English speech* (Vol. 53). Philadelphia/Amsterdam: John Benjamins Publishing.

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

Jackson, D. O., & Suethanapornkul, S. (2013). The Cognition Hypothesis: A synthesis and meta-analysis of research on second language task complexity. *Language Learning, 63*, 330–367.

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

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, MI: University of Michigan Press.

MacIntyre, P. D., & Gardner, R. C. (1994). The subtle effects of language anxiety on cognitive processing in the second language. *Language Learning, 44*, 283–305.

McGraw, K., & Wong, S. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46.

Pawley, A., & Syder, F. (2000). The One-Clause-at-a-Time Hypothesis. In H. Riggenbach (Ed.), *Perspectives on Fluency* (pp. 163-199). Ann Arbor: University of Michigan Press.

Préfontaine, Y. (2013a). *Fluency in French: A psycholinguistic study of second language speech production and perception.* (Unpublished doctoral dissertation). Lancaster University. Lancaster, UK.

Préfontaine, Y. (2013b). Perceptions of French fluency in second language speech production. *Canadian Modern Language Review*, *69*, 324–348.

Préfontaine, Y., & Kormos, J. (forthcoming). A qualitative analysis of perceptions of fluency in second language French. *International Review of Applied Linguistics*

Préfontaine, Y., & Kormos, J. (2015). The relationship between task difficulty and second language fluency in French: A mixed methods approach. *Modern Language Journal,*

R Core Team. (2014). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Raupach, M. (1987). Procedural learning in advanced learners of a foreign language. In J. A. Coleman & R. Towell (Eds.), *The advanced language learner* (pp. 123–155). London: CILT.

Riazantseva, A. (2001). Second language proficiency and pausing: A study of Russian speakers of English. *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 non-native speakers of English. *Canadian Modern Language Review, 65*, 395–412.

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

Towell, R. (2002). Relative degrees of fluency: A comparative case study of advanced learners of French. *International Review of Applied Linguistics in Language Teaching, 40*, 117–150.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*, 84–119.

Wennerstrom, A. (2001). *The music of everyday speech: Prosody and discourse analysis*. Oxford: Oxford University Press.

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence and classroom applications*. London/New York: Continuum.