What Features Best Characterize Adult Second Language Utterance Fluency and What Do They Reveal About Fluency Gains in Short-Term Immersion?

Norman Segalowitz
Concordia University
Queensland University of Technology (QUT)

Leif French
Sam Houston State University

Jean-Daniel Guay
Université Laval

Abstract

This study reports on how one can examine a second language (L2) speech corpus in order to define which of many possible features of L2 utterance fluency (i.e., speech fluidity) should be the focus of an L2 fluency gains investigation. Participants were 100 adult English-speakers enrolled in a French immersion program. Data from 50 randomly selected participants were assigned to Sample A for Analysis 1 and the remainder to Sample B for Analysis 2. In Analysis 1, 23 candidate speech features, drawn from the literature at large, were examined in Sample A through a series of logical and statistical steps and systematically reduced to four features as constituting a core set of L2 utterance fluency features. In Analysis 2, these four features were examined in the Sample B corpus for gains after 5 weeks of immersion. Results indicated strong gains on all four. In Analysis 3, by way of replication, we reversed the process by using the Sample B data to first define the target fluency features and then the Sample A data to test for fluency gains. The main results replicated those of Analyses 1 and 2. The four features that emerged as core L2 utterance fluency features were mean syllable run length and mean phonation run length between silent pauses, and mean syllable duration and mean silent pause duration. Mean filled pause duration did not meet the criteria for belonging to the same fluency construct. Overall, the results showed that it is possible (a) to operationally define L2 fluency markers without reference to fluency gains, and (b) to then use these fluency markers to study L2 fluency gains without the gains data having shaped the operational definition of fluency in the first place, thereby avoiding the circularity of post hoc identification of relevant variables.

Résumé

Cette étude rapporte une méthode qui peut être utilisée pour examiner un corpus de parole en langue seconde dans le but de déterminer les variables parmi toutes celles présentées dans la littérature sur l'aisance à l'oral énonciative en langue seconde (c.-à-d. la fluidité de la parole) qui devraient être au centre des recherches portant sur le développement de l'aisance à l'oral en L2. Les participants étaient 100 adultes anglophones qui ont complété

un programme d'immersion française. Les données de 50 participants sélectionnés au hasard ont été assignées à l'Échantillon A et celles des autres 50 participants à l'Échantillon B. En lien avec la première analyse, 23 variables de parole candidates, tirées de la littérature sur le sujet, ont été examinées dans l'Échantillon A à travers une série d'analyses logiques et statistiques et ont été systématiquement réduites à 4 variables fondamentales pour représenter l'aisance à l'oral énonciative en langue seconde. En lien avec la deuxième analyse, ces 4 variables ont été examinées dans le corpus de l'Échantillon B pour observer les gains après 5 semaines d'immersion. Les résultats indiquent des gains robustes pour les 4 variables. En lien avec la troisième analyse, en utilisant la réplication, nous avons renversé le processus en sélectionnant les données de l'échantillon B pour déterminer les variables fondamentales représentant l'aisance à l'oral énonciative en langue seconde et celles de l'échantillon A pour observer les gains en aisance à l'oral. Les résultats principaux ont répliqué ceux des deux premières analyses. Les 4 variables qui ont émergé des analyses comme étant fondamentales sont la longueur moyenne de l'énoncé en syllabe, le temps moyen de phonation entre les pauses silencieuses, la durée moyenne de la syllabe et la durée moyenne de la pause silencieuse. La durée moyenne des pauses remplies n'a pas répondu aux critères pour appartenir au même construit d'aisance à l'oral. De façon général, les résultats indiquent qu'il est possible (a) de définir de façon opérationnelle les variables qui représentent l'aisance à l'oral en langue seconde sans référer aux gains, et (b) d'utiliser ces variables pour étudier les gains en aisance à l'oral en langue seconde par la suite sans que ces derniers influencent la conception de la définition opérationnelle de l'aisance à l'oral au départ dans le but d'éviter la circularité des analyses post hoc pour identifier les variables pertinentes.

What Features Best Characterize Adult Second Language Utterance Fluency and What Do They Reveal About Fluency Gains in Short-Term Immersion?

Introduction

A challenging task for second language (L2) speech researchers is determining what features of oral proficiency best characterize fluency in adult L2 speech. Correctly identifying such features is clearly important for studying L2 development and assessing gains that can be attributed to particular language learning experiences or forms of teaching. A major challenge, however, in operationally defining L2 fluency is deciding which features to look at. Generally speaking, people use *fluency* to refer to a broad range of phenomena, including various aspects of speech delivery but also to refer to general language knowledge and proficiency in specific language skills such as public speaking, writing, reading, et cetera. We define fluency here more narrowly in terms of temporal and hesitation phenomena that characterize the *fluidity* of speech delivery. These phenomena are also known as features of "utterance fluency" (Segalowitz, 2010, 2016). This narrower definition does help to reduce, somewhat, the scope of what the term fluency might refer to, but even so, researchers still face the task of figuring out on which of many potential features to focus (e.g., see De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012; Kormos, 2006; Segalowitz, 2010).

Our goals in this study were to operationally define *fluency*, as speech fluidity, in a way that narrows down the potentially large pool of features to a smaller core set and to investigate gains in fluency as defined by this smaller core set. We attempt here to operationally identify core features of L2 fluency prior to studying fluency gains. By having such a core set of features in hand from the outset, one can then proceed to study L2 fluency gains in a way that avoids circularity due to after-the-fact selection of just those features that happen to show gains over time in a particular study or in a particular sample.

In this study, we conducted three sets of analyses based on data from a large pool of participants. In Analysis 1, we used data from learner sample A (half the learner pool, randomly selected) and we identified which speech features should be considered as potentially reflective of fluency, independently of (i.e., prior to) analyzing any developmental data. This analysis ensured that the developmental data did not shape the operational definition of fluency itself, thereby avoiding a potential source of circularity. In Analysis 2, we applied the results of Analysis 1 to fluency development in learner sample B (the other half of the original learner pool). Because this analysis involved a different set of learners, the study avoided the possibility of finding a spurious connection between the fluency measures chosen and fluency gains that might be true only of one particular set of individuals. In Analysis 3, we examined fluency gains in sample A, as a replication of the fluency gain analysis with sample B. Thus, in summary, we attempted to operationalize L2 speech fluency separately from—and prior to—analyzing the fluency gain data themselves, in contrast to simply looking for gains across a wide spectrum of speech features to see which features happen to yield gains and which do not, and we included a built-in opportunity to replicate the fluency gain aspect of the study.

We describe the study in terms of General Methodology (how the data were collected), Analysis 1 (operationally defining L2 fluency), Analysis 2 (investigating fluency gains), and Analysis 3 (replication of Analysis 2), concluding with a General Discussion.

General Methodology

Participants

One hundred participants provided the data analyzed in this study (*M* age = 23.90 [7.65] years, *range* = 18-54; 66 females, 34 males). All were first language (L1) speakers of English and L2 learners of French enrolled in an immersion program for that purpose. Of these, 50 were chosen randomly (Sample A) for purposes of Analysis 1 and the remainder (Sample B) for Analysis 2. Data from both samples were used for purposes of Analysis 3.

A group of 23 L1 speakers of French also participated in this study (*M* age = 25.78 [8.60] years, *range* = 19-54; 14 females, nine males).

Materials

The speech elicitation task used was the "Suitcase Story" (Derwing, Rossiter, Munro, & Thomson, 2004), an 8-frame cartoon story about a woman and a man accidentally bumping into each other, dropping their suitcases as a result, mistakenly picking up the wrong suitcases before going on their way, and then being surprised when they discovered having the wrong suitcases. This task has been used in a variety of different contexts to study oral production phenomena (e.g., Rossiter, Derwing, & Jones, 2008).

Procedure

The L2 learners completed a consent form; a language background and biographical information questionnaire; tests of French knowledge (grammar, vocabulary) and cognitive measures not analyzed here; and the "Suitcase Story" speech-elicitation task. Participants had about 1 minute to prepare their description of the story and up to 3 minutes to tell it while still viewing the pictures. A French-speaking researcher conducted all testing sessions. The L1 French speakers did the same tests as the L2 learners except for the test of French knowledge.

All participants were tested at the beginning of the program (Time 1) and the learners again at the end after 5 weeks (Time 2).

