L2 Utterance Fluency Development Before, During, and After Residence Abroad: A Multidimensional Investigation

AMANDA HUENSCH and NICOLE TRACY-VENTURA
University of South Florida, Department of World Languages, 4202 E Fowler Ave, CPR 107,
Tampa, FL 33620 Email: huensch@usf.edu

ACKNOWLEDGMENTS

LANGSNAP (http://langsnap.soton.ac.uk/) was funded by the ESRC (award number RES-062-23-2996). We are grateful to our colleagues Rosamond Mitchell, Kevin McManus, Laurence Richard, Patricia Romero de Mills, Karen Ruebens, the participants, and our transcribers and research assistants for their contribution to this research. Finally, we wish to thank three anonymous reviewers for their comments.

Revised manuscript submitted January 10, 2017

huensch@usf.edu

ABSTRACT

This study investigated L2 fluency development over a nearly 2-year period which included an academic year abroad and the year immediately following once participants had returned to the home university to complete their degree. Data from 24 English L1 learners of Spanish were collected 6 times: once before, 3 times during, and 2 times after a 9-month stay abroad. Participants were recorded orally retelling a picture-based narrative, and data were coded for 9 measures of utterance fluency. Results indicated different developmental trends: Gains in speed fluency appeared quickly and were maintained after return from study abroad, whereas gains in breakdown fluency often took longer and were more sensitive to attrition after return home. There were no changes over time in repair fluency. These results appear to indicate that some fluency improvements are more robust and less likely to be affected by the change in context (study abroad vs. home country). The findings fill a gap in our understanding of the relationship between oral fluency development and L2 speech production processes, and have implications for study abroad researchers as well as post-study abroad instruction.

Keywords: study abroad; Spanish; acquisition/learning/development; L2 fluency

Research investigating language learning during residence/study abroad has shown that time abroad does not lead to equal improvement in all areas of language (Collentine, 2009; Kinginger, 2009; Llanes, 2011). One of the most consistent results has been that the largest gains are found in the area of oral fluency (Du, 2013; Freed, 1995; Kim, Dewey, Baker-Smemoe, Ring, Westover, & Eggett, 2015; Mora & Valls-Ferrer, 2012; Segalowitz & Freed, 2004; Towell, Hawkins, & Bazergui, 1996). And when context of learning is compared—for example, study abroad and at-home instruction—study abroad learners typically make significantly more gains in fluency than at-home learners receiving traditional instruction (Freed, Segalowitz, & Dewey, 2004; Serrano, Llanes, & Tragant, 2011). Yet despite evidence that residence abroad is beneficial for second language (L2) fluency, less is known about the trajectory of that development while learners are abroad, particularly for stays of longer than one semester. Additionally, because most studies follow a pre-posttest design and lack a delayed posttest, it is also currently unknown to what extent gains made in L2 fluency are maintained once learners are no longer living in a target-speaking environment but are still enrolled in language programs back in their home universities.

The current research analyzed Spanish L2 learners' longitudinal development of oral fluency over a nearly two-year period to examine whether different aspects of utterance fluency (e.g., speed, breakdown, and repair) change similarly over time. The longitudinal data include a pretest, three data collection points while abroad (approximately at 2 months, 5 months, and 9 months) and two points after returning to the home university (approximately 4 months and 8 months after returning). The findings fill a gap in our understanding of the relationship between oral fluency development and L2 speech production processes, and provide implications for study abroad researchers as well as post-study abroad instruction.

BACKGROUND

Defining and Measuring Fluency

Oral language fluency can be defined in a variety of ways but is most often connected to general proficiency, in a broad definition, or the smoothness and fluidity of speech, in a narrow definition (Lennon, 1990). The current study is concerned with this narrow definition of fluency. Fluency can be further categorized into cognitive fluency (i.e., processing efficiency), utterance fluency (i.e., temporal measures of speech), and perceived fluency (i.e., listener judgement of fluency) (Segalowitz, 2010). The current study investigated the development of utterance fluency during and after residence abroad. Operationalization of utterance fluency is not, however, straightforward, and a variety of measurements have been used in the literature. One way of choosing measurements is to opt for those that correspond to listeners' perceptions of fluent speech. Research comparing perceived fluency to utterance fluency (see e.g., Cucchiarini, Strik, & Boves, 2000, 2002; Kormos & Dénes, 2004; Préfontaine, Kormos, & Johnson, 2016; Rossiter, 2009) has shown, for example, that measurements such as speech rate and mean length of run correlate to/explain variance of raters' judgements; however, the findings from this research are not consistent, and factors such as task type (read vs. spontaneous speech) and linguistic accuracy have been found to influence results. It is also important to consider, as noted in De Jong et al. (2013), that perceived fluency ratings are dependent on the definitions/instructions given to raters or the notions that non-instructed raters have about fluency.

Another way of choosing measurements is to opt for those that correspond to the sub-dimensions of utterance fluency put forth by Skehan (2003) and Tavakoli & Skehan (2005) who suggested that the variety of utterance fluency measurements adopted in previous research could be categorized into three sub-dimensions: (a) speed (e.g., speech rate, articulation rate, mean length of run), (b) breakdown (e.g., mean duration of silent pauses, number of filled pauses per second), and (c) repair (e.g., number of corrections per second, number of repetitions per second). The categorization of measurements into the sub-dimensions is not fully agreed upon in the literature because some measures blend aspects of speed and breakdown fluency. For example, mean length of run has been called a composite measure (Skehan, 2014; Tavakoli, 2016), and speech rate has been argued to be confounded with breakdown fluency (De Jong et al., 2013). Nevertheless, adopting the sub-dimensions has been a useful starting point for a number of recent studies (e.g., De Jong et al., 2013; De Jong et al. , 2015; Kahng, 2014), including Bosker et al. (2013) who also investigated how utterance fluency measures from each of the three sub-dimensions best aligned with measures of perceived fluency. Bosker et al. conducted linear regression analyses that showed that speed (R^2 = .59) and breakdown (R^2 = .54) measures best predicted perceived fluency ratings. Repair fluency contributed to the model, but not as much (R^2 = .16). The best model included all aspects (R^2 = .84). Studies like this demonstrate how each sub-dimension contributes differentially to perceived fluency. To date no study has examined how the different sub-dimensions of utterance fluency develop over time, leaving unresolved the question whether there are different developmental paths for the three sub-domains of utterance fluency.

An additional consideration when choosing measures to track the development and maintenance of oral fluency relates to the location of breakdowns in speech. When choosing breakdown fluency measures, or those that investigate different types of pausing phenomena, one can consider location, frequency, and duration. In their 2013 studies, Bosker and colleagues used measures that represent the frequency and duration of pauses, but not the location. Kahng (2014) compared the utterance fluency of L1 English speakers and L2 English speakers with L1 Korean and related L2 utterance fluency to SPEAK test scores. She focused on pause distribution using both analysis of speech units (ASUs - Foster, Tonkyn, & Wigglesworth, 2000) and clause boundaries. The ASU was chosen because of its appropriateness for segmenting oral, as opposed to written, data. Findings indicated that, while some group differences were found with pause duration and frequency, none of the pausing measures correlated to speaking scores until location was considered. In later work, De Jong et al. (2015) included two measures which considered location: (a) the mean length of silent pauses within ASUs and (b) the mean length of silent pauses between ASUs. Thus, in the current study, breakdown fluency measures that take into account pause location were included.

Relating Utterance Fluency Measurements to Models of Speech Production

Approaching utterance fluency as a multi-dimensional construct allows us to consider how the difference pieces contribute to aspects of L2 speech production. The most widely adopted model of speech production is based on the work of Levelt (1989, 1999) and De Bot (1992) and was further developed in Segalowitz (2010) to include fluency vulnerability points. This model of L2 speech production includes three main parts: a conceptualizer, a formulator, and an articulator. Speech is initiated as a preverbal message in the conceptualizer. Then the formulator uses grammatical and phonological encoding of lemmas from the lexicon to form the message. Finally, in the last stage, the articulator converts the phonetic plan to actual speech. Difficulties at multiple points in the model can be linked to disfluency in speech (i.e., fluency vulnerability points). Towell et al. (1996) attempted to associate gains made in utterance fluency measurements to improvements in different aspects of the model. For example, they argued that speech rate (a measure that includes information about speed and pausing) is representative of all three stages of the model such that improvements on this measure indicate that proceduralization has occurred. On the other hand, improvements in articulation rate (which excludes pausing) are argued to provide evidence of improvements with the articulator only. Measures of pausing are argued to be less straightforward to interpret in that difficulties might be connected to the conceptualizer or the formulator. Increases in mean length of run might provide evidence for increased proceduralized knowledge in the formulator but again, may interact with conceptualization if longer pauses exist for more planning. Based on a comparison of NS and NNS speech, Skehan, Foster, & Shum (2016) argued that pausing which occurs at ASU boundaries indicates macro-planning and is connected to the conceptualizer, whereas pausing which occurs within ASUs indicates lexical choices and is connected to the formulator and/or articulator stages. In order to better understand the relationship between utterance fluency measures and what they might represent in terms of models of L2 speech production, it would be beneficial to compare the development of measures from each of the sub-dimensions of fluency over time to investigate whether there is evidence of different developmental trajectories.

Fluency and Study Abroad

One of the most popular areas of research investigating fluency development is the subdomain of SLA known as study abroad. A variety of research designs have been adopted in this line of research, most longitudinal in nature to some extent, but studies which include any kind of delayed posttest are rare. A number of studies have examined the effect of context of learning by comparing the fluency development of learners who studied abroad versus those who stayed at home in traditional languages classes (e.g., Freed, 1995; García–Amaya, 2009) as well as those in domestic immersion programs (e.g., Freed et al., 2004; Serrano et al., 2011). In general, those students participating in study abroad often make significantly more gains in fluency when measured immediately after returning home, compared to those in traditional at-home language classes. The few studies which included domestic immersion groups also demonstrated significant gains for this group. For example, in Freed et al. the domestic immersion group, who took advantage of more out-of-class L2 contact, made significantly more gains than the SA group on four measures of utterance fluency: total number of words, length of the longest turn, speech rate, and a composite fluidity score.

