Second language fluency: speaking style or proficiency? Correcting measures of second language fluency for first language behavior

de Jong, N.H.; Groenhout, R.; Schoonen, R.; Hulstijn, J.H.

DOI
10.1017/S0142716413000210

Publication date
2015

Document Version
Final published version

Published in
Applied Psycholinguistics

Applied Psycholinguistics **36** (2015), 223–243
doi:10.1017/S0142716413000210

Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior

NIVJA H. DE JONG
Utrecht University

RACHEL GROENHOUT, ROB SCHOONEN, and JAN H. HULSTIJN
University of Amsterdam

Received: April 1, 2012 Accepted for publication: October 22, 2012

ADDRESS FOR CORRESPONDENCE
Nivja de Jong, Utrecht Institute of Linguistics OTS, Utrecht University, Trans 10, Utrecht 3512 JK, The Netherlands. E-mail: n.dejong@uu.nl

ABSTRACT
In second language (L2) research and testing, measures of oral fluency are used as diagnostics for proficiency. However, fluency is also determined by personality or speaking style, raising the question to what extent L2 fluency measures are valid indicators of L2 proficiency. In this study, we obtained a measure of L2 (Dutch) proficiency (vocabulary knowledge), L2 fluency measures, and fluency measures that were corrected for first language behavior from the same group of Turkish and English native speakers ($N = 51$). For most measures of fluency, except for silent pause duration, both the corrected and the uncorrected measures significantly predicted L2 proficiency. For syllable duration, the corrected measure was a stronger predictor of L2 proficiency than was the uncorrected measure. We conclude that for L2 research purposes, as well as for some types of L2 testing, it is useful to obtain corrected measures of syllable duration to measure L2-specific fluency.

Imagine two immigrants who have learned to speak Dutch as their second language (L2), let us say Oscar and Mark. Both have acquired Dutch after adolescence and find themselves at an intermediate level of oral proficiency. However, they happen to differ in their level of speaking fluency: Compared to other speakers, Mark seems to use many filled pauses ("uhms") in his L2 speech, whereas Oscar only uses a few filled pauses. Should we therefore conclude that Oscar and Mark operate at different levels of oral proficiency, evidenced by the different levels of speaking fluency? Now imagine that we also know how they speak in their native (first) language (L1). It turns out that Oscar rarely uses a filled pause in his L1 speech, whereas Mark exhibits many filled pauses in his L1. Can we still conclude that

the two speakers operate at two different levels of L2 oral proficiency? Should we take L1 behavior into account to gauge L2-specific fluency? It seems unlikely that someone like Mark, who uses many filled pauses in his L1, will speak with few filled pauses in his L2 (or any language).

This paper deals with the question whether L2 measures of oral fluency, such as number of filled pauses, should be adjusted for L1 fluency behavior to reflect L2-specific processing. It may be the case that the original L2 measures of fluency reflect a combination of personal speaking style and L2-specific skills and that correcting the original measures for L1 behavior will lead to more precise measures of L2-specific processing. We argue that if these adjusted measures better reflect L2-specific skills, L2 acquisition research and (some types of) L2 testing stand to benefit from measuring both L1 and L2 fluency measures in order to gauge the L2-specific measures of fluency.

Here we are concerned with fluency as an aspect of overall speaking proficiency, also described as fluency in the narrow sense (Chambers, 1997; Lennon, 1990) and often contrasted with the linguistic complexity and the accuracy of the linguistic forms that speakers use (Housen & Kuiken, 2009). Fluency in the narrow sense is usually described in terms of speedy and smooth delivery of speech without (filled) pauses, repetitions, and repairs. In L2 testing, speaking fluency is a component construct in the evaluation of speaking proficiency. For example, the Common European Framework of Reference for Languages takes fluency as a component for describing overall proficiency (Council of Europe, 2001). The descriptors in the global scale (p. 24) state that speakers at level B2 can communicate "with a degree of fluency"; at C1 speakers can express themselves "fluently" and at C2, "very fluently."

In the broad sense of fluency (overall global proficiency), Fillmore (1979) describes how native speakers can show individual differences. The narrow notion of fluency is almost exclusively used for nonnative speakers of a language (with the exception of individuals with speech disorders). Native speakers supposedly talk fluently by default (Riggenbach, 1991). However, differences between native speakers with respect to temporal aspects of speech have also been documented (e.g., Goldman-Eisler, 1968; Shriberg, 1994). Moreover, these differences between individuals have been shown to relate to individual characteristics such as extraversion (Ramsay, 1968). This leads to the question: Is it at all valid to evaluate nonnative speakers with respect to differences in L2 fluency?