Analysis 1: Operationally Defining L2 Fluency

In this analysis, we focused on speech features that other researchers have used as potential markers of L2 speech disfluency, plus a few additional features that could reasonably be considered alternates for some of them. *Speech fluency*, as used here, refers to the fluidity of speech, not general speech proficiency that includes much more than fluidity, such as vocabulary size and breadth, knowledge of formulaic utterances, syntax knowledge, et cetera, although all these may be related to fluency in important ways. We identified 23 potential fluency features, a large number that immediately raises the question of whether all are really necessary and useful for operationally defining L2 speech fluency. On the one hand, it is possible, of course, that each does reveal something unique and important about L2 fluency, thereby justifying retaining a focus on all 23. On the other hand, some may be redundant, that is, mere transforms of others and so justifiably dropped

from consideration. Still others, while at first seeming to possibly reflect some interesting aspect of fluency, may turn out upon further examination not to be empirically interesting. Thus, the goal of this first analysis was to see whether the initially identified 23 potential features of L2 fluency could be reduced in a principled way to a smaller core set and whether this core set could be said to operationally define a general L2 speech fluency construct. Such a finding would not necessarily mean that each of the retained features reflected exactly and only the same thing about fluency as the other retained features, but it would support the idea that speech fluency itself is a coherent construct, even if future research were to reveal differences among the different features of speech fluency (e.g., some features may typically develop earlier or different cognitive processes may underlie mastery of particular features).

To reduce the list of 23 potential fluency markers to a more manageable number in a principled way, we proceeded as follows. First, we first identified those features that could be considered *basic* in the sense that, while not qualifying as core features, they are essential for defining other features that could be considered core; in other words, these would serve as basic building blocks for defining higher order features. Such basic features included, for example, total speech time duration (the time the speaker took to perform the speech elicitation task), the total number of syllables uttered, the amount of time a person actually spoke (phonation time) as opposed to remaining silent, and the amount of pausing (number of hesitations interrupting phonation). These features cannot themselves be used individually to operationally define speech fluency because fluency (fluidity) does not necessarily change just because a person produces a longer speech sample. However, these basic features can be used to define higher order features, such as syllable utterance rate by taking the ratio of the total number of syllables spoken to total speech time duration, where rate does reflect an aspect of speech fluidity.

After eliminating basic features from the initial set of 23, we continued to reduce the list by eliminating additional features on *logical* grounds, that is, features whose measures were just mathematical transforms of other measures (e.g., "syllables per second" vs. "milliseconds per syllable"). This procedure was responsible for the greatest reduction of the set.

Next, additional features were eliminated on *statistical* grounds, either because their measures correlated too well or too poorly with other measures. For example, one measure might be highly collinear with another and thus redundant insofar as the two may be alternative measures of the same thing. In this case, a choice would have to be made regarding which of the two collinear measures to keep and which to drop. It is also possible that a given measure might not correlate very well at all with any other measures. While this would not necessarily mean that the feature in question is inherently uninteresting or irrelevant to an understanding of L2 speech performance, it would mean that this feature is not related to the same aspect of fluidity as are the other features. Finally, having narrowed down the list of potential core features with these procedures, we then looked to see if the surviving set plausibly reflected a fluency construct.

We now present a brief background description of the 23 speech features that were considered. To facilitate matters for the reader, we have sequenced the items in the order they are discussed in the text that follows (see also Table 1).

1. *Total Duration (totDur)*—the total time to complete the speech elicitation task, including time speaking (*phonation*) plus all intervening silent and filled pauses.

See Ginther, Dimova, and Yang (2010) and Hilton (2009) for examples. A variant of this measure is pruned total duration, with self-repetitions, repairs, and other language words removed.

2. *Total Silent Pause Duration (totSilPauseDur)*—the sum of all silent pause times. For this feature, a minimum silence threshold had to be set for defining what counts as fluency-related silence as opposed to linguistically-related silence not reflective of fluency as such (e.g., silence normally occurring after stop consonants; Zellner, 1994). See also De Jong and Bosker (2013) and Segalowitz (2016) for more on this issue. The minimum threshold used here was 400 ms, based on past practice (Freed, Segalowitz, & Dewey, 2004; Ginther et al., 2010; Iwashita, Brown, McNamara, & O'Hagan, 2008; Lennon, 1990). De Jong and Bosker (2013) discussed this issue more fully.
3. *Total Filled Pause Time (totFilPauseDur)*—the sum of all filled pause times, pauses where speakers use *um, uh, er, euh, hmmm,* et cetera. See Bosker, Quené, Sanders, and De Jong (2014), Ginther et al. (2010), Iwashita et al. (2008), and Lennon (1990).
4. *Number of Silent Pauses (nSilPauses)*. See Cucchiarini, Strik, and Boves (2000); Derwing et al. (2004); Freed, (2000); Ginther et al. (2010); Segalowitz et al. (2004); and Trenchs-Parera (2009).
5. *Number of Filled Pauses (nFilPauses)*. See Ginther et al. (2010), Iwashita et al. (2008), and Lennon (1990).
6. *Pruned Number of Syllables (nSyl)*—number of syllables after removing (pruning) self-repetitions, repairs, and other language words. See Ginther et al. (2010).
7. *Phonation Time (PhonTime)*—total time actually speaking. See Ginther et al. (2010).
8. *Speech Time Ratio (SpTimeRatio)*—phonation time as a proportion of total duration. See Cucchiarini, Strik, and Boves (2002); Ginther et al. (2010); Kormos and Dénes (2004); Mora and Valls-Ferrer (2012); and Towell, Hawkins, and Bazergui (1996).
9. *Silent Pause Ratio (SilPauseRatio)*—silent pause time as a proportion of total duration. See Ginther et al. (2010).
10. *Filled Pause Ratio (FilPauseRatio)*—duration of filled pauses as a proportion of total duration. See Ginther et al. (2010).
11. *Pruned Speech Rate (SpRate)*—the number of pruned syllables per minute of total duration. See Derwing et al. (2004), Freed (2000), García-Amaya (2009), Iwashita et al. (2008), Lennon (1990), and Llanes and Muñoz (2009).
12. *Pruned Articulation Rate* or *Syllables Per Phonation Minute (SylPerPhonMin)*—similar to *SpRate* but based on phonation time. See Cucchiarini et al. (2000), Ginther et al. (2010), Kormos and Denes (2004), Llanes and Muñoz (2009), Mora and Valls-Ferrer (2012), and Towell et al. (1996).
13. *Silent Pause Rate (SilPauseRate)*— silent pauses per minute of the total duration.
14. *Silent Pause Rate (phonation-based) [SilPauseRatePerPhonMin]*—number of silent pauses per minute of speech, similar to *SilPauseRate* [13] but uses phonation time as the time divisor instead of total duration.
15. *Silent Pause Rate per 100 Syllables (SilPauseRatePer100Syl)*— number of pauses per 100 syllables spoken.
16. *Filled Pause Rate (FilPauseRate)*—rate of occurrence of silent pauses over the total duration. See Iwashita et al. (2008) and Kormos and Denes (2004).

17. *Syllable Run All Pauses (SylRunAllPauses)*—run length or mean number of syllables spoken before all pause interruptions, filled or silent. See Cucchiarini et al. (2002); Derwing et al. (2004); Garcia-Amaya (2009); Ginther et al. (2010); Kormos and Denes (2004); Lennon (1990); Mora and Valls-Ferrer (2012); O'Brien, Segalowitz, Freed, and Collentine (2007); Segalowitz and Freed (2004); and Towell et al. (1996).
18. *Phonation Run All Pauses (PhonRunAllPauses)*—phonation duration divided by number of silent & filled pauses, a phonation-based counterpart to *SylRunAllPauses* [17].
19. *Syllable Run (SylRun)*—similar to *SylRunAllPauses* [17] but based on silent pauses only.
20. *Phonation Run (PhonRun)*— similar to *PhonRunAllPauses* [17] but calculated with reference to silent pauses only.
21. *Pruned Syllable Duration (SylDur)*—syllable duration (the inverse of *SylPerPhonMin*).
22. *Mean Silent Pause Duration (SilPauseDur)*—mean length of silent pauses = total silent pause duration divided by number of silent pauses. See De Jong, Schoonen, & Hulstijn (2009), Ginther et al. (2010), Hilton (2009), and Kormos and Denes (2004).
23. *Mean Filled Pause Duration (FilPauseDur)*—mean filled pause duration = total filled pause duration divided by number of filled pauses (Ginther et al., 2010; Hilton, 2009).