Another body of research has tested the same group of participants before and immediately after study abroad, the difference in time most often an academic semester abroad, similar to the context of learning studies. For example, after a six-month stay in France, the twelve English L1 learners of French in Towell et al. (1996) made significant improvements in speech rate (syllables/total time including pausing), articulation rate (syllables/phonation time), mean length of run (average syllables produced in utterances between pauses greater than 280 ms), and average length of silent pauses. However, there was no change over time with regard to phonation time ratio (time speaking/total time including pausing). Kim et al. (2015) investigated Chinese L2 learners (English L1) before and after a semester in China and found significant improvements in speech rate (words/minute) and mean silent pause length. Interestingly, learners produced significantly more filled and silent pauses/minute at the posttest, but the authors do not comment on potential reasons for these changes. Mora and Valls-Ferrer (2012) investigated the development of fluency of the same group of Catalan/Spanish bilinguals learning L2 English (*n* = 30) over a period of 2 years. Data were collected twice before students went abroad during formal instruction at their home institution, and then once after they returned from a 3-month stay abroad. The results pointed to significant gains in fluency ("speech rate, mean length of run, pause frequency and duration, and a composite frequency fluency index" p. 637) after study abroad only.

Di Silvio, Diao, and Donovan (2016) compared the fluency development of three L2 groups after a semester abroad: Mandarin, Russian, and Spanish. In general results of their study, which compared measures of fluency across four tasks included as part of the SOPI, support the trend in fluency development as a result of study abroad, yet there were some differences between the language groups. For example, the Mandarin and Spanish groups made more gains than the Russian group. The Spanish group in particular demonstrated significant gains in most of the fluency measures but the results across tasks were not consistent. Speech rate (based on the number of words/minute) was the only measure in which significant gains were made on all four tasks. The number of filled pauses per minute decreased significantly on two tasks (including picture narration), whereas the number of repairs per minute did not change after a semester abroad.

Another group of studies collected data multiple times while participants were abroad but included only an immediate post-test. For example, Du (2013) tracked the oral fluency development of English learners of Mandarin using monthly semi-structured interviews during a 4-month stay abroad. Temporal measures included speech rate, longest turn, and number of characters. Results indicated that the largest improvements occurred during the first month for speech rate and total characters, and these gains were maintained for the remaining few months of the stay abroad. Similar results were found in Serrano, Tragant, & Llanes (2012) who investigated fluency, operationalized as number of pruned syllables/minute (which excludes false starts, repetitions, corrections, and L1 use), along with accuracy and complexity in an oral narrative at three times during Spanish L1 students' academic year abroad in the UK: at the start in September (T1), in December before break (T2), and at the end of their stay in May (T3). Significant improvements in oral fluency were found between T1 and T2 only. In sum, research in this area suggests that gains in fluency occur relatively early during study abroad and are maintained over time while participants are still abroad. However, what happens to the gains made once participants return home is currently unknown as previous research has not included delayed posttests.

It is a rather robust finding now that participation in study abroad leads to significant gains in oral fluency, particularly for measures such as speech rate, when measured at the end of a stay abroad (Di Silvio et al., 2016; Du, 2013; Mora & Valls-Ferrer, 2012; Segalowitz & Freed, 2004; Towell et al., 1996). However, less is known about the development of oral fluency for stays longer than the typical academic semester or after participants return to their home universities. To better understand fluency development, it is beneficial to investigate the trajectory of development of the sub-dimensions of utterance fluency (speed, breakdown, and repair). By doing so, it becomes possible to investigate whether the gains made during study abroad are attributable to improvements in all three stages of speech production or in particular areas only (e.g., the articulator). Based on this review of the literature, the current study examined the following research questions:

RQ1. To what extent does the speed, breakdown, and repair fluency of L2 learners of Spanish change during and after residence abroad?

RQ2. To what extent do the gains made in oral fluency demonstrate improvement in all three stages of speech production or primarily in individual stages?

METHODS

Participants

The data in the current study come from the Spanish subset of the Languages and Social Networks Abroad Project (LANGSNAP: Mitchell, Tracy-Ventura, & McManus, 2017). All of the data are publically available and can be downloaded from the project website: http://langsnap.soton.ac.uk. Participants included 27 British undergraduate students majoring in Spanish who were required to spend their third year of a 4-year program abroad. Because in this context all students majoring in languages are required to go abroad, there is no comparable at-home group of students. Three participants' data were excluded because they were either non-native speakers of English (participants 158 and 165) or the quality of their sound files was such that it could not be analyzed (participant 150's pre-sojourn recording). Of the remaining 24 participants (17 female, 7 male), average age at the beginning of data collection was 20.6 (*SD* = 1.2 years). While the participants were all at the same instructional level of Spanish (end of second year of their bachelor's degree), their previous classroom learning experience varied (*M* = 5.9 years, *SD* = 3.2 years). Participants also completed a proficiency test pre-sojourn in the form of an elicited imitation test, which has been shown to be a valid and reliable measure of L2 proficiency (Ortega, 2000; Bowden, 2016). The average score was 84.0 (*SD* = 11.7) out of 120 or 70%, and ranged from 59-108.

Typical of the British "year abroad" experience, the LANGSNAP participants spent their year spread out in various cities and small towns, undertaking a placement type of their choosing: (a) as an exchange student at a partner university, (b) as an English teaching assistant at a school, or (c) as a work intern (for more information see Mitchell, McManus, & Tracy-

Ventura, 2015). They were most often responsible for finding their own accommodation and accomplishing other tasks related to moving abroad (e.g., opening a bank account, getting a mobile phone set-up). In other words, there was very little involvement by the home university in planning the students' experiences abroad beyond helping to negotiate the placement. Exchange students usually took content classes with local students. The teaching assistants worked in a school for at least 12 hours a week, most often assisting the main teacher, although occasionally they taught lessons on their own or coordinated conversation classes. The work interns (not included in this study, participants 150 & 158 - see above) were responsible for coordinating their own work placement and typically worked full-time (see also Mitchell et al., 2017).

Once they returned from their year abroad and began the final year of their degree program, the LANGSNAP participants were required to take at least 3 hours of Spanish language classes per week. They also took content classes related to their degree (e.g., history of Spain), but these classes were typically not taught in the target language so as not to limit enrollment.

Procedure and Materials

As part of LANGSNAP, a variety of data were collected at six time points over approximately a 2-year period: before, during, and after participants resided abroad as displayed in Table 1. Pre-sojourn assessments were administered at the end of the participants' second year of university. Their third year was spent abroad in the same location (most participants were the only one in their city/town – see Participants section). Participants were visited by a member of the research team three times while abroad. After they returned to their home university, during the fourth and final year of their degree program, delayed posttests were administered twice while all participants were still enrolled in Spanish language classes.

<INSERT TABLE 1 ABOUT HERE>

TABLE 1

Timeline and Tasks Administered at the Six Data Collection Points

Location	Time	Task
Pre-sojourn (at home university)	May 2011	Cat Story
In-sojourn 1 (abroad)	November 2011	Sister Story
In-sojourn 2 (abroad)	February 2012	Brothers Story
In-sojourn 3 (abroad)	May 2012	Cat Story
Post-sojourn 1 (at home university)	October 2012	Sister Story
Post-sojourn 2 (at home university)	February 2013	Brothers Story

Oral data were gathered via three picture-based narration tasks that were each administered twice in sequence, approximately 1 year apart to reduce potential task familiarity effects (see Table 1). Task comparability between stories was examined using native speaker data (n = 8). Friedman tests indicated that eight out of nine measures were not significantly different (all p values greater than .095), with the only significant difference being the post hoc comparison of the measure of mean silent pause duration between ASU for the Cat Story and the Sister Story (p = .012, d = .61). Thus, any changes in fluency in the current data set were attributed to time.

At each data collection session, participants were instructed to narrate the story presented in a series of related pictures. The stories were provided as a paper handout and were approximately 15 pages in length. The Cat Story and the Sister Story were borrowed from Dominguez, Tracy-Ventura, Arche, Mitchell, & Myles (2013). The Brothers Story was based on the children's book *I Very Really Miss you* (Langley, 2006) and was created by the LANGSNAP research team to mirror the same plot structure as the other two stories. All stories began with background information about what the characters would normally do before the main events of the story began. More information about the picture stories as well as all transcripts and audio files are available on the LANGSNAP website and described in Tracy-Ventura, Mitchell, & McManus (2016).

Participants were first given approximately a minute to look through the pictures before beginning, and they could continue to use the handout while they narrated the story. Participants were not given a time limit in narrating the story; thus the length of narration time varied.[1] The entire data set consists of approximately nine and a half hours of talk. Table 2 provides the means and standard deviations for the duration of each narration at each data collection point.