This question has lately become even more relevant because the differences with respect to temporal aspects of speech that exist between native speakers have been shown to carry over to similar differences in an L2. Derwing, Munro, Thomson, & Rossiter (2009) explored to what extent L1 fluency measures are related to L2 fluency measures for Slavic and Mandarin speakers of English. Sixteen Mandarin and 16 Slavic speakers performed the same task in their L1 and in their L2 (English). Derwing et al. found that there was a significant correlation between the L1 and L2 behavior with respect to number of pauses per second, speech rate, and pruned syllables per second.

Because L1 speech is not fluent by default and because L2 fluency behavior is partly carried over from L1 fluency behavior, one could argue that when judging L2

speakers on their fluency, only those disfluencies that are related to L2 proficiency and automaticity of L2 processing should play a role. Segalowitz (2010) adopts Levelt's (Levelt, 1989; Levelt, Roelofs, & Meyer, 1999) "blueprint" of the speaker and uses De Bot's (1992) remarks regarding the bilingual speaker to indicate specific points during the speech process at which an L2 speaker is disfluent due to developing lexical and grammatical L2 knowledge and skills (see also Kormos, 2006). In this cognitive viewpoint, L2 disfluencies come about because the formulation and encoding of speech lags behind the articulation of previously formulated speech (Howell & Au-Yeung, 2002). Because of nonnative speakers' slower processing during formulation and articulation, and due to developing lexical and grammatical knowledge in the L2, nonnative speakers are more likely to lag behind in formulating speech than are native speakers; they are therefore likely to display more disfluencies in speech.

Segalowitz (2010) proposes that to measure aspects of L2-specific fluency, one should gather both L1 data and L2 data to take speakers' L1 fluency into account. He proposes calculating corrected fluency measures by partialing out the L1 variance from the L2 measures. In this way, disfluencies that are specifically related to the use of an L2 are distinguished from disfluencies as they appear in L1 speech. If it is the case that such corrected scores of L2 fluency better reflect disfluencies that L2 speakers exhibit because of L2-specific difficulties in formulating and articulating L2 speech, it follows that such corrected scores are better predictors of L2 proficiency than are the uncorrected measures.

One should note that, under this view, cognitive sources of L2 fluency are considered. This is what Segalowitz (2010) calls *cognitive fluency*: the ability of the speaker to smoothly translate thoughts to speech. However, this ability cannot be measured directly. Therefore, researchers use measures of *utterance fluency* to gauge speech-planning difficulties that surface in utterances by counting the number of filled pauses, corrections, and repairs, and by measuring the duration of pauses. Yet another sense of fluency is *perceived fluency*, which pertains to the inference listeners (raters) make on the basis of the utterance about speakers' ability (about speakers' cognitive fluency).

THE PRESENT STUDY

In this study, we focused on L2 *utterance fluency* and operationalized it in two ways: uncorrected measures and corrected measures that are adjusted for L1 behavior. We then related both types of objective measures to a measure of L2 proficiency (vocabulary knowledge) to find out whether the corrected measures better reflect L2 cognitive fluency as indicators of L2 speaking proficiency. To obtain the corrected measures, we saved the residuals from models predicting L2 measures from L1 measures; in doing so, we ascertained to what extent L2 fluency behavior is related to L1 fluency behavior.

Finally, to be able to generalize our results to different language groups, we tested two typologically different languages as L1 and Dutch as an L2. The two different L1s were chosen to be typologically close (English) and typologically distant (Turkish) from the L2 (Dutch). This allowed us to determine possible differences in the relation between L1 and L2 fluency behavior for the English

native speakers compared to the Turkish native speakers. It might be the case that the relation between L1 and L2 fluency is different for different L1s and L2s—for instance, because the two L1s show cross-linguistic differences. There is little research investigating such cross-linguistic differences. One exception is Riazantseva (2001) who studied cross-linguistic differences between Russian and English. In a study investigating the fluency behavior of 14 intermediate and 16 advanced learners of English (with L1 Russian), she also compared the results of a control group of 20 English native speakers who performed L1 speaking tasks with results from the Russian participants who performed similar tasks in their L1. She found that pause durations in L1 Russian were on average longer than in L1 English.