Table 1
*The 23 Speech Features and Their Inclusion (✔) or Exclusion (**X**) Status as Core Fluency Features*

	Abbreviation	Feature	Operational Definition	Status[a]
1	*totDur*	Total Duration	Total duration (includes pausing, etc.). A variant is pruned total duration which excludes repairs, self-repetition, other language words	**X** L
2	*totSilPauseDur*	Total Silent Pause Duration	Total time of silent pauses	**X** L
3	*totFilPauseDur*	Total Filled Pause Duration	Total time of filled pauses	**X** L
4	*nSilPauses*	Number of Silent Pauses	Number of silent pauses ≥ 400 ms	**X** L
5	*nFilPauses*	Number of Filled Pauses	Number of filled pauses	**X** L
6	*nSyl*	Pruned Number of Syllables	Excludes repairs, self-repetition, other language words	**X** L
7	*PhonTime*	Phonation Time (seconds)	= *totDur* – *totSilPauseDur* – *totFilPauseDur*	**X** L
8	*SpTimeRatio*	Speech Time Ratio	= *PhonTime* / *totDur*	**X** L
9	*SilPauseRatio*	Silent Pause Ratio	= *SilPauseDur* / *totDur*	**X** L
10	*FilPauseRatio*	Filled Pause Ratio	= *FilPauseDur* / *totDur*	**X** L
11	*SpRate*	Pruned Speech Rate	= *nSyl* / *totDur* (using pruned data)	**X** L
12	*SylPerPhonMin*	Pruned Articulation Rate (syllables per phonation minute)	= 60 * *nSyl* / *PhonTime*	**X** L
13	*SilPauseRate*	Silent Pause Rate (per minute of total duration)	= 60 * *nSilPauses* / *totDur*	**X** L
14	*SilPauseRatePer PhonMin*	Silent Pause Rate (per phonation minute)	= 60 * *nSilPauses* / *PhonTime*	**X** L
15	*SilPauseRatePer100Syl*	Silent Pause Rate (per 100 syllables)	= 100 * *nSilPauses* / *nSyl*	**X** L
16	*FilPauseRate*	Filled Pause Rate (phonation based)	= 60 * *nFilPauses*/*PhonTime*	**X** S
17	*SylRunAllPauses*	Syllable run length = number of syllables between all pauses	= *nSyl* / (*nSilPauses* + *nFilPauses*)	**X** L/S
18	*PhonRunAllPauses*	Seconds of phonation between silent and filled pauses	= *PhonTime* / (*nSilPauses* + *nFilPauses*)	**X** S
19	*SylRun*	Syllable run length = number of syllables between silent pauses	= *nSyl* / (*nSilPauses*)	✔
20	*PhonRun*	Seconds of phonation between silent pauses	= *PhonTime* / (*nSilPauses*)	✔
21	*SylDur*	Pruned Articulated Syllable Duration (ms)	= 60,000 / *SylPerPhonMin*	✔
22	*SilPauseDur*	Mean Silent Pause Duration	= *totSilPauseDur* / *nSilPauses*	✔
23	*FilPauseDur*	Mean Filled Pause Duration	= *totFilPauseDur* / *nFilPauses*	**X** S

[a]Included (✔) or excluded (X) as a core fluency feature based on logical (L) or statistical (S) grounds.

Method

Participants. Fifty randomly selected participants from the original 100 learners, here designated as Sample A, provided the data analyzed in this phase of the study (*M* age = 25.24 [9.42] years, *range* = 18-54; 38 females, 12 males). As described earlier, all were L1 speakers of English and L2 learners of French enrolled in an immersion program. Twenty-three L1 French speakers provided the native speaker data for this study (*M* age = 25.78 [8.60] years, *range* = 19-54; 14 females, nine males).

Materials and procedure. These were as described earlier under General Methodology.

Results

All the speech data were manually segmented by hand, with the aid of Praat (Boersma & Weenink, 2007) to visualize the speech waveforms, and segment durations were calculated using an automated script (Kawahara, 2010). For the purposes of this study, we obtained measures of the 23 speech features listed earlier (Table 1). These measures were subjected to two kinds of analyses. The first examined the logical status of each feature to see if it should be eliminated without additional empirical consideration (e.g., if it is simply a mathematical transform of some other feature). The second analysis involved a statistical examination of the suitability of each remaining feature as a potential core utterance fluency feature.

Logical analyses. The logical analyses focused on fluency as a reflection of the flow of speech—its fluidity—as opposed to other aspects of oral performance sometimes included in studies of L2 fluency (e.g., amount of speech). These logical analyses led us to eliminate 15 of the 23 measures as inappropriate for retention as core fluency features, as follows.

The first measure considered was Total Duration (*totDur* [1] in Table 1). This measure is basic to the study of fluency in that it is required for computing values of other "higher order" measures (e.g., speech rate). However, it reflects how long the person spoke for and is therefore confounded with how talkative speakers are and/or knowledgeable about the topic at hand—factors not directly associated with L2 fluency as such. Moreover, total duration does not reflect the flow or fluidity of speech. For these reasons, Total Duration was deemed not appropriate to retain as a core feature of L2 utterance fluency.

The following six measures, though also required to compute higher order measures and therefore also basic measures, do not qualify as fluency features because they are all confounded with Total Duration [1]: total silent pause duration (*totSilPauseDur* [2]), total filled pause duration (*totFilPauseDur* [3]), number of silent pauses (*nSilPauses* [4]), number of filled pauses (*nSilPauses* [5]), number of syllables (*nSyl* [6]), and phonation time (*PhonTime* [7]).

Another set of measures, reflective of L2 speech proficiency (general ability) but not specifically of speech fluidity, are speech/time ratio (*SpTimeRatio* [8]), silent pause ratio (*SilPauseRatio* [9]), and filled pause ratio (*FilPauseRatio* [10]). All these provide different ways of measuring the volume of speech produced, corrected for total speaking time, without directly reflecting the actual fluidity or flow of oral production as such.

SpTimeRatio focuses on speech/phonation where *SilPauseRatio* and *FilPauseRatio* focus on its complement, the nonspeech/nonphonation aspects of production.

Next we considered several measures of speech rate, all of which do reflect speech fluidity. Pruned speech rate (*SpRate* [11]) and syllables per phonation minute (*SylPerPhonMin* [12]) are very similar to each other. Both involve pruned syllables but the former uses total duration (i.e., including silences) as the time divisor whereas the latter uses phonation time only. Phonation time more accurately reflects time spent speaking than does total duration and so we dropped speech rate [11] in favour of *SylPerPhonMin* [12]. *SylPerPhonMin* [12], however, is the logical inverse of pruned syllable duration (*SylDur* [21], discussed below), and so is redundant with it. Given this analysis, we decided to drop *SylPerPhonMin* [12] and provisionally kept *SylDur* [21] as a potential core feature.

There are several measures of silent pause production that can also be eliminated on logical grounds. Rate of silent pause production (*SilPauseRate* [13]) confounds frequency of silent pause production with silent pause duration because the time divisor is total duration, which itself is affected by silent pause duration. Silent pause rate per phonation minute (*SilPauseRatePerPhonMin* [14]), however, avoids this confound because it is based on phonation time instead of total duration. But, it is logically the inverse of seconds of phonation between silent pauses (*PhonRun* [20]), which reflects the length of phonation "bursts" between silent pause interruptions. We therefore dropped *SilPauseRatePerPhonMin* [14] and provisionally kept *PhonRun* [20] (discussed below). Similarly, *SilPauseRatePer100Syl* [15]) is the logical inverse of number of syllables between silent pauses (*SylRun* [19]) and so it too was dropped and we provisionally kept *SylRun* [19].

Statistical analyses. At this point, we had eliminated features [1] to [15] in Table 1 on logical grounds, leaving eight features provisionally retained as potential core fluency features: syllable duration (*SylDur* [21]), syllable run length (*SylRunAllPauses* [17]), number of syllables between silent pauses (*SylRun* [19]), seconds of phonation between silent and filled pauses (*PhonRunAllPauses* [18]), seconds of phonation between silent pauses (*PhonRun* [20]), mean silent pause duration (*SilPauseDur* [22]), filled pause rate per phonation minute (*FilPauseRate* [16]), and mean filled pause duration (*FilPauseDur* [23]). In the next step, we proceeded to statistically examine these eight items, primarily by looking for those that were either very highly collinear (Spearman correlation $r_s \geq |.90|$) or hardly related at all to the others ($r_s \geq |.30|$ on fewer than 30% of the items).