<INSERT TABLE 2 ABOUT HERE>

TABLE 2

Means (and Standard Deviations) of Recording Length for Each Narrative Task at the Six Data Collection Points

Location	Time	Task
Pre-sojourn	*M* = 4:23 (*SD* = 1:47)	Cat Story
In-sojourn 1	*M* = 4:52 (*SD* = 1:44)	Sister Story
In-sojourn 2	*M* = 3:05 (*SD* = 1:27)	Brothers Story
In-sojourn 3	*M* = 3:14 (*SD* = 1:36)	Cat Story
Post-sojourn 1	*M* = 3:56 (*SD* = 1:25)	Sister Story
Post-sojourn 2	*M* = 2:34 (*SD* = 1:19)	Brothers Story

Data Coding

Data were first transcribed following CHAT (Codes for the Human Analysis of Transcripts) conventions for analysis in CLAN (Computerized Language ANalysis) (MacWhinney, 2000). Instances of repetitions (e.g., *el* [/] *el gato* 'the [/] the cat'), corrections (e.g., *la* [//] *el gato* 'the$_{\text{FEM}}$ [//] the$_{\text{MASC}}$ cat$_{\text{MASC}}$'), and filled pauses (e.g., um, eh) were coded in the transcription to allow for automatic counting. Transcript accuracy, including ASU placement, was checked by multiple members of the research team. Syllables were counted manually by members of the research team and inter-rater reliability was calculated using Cronbach's alpha because the data were continuous. The result was high at .99. Next, data were segmented and annotated in Praat (Boersma & Weenik, 2015) for instances of speech and silent pauses. A silent pause length threshold was set at 250 ms based on research (De Jong & Bosker, 2013) demonstrating that at this length, as opposed to shorter (e.g., 100 ms) or longer (e.g., 400 ms) durations, pause rate correlated strongest with L2 proficiency. The annotated TextGrids in Praat were then used to gain duration information about speech and silent pauses. Because of the sound quality of some of the files, manual checking of Praat's automated function was required. Data were coded by members of the research team. Because these data were nominal (i.e., existence of a pause or sound), inter-rater reliability was calculated on a subset of the data using Cohen's kappa. The result was high at .99.

Fluency measurements for each of the sub-dimensions of fluency were calculated. For comparison purposes, the seven utterance fluency measures from De Jong et al. (2015) were used in the current study, along with two other commonly used measures in the study abroad literature that also have been shown to correlate with perceived fluency (speech rate and mean length of run). Three measurements of speed fluency were included: mean syllable duration (the inverse of articulation rate), speech rate, and mean length of run. Mean syllable duration was calculated by dividing phonation time (the total speaking time excluding pauses) by the total number of syllables. Speech rate was calculated by dividing the total speaking time including pauses by the total number of syllables. Mean length of run was calculated by dividing the total number of syllables by the total number of utterances (i.e., the number of speaking turns between two silent pauses of 250 ms or more). Four measurements of breakdown fluency were included, taking into consideration frequency, location, and duration of pauses. The number of silent pauses per second was calculated by dividing the total number of silent pauses greater than 250 ms by the total speaking time excluding pauses. The number of non-lexical filled pauses per second was calculated by dividing the total number of filled pauses of any length by the total speaking time excluding pauses. Mean silent pause duration within ASU represents the average length of pauses greater than 250 ms within ASUs, and mean silent pause duration between ASU represents the average length of pauses greater than 250 ms between ASUs. Two measures of repair fluency were included: The number of repetitions per second was calculated by dividing the total number of repetitions by the total speaking time excluding pauses, and the number of corrections per second was calculated by dividing the total number of corrections by the total speaking time excluding pauses.

RESULTS

Before addressing the research questions, descriptive statistics are presented for each of the measures of utterance fluency at each data collection point in Table 3. Because not all the data were normally distributed, both means with standard deviations and medians are provided for each fluency measure.

<INSERT TABLE 3 ABOUT HERE>

TABLE 3

Means (and Standard Deviations) and Medians of Temporal Fluency Measures from Pre-sojourn, In-sojourn 1, In-sojourn 2, In-sojourn 3, Post-sojourn 1, and Post-sojourn 2

	Pre-sojourn	In-sojourn 1	In-sojourn 2	In-sojourn 3	Post-sojourn 1	Post-sojourn 2
Speed Fluency						
Mean syllable duration	350	242	246	237	228	253
(ms)	(85)	(44)	(56)	(40)	(39)	(53)
Median	340	231	234	233	230	238
Speech Rate	1.81	2.59	2.85	2.85	2.94	2.74
	(0.51)	(0.62)	(0.74)	(0.69)	(0.74)	(0.74)
Median	1.78	2.53	2.96	2.89	2.87	2.67
Mean Length of Run	4.53	6.08	7.40	7.25	7.70	6.94
	(1.58)	(1.87)	(2.83)	(2.50)	(2.75)	(2.48)
Median	4.06	6.32	7.25	7.04	7.40	6.66
Breakdown Fluency						
Mean silent pause duration						
Within ASU (ms)	803	641	632	623	592	636
	(158)	(101)	(117)	(101)	(99)	(124)
Median	770	632	609	592	573	635
Between ASU (ms)	1171	1227	1199	1024	1129	1096
	(295)	(301)	(345)	(195)	(256)	(268)
Median	1123	1235	1201	985	1057	1077
Number of						
Silent pauses/second	0.76	0.73	0.63	0.64	0.63	0.66
	(0.18)	(0.17)	(0.17)	(0.20)	(0.20)	(0.20)
Median	0.76	0.72	0.60	0.65	0.63	0.64
Filled pauses/second	0.29	0.19	0.13	0.11	0.18	0.19
	(0.17)	(0.13)	(0.11)	(0.11)	(0.15)	(0.13)

Median	0.35	0.17	0.10	0.07	0.16	0.18
Repair Fluency						
Repetitions/second	0.08	0.08	0.08	0.09	0.07	0.07
	(0.05)	(0.05)	(0.06)	(0.06)	(0.04)	(0.06)
Median	0.07	0.06	0.07	0.08	0.06	0.06
Corrections/second	0.07	0.08	0.09	0.08	0.08	0.07
	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Median	0.06	0.08	0.08	0.07	0.08	0.08

RQ1 examined to what extent the different sub-dimensions of fluency (speed, breakdown, repair) change over time. Before conducting inferential statistics to test whether changes were statistically significant, the assumptions of the parametric repeated-measures ANOVA (e.g., normality, detection of outliers) were checked for each utterance fluency measure at each data collection point. The assumptions were not always met (e.g., some data were not normally distributed and outliers were found) and transformations did not result in normally distributed data nor did they remove outliers. Therefore, the original data were retained and the non-parametric Friedman test was chosen (Field, 2013; Larson–Hall, 2010) and conducted using SPSS v.22. In the case of significant differences, pairwise comparisons with significance levels adjusted using a Bonferroni comparison (SPSS, 2013) were analyzed. Effect sizes are reported with the post hoc tests because the Friedman test is an omnibus test and therefore it is more appropriate to report effect sizes with the post hoc tests (Larson–Hall, 2010). Effect sizes are reported as absolute values and interpreted following the recommendations of Plonsky and Oswald (2014) for within-contrasts: small effect (*d* = .60), medium effect (*d* = 1.00), and large effect (*d* = 1.40).

As shown in Table 4, results of the Friedman tests for the measures of speed fluency (mean syllable duration, speech rate, and mean length of run) indicated statistically significant differences among data collection rounds. All three measures demonstrated improvement over time from the pre-sojourn, with mean syllable duration scores decreasing over time, and speech rate and mean length of run scores increasing over time. Figure 1 presents the boxplots for the speed fluency results.

<INSERT TABLE 4 ABOUT HERE>

TABLE 4

Chi Square Test Statistics and Their Associated *p* Values as a Result of the Friedman Tests

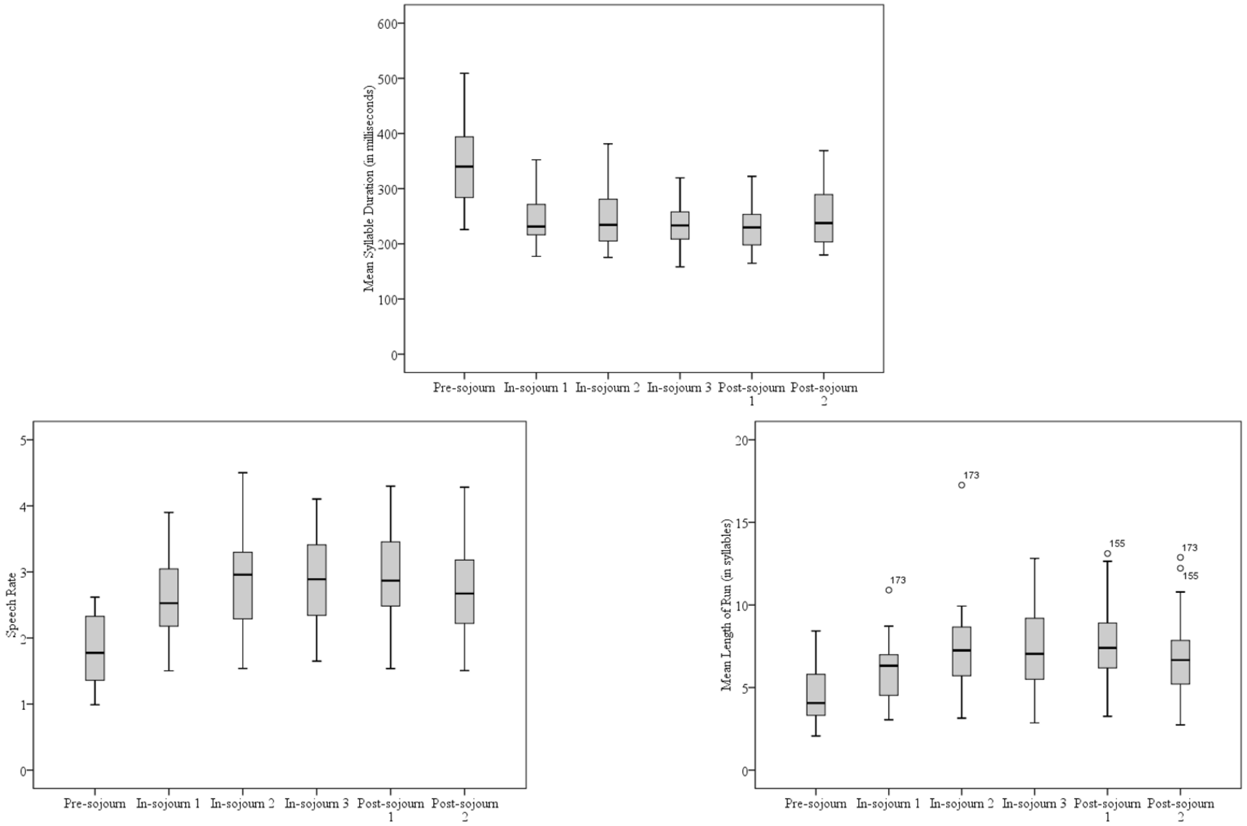
	Friedman Test
Speed Fluency	
Mean Syllable Duration (ms)	$\chi^2(5) = 64.83$, $p < .001$*
Speech Rate	$\chi^2(5) = 71.83$, p < .001*
Mean Length of Run	$\chi^2(5) = 72.67$, p < .001*
Breakdown Fluency	
Mean Silent Pause duration Within ASU (ms)	$\chi^2(5) = 39.57$, $p < .001$*
Between ASU (ms)	$\chi^2(5) = 17.67$, $p = .003$*
Number of Silent Pauses/second	$\chi^2(5) = 33.33$, $p < .001$*
Filled Pauses/second	$\chi^2(5) = 50.02$, $p < .001$*
Repair Fluency	
Number of Repetitions/second	$\chi^2(5) = 2.62$, $p = .758$
Corrections/second	$\chi^2(5) = 9.14$, $p = .103$