The previous discussion leads to the following three research questions:

1. To what extent can different measures of L2 fluency (e.g., length of pauses or speed of speech) be predicted from the equivalent measures in L1?
2. Are L2 fluency measures that are corrected for L1 fluency behavior better predictors of overall L2 proficiency than are uncorrected L2 measures?
3. Is the predictive value of (corrected) measures of L2 fluency dependent on typological distance between L1 and L2?

To answer these research questions, native speakers of Turkish or English performed tasks in their L1 and very similar tasks in their L2, Dutch. Because we wanted to exclude possible repetition effects by using the same tasks, the tasks in the L1 and the L2 were constructed in such a way that they would be maximally comparable, without any literal repetitions. We chose to use many (eight) tasks in each language to ensure a large enough sample of L1 and L2 speech data for each participant. In this way, we could measure L2 fluency (related to L2 proficiency) and L1 fluency (reflecting personal speaking style).

Several measures of fluency were taken from the L1 and L2 speaking performances. Skehan (2003) and Tavakoli and Skehan (2005) noted that fluency has several aspects: breakdown fluency, speed fluency, and repair fluency. In this study, measures of fluency were chosen such that they were minimally related to other measures and such that they measured only one aspect at a time. Previous research has often reported on measures such as speech rate or pruned syllables per second (e.g., Derwing et al., 2009; Freed, 1995; Lennon, 1990; Riggenbach, 1991; Towell, 2002). These are global measures of fluency that capture several aspects at once. For instance, speech rate is measured as number of syllables divided by total time, including silent pausing time, and therefore incorporates speed of speech and pausing in speech at the same time. In this study, theoretically unrelated measures were chosen to investigate possible differences between these measures (with respect to the research questions). For *breakdown fluency*, the number and length of silent pauses were measured, as well as the number of nonlexical filled pauses. For *speed fluency*, the mean duration of syllables was measured. Finally, for *repair fluency*, the number of repetitions and the number of corrections were measured.

To estimate L2 proficiency separately, all participants completed an L2 productive vocabulary test.

METHOD

Participants

Twenty-nine native speakers of English and 24 native speakers of Turkish were paid to take part in our experiment. This research was part of a larger project with, for some participants, more tasks than are reported on here. Depending on how many tasks the participants completed, they were paid between 30 and 50 euros. The Turkish (7 male, 17 female; mean age = 32, range = 23–48) and English participants (11 male, 17 female; mean age = 31, range = 23–43) had come to the Netherlands between the ages of 18 and 40 (English range = 22–40, Turkish range = 18–35). Most participants had lived in the Netherlands for fewer than 10 years (English mean = 4.5 years, range = 1 month to 21 years; Turkish mean = 7 years, range = 9 months to 20 years). All participants were at an intermediate to advanced level of Dutch as an L2 and most participants were taking intermediate or advanced level Dutch courses to prepare for enrollment at the University of Amsterdam.

Vocabulary task

Because we wanted overall L2 proficiency to be assessed separately from the speaking performances, we chose to use a productive vocabulary task for this measurement. Vocabulary knowledge has been shown to be a good predictor of overall proficiency (Beglar & Hunt, 1999; Zareva, Schwanenflugel, & Nikolova, 2005). Moreover, in two recent papers, the same vocabulary test as used in the present study has been shown to be a strong predictor of overall speaking proficiency. In De Jong, Steinel, Florijn, Schoonen, and Hulstijn (2012a), using structural equation modeling, the vocabulary knowledge score showed a strong relation to ratings of overall speaking proficiency ($r = .79$). In Hulstijn, Schoonen, De Jong, Steinel, and Florijn (2012), speakers from De Jong et al. (2012a) who were operating at B1 and B2 levels, based on Common European Framework of Reference for Languages ratings (Council of Europe 2001) of overall speaking proficiency, were selected. In that study, the vocabulary scores were strongly related (between $r = .79$ and $r = .89$) to the discriminant function that could differentiate between B1 and B2 performances on the different tasks that were used.

The paper-and-pencil task (with instructions in the L2, Dutch) elicited knowledge of single words (90 items) and of multiword units (26 items). We used the format suggested by Laufer and Nation (1999): For each item a meaningful sentence was presented with the target word omitted, except for its first letter(s). When alternative words beginning with the same letter could also be appropriately used, more letters were given. Although this format capitalizes on testing productive knowledge of the target word, comprehension skills are also tested. To maximize testing productive knowledge of the targeted word, most words in the carrier sentence were chosen such that their frequency was higher than the frequency of the omitted target word.