The statistical analyses revealed that number of filled pauses per minute (*FilPauseRate* [16]) failed to correlate with $r_s \geq |.30|$ with any of the other seven measures. For this reason, this feature was dropped because it was not sufficiently related to other fluency measures to justify considering it part of the same construct.

The analyses also revealed that number of syllables between silent and filled pauses (*SylRunAllPauses* [17]) was collinear with number of syllables between silent pauses (*SylRun* [19]) ($r_s = .96$, $p < .001$, [.93 .98]), requiring that one of the measures be retained and the other dropped. Because *SylRunAllPauses* is based on syllable runs between all pauses, including filled pauses, we decided to drop that measure and retain *SylRun* [19], which involves only silent pauses. Filled pauses can sometimes reflect more than simple disfluency disruptions in the speech flow (Clark & Fox Tree, 2002) by expressing communicative intent (e.g., signaling intent to continue speaking). Filled and silent pauses may thus differ from each other in communicatively important ways and may not,

therefore, be simply different forms of the same thing (speech flow interruptions). A similar issue arises regarding phonation between silent and filled pauses (*PhonRunAllPauses* [18]) and phonation between silent pauses only (*PhonRun* [20]). In this case, these two were not as strongly collinear (r_s = .56, p < .001, [.33 .73]) as were the two-syllable run measures. However, on logical grounds, the two features are conceptually highly similar and so ideally one of the two should be dropped. The only difference between them is that *PhonRunAllPauses* has the basic feature *nFilPauses* [5] as an underlying component whereas *PhonRun* does not. Correlation analyses (Spearman) revealed that this basic feature did not correlate significantly with any of the four other fluency measures retained so far (*SylRun* [19], *PhonRun* [20], *SylDur* [21], *SilPauseDur* [22], and *FilPauseDur* [23]; all $r_s \leq .27$; all $p > .05$; all lower 95% CI bounds ≤ -.31; all upper 95% CI bounds ≥ 0). For this reason, *PhonRunAllPauses* was regarded as being less related to fluency in the way the other retained measures were and so was dropped, and *PhonRun* [20] was retained.

At this point, there remained only five potential fluency features out of the original set of 23: mean number of syllables between silent pauses [*SylRun* [19]), mean phonation duration between silent pauses [*PhonRun* [20]), mean pruned articulated syllable duration (*SylDur* [21]), mean silent pause duration (*SilPauseDur* [22]), and mean filled pause duration (*FilPauseDur* [23]). The final step in this round of analyses was to determine to what extent these five could be considered to reflect a fluency construct.

On the surface, it would seem that the five retained measures do merit being viewed as reflecting a common underlying fluency construct. The set includes two measures of speech run lengths between silent pauses and three measures of element duration (syllable, silent pauses, and filled pause durations). The run features *SylRun* [19] and *PhonRun* [20] reflect the size of speech bursts between disfluent interruptions, and the duration measures *SylDur* [21], *SilPauseDur* [22], and *FilPauseDur* [23] reflect the size of elements making up the speech flow. However, in the present analysis, these features remain only by default, that is, by virtue of having not (yet) been eliminated. The next question, therefore, was whether positive evidence for retaining any or all of these five features as reflecting a common fluency construct could be found.

To address this, we conducted two additional analyses. First, we looked at the zero-order intercorrelation among these features. Table 2 shows these with their significance levels (p values) and 95% confidence intervals (range of uncertainty associated with each correlation). As can be seen, these five fluency measures correlated significantly with each other, ranging from r_s = |.34| to r_s = |.90| (the *SpTimeRatio* [8] measure in Table 2 is discussed below). Of course, *SylRun* and *PhonRun* were highly collinear (r_s = .90) because both reflect the amount of speech in the run, and it is difficult a priori to decide which of these would be better to retain and which to drop. The levels of intercorrelations in Table 2 are what one might expect if the features together reflected an underlying fluency construct, although it is true that some appear to be more strongly associated than others. It would be useful, however, to see whether these five features are also associated in a meaningful way to some other measure of oral performance that is not itself just a transform or "repackaging" of the fluency measures, given that both oral proficiency and fluency are typically gained through practice. That is the focus of the next analysis.

For this second analysis, we looked at speech-time ratio (*SpTimeRatio* [8]). This measure is the proportion of total time to perform the task that is actually speech as opposed to silence. It reflects speaking ability while not reflecting speech fluency or flow as such (i.e., how the speech is packaged into runs and interruptions). However, even

though *SpTimeRatio* does not directly reflect fluency, one would nevertheless expect that speakers with higher speech-time ratio measures of speech would also speak more fluently. Thus, we expected that if the five fluency measures retained so far were related to a fluency construct, in addition to correlating among themselves they would also correlate with *SpTimeRatio.* As can be seen from Table 2, this is indeed the case; all correlated relatively highly and significantly with *SpTimeRatio*.

Table 2
Zero-Order Intercorrelations (Spearman) [and 95% Confidence Intervals] Between the Oral Proficiency Measure, the Four Retained Core Utterance Fluency Measures, and One Retained Additional Utterance Measure, Based on the Data from the 50 Second Language Learners in Sample A

	Speech Time Ratio	Syllable Run	Phonation Run	Syllable Duration	Silent Pause Duration	Filled Pause Duration
Oral Proficiency Measures						
Speech Time Ratio *SpTimeRatio* [8][b]	—	.79*** [.66 .88]	.87** [.78 .92]	.58** [.35 .74]	.82** [.70 .89]	.39** [.12 .60]
Utterance Fluency Measures						
Syllable Run[a] *SylRun* [19]		—	.90** [.83 .94]	.86** [.77 .92]	.50*** [.25 .68]	.51*** [.27 .69]
Phonation Run[a] *PhonRun* [20]			—	.62** [.42 .77]	.49** [.25 .68]	.34* [.07 .56]
Syllable Duration[a] *SylDur* [21]				—	.40** [.14 .61]	.53** [.29 .70]
Silent Pause Duration[a] *SilPauseDur* [22]					—	.35* [.08 .57]
Filled Pause Duration *FilPauseDur* [23]						—

Note. All data on a given measure were first transformed to z scores and the data for the three duration measures were adjusted by reversing the scale order, so that higher values would indicate higher fluency.
[a]Measures ultimately selected as core fluency measures.
[b]Numbers in square brackets refer to the entry numbers in Table 1.
*$p < .05$. **$p < .01$. ***$p < .001$.

It is important to note, however, that this outcome could be due to an artifact, namely to the fact that the fluency measures are derived from one or more basic measures also involved in the derivation of *SpTimeRatio* [8]. It may be, therefore, that this confound of overlapping basic measures is driving the correlations rather than a connection between fluency and proficiency as such. For example, *SpTimeRatio* is calculated in terms of three basic duration measures, namely as (*totDur* [1] – *totSilPauseDur* [2] – *totFilPauseDur* [3]) / (*totDur* [1]), that is, total phonation time divided by total time. Four of the five fluency measures retained up to this point contain one or more of these duration measures in their derivation (*SylRun* is the exception). For example, *PhonRun* [20] and *SylDur* [21] each

involve total phonation time in their derivation, where phonation time is calculated by subtracting out *totSilPauseDur* and *totFilPauseDur* from *totDur*. Thus, correlational analyses involving *SpTimeRatio* and *PhonRun* would involve a computational confound, namely that part of *PhonRun* (i.e., the duration measure) is also contained within *SpTimeRatio* and it could be this that drives the correlation between them.

However, the five fluency measures retained so far also involve other basic measures as underlying components not in the makeup of *SpTimeRatio*. These are the number measures: number of syllables (*nSyl* [6]), number of silent pauses (*nSilPauses* [4]), and number of filled pauses (*nFilPauses* [5]). These number measures are logically independent of the duration measures underlying *SpTimeRatio.* They are logically independent because an utterance's duration can be composed of any number of syllables depending on how slowly or quickly the person is speaking (and similarly for episodes of silence). These number measures themselves are not useful as fluency measures because they are confounded with how long a person spoke (an individual can produce short or long utterances with the exact same level of fluency). However, if these number measures are themselves significantly correlated with *SpTimeRatio*, a measure that is corrected for overall duration*,* then any fluency measures derived from them can also be said to be related to *SpTimeRatio* for reasons other than possessing shared, underlying components.