* Significant differences at the $p < .05$ level.

<INSERT FIGURE 1 ABOUT HERE>

FIGURE 1

Boxplots for Speed Fluency Measures at Each Data Collection Point



As shown in Table 5,[2] pairwise comparisons for the speed fluency measures indicated significant differences between the pre-sojourn and all other rounds, (except for the pre-sojourn and in-sojourn 1 for mean length of run in syllables which approached significance, *p* = .065), with medium to large effect sizes (*d* = 1.16 – 1.85). These results suggest that the gains in speed fluency recorded at the first visit abroad were retained over time and most notably, after return to the home university. For mean syllable duration, a statistically significant difference was also found between post-sojourn 1 and post-sojourn 2, with a small effect size (*d* = 0.53), which suggests that participants may have been starting to show signs of attrition on this measure after 8 months back in their home country. For both speech rate and mean length of run, there were also statistically significant differences between in-sojourn 1 and all subsequent rounds except for post-sojourn 2, with small effect sizes (*d* = 0.38 – 0.69). These results suggest that participants' speech rate and mean length of run peaked at in-sojourn 2 and gains were maintained at least initially after they returned home. However, by post-sojourn 2 (8 months after returning home), they were possibly starting to show signs of attrition in these areas.

<INSERT TABLE 5 ABOUT HERE>

TABLE 5

Statistically Significant Post Hoc Pairwise Comparisons and Their Associated *p* Values and Effect Sizes for the Measures of Speed Fluency

	Pairwise Comparison		
Rounds Compared	Mean Syllable Duration	Speech Rate	Mean Length of Run
Pre-sojourn vs. In-sojourn 1	*z* = 5.09, *p* < .001*, *d* = 1.60	*z* = −3.24, *p* = .018*, *d* = 1.36	*z* = −2.86, *p* = .065, *d* = 0.89
Pre-sojourn vs. In-sojourn 2	*z* = 5.02, *p* < .001*, *d* = 1.45	*z* = −6.12, *p* < .001*, *d* = 1.62	*z* = −6.33, *p* < .001*, *d* = 1.25
Pre-sojourn vs. In-sojourn 3	*z* = 5.63, *p* < .001*, *d* = 1.72	*z* = −6.33, *p* < .001*, *d* = 1.70	*z* = −4.86, *p* < .001*, *d* = 1.30
Pre-sojourn vs. Post-sojourn 1	*z* = 7.72, *p* < .001*, *d* = 1.85	*z* = −7.33, *p* < .001*, *d* = 1.77	*z* = −7.33, *p* < .001*, *d* = 1.41
Pre-sojourn vs. Post-sojourn 2	*z* = 4.32, *p* < .001*, *d* = 1.37	*z* = −4.71, *p* < .001*, *d* = 1.45	*z* = −4.94, *p* < .001*, *d* = 1.16
In-sojourn 1 vs. In-sojourn 2	*z* = −0.08, *p* = .939, *d* = 0.08	*z* = −2.93, *p* = .051, *d* = 0.38	*z* = −3.47, *p* = .008*, *d* = 0.55
In-sojourn 1 vs. In-sojourn 3	*z* = 0.54, *p* = 1.000, *d* = 0.14	*z* = −3.09, *p* = .030*, *d* = 0.40	*z* = −3.01, *p* = .039*, *d* = 0.53
In-sojourn 1 vs. Post-sojourn 1	*z* = 2.62, *p* = .131, *d* = 0.34	*z* = −4.09, *p* = .001*, *d* = 0.52	*z* = −4.48, *p* < .001*, *d* = 0.69
Post-sojourn 1 vs. Post-sojourn 2	*z* = −3.40, *p* = .010*, *d* = 0.53	*z* = 2.62, *p* = .131, *d* = 0.27	*z* = 2.39, *p* = .252, *d* = 0.29

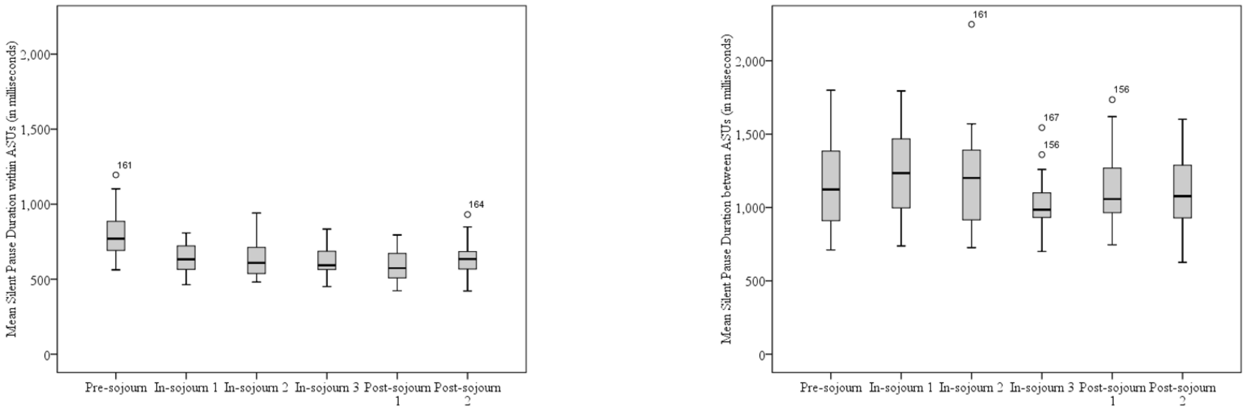
* Significant differences at the *p* < .05 level.

Figure 2 presents the boxplots for the breakdown fluency measures of mean silent pause duration between and within ASUs at each data collection point. As shown in Table 4, results of the Friedman tests for these measures indicated statistically significant differences among data collection rounds. For the mean silent pause duration within ASUs, post hoc pairwise comparisons indicated significant differences between the pre-sojourn and all other data collections rounds, with medium to large effect sizes (*d* = 1.18 – 1.60). No other statistically significant differences between rounds were found. For the measure of the mean silent pause duration between ASUs, post hoc pairwise comparisons indicated significant differences between in-sojourn 3 and both in-sojourn 1 and in-sojourn 2, with medium effect sizes (*d* = 0.80 and *d* = 0.63). No other statistically significant differences between rounds were found. Table 6 presents the results of the pairwise comparisons for these two measures of breakdown fluency that indicated significant differences and the complete table can be found in Appendix B.

<INSERT FIGURE 2 ABOUT HERE>

FIGURE 2

Boxplots for Mean Silent Pause Duration Within and Between ASUs at Each Data Collection Point



<INSERT TABLE 6 ABOUT HERE>

TABLE 6

Statistically Significant Post Hoc Pairwise Comparisons and Their Associated *p* Values and Effect Sizes for Mean Silent Pause Duration Within and Between ASUs

	Pairwise Comparison	
Rounds Compared	Mean Silent Pause Duration Within ASU	Mean Silent Pause Duration Between ASU
Pre-sojourn vs. In-sojourn 1	$z = 4.63$, $p < .001$*, $d = 1.23$	$z = -1.23$, $p = 1.000$, $d = 0.19$
Pre-sojourn vs. In-sojourn 2	$z = 4.48$, $p < .001$*, $d = 1.24$	$z = -0.69$, $p = 1.000$, $d = 0.09$
Pre-sojourn vs. In-sojourn 3	$z = 4.24$, $p < .001$*, $d = 1.36$	$z = 2.32$, $p = .310$, $d = 0.59$
Pre-sojourn vs. Post-sojourn 1	$z = 5.79$, $p < .001$*, $d = 1.60$	$z = 0.46$, $p = 1.000$, $d = 0.15$
Pre-sojourn vs. Post-sojourn 2	$z = 3.55$, $p = .006$*, $d = 1.18$	$z = 1.46$, $p = 1.000$, $d = 0.27$
In-sojourn 1 vs. In-sojourn 3	$z = -0.39$, $p = 1.000$, $d = 0.18$	$z = 3.55$, $p = .006$*, $d = 0.80$
In-sojourn 2 vs. In-sojourn 3	$z = -0.23$, $p = 1.000$, $d = 0.08$	$z = 3.01$, $p = .039$*, $d = 0.63$

* Significant differences at the $p < .05$ level.