The multiword units consisted of prepositional phrases and verb–noun collocations; the preposition or main verb was omitted and the gap had to be filled in. For

these multiword items, no first letter(s) were given, because the carrier sentences always narrowed down the possible candidates to the target word(s).

The total score for each participant was calculated as the total number of correct responses. Spelling mistakes and inflectional variants were counted as correct answers. There was no time pressure to do the vocabulary task and participants took between 20 min and 1 hr to finish the task.

Speaking tasks

Materials. To elicit L2 speech, we used the eight speaking tasks as described in De Jong et al. (2012a) and Hulstijn et al. (2012). The speaking tasks in the two languages were matched to each other on three parameters: complexity (simple vs. complex), formality (informal vs. formal), and discourse mode (descriptive vs. persuasive). De Jong et al. (2012a) operationalized this three-way distinction in task type by the contents of the task and by the instructions given in each task. These instructions contained specific information about the speaking task itself, which was provided by one or several visual-verbal cues. No additional knowledge about the topic beyond the information provided in the tasks was needed to successfully complete each speaking task. Information about the purpose and audience of the task was also provided.

Mirroring each Dutch task, we constructed a similar task in English. For example, for the Dutch task in which the participants had to describe a graph depicting unemployment figures for men and women in the last 12 years to a good friend (informal descriptive complex task), we constructed a task in English that required describing a graph depicting house sales in rural and urban areas to a good friend. Appendix A gives a short description of the Dutch and English tasks. Finally, a native speaker of Turkish with a good command of English translated the English tasks to Turkish, discussing the translations with the first and second authors.

Procedure. Participants completed the eight L2 (Dutch) tasks with a native Dutch-speaking experimenter present. The eight L1 tasks were performed with other experimenters present, a Turkish–Dutch bilingual or a native English speaker. The tasks were presented in Authorware (version 7) on a PC laptop, and the participants' speech was recorded. Participants navigated the experiment and instructions with a (computer) mouse. Each task consisted of several screens containing specific information about the task in the L2 for the L2 tasks and in the L1 for the L1 tasks, and pictures giving additional information. For each task, participants had 30 s of preparation time and 120 s of speaking time, which was shown by a status bar at the bottom of the screen. Participants could press a "finished" button if they finished the task before the 120 s had gone by. Due to constraints in using the recording studio that was available, the setting in the Dutch session was slightly different from that used in the English or the Turkish sessions. The L1 tasks were recorded in a recording studio, with the experimenter recording the participant remotely from an observation room, and the L2 session was performed in an office with the experimenter in the same office.

Participants completed both the L1 and the L2 versions of the eight tasks at their own pace with no breaks in between tasks. Total time varied between participants, but the average time used for completing all eight L1 tasks was 20 min, and the

time used for completing all eight L2 tasks was 30 min. Each participant performed all speaking tasks in Dutch in the first session and approximately 1 to 4 weeks later performed all tasks in English or Turkish in a second session. In a separate session, prior to the L1 speaking task session, the participants performed the vocabulary task.

Obtaining fluency measures

All speech recordings (53 participants, eight tasks in L2 Dutch, eight in L1 Turkish or English; totaling 19 hr and 51 min of speech materials) were transcribed and annotated by two research assistants, a native speaker of Turkish who had a good command of both English and Dutch, and a native speaker of Dutch who had a good command of English. They worked in close collaboration with each other to ensure the Turkish, English, and Dutch speaking performances were transcribed, annotated, and measured as similarly as possible. For 2 (English) participants, more than half of the performances in the L2 were not recorded well (interference with the computer caused a strong hum). Therefore, these 2 participants were discarded from further analyses. In addition, eight recordings in the L2 (from 6 participants) were not recorded with sufficient quality to make precise transcriptions. These recordings were also discarded. We then also discarded the eight recordings in the L1 from these 6 participants. In this way, we ensured that for each participant, the amount and type of speech in the L1 and in the L2 would be maximally similar. We thus obtained from 51 participants speech performances from at least six, but for most participants eight, tasks in their L1 and their L2. For the analyses, the six to eight speaking performances in the L1 were combined into one stretch of speech. The same was done for the six to eight speaking performances in the L2. Combining the L1 performances led to a minimum of 4 min of transcribed L1 speech, and combining the performances from L2 tasks resulted in a minimum of 6 min of transcribed L2 speech for each participant. The maximum for both L1 and L2 speech was the maximum allotted time (i.e., 16 min in each language).