Based on the above reasoning, we conducted the following regression analysis. The dependent measure was *SpTimeRatio* [8], the index of oral proficiency. The independent measures were *nSyl* [6], *nSilPauses* [4], and *nFilPauses* [5]. Outlier data (≥ 3 *SD* from the mean) from two participants were removed and the remaining data again normalized on each variable. The data set (N = 48) met all assumptions of heteroscedacity, skewness, and kurtosis for regression. Table 3 summarizes the results. The independent variables accounted for 51% of variance, but only *nSyl* [6] and *nSilPauses* [4] had β values (0.71 and -0.53, respectively), indicating that they were significantly related to the oral proficiency measure whereas *nFilPauses* [5] was not. This pattern supports the conclusion that four of the retained fluency measures—*SylRun* [19], *PhonRun* [20], *SylDur* [21], and *SilPauseDur* [22]— are significantly related to oral proficiency (*SpTimeRatio* [8]) for reasons beyond the sharing of underlying components with oral proficiency. In contrast, the measure *FilPauseDur* [23] appears to be related to oral proficiency only because of having shared components. For this reason, we concluded on statistical grounds that *FilPauseDur* [23] does not meet the criterion of being correlated with the measure of oral proficiency, *SpTimeRatio*, as do the other measures. In sum, of 23 potential fluency measures, four—*SylRun* [19], *PhonRun* [20], *SylDur* [21], and *SilPauseDur* [22]—can be retained as core measures, based on logical and statistical considerations. *FilPauseDur* [23], while remaining an utterance measure of interest that is conceptually related to some aspect of speech flow, appears to be related to fluency differently from the other measures retained as part of a final set of core features.

Table 3
Regression Analysis of Components Specific to Measures of Second Language Fluency (Number of Syllables, Number of Silent Pauses, Number of Filled Pauses) Predicting Oral Proficiency (Speech Time Ratio) With Data From 48 Participants in Study 1 at Time1

Dependent Variable	β	*SE* β	*t*
Speech time ratio (proficiency) (*SpTimeRatio*) [8][a]			
Independent Variables			
Number of syllables (*nSyl*) [6]	0.71	0.11	6.21 ***
Number of silent pauses (*nSilPauses*) [4]	-0.53	1.24	-4.23 ***
Number of filled pauses (*nFilPauses*) [5]	-0.04	0.12	-0.32

Note. R^2 = .51; adjusted R^2 = .48; F (3,44) for ΔR^2 = 15.3***.
[a]Numbers in square brackets refer to the entry numbers in Table 1.
***p < .001.

As a final step in this process, we examined how well the retained measures distinguished learners from native speakers, as one would expect fluency measures to do successfully. Table 4 reports the means for both learner and native speakers, the *p* values associated with *t* tests of differences between groups, and *Hedge's g* measure of effect size.[1] In addition to reporting the data for the retained measures, the table also reports data for the seven basic measures, the proficiency measure, the four core fluency measures finally retained, plus the measure of mean filled pause duration. The last measure was included because, despite having been excluded as a core fluency measure on statistical grounds in the previous step, many researchers do continue to use it as a reflection of fluency and it retains an intuitive appeal as an indicator of speech flow. It remains interesting, therefore, to see how this measure compares to those that have been retained as core fluency measures.

As shown in Table 4, except on the basic measure phonation time (*PhonTime* [7]), the learners performed less well than the native speakers on all measures, with large effect sizes on all the retained core fluency measures.

The proficiency measure discussed earlier, speech time ratio (*SpTimeRatio* [8]), revealed a very large difference between learners and native speakers (effect size of 2.87), larger than for any of the basic measures, with native speakers having a mean *SpTimeRatio* of .82 and learners only .54. This result underscores the usefulness of *SpTimeRatio* as a measure of proficiency (beyond the fact that it also does not confound performance with total task duration as do the basic measures). It is thus an appropriate benchmark for evaluating the fluency measures. Table 2 shows that the four retained fluency measures correlated significantly with this proficiency measure. It should be noted that mean filled pause duration also correlated significantly with the proficiency measure but did so more weakly than did the other fluency measures.

Table 4
Means (SD) for Basic, Oral Proficiency, and Core Fluency Measures Plus One Additional Utterance Measure From the Data of 50 Second Language (L2) Learners (Sample A) and 23 Native Speakers of French, Showing p Values and Hedge's g Effect Sizes [95% CI] for Group Differences

	L2 Learners *M* (*SD*)		Native Speakers *M* (*SD*)		*p*	Effect Size	[95% CI]	
Basic Measures								
Total Duration (s) [1][a]	81.85	(37.02)	57.84	(20.88)	< .001	0.72**	[0.20	1.25]
Total Silent Pause Duration (s) [2]	33.70	(20.15)	8.54	(3.97)	< .001	1.47***	[0.91	2.04]
Total Filled Pause Duration (s) [3]	5.06	(3.88)	2.14	(2.04)	< .001	0.84***	[0.31	1.37]
Number of Silent Pauses [4]	28.68	(14.59)	11.70	(4.28)	<.001	1.36***	[0.80	1.92]
Number of Filled Pauses [5]	10.38	(7.63)	5.91	(5.20)	.005	0.63**	[0.11	1.16]
Pruned Number of Syllables [6]	117.36	(62.37)	219.48	(73.80)	< .001	-1.53***	[-2.10	-0.96]
Phonation Time (s) [7]	43.09	(20.48)	47.16	(17.10)	.04	-0.21	[-0.72	0.30]
Oral Proficiency Measure								
Speech Time Ratio [8]	.54	(0.13)	.82	(0.50)	< .001	-2.87***	[-3.15	-1.82]
Core Fluency Measures								
Syllable Run (number) [19]	4.60	(2.69)	20.06	(7.02)	< .001	-3.40***	[-4.17	-2.63]
Phonation Run (s) [20]	1.61	(0.63)	4.27	(1.39)	< .001	-2.82***	[-3.52	-2.12]
Syllable Duration (ms) [21]	434.10	(111.02)	224.83	(26.97)	< .001	2.22***	[1.59	2.86]
Silent Pause Duration (s) [22]	1.18	(0.41)	0.71	(0.17)	< .001	1.21***	[0.66	1.76]
Additional Utterance Measure								
Filled Pause Duration (s) [23]	0.49	(0.12)	0.38	(0.12)	< .001	0.90***	[0.37	1.43]

Note. Effect size: *"small," **"medium," ***"large."
[a]Index in square brackets refers to measures listed in Table 1.

Discussion

The main focus of Analysis 1 was to operationalize L2 utterance fluency without at the same time relying on other data (such as fluency gain data) whose relationship to utterance fluency one might later want to investigate. To accomplish this, we looked at L2 learners' speech data with respect to 23 different ways of defining speech features that could plausibly serve as markers of L2 fluency. Through a series of logical and statistical analyses, it was possible to reduce the set of 23 to four core measures. In addition, these measures were found to correlate well with each other and with a measure of oral proficiency that itself is not confounded in the way it is calculated with fluency. Finally, the L2 learners also performed significantly less well on each of these measures than did native speakers, with large effect sizes in every case (the L2 learners and native speakers also differed, unsurprisingly, on six of the seven basic measures which, as pointed out earlier, are not fluency measures as such). Together, these results demonstrate that prior to directly investigating fluency gains or comparing groups on fluency attainment it is possible to reduce the large array of possible markers of L2 disfluency down to four. In this respect, the results should be useful for future research on fluency by providing advance guidance as to which features to look at, thereby avoiding "fishing expeditions" to find appropriate features on which to focus.

There are some important limitations to this analysis, which should be addressed by future research. First, there were no L1 data from the L2 learners. These data would be useful for controlling for general individual differences in speaking that are unrelated to L2 fluency, such as general tendencies to hesitate, to speak slowly, et cetera (De Jong et al., 2009; Segalowitz, 2010).

Second, the speech elicitation task used here is only one of many that could have been used (Segalowitz, 2010), and it remains therefore an open question whether other tasks involving different levels of complexity, availability of planning time, opportunity for spontaneity in the communicative task, and so on, would have yielded similar patterns of results.