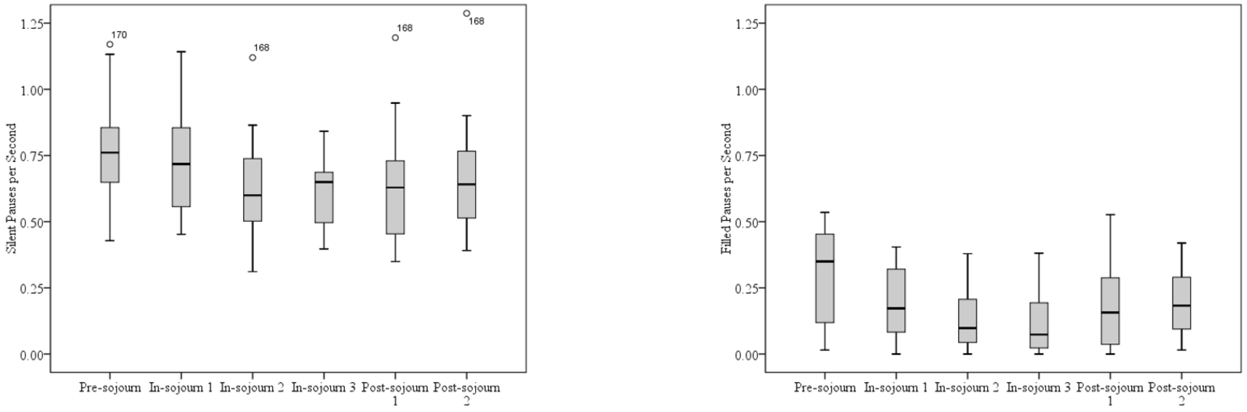
Figure 3 presents the boxplots for the breakdown fluency measures of number of silent pauses per second and the number of filled pauses per second at each data collection point. The Friedman tests for these measures also indicated statistically significant differences among data collection rounds. For the number of silent pauses per second, post hoc pairwise comparisons indicated significant differences between the pre-sojourn, in-sojourn 2, in-sojourn 3, and post-sojourn 1, with small effect sizes (*d* = 0.65 – 0.77), as well as significant differences between in-sojourn 1 and in-sojourn 2, in-sojourn 3, and post-sojourn 1, with small effect sizes (*d* = 0.47 – 0.58). For the number of non-lexical filled pauses per second, post hoc pairwise comparisons indicated significant differences between the pre-sojourn, in-sojourn 2, in-sojourn 3 with medium effect sizes (*d* = 1.10 and *d* = 1.26), as well as significant differences between in-sojourn

1 and in-sojourn 3, in-sojourn 2 and post-sojourn 2, and in-sojourn 3 and both post-sojourn 1 and post-sojourn 2 with small effect sizes (*d* = 0.10 – 0.72). Table 7 presents the results of the statistically significant pairwise comparisons for these two measures of breakdown fluency. The complete table can be found in Appendix C.

<INSERT FIGURE 3 ABOUT HERE>

FIGURE 3

Boxplots for Silent and Filled Pauses per Second at Each Data Collection Point



<INSERT TABLE 7 ABOUT HERE>

TABLE 7

Statistically Significant Post Hoc Pairwise Comparisons and Their Associated *p* Values and Effect Sizes for the Number of Silent and Filled Pauses per Second

	Pairwise Comparison	
Rounds Compared	Number of Silent Pauses per Second	Number of Filled Pauses per Second
Pre-sojourn vs. In-sojourn 2	$z = 4.09$, $p = .001$*, $d = 0.77$	$z = 5.09$, $p < .001$*, $d = 1.10$
Pre-sojourn vs. In-sojourn 3	$z = 3.47$, $p = .008$*, $d = 0.65$	$z = 6.17$, $p < .001$*, $d = 1.26$
Pre-sojourn vs. Post-sojourn 1	$z = 3.78$, $p = .002$*, $d = 0.67$	$z = 2.82$, $p = .073$, $d = 0.72$
In-sojourn 1 vs. In-sojourn 2	$z = 3.94$, $p = .001$*, $d = 0.58$	$z = 2.51$, $p = .182$, $d = 0.50$
In-sojourn 1 vs. In-sojourn 3	$z = 3.32$, $p = .014$*, $d = 0.47$	$z = 3.59$, $p = .005$*, $d = 0.68$
In-sojourn 1 vs. Post-sojourn 1	$z = 3.63$, $p = .004$*, $d = 0.50$	$z = 0.23$, $p = 1.000$, $d = 0.10$
In-sojourn 2 vs. Post-sojourn 2	$z = -1.70$, $p = 1.000$, $d = 0.20$	$z = -3.24$, $p = .018$*, $d = 0.53$
In-sojourn 3 vs. Post-sojourn 1	$z = 0.31$, $p = 1.000$, $d = 0.04$	$z = -3.36$, $p = .012$*, $d = 0.53$
In-sojourn 3 vs. Post-sojourn 2	$z = -1.08$, $p = 1.000$, $d = 0.13$	$z = -4.32$, $p < .001$*, $d = 0.72$

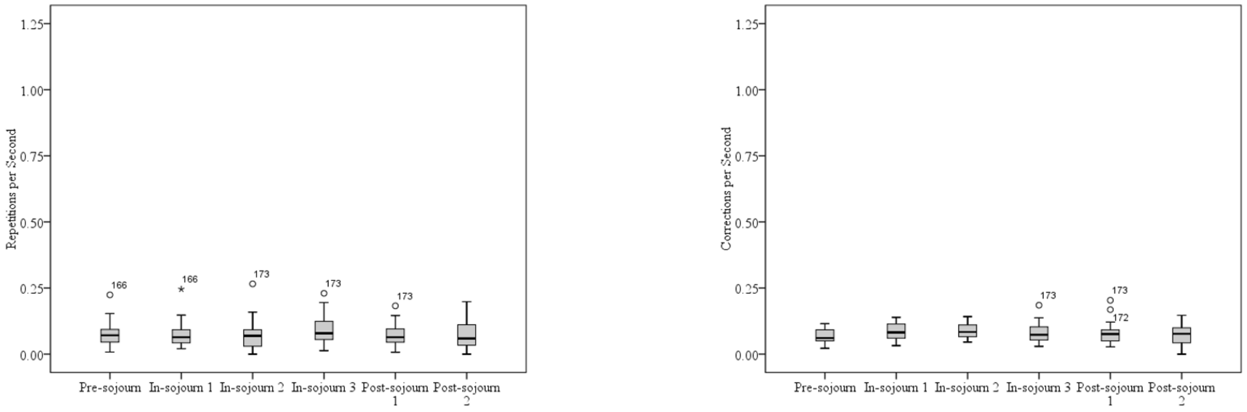
* Significant differences at the $p < .05$ level.

Unlike the results just presented for speed and breakdown measures, results of the Friedman tests comparing repetitions and corrections over time were not significant. Figure 4 presents the boxplots for the measures of repair fluency at each data collection point, demonstrating no change during or after residence abroad.

<INSERT FIGURE 4 ABOUT HERE>

FIGURE 4

Boxplots for the Number of Silent and Filled Pauses per Second at Each Data Collection Point



DISCUSSION AND IMPLICATIONS

The purpose of this study was to investigate the longitudinal development of oral fluency before, during, and after a 9-month stay abroad. Utterance fluency was operationalized using measures from each of the different sub-dimensions proposed by Skehan (2003) and Tavakoli and Skehan (2005): speed, breakdown, and repair. Unlike previous research investigating fluency development during study abroad, data from English L1 learners of Spanish were collected six times over a 2-year period (pre-sojourn, in-sojourn 1, in-sojourn 2, in-sojourn 3, post-sojourn 1, and post-sojourn 2), including two delayed posttests that occurred 4 and 8 months after returning home.

RQ1 investigated to what extent speed, breakdown, and repair fluency changed during and after residence abroad. In this study, three measures of speed fluency were included: (a) mean syllable duration, (b) speech rate, and (c) mean length of run. A statistically significant result was found for change over time for all three measures. Post hoc tests for mean syllable duration demonstrated a significant change from the pre-sojourn to in-sojourn 1 which was maintained over time. The only other significant difference was between post-sojourn 1 and post-sojourn 2, suggesting that participants may have started to show signs of attrition. Results of the post hoc tests for speech rate and mean length of run were similar to those for mean syllable duration but showed more gradual development that peaked after 5 months abroad (in-sojourn 2). By post-sojourn 2, participants were also possibly starting to show signs of attrition in speech rate and mean length of run. The current result for speech rate contrasts with Du (2013) who found a significant change in speech rate quite early during participants' stay abroad, after just 1 month, and this gain was maintained for the rest of the stay abroad (1 semester). Serrano et al. (2012) also analyzed speech rate and, similar to the current results, found that significant improvements in speech rate occurred between T1 and T2, approximately 4 months apart, with little change from T2 to T3. As a reminder, speech rate includes information about pausing, whereas mean syllable duration (inverse articulation rate) does not. This difference likely explains the results of the current study in which mean syllable duration peaked at in-sojourn 1, after 2 months abroad. Thus, it is possible that the quick improvements for speech rate in Du (2013) ultimately were driven by improvements in articulation rate. The calculations of mean syllable duration and speech rate are quite similar, the main difference being that mean syllable duration does not include pausing time. In a speech sample, the percentage of pausing time will often be small compared to the percentage of phonation time, which is why if improvements are made in articulation, the difference would be more noticeable than if the improvements were made in pausing.