A third limitation is that silent pauses were defined here in terms of a minimum threshold of 400 ms. This particular choice is supported in the literature, but it is not the only possible choice (De Jong & Bosker, 2013; Segalowitz, 2016). It is not obvious to us that lowering the threshold to, say, 250 ms from 400 ms would have changed the results meaningfully in this study, but it is important to use an appropriate threshold for defining silent pauses. Unfortunately, the choice of threshold for silent pauses is still often justified more by custom than by empirical criteria. It remains an important goal of future research, therefore, to establish a principled basis for defining what the minimal threshold should be for defining silent pauses (see especially De Jong & Bosker, 2013).

Fourth, no distinction was made in this research between features of disfluency that occurred between grammatical structures (e.g., between clauses) and within grammatical structures, a distinction that is often a focus of fluency research.

Finally, there are other important forms of disfluency that are not necessarily temporal as such—for example, reformulations, false starts, replacements, and repetitions (Skehan, 2003; Skehan, Foster, & Shum, 2016; Tavakoli & Skehan, 2005), discussion of which is beyond the scope of this paper.

The narrowing down of the number of temporal measures on which to focus is important because it will facilitate research aimed at linking temporal aspects of cognitive

fluency, the speed and efficiency of the executive control processes underlying speech production, to utterance fluency, the fluidity of actual speech (Segalowitz, 2016). The discovery of specific links between cognitive fluency and utterance fluency will ultimately contribute to a broader understanding of L2 fluency and how to overcome the challenges learners face in fluency attainment (Segalowitz, 2010).

Analysis 2: Fluency Gains

This analysis aimed to build on the findings reported in Analysis 1 by examining fluency gains from Time 1 to Time 2 on the four operationally defined core fluency features in a different and independent sample of L2 learners. To our knowledge this has not been done before; researchers typically use intuitions or past fluency gain results to decide what measures to look at when studying fluency gains in some other context.

We were also interested in whether significant utterance fluency gains could be achieved in as little as 5 weeks in an immersion program. Immersion programs, whether in domestic or study abroad settings, typically last at least a semester (10-13 weeks) and much of the research on immersion programs has addressed such longer-term exposure to the L2. It would be interesting, therefore, both from a research and from a policy perspective, to see what gains are possible after such a short-term immersion experience.

Method

Participants. These were the remaining 50 French-language learning participants, here designated as Sample B, from the original set of 100 L2 learners who were not selected by the random process used for selecting Sample A in Analysis 1.

Materials. The same materials were used as described in the General Methodology.

Procedure. The procedure was as described under General Methodology, with data collected at both Time 1 and Time 2, 5 weeks apart.

Results

The results of primary interest were changes in utterance fluency from Time 1 to Time 2, based on the features identified in Study 1. Table 5 shows means, (SDs), effect sizes and the 95% confidence intervals for the basic measures, the oral proficiency measure, each of the four core features as measured at Time 1 and Time 2, and for the reasons given earlier, also the additional utterance measure mean filled pause duration. As the table shows, there were significant gains in the four core measures, with effect sizes ranging from |0.49| to |0.80|. Filled pause duration also showed significant gains (effect size = |0.55|). There were significant gains in proficiency but in only three of the seven basic measures, all related to total amount of speech or silence.

Table 5

Means (SD) for Basic, Oral Proficiency, and Core Fluency Measures Plus One Additional Utterance Measure From the Data of 50 Second Language Learners (Analysis 2, Sample B) Taken at Times 1 and 2, 5 Weeks Apart, Showing p Values and Hedge's g Effect Sizes [95% CI] for Time Differences

	Time 1 *M* (*SD*)		Time 2 *M* (*SD*)		*p*	*Hedges g* Effect Size	[95% CI]
Basic Measures							
Total Duration (s) [1][a]	76.21	(25.88)	82.61	(21.43)	.15	0.27	[-0.14 0.67]
Total Silent Pause Duration (s) [2]	32.05	(16.51)	24.90	(12.37)	< .01	-0.49 *	[-0.89 -0.08]
Total Filled Pause Duration (s) [3]	4.60	(3.32)	4.78	(3.44)	.70	0.05	[-0.35 0.45]
Number of Silent Pauses [4]	26.44	(11.75)	28.34	(10.13)	.33	-0.17	[-0.23 0.57]
Number of Filled Pauses [5]	9.68	(6.88)	10.52	(6.85)	.39	0.12	[-0.28 0.52]
Pruned Number of Syllables [6]	113.00	(62.95)	159.80	(62.91)	< .001	0.74**	[0.32 1.15]
Phonation Time (seconds) [7]	39.56	(17.24)	52.93	(15.23)	< .001	0.81***	[0.40 1.23]
Oral Proficiency Measure							
Speech Time Ratio [8]	.51	(.16)	.65	(.12)	< .001	0.94***	[0.51 1.36]
Core Fluency Measures							
Syllable Run (number) [19]	4.66	(3.02)	6.46	(3.72)	< .001	0.53**	[0.12 0.93]
Phonation Run (s) [20]	1.58	(0.70)	2.07	(0.89)	< .001	0.60**	[0.19 1.02]
Syllable Duration (ms) [21]	438.00	(127.80)	383.20	(91.19)	< .001	-0.49*	[-0.89 -0.08]
Silent Pause Duration (s) [22]	1.30	(0.66)	0.88	(0.32)	< .001	-0.80***	[-1.22 -0.39]
Additional Utterance Measure							
Filled Pause Duration (s) [23]	0.49	(0.12)	0.43	(0.11)	< .01	-0.55**	[-0.96 -0.14]

Note. Effect size: *"small," **"medium," ***"large."

[a]Index in square brackets refers to measures listed in Table 1.

Discussion

The data indicate fluency gains in speech syllable run (*SylRun* [19]), phonation run (*PhonRun* [20]), syllable duration (*SylDur* [21]), and silent pause duration (*SilPauseDur* [22]). This is encouraging news for short-term intensive language training programs such as the immersion program these participants attended. The data also showed meaningful gains on filled pause duration (*FilPauseDur* [23]). Recall that this measure failed to attain core fluency feature status in Analysis 1. It is possible, of course, for learners to make gains on this utterance measure even if there are reasons to exclude it as a core fluency (fluidity) measure as such. Research on filled pauses in the L1 has shown, for example, that these filled pauses can have communicative functions and therefore may not reflect disfluency in the same way as do silent pauses (Clark & Fox Tree, 2002). This may, therefore, also be true for L2 speakers after gaining a certain amount of experience using the L2; more research is needed on this topic (see Bosker et al., 2014).

It is interesting to look at the fluency gains or lack thereof in the basic, non-fluency measures (measures 1-7, Table 5). Of the seven basic measures, there were gains only in three—a decrease in total silent pause duration (less silence), and an increase in number of syllables and phonation time (more speech). The features for which gains were lacking were total duration, mean duration of filled and silent pauses, and number of filled pauses. This was true despite there being room for gains on these measures, as was evident from the strong learner-native speaker differences revealed at Time 1 (Table 4). In contrast, as noted above, there were gains from Time 1 to Time 2 in all four core fluency measures and fewer gains on basic measures (gains being confined to generally more speech and less silence). This pattern difference supports the construct validity of the core measures used for operationalizing L2 fluency.

It is interesting that such fluency gains, over and above there simply being more speech and less silence, were achieved in such a relatively short time (5 weeks). Segalowitz and Freed (2004) found that after 13 weeks, L2 Spanish learners in a study abroad program (in Spain) also showed gains on somewhat similar measures whereas learners in a regular program at home (in the United States) did not. The learners in our study (and in Segalowitz & Freed, 2004) would have had relatively intensive exposure to the target language under conditions of genuine social communication as a result of living with host families and participating in daily sociocultural activities throughout the local community. Towell et al. (1996) and Segalowitz (2010) have suggested that such exposure would involve contact with native speakers that requires the learners to massively repeat target expressions and constructions, resulting in automatizing L2 speech production and in fluency gains. Such social contact should lead not only to faster and less hesitant speech, but also to more nativelike fluency in which speakers use fixed expressions and "lexicalized" sentence stems, that is, multi-element speech processed as single units (Pawley & Syder, 1983). Fuller and more qualitative analyses of the learners' speech, especially their spontaneous speech, could reveal if automatization and lexicalization of sentence stems underlie the fluency gains reported here (such analysis is beyond the scope of this paper).

Analysis 3: Replications

So far, in this study we used the data from Sample A (50 participants selected randomly from an initial set of 100) to operationally define core utterance fluency features and then we used the data from Sample B (the 50 remaining participants) to investigate fluency gains on these core fluency features. This ensured that the determination of core fluency features was not itself influenced by the fluency gain data and vice versa. The goal of the present analysis was to conduct a replication investigation. For this we reversed the roles of samples A and B, using Sample B data to operationally define core fluency features and Sample A data to investigate fluency gains. The descriptions of participants and methods have already been presented. Next, we report the results and discussion from each analysis. First, we look at the data from Sample B for operationally defining core fluency features and then we look at the data from Sample A for evidence of fluency gains on these core features.

Results and Discussion

Core fluency features. We analyzed the Sample B data exactly as in Analysis 1. Of interest was whether these analyses would replicate the patterns obtained earlier. Recall that the logical analysis alone enabled us to reduce the initial set of 23 measures to eight measures, so the question now concerned which of these eight measures to retain. Here, in the Sample B data, *FilPauseDur* [23] failed to correlate with $r_s \geq |.30|$ with any of the other seven measures. The measure *SylRunAllPauses* [17] was again collinear with *SylRun* [19] ($r_s = .95$, $p < .001$, [.91 .97]) and so was dropped for the reasons given earlier. As in Analysis 1, *PhonRunAllPauses* [18] was significantly but not highly correlated with *PhonRun* [20] ($r_s = .43$, $p < .002$, [.17 .63]). However, because *PhonRunAllPauses* and *PhonRun* are conceptually so similar, *PhonRunAllPauses* was dropped in favour of retaining *PhonRun* for the same reasons as given in Analysis 1. This left five measures surviving the logical and statistical triage up to this point. Table 6 shows the zero-order intercorrelations among these five together with the proficiency measure speech-time ratio (*SpTimeRatio* [8]). Overall, they show a high degree of interrelatedness. The general pattern is similar to that in Table 2, except that here in the Sample B data *FilPauseDur* did not correlate significantly with *PhonRun*.

Table 6
Zero-Order Intercorrelations (Spearman) [and 95% CI] Between the Oral Proficiency Measure, the Four Retained Core Utterance Fluency Measures, and One Retained Additional Utterance Measure, Based on the Data From the 50 Second Language Learners in Sample B

	Speech Time Ratio	Syllable Run	Phonation Run	Syllable Duration	Silent Pause Duration	Filled Pause Duration
Oral Proficiency Measures						
Speech Time Ratio		.86***	.89***	.66***	.89***	.45***
SpTimeRatio [8][b]	—	[.77 .92]	[.82 .94]	[.47 .80]	[.70 .89]	[.20 .65]
Utterance Fluency Measures						
Syllable Run [a]			.93***	.85***	.64***	.33*
SylRun [19]		—	[.88 .96]	[.74 .91]	[.44 .78]	[.05 .56]
Phonation Run [a]				.62***	.65***	.21
PhonRun [20]			—	[.42 .72]	[.46 .79]	[-.07 .46]
Syllable Duration [a]					.50***	.40**
SylDur [21]				—	[.26 .68]	[.14 .61]
Silent Pause Duration [a]						.62***
SilPauseDur [22]					—	[.41 .76]
Filled Pause Duration						—
FilPauseDur [23]						

Note. All data on a given measure were first transformed to *z* scores and the data for the three duration measures were adjusted by reversing the scale order, so that higher values would indicate higher fluency.
[a]Measures ultimately selected as core fluency measures.
[b]Numbers in square brackets refer to the entry numbers in Table 1.
*$p < .05$. **$p < .01$. ***$p < .001$.

Next, we turned to the construct validity of these five features, defined in terms of strong relationships with an oral proficiency measure that is different from fluency as such. All five measures did correlate strongly and significantly with the oral proficiency measure (Table 6), but this could have been due to the presence of shared underlying components. We therefore again conducted a regression analysis, as in Analysis 1, to see whether the derivational components present in the fluency measures that were absent from the oral proficiency measure nevertheless were meaningfully and significantly related to oral proficiency. The analysis revealed no outliers ($N = 50$) and the assumptions of heteroscedacity, skewness, and kurtosis for regression were met. Table 7 shows the results of this analysis. As with Sample A, only number of syllables (*nSyl* [6]) and number of silent pauses (*nSilPauses* [5]) were significantly related to *SpTimeRatio* [8], with $\beta = 0.91$ and -0.30 respectively. Number of filled pauses (*nFilPauses* [4]) again did not yield a significant association with *SpTimeRatio.* Because *nFilPauses* [4] is the only underlying component of *FilPauseDur* [23] that is not confounded with *SpTimeRatio* [8], the conclusion drawn from this result, as in Analysis 1, is that on statistical grounds, *FilPauseDur* should not be considered a core fluency feature. The results of Analysis 3 thus replicate those of Analysis 1; based on logical and statistical grounds, the 23 candidate features can be reduced to four—*SylRun* [19]), *PhonRun* [20], syllable duration *SylDur*

[21], and *SilPauseDur* [22]—and these can reasonably be understood as core features of an L2 fluency construct. The feature *FilPauseDur* [22], while remaining an utterance feature of interest, did not qualify by the criteria used here for inclusion as a core L2 fluency feature.

Table 7
Regression Analysis of Components Specific to Measures of Second Language Fluency (Number of Syllables, Number of Silent Pauses, Number of Filled Pauses) Predicting Oral Proficiency (Speech Time Ratio) Based on the Time 1 Data from Sample B Participants (N = 50)

	β	*SE* β	*t*
Dependent variable			
Speech time ratio (proficiency) (*SpTimeRatio*) [8][a]			
Independent variables			
Number of syllables (*nSyl*) [6]	0.91	0.08	11.70 ***
Number of silent pauses (*nSilPauses*) [5]	-0.30	0.09	-3.40 **
Number of filled pauses (*nFilPauses*) [4]	0.06	0.08	0.70

Note. R^2 = .75. Adjusted R^2 = .73. F (3,46) for ΔR^2 = 45.5 ***
[a]Numbers in square brackets refer to the entry numbers in Table 1.
*p < .05. **p < .01. ***p < .001.

Fluency gains. Next, we examined fluency gains from Time 1 to Time 2 in Sample A for the core fluency features just re-identified. Table 8 presents the fluency gains for these measures. The table also presents Time 1 and Time 2 data for the seven basic measures, the oral proficiency measure, and the additional utterance measure (mean filled pause duration) to allow comparison with Table 5 from Analysis 2. The results largely replicate the findings reported in Analysis 2. First, as before, there were gains in oral proficiency. More importantly, there were gains across the board on all four core fluency measures. Unlike in the previous analysis, however, the non-core utterance feature *FilPauseDur* [23] did not show gain over time. As for the seven basic measures, there were gains in the same three as in Analysis 2—in total silent pause duration, total number of syllables, and total phonation time—once again reflecting reduced silence and more speech at Time 2 compared to Time 1. Thus overall, the results of Analysis 3 replicated the main findings of Analyses 1 and 2.