Measures related to pausing are part of breakdown fluency. In this study four measures of breakdown fluency were adopted: (a) mean silent pause duration within ASU, (b) mean silent pause duration between ASU, (c) number of silent pauses per second, and (d) number of non-lexical filled pauses per second (see De Jong et al., 2015). Mean silent pause duration within

ASUs improved by in-sojourn 1 and the gains were maintained through post-sojourn 2. Improvement on this measure is likely the result of fewer processing difficulties (Kahng, 2014; Pawley & Syder, 2000) which participants appear to benefit from within the first few months of living abroad. Pausing within ASUs is more characteristic of L2 learners compared to native speakers who tend to pause more often between ASUs (Kahng, 2014). In this study there was little change over time for the mean duration of silent pauses between ASUs. Significant differences were found between in-sojourn 1 and in-sojourn 3, and in-sojourn 2 and in-sojourn 3 only. Significant improvement in the number of silent and filled pauses per second peaked by in-sojourn 2, suggesting that improvement in these areas of breakdown fluency are more gradual than mean syllable duration and mean silent pause duration within ASU; however, they both returned to pre-sojourn levels by post-sojourn 2 demonstrating that they were also more sensitive to attrition. Why this might be the case is an interesting question for future research. It is likely that examining the location and duration of filled pauses would be useful, as was done with silent pauses. In the current study the mean duration of silent pauses within and between ASUs was investigated, but more fine-grained analyses at the clausal and phrasal levels would also contribute to our understanding of why breakdowns occur (see e.g., De Jong, 2016; Skehan et al., 2016).

Results focusing on repair fluency demonstrated no change over time for either measure (number of repetitions and corrections per second) a finding similar to Di Silvio et al. (2016) for their L2 Spanish group. A possible explanation for this finding is that increases in learners' proficiencies (see Huensch & Tracy-Ventura, 2016; Mitchell et al., 2017) allowed for an increased ability to monitor and/or repair their speech. Thus, while it might be predicted that repetitions and/or corrections decrease as proficiency increases, this would not be the case if learners are simultaneously more able to notice mistakes in their speech. In fact, previous research has shown that repair fluency often does not distinguish learners of different proficiencies (Baker–Smemoe, et al. 2014; Kormos & Dénes, 2004; Tavakoli & Skehan, 2005). A lack of correlation between repair fluency and L2 proficiency across five language groups led Baker–Smemoe et al. (2014) to hypothesize that measures of repair fluency might better reflect the trait of an individual speaker. Kahng (2014) also argued that individual differences appear to play more of a role for measures of repair fluency. Therefore, it is likely the case, as described in Ellis & Barkhuizen (2005), that repair "indicates the extent to which the learner is oriented towards accuracy," (pp. 149–151) rather than being solely a dimension of fluency. Additional research is needed in which the repetitions and corrections initiated by the learners are the subject of an in-depth analysis to determine what aspects of language are self-corrected during oral production (see Kormos, 2000).

In sum, the results from RQ1 suggest that different aspects of fluency show different developmental and retention patterns before, during, and after study abroad. Considering these results, our second research question examines to what extent the gains made in oral fluency demonstrate improvement in all stages of L2 speech production or primarily in individual stages. Towell et al. (1996) and others have argued that improvements in articulation rate (a measure of speed fluency) represent improvements within the articulator only, whereas improvements in different aspects of pausing (measures of breakdown fluency) are more difficult to interpret because they can represent improvements with formulation and conceptualization. In the current study, the early gains in mean syllable duration which were less subject to attrition over time—in contrast to the more varying developmental trends in breakdown fluency which were slower to emerge and more subject to attrition over time—appear to support the argument that mean syllable duration does indeed reflect articulation only. In Segalowitz's model (2010, p. 9) of L2 speech production based on the work of Levelt (1989, 1999) and De Bot (1992), only one fluency vulnerability point exists at the articulation stage, whereas multiple fluency vulnerability points exist among the stages involved in formulation (e.g., grammatical encoding, morpho-phonological encoding). Thus, quicker and more robust improvements in mean syllable duration as opposed to measures that include pausing may represent the fact that fewer stages or processes are involved. Varying trends within breakdown fluency (e.g., gains in mean pause duration within ASU were evident by in-sojourn 1, but gains for the number of silent and filled pauses were slower to develop) may indicate that different measurements of breakdown fluency are connected to different fluency vulnerability points related to formulation. Alternatively, the fact that there are multiple ways to evidence breakdown could also explain why there are different developmental trends. It may be that a learner shows evidence of breakdown more often with one type versus another as a result of individual differences (e.g., preferring use of filled pauses over longer silent pauses), but over time the constellation of those patterns changes as development occurs. A thorough analysis of each learner's amount and type of breakdown fluency would be needed to investigate this possibility further and may be a promising direction for future research.

When considering speech rate and mean length of run, which have been used as measures of fluency in a wide variety of previous research, it is important to recall that these measures include information about pausing (breakdown fluency) either directly or indirectly. For example, the calculation for speech rate is based on total speaking time, including pausing, and mean length of run is often based on the number of syllables between two silent pauses. Thus, the number of fluency vulnerability points in speech production increases with these measures compared to articulation rate. In the current study, speech rate and mean length of run had similar developmental trajectories and ones that were different from mean syllable duration. The results of the current study appear to suggest that speech rate and mean length of run might be better categorized as composite measures rather than part of the speed fluency dimension. Further evidence for the composite nature of these two measures can be seen if the relationship among the measures of fluency, as indicated by correlations, is considered, as has been done in previous research (Bosker et al., 2013 and Préfontaine et al., 2016). In the current data, a correlation analysis using fluency measurements from the pre-sojourn data collection point (see Appendix D) indicated that speech rate correlated with most other measures of fluency, except average silent pause between ASU and measures of repair. Similar results were found for mean length of run, whereas mean syllable duration only correlated with filled pauses per second (in addition to mean length of run and speech rate). Thus, these composite measures (speech rate and mean length of run) appear to be more useful as general measures of fluency development. Considering them as a separate category and not a part of speed fluency will likely be more informative when exploring the connections between utterance fluency and L2 speech production systems.

Results of the current study for the four measures of breakdown fluency also show different developmental trajectories. For example, comparing measures that included information about location of the breakdown, differences were found such that within ASUs, improvement occurred by in-sojourn 1 and gains were maintained through post-sojourn 2, whereas there was little change over time between ASUs. Yet another pattern was found with those measures that only included information about frequency (number of silent/filled pauses per second). Again, when considering the relationship between measures of breakdown fluency via a correlation analysis of data at pre-sojourn (see Appendix D), results indicated few correlations between these measures. In fact, the only breakdown fluency measures that correlated were the mean silent pause duration between ASUs and the mean silent pause duration within ASUs (*r* = .456) and the number of filled pauses (*r* = –.509), both medium correlations. Given the fact that these breakdown measures neither correlated nor developed similarly, that might be evidence that they should not necessarily be categorized together. In fact, recent research (Skehan et al., 2016) has argued for a different way of approaching the dimensions of fluency such that instead of a three-way differentiation between speed, breakdown, and repair, differences should be considered between discourse-level and clause-level, which Skehan et al. proposed connect to conceptualization and formulation/articulation, respectively, with speed remaining its own dimension. However, given the operationalization of measurements in Skehan et al. and the current study, it appears that a combination of these ways of approaching fluency might make the most sense: speed fluency is represented by mean syllable duration/inverse articulation rate and connects to issues of articulation; instead of a single breakdown fluency dimension, information about location should necessarily be included to differentiate between those pauses (silent and filled) that occur within and between clause boundaries, which would then connect to formulation and conceptualization, respectively. Repair fluency is represented by measures of repetition and correction and may be connected to processes of self-monitoring. Finally, measures of speech rate and mean length of run represent composite dimensions which combine information from multiple sub-dimensions. Of course, this is an area that requires future investigation, but it appears that including information about location (clausal and phrasal) is critical for better understanding breakdown fluency measurements.

Another consideration when interpreting the longitudinal results of the current study is that the participants in this data set share a single L1 and L2. Some research has indicated that crosslinguistic differences between an L1 and L2 might influence the potential gains a learner may be able to make (see e.g., De Jong et al., 2015; Huensch & Tracy-Ventura, 2016; Riazantseva, 2001). For example, Riazantseva (2001) demonstrated crosslinguistic differences between the pausing patterns of native speakers of Russian and English. Huensch & Tracy-Ventura (2016) reported significant differences between native speakers of English, French, and Spanish on a few measures of fluency (e.g., mean syllable duration) and demonstrated that for those measures where differences existed between native speakers, L1-L2 language pairing was a significant predictor in a regression model predicting L2 fluency from L1 fluency, L1-L2 language pairing, and proficiency, but only after 5 month's residence abroad. To fully answer questions related to how different language pairings might influence the potential gains learners can make, more work is needed that compares different L1 and L2 language pairings in various contexts and at various levels of proficiency.

CONCLUSION

One of the goals of this study was to investigate how different aspects of fluency develop over time. Recall that Bosker et al. (2013) showed different predictive values of speed, breakdown, and repair fluency. The current study has demonstrated different development and retention patterns for these different sub-dimensions of fluency. Combined, the evidence suggests that there are different sub-dimensions of fluency which represent different speech processes. Thus, the trend evident in the current data is that those aspects of fluency which improve within the first 3 months abroad (e.g., mean syllable duration and mean length of pause within ASU) are those that are retained even 8 months after returning home. In contrast, those which show early signs of attrition are the ones which are slower to develop. In sum, it appears that some fluency improvements are more robust and less likely to be affected by the change in context (study abroad vs. home country).

Results of the current study have implications for study abroad research and for instruction post-study/residence abroad. First, although the results suggest that gains in oral fluency happened quite quickly and peaked within 5 months, the results also showed that continued time abroad is likely needed to maintain those gains. In other words, a semester abroad is good for improving oral fluency, and staying for even longer is beneficial as well.

In the current study a variety of utterance fluency measures were adopted, two which can be considered composite measures (speech rate and mean length of run) and seven that more directly represent sub-dimensions of speed, breakdown, and repair fluency. While these composite measures are helpful for certain lines of inquiry, they mask the interesting results that are uncovered by investigating articulation rate and pausing behavior on their own. Thus, in future research focusing on fluency development, we would like to suggest that, at least for now until more detailed work has been carried out on the contribution of the numerous utterance fluency measures, that researchers include both kinds of measures and consider location of pauses (clausal and phrasal) as well. Obviously, it will be more labor intensive but we believe it will be beneficial in the long run for understanding the development of L2 speech production. Additionally, one limitation of the current study is that it was based on planned, semi-guided, elicited oral data (i.e., picture-based narration tasks). Future research should compare the development of oral fluency in non-planned, spontaneous speech using these and other utterance measures to corroborate current findings.

Regarding implications for instruction post-residence abroad, the findings of the two delayed posttests administered in this study showed that participants began to show different rates of attrition once they returned to their home university depending on the measure of fluency analyzed. Some aspects of their fluency were attriting slowly (e.g., speed fluency), whereas others were affected much quicker (e.g., number of filled and silent pauses). As mentioned in the methodology section, as Spanish degree students, the LANGSNAP participants were required to take at least 3 hours of Spanish language classes per week, in addition to content classes related to their degree (e.g., history of Spain). The content classes were rarely taught in the target language; thus, at university participants were likely speaking little Spanish which could have affected their fluency. In interviews conducted during the same time, participants mentioned maintaining contact with friends made abroad and host family members via social media and Skype. They also mentioned befriending some Erasmus and other exchange students who came to their UK university. Therefore, they continued to have other types of L2 contact in addition to formal coursework. Whether the total amount of L2 contact during their final year at university was less than their year abroad is an empirical question that is worthy of future research. In LANGSNAP, language contact data were only collected while the participants were abroad. In our current work we are investigating to what extent language contact and social networking predict gains made in fluency during their stay abroad. More research is also needed that focuses on students' experiences in language classes after returning from residence/study abroad and how their fluency (as well as accuracy and complexity) are affected depending on the classes they take. Examining learners' motivations for maintaining the gains made abroad as well as their self-perceptions of language maintenance would be interesting areas to explore in future research to better understand to what extent we are meeting the needs of these students in order to maintain the gains they made while abroad.

NOTES

[1] Because of the variation in narration time, we tested whether participants who spoke longer had more disfluencies as they searched for more to say. The first half of the file and the second half of the file were compared for anyone who spoke longer than 4 ½ minutes. Results of non-parametric Wilcoxon Signed Rank tests indicated no significant differences between the samples on any of the measures of fluency. Thus, speaking longer did not result in more disfluent behavior.

[2] For all post hoc tests, only significant differences are presented in separate tables in the text. The complete tables including pairwise comparisons for all rounds (significant and nonsignificant results) can be found in Appendices A, B, and C.

REFERENCES

Baker-Smemoe, W., Dewey, D. P., Bown, J., & Martinsen, R. A. (2014). Does measuring L2 utterance fluency equal measuring overall L2 proficiency? Evidence from five languages. *Foreign Language Annals*, *47*, 707–728.

Boersma, P., & Weenink, D. (2015). Praat: Doing phonetics by computer (Version 5.4.18) [Computer program]. Retrieved from http://www.praat.org/

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

Bowden, H. W. (2016). Assessing second language oral proficiency for research. *Studies in Second Language Acquisition*, 1–29.

Collentine, J. G. (2009). Study abroad research: Findings, implications and future directions. In M. H. Long & C. J. Catherine (Eds.), *The handbook of language teaching* (pp. 218–233). New York: Wiley.

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America, 107,* 989–999.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, *111*, 2862-2873.

De Bot, K. (1992). A bilingual production model: Levelt's "Speaking" model adapted. *Applied Linguistics, 13,* 1–24.

De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching, 54,* 113-132.

De Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In R. Eklund (Ed.) *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech (DiSS),* 17–20. Stockholm, Sweden: Royal Institute of Technology.

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

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. N. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics, 34,* 893–916.

Di Silvio, F., Diao, A., & Donovan, W. (2016). The development of L2 fluency during study abroad: A cross-language study. *Modern Language Journal, 100,* 610–624.

Dominguez, L., Tracy-Ventura, N., Arche, M., Mitchell, R., & Myles, F. (2013). The role of dynamic contrasts in the L2 acquisition of Spanish past tense morphology. *Bilingualism: Language and Cognition, 16,* 558–577.

Du, H. (2013). The development of Chinese fluency during study abroad in China. *Modern Language Journal*, *97*, 131–143.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language.* Oxford: Oxford University Press.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London: Sage.

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

Freed, B. F. (1995). Do students who study abroad become fluent? In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 123–148). Philadelphia, PA: John Benjamins.

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*, 275–301.

García–Amaya, L. (2009). New findings on fluency measures across three different learning contexts. In J. Collentine, M. García, B. Lafford, & F. Marcos–Marín (Eds.), *Selected proceedings of the 11th Hispanic linguistics symposium* (pp. 68–80). Somerville, MA: Cascadilla Press.

Huensch, A., & Tracy-Ventura, N. (2016). Understanding second language fluency behavior: The effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. *Applied Psycholinguistics*, pp. 1–31. DOI: https://doi.org/10.1017/S0142716416000424

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning, 64,* 809–854.

Kim, J., Dewey, D. P., Baker–Smemoe, W., Ring, S., Westover, A., & Eggett, D. L. (2015). L2 development during study abroad in China. *System, 55*, 123–133.

Kinginger, C. (2009). *Language learning and study abroad: A critical reading of research.* Basingstoke, UK: Palgrave Macmillan.

Kormos, J. (2000). The role of attention in monitoring second language speech production. *Language Learning, 50*, 343–384.

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

Langley, J. (2006). *I Very Really Miss You.* (illustrations). London: Francis Lincoln.

Larson–Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

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

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

Levelt, W. (1999). Producing spoken language: A blueprint of the speaker. In C. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 83–122). Oxford: Oxford University Press.

Llanes, À. (2011). The many faces of study abroad: An update on the research on L2 gains emerged during a study abroad experience. *International Journal of Multilingualism, 8,* 189–215.

Mitchell, R., McManus, K., & Tracy-Ventura, N. (2015). Comparison of language development during different residence abroad programmes. In R. Mitchell, N. Tracy-Ventura, & K. McManus (Eds.), *Social interaction, identity and language learning during residence abroad* (pp.115–138). EUROSLA Monograph Series 4.

Mitchell, R., Tracy-Ventura, N., & McManus, K. (2017). *The Anglophone student abroad: Identity, social relationships and language learning*. New York: Routledge.

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum.

Mora, J. C., & Valls–Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly*, *46*, 610–641.

Ortega, L. (2000). *Understanding syntactic complexity: The measurement of change in the syntax of instructed L2 Spanish learners*. (Unpublished doctoral dissertation). University of Hawaii, Honolulu, HI.

Pawley, A., & Syder, F. (2000). The one clause at a time hypothesis. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 163–191). Ann Arbor, MI: The University of Michigan Press.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.

Préfontaine, Y., Kormos, J., & Johnson, D. E. (2016). How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing, 33,* 53-73.

Riazentseva, A. (2001). Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition, 23,* 497–526.

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.

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*, 173–199.

Serrano, R., Llanes, À., & Tragant, E. (2011). Analyzing the effect of context of second language learning: domestic intensive and semi-intensive courses vs. study abroad in Europe. *System*, *39*, 133–143.

Serrano, R., Tragant, E., & Llanes, À. (2012). A longitudinal analysis of the effects of one year abroad. *The Canadian Modern Language Review*, *68*, 138–163.

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

Skehan, P. (2014). The context for researching a processing perspective on task performance. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 2–36). Philadelphia/Amsterdam: John Benjamins.

Skehan, P., Foster, P., & Shum, S. (2016). Ladders and snakes in second language fluency. *International Review of Applied Linguistics in Language Teaching*, *54*, 97–111.

SPSS (2013). IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching*, *54*, 133–150.

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. 239–276). Philadelphia/Amsterdam: John Benjamins.

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

Tracy-Ventura, N., Mitchell, R., & McManus, K. (2016). The LANGSNAP longitudinal learner corpus: Design and use. In M. Alonso Ramos (Ed.), *Spanish learner corpus research: State of the art* (pp. 117–142). Philadelphia/Amsterdam: John Benjamins.