Table 8

Means (SD) for Basic, Oral Proficiency, and Core Fluency Measures Plus One Additional Utterance Measure From the Data of 50 Second Language Learners (Analysis 3, Sample A) Taken at Times 1 and 2, 5 Weeks Apart, Showing p values and Hedge's g Effect Sizes [95% CI] for Time Differences

	Time 1 Mean (*SD*)		Time 2 Mean (*SD*)		*p*	*Hedge's g* Effect Size	[*95%* CI]
Basic Measures							
Total Duration (s) [1]	81.85	(37.02)	84.75	(24.10)	.52	0.09	[-0.31 0.49]
Total Silent Pause Duration (s) [2]	33.70	(20.15)	24.37	(10.91)	< .001	-0.57 **	[-0.98 -0.16]
Total Filled Pause Duration (s) [3]	5.06	(3.88)	5.41	(3.32)	.58	0.09	[-0.31 0.50]
Number of Silent Pauses [4]	28.68	(14.59)	28.64	(11.31)	.98	-0.003	[-0.40 0.40]
Number of Filled Pauses [5]	10.38	(7.63)	11.48	(6.78)	.35	0.15	[-0.25 0.55]
Pruned Number of Syllables [6]	117.36	(62.37)	165.14	(60.01)	< .001	0.77 **	[0.36 1.19]
Phonation Time (s) [7]	43.09	(20.48)	54.98	(16.18)	< .001	0.64 **	[0.23 1.05]
Oral Proficiency Measure							
Speech Time Ratio [8]	.54	(0.13)	.66	(.10)	< .001	1.02***	[0.60 1.45]
Core Fluency Measures							
Syllable Run (number) [19]	4.60	(2.69)	6.51	(3.08)	< .001	0.66**	[0.25 1.07]
Phonation Run (s) [20]	1.61	(0.63)	2.11	(0.74)	< .001	0.72**	[0.31 1.34]
Syllable Duration (ms) [21]	434.10	(111.02)	386.83	(102.69)	< .001	-0.44*	[-0.85 -0.04]
Silent Pause Duration (s) [22]	1.18	(0.41)	0.85	(0.25)	< .001	-0.89***	[-1.31 -0.47]
Additional Utterance Measure							
Filled Pause Duration (s) [23]	0.49	(0.12)	0.46	(0.11)	<.001	-0.24	[-0.64 0.17]

Note. Effect size: *"small," **"medium," ***"large."

[a]Index in square brackets refers to measures listed in Table 1.

Summary and Conclusions

This research had three aims. The first was to develop a process for identifying a set of core speech features to be used to operationally define L2 fluency (speech fluidity). An important aspect of this process was that it had to be carried out without reference to developmental or other external data to ensure that the emerging fluency construct would be independent from the fluency phenomena to be subsequently investigated. Analysis 1, using data from participant Sample A, revealed through a series of logical and statistical analyses that 23 candidate measures, most of which have been used in prior research on L2 fluency, could be reduced to a set of four core features reflecting an L2 fluency construct. Two measures were mean run length defined either as syllable or phonation run between silent pauses, and two other measures were mean duration of syllables and silent pauses. By operationally defining fluency measures this way, it is possible for researchers to avoid relying on intuition or tradition when deciding which features to use for investigating fluency development or other fluency issues.

The second aim was to examine L2 fluency gains in adult learners participating in a French-language immersion program, where fluency had been operationalized independently and in advance of the study of gains. Analysis 2, using data from Sample B, showed that after 5 weeks in the immersion program learners made gains on all four core fluency features. This finding is important because it demonstrates that in as little as 5 weeks, adult learners in an immersion program were able to make significant and meaningful gains in fluency.

The third aim was to replicate the major results. This was accomplished in Analysis 3, where data for operationalizing core utterance fluency features now came from Sample B and fluency gain data now came from Sample A. The results yielded the same four core fluency features identified earlier, and the fluency gains analyses yielded the same general pattern. The one difference between the two sets of developmental results was that learners in Sample A showed a significant reduction in filled pause duration (but not in number of filled pauses) whereas learners in Sample B did not show such significant reductions. Overall, results from both developmental analyses suggest that filled pause measures do not belong in the same construct as the four measures identified as core features of fluency.

There are two main take-home messages from this research. The first is that it is possible to operationally define an L2 fluency construct in a non-arbitrary way. Importantly, this can be done independently of whatever other fluency phenomenon one ultimately wants to study, such as fluency development over time or the relationship between utterance fluency and cognitive fluency (Segalowitz, 2010). In this way, one avoids the circularity that arises when the outcome variable (say, fluency gains) is used to select the predictor variable (the fluency measure that is expected to show the gains). Having an independent rationale for focusing on particular fluency measures means that researchers can avoid this circularity. Hopefully, this will lead to a consensus on what features of L2 fluency are important to investigate, making it easier to directly compare results across studies. The second take-home message is that in a relatively short, 5-week period, there were meaningful gains in L2 learners' overall speaking ability as seen in speech-time ratio (a non-fluency proficiency gain) as well as in the four core measures of utterance fluency or speech fluidity, that is, in both syllable and phonation run length between silent pauses, and in reduced syllable and silent pause durations. This evidence of L2 fluency gains is all the more noteworthy because it was obtained using an operational

definition of fluency that was established prior to and independently of the analysis of the developmental data.

Correspondence should be addressed to Norman Segalowitz.
Email: norman.segalowitz@concordia.ca

Acknowledgements

The authors would like to thank José Simard and Ariane Tremblay at the *École de langue française et de culture québécoise* for their logistic support during different phases of the project. We would also like to acknowledge the support for this research from the Social Sciences and Humanities Research Council of Canada to Leif French (PI), Norman Segalowitz and Elizabeth Gatbonton.

Notes

[1]*Hedge's g* is the difference between the two means expressed in standard deviation units to allow comparisons across conditions and studies, using *N-1* in its computation to correct for small sample bias. By convention, effect sizes of 0.2, 0.5, and 0.8 are considered "small," "medium," and "large," respectively, and below 0.2 is considered "trivial," regardless of statistical significance.

References

Boersma, P., & Weenink, D. (2007). Praat (Version 4.5.25) [Software]. Retrieved from www.praat.org

Bosker, H. R., Quené, H., Sanders, T., & De Jong, N. H. (2014). Native "um"s elicit prediction of low-frequency referents, but non-native "um"s do not. *Journal of Memory and Language*, *75*, 104-116. http://doi.org/10.1016/j.jml.2014.05.004

Clark, H. H., & Fox Tree, J. E. F. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73-111.

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*(2), 989-999.

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

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

De Jong, N. H., Schoonen, R., & Hulstijn, J. H. (2009, July). *Fluency in L2 is related to fluency in L1*. Paper presented at the Seventh International Symposium on Bilingualism (ISB7), Utrecht, Netherlands.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(01), 5-34. http://doi.org/10.1017/S0272263111000489

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*(4), 655-679.

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: The University of Michigan Press.

Freed, B. F., Segalowitz, N., & Dewey, D. P. (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*(02), 275-301.

García-Amaya, L. (2009). New findings on fluency measures across three different learning contexts. In J. Collentine, B. A. Lafford, M. García, & F. Marcos Marín (Eds.), *Selected Proceedings of the 11th Hispanic Linguistics Symposium.* Somerville, MA: Cascadilla Proceedings Project. Retrieved from www.lingref.com, document #2203

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*(3), 379-399.

Hilton, H. (2009). Annotation and analyses of temporal aspects of spoken fluency. *Calico Journal*, *26*(3), 644-661.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24-49.

Kawahara, S. (2010). Praat script - Calculates the duration of all intervals of all the files in a specified folder. Retrieved from http://user.keio.ac.jp/~kawahara/scripts/duration_getter.praat

Kormos, J. (2006). *Speech production and second language acquisition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145-164.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 387-417.

Llanes, À., & Muñoz, C. (2009). A short stay abroad: Does it make a difference? *System*, *37*(3), 353-365. http://doi.org/10.1016/j.system.2009.03.001

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly*, 610-641.

O'Brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition*, *29*(04), 557-581.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and natively fluency. In J. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191-226). London, United Kingdom: Longman.

Rossiter, M. J., Derwing, T. M., & Jones, V. M. L. O. (2008). Is a picture worth a thousand words? *TESOL Quarterly, 42*(2), 325-329.

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

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 79-95. http://doi.org/10.1515/iral-2016-9991

Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, *26*(02), 173-199.

Segalowitz, N., Freed, B., Collentine, J., Lafford, B., Lazar, N., & Díaz-Campos, M. (2004). A comparison of Spanish second language acquisition in two different learning contexts: Study abroad and the domestic classroom. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *10*, 1-18.

Skehan, P. (2003). Task based instruction. *Language Teaching, 36,* 1-14.

Skehan, P., Foster, P., & Shum, S. (2016). Ladders and snakes in second language fluency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 97-111. http://doi.org/10.1515/iral-2016-9992

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

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, *17*(1), 84-119.

Trenchs-Parera, M. (2009). Effects of formal instruction and a stay abroad on the acquisition of native-like oral fluency. *Canadian Modern Language Review*, *65*(3), 365-393.

Zellner, B. (1994). Pauses and the temporal structure of speech. In E. Keller (Ed.), *Fundamentals of speech synthesis and speech recognition* (pp. 41-62). Chichester, United Kingdom: John Wiley.