APPENDIX A

Post Hoc Pairwise Comparisons and Their Associated *p* Values and Effect Sizes for the Measure of Speed Fluency

		Pairwise Comparison	
Comparison	Mean Syllable Duration	Speech Rate	Mean Length of Run
Pre-sojourn vs. In-sojourn 1	$z = 5.09$, $p < .001$, $d = 1.60$	$z = -3.24$, $p = .018$, $d = 1.36$	$z = -2.86$, $p = .065$, $d = 0.89$
Pre-sojourn vs. In-sojourn 2	$z = 5.02$, $p < .001$, $d = 1.45$	$z = -6.12$, $p < .001$, $d = 1.62$	$z = -6.33$, $p < .001$, $d = 1.25$
Pre-sojourn vs. In-sojourn 3	$z = 5.63$, $p < .001$, $d = 1.72$	$z = -6.33$, $p < .001$, $d = 1.70$	$z = -4.86$, $p < .001$, $d = 1.30$
Pre-sojourn vs. Post-sojourn 1	$z = 7.72$, $p < .001$, $d = 1.85$	$z = -7.33$, $p < .001$, $d = 1.77$	$z = -7.33$, $p < .001$, $d = 1.41$
Pre-sojourn vs. Post-sojourn 2	$z = 4.32$, $p < .001$, $d = 1.37$	$z = -4.71$, $p < .001$, $d = 1.45$	$z = -4.94$, $p < .001$, $d = 1.16$
In-sojourn 1 vs. In-sojourn 2	$z = -0.08$, $p = .939$, $d = 0.08$	$z = -2.93$, $p = .051$, $d = 0.38$	$z = -3.47$, $p = .008$, $d = 0.55$
In-sojourn 1 vs. In-sojourn 3	$z = 0.54$, $p = 1.000$, $d = 0.14$	$z = -3.09$, $p = .030$, $d = 0.40$	$z = -3.01$, $p = .039$, $d = 0.53$
In-sojourn 1 vs. Post-sojourn 1	$z = 2.62$, $p = .131$, $d = 0.34$	$z = -4.09$, $p = .001$, $d = 0.52$	$z = -4.48$, $p < .001$, $d = 0.69$
In-sojourn 1 vs. Post-sojourn 2	$z = -0.77$, $p = 1.000$, $d = 0.22$	$z = -1.47$, $p = 1.000$, $d = 0.22$	$z = -2.08$, $p = .559$, $d = 0.39$
In-sojourn 2 vs. In-sojourn 3	$z = 0.62$, $p = 1.000$, $d = 0.20$	$z = -0.15$, $p = 1.000$, $d = 0.00$	$z = 0.46$, $p = 1.000$, $d = 0.06$
In-sojourn 2 vs. Post-sojourn 1	$z = 2.70$, $p = .104$, $d = 0.38$	$z = -1.16$, $p = 1.000$, $d = 0.13$	$z = -1.00$, $p = 1.000$, $d = 0.10$
In-sojourn 2 vs. Post-sojourn 2	$z = -0.69$, $p = 1.000$, $d = 0.12$	$z = 1.47$, $p = 1.000$, $d = 0.14$	$z = 1.39$, $p = 1.000$, $d = 0.18$
In-sojourn 3 vs. Post-sojourn 1	$z = 2.08$, $p = .559$, $d = 0.21$	$z = -1.00$, $p = 1.000$, $d = 0.13$	$z = -1.47$, $p = 1.000$, $d = 0.17$
In-sojourn 3 vs. Post-sojourn 2	$z = -1.31$, $p = 1.000$,	$z = 1.62$, $p = 1.000$,	$z = 0.93$, $p = 1.000$,

	$d = 0.35$	$d = 0.15$	$d = 0.13$
Post-sojourn 1 vs. Post-sojourn 2	$z = -3.40$, $p = .010$, $d = 0.53$	$z = 2.62$, $p = .131$, $d = 0.27$	$z = 2.39$, $p = .252$, $d = 0.29$

* Significant differences at the $p < .05$ level.

APPENDIX B

Post Hoc Pairwise Comparisons and Their Associated *p* Values and Effect Sizes for the Measures of Mean Silent Pause Duration Within and Between ASUs

	Pairwise Comparison	
Comparison	Mean Silent Pause Duration Within ASU	Mean Silent Pause Duration Between ASU
Pre-sojourn vs. In-sojourn 1	$z = 4.63$, $p < .001$, $d = 1.23$	$z = -1.23$, $p = 1.000$, $d = 0.19$
Pre-sojourn vs. In-sojourn 2	$z = 4.48$, $p < .001$, $d = 1.24$	$z = -0.69$, $p = 1.000$, $d = 0.09$
Pre-sojourn vs. In-sojourn 3	$z = 4.24$, $p < .001$, $d = 1.36$	$z = 2.32$, $p = .310$, $d = 0.59$
Pre-sojourn vs. Post-sojourn 1	$z = 5.79$, $p < .001$, $d = 1.60$	$z = 0.46$, $p = 1.000$, $d = 0.15$
Pre-sojourn vs. Post-sojourn 2	$z = 3.55$, $p = .006$, $d = 1.18$	$z = 1.46$, $p = 1.000$, $d = 0.27$
In-sojourn 1 vs. In-sojourn 2	$z = -0.15$, $p = 1.000$, $d = 0.08$	$z = 0.54$, $p = 1.000$, $d = 0.09$
In-sojourn 1 vs. In-sojourn 3	$z = -0.39$, $p = 1.000$, $d = 0.18$	$z = 3.55$, $p = .006$, $d = 0.80$
In-sojourn 1 vs. Post-sojourn 1	$z = 1.16$, $p = 1.000$, $d = 0.49$	$z = 1.70$, $p = 1.000$, $d = 0.35$
In-sojourn 1 vs. Post-sojourn 2	$z = -1.08$, $p = 1.000$, $d = 0.04$	$z = 2.70$, $p = .104$, $d = 0.46$
In-sojourn 2 vs. In-sojourn 3	$z = -0.23$, $p = 1.000$, $d = 0.08$	$z = 3.01$, $p = .039$, $d = 0.63$
In-sojourn 2 vs. Post-sojourn 1	$z = 1.31$, $p = 1.000$, $d = 0.37$	$z = 1.16$, $p = 1.000$, $d = 0.23$
In-sojourn 2 vs. Post-sojourn 2	$z = -0.93$, $p = 1.000$, $d = 0.04$	$z = 2.16$, $p = .461$, $d = 0.33$
In-sojourn 3 vs. Post-sojourn 1	$z = 1.54$, $p = 1.000$, $d = 0.31$	$z = -1.85$, $p = .961$, $d = 0.46$
In-sojourn 3 vs. Post-sojourn 2	$z = -0.69$, $p = 1.000$, $d = 0.12$	$z = -0.85$, $p = 1.000$, $d = 0.31$
Post-sojourn 1 vs. Post-sojourn 2	$z = -2.24$, $p = .379$, $d = 0.39$	$z = 1.00$, $p = 1.000$, $d = 0.13$

* Significant differences at the $p < .05$ level.

APPENDIX C

Post Hoc Pairwise Comparisons and Their Associated p Values and Effect Sizes for the Measures of the Number of Silent Pauses per Second and the Number of Filled Pauses per Second

	Pairwise Comparison	
Comparison	Number of Silent Pauses per Second	Number of Filled Pauses per Second
Pre-sojourn vs. In-sojourn 1	$z = 0.15, p = 1.000,$ $d = 0.20$	$z = 2.59, p = .146,$ $d = 0.65$
Pre-sojourn vs. In-sojourn 2	$z = 4.09, p = .001,$ $d = 0.77$	$z = 5.09, p < .001,$ $d = 1.10$
Pre-sojourn vs. In-sojourn 3	$z = 3.47, p = .008,$ $d = 0.65$	$z = 6.17, p < .001,$ $d = 1.26$
Pre-sojourn vs. Post-sojourn 1	$z = 3.78, p = .002,$ $d = 0.67$	$z = 2.82, p = .073,$ $d = 0.72$
Pre-sojourn vs. Post-sojourn 2	$z = 2.39, p = .252,$ $d = 0.51$	$z = 1.85, p = .961,$ $d = 0.65$
In-sojourn 1 vs. In-sojourn 2	$z = 3.94, p = .001,$ $d = 0.58$	$z = 2.51, p = .182,$ $d = 0.50$
In-sojourn 1 vs. In-sojourn 3	$z = 3.32, p = .014,$ $d = 0.47$	$z = 3.59, p = .005,$ $d = 0.68$
In-sojourn 1 vs. Post-sojourn 1	$z = 3.63, p = .004,$ $d = 0.50$	$z = 0.23, p = 1.000,$ $d = 0.10$
In-sojourn 1 vs. Post-sojourn 2	$z = 2.24, p = .379,$ $d = 0.33$	$z = -0.73, p = 1.000,$ $d = 0.01$
In-sojourn 2 vs. In-sojourn 3	$z = -0.62, p = 1.000,$ $d = 0.07$	$z = 1.08, p = 1.000,$ $d = 0.20$
In-sojourn 2 vs. Post-sojourn 1	$z = -0.31, p = 1.000,$ $d = 0.03$	$z = -2.28, p = .343,$ $d = 0.36$
In-sojourn 2 vs. Post-sojourn 2	$z = -1.70, p = 1.000,$ $d = 0.20$	$z = -3.24, p = .018,$ $d = 0.53$
In-sojourn 3 vs. Post-sojourn 1	$z = 0.31, p = 1.000,$ $d = 0.04$	$z = -3.36, p = .012,$ $d = 0.53$
In-sojourn 3 vs. Post-sojourn 2	$z = -1.08, p = 1.000,$ $d = 0.13$	$z = -4.32, p < .001,$ $d = 0.72$
Post-sojourn 1 vs. Post-sojourn 2	$z = -1.39, p = .379,$ $d = 0.16$	$z = -0.96, p = 1.000,$ $d = 0.12$

* Significant differences at the $p < .05$ level.

APPENDIX D

Correlations Between Fluency Measures at the Pre-Sojourn

		Mean Syllable Duration	Mean Length of Run	Speech Rate	Mean Silent Pause Between ASU	Mean Silent Pause Within ASU	Number of Silent pauses	Number of Filled Pauses	Number of Repetitions	Number of Corrections
Speed Fluency	Mean Syllable Duration	1.00	–0.772*	–0.912*	–.096	.402	.361	0.521*	.153	–.293
	Mean Length of Run		1.00	0.889*	–.083	–0.521*	–0.777*	–.381	–.157	.043
	Speech rate			1.00	–.144	–0.627*	–0.559*	–0.427*	–.054	.302
Breakdown Fluency	Mean Silent Pause Between ASU				1.00	0.456*	.074	–0.509*	–.245	–.333
	Mean Silent Pause Within ASU					1.00	.295	–.050	–.222	–.327
	Number of Silent pauses						1.00	.361	.108	.147
	Number of Filled Pauses							1.00	0.415*	.144
Repair Fluency	Number of Repetitions								1.00	0.677*

Number of Corrections	1.00

* Significant at the $p < .05$ level.