The transcriptions were made in CLAN (MacWhinney, 2000). Besides orthographic transcriptions, the two research assistants inserted information relevant for measuring aspects of fluency. Silent pauses were detected by careful listening and by using the waveform (as shown in CLAN), and measured manually in milliseconds. The transcriptions were also split up into so-called analysis of speech units (ASU). Foster, Tonkyn, and Wigglesworth (2000) have shown that using the AS-unit is the optimal way of dividing transcribed data into analyzable units. As defined by Foster et al. (2000, p. 365), an AS-unit is "a single speaker's utterance consisting of an independent clause, or a subclausal unit, together with any subordinate clause(s) associated with either." Silent pauses were categorized as being either between or within ASU. Furthermore, the transcripts were annotated with nonlexical filled pauses (such as "uh," "uhm," "er," "mm"), corrections (false starts, reformulations, and self-corrections), repetitions (repetitions of exact words, syllables, or phrases), and syllable counts.

Using the annotations as described above, fluency measures for three aspects of fluency were computed (by using a GNU Awk-script written for that purpose; http://www.gnu.org/software/gawk). For *speed fluency*, mean syllable duration in L1 and L2 (i.e., inverse articulation rate) was computed by dividing total speaking

time (total time excluding silent pauses) by total number of syllables. *Breakdown fluency* was computed as mean length of silent pauses within and mean length of pauses between ASU per participant, in L1 and in L2. The lower bound of silent pauses was 250 ms (following Goldman-Eisler, 1968). In addition, number of silent pauses per second speaking time and number of nonlexical filled pauses per second speaking time in L1 and L2 were also computed. *Repair fluency* was computed as number of repetitions per second speaking time and number of corrections per second speaking time in L1 and L2. For all frequency measures of fluency (number of silent pauses, filled pauses, repetitions, and repairs), we divided the number by total speaking time, excluding pausing time. In this way, the frequency measures and the duration measures are theoretically less confounded than when they are calculated as divided by total time (including pausing time).

RESULTS

In this section, descriptive statistics of the L2 vocabulary measure for both participant groups (English vs. Turkish native speakers) and of the fluency measures for both groups in both languages (L1 vs. L2) are presented and are tested for differences. Following these analyses of group means, correlations between the measures of fluency and regression analyses are presented to examine how much variance in the L2 fluency measures can be predicted merely on the basis of the same fluency measures in the L1, as well as on the L1 group (English vs. Turkish). Finally, Segalowitz's (2010) claim is tested, and the predictive value of corrected scores of L2 fluency (adjusted for L1 fluency behavior) for L2 proficiency is compared to the predictive value of the uncorrected L2 fluency measures.

Descriptive statistics and group differences

The Turkish and English native speakers performed the vocabulary task equally well ($t < 1$). The Turkish native speakers scored on average 55.1 out of a total score of 116 ($SD = 24.7$; range = 17–103), and the English native speakers scored on average 56.7 ($SD = 22.0$; range = 8–88). Comparing these scores to the scores on the same vocabulary test by speakers rated as being B1 and B2 level speakers in the study by Hulstijn et al. (2012), one can conclude that the Turkish and English native speakers of this study scored within the same range. Table 1 shows the means and standard deviations for both English and Turkish participants in their L1 and L2 (Dutch) for all measures of fluency.

Before running any inferential analyses, it was tested whether the L1 and L2 measures could be assumed to be normally distributed, by running Shapiro–Wilk normality tests. For the measures of duration (syllables duration and silent pause duration within ASU and between ASU), a log-transformation was needed. This transformation resulted in Ws > 0.97, indicating that the log-transformed measures could be assumed to be reasonably normal. For the measures number of filled pauses, number of repetitions, and number of corrections, taking the square root was necessary (resulting in Ws > 0.96).

Mixed between-within analyses of variance for each (transformed) measure of fluency were performed, with language group (Turkish or English) as the

Table 1. *Means (and standard deviations) in first language (English/Turkish) and in second language (Dutch) for all measures of fluency*

	English Group ($N = 27$)		Turkish Group ($N = 24$)	
	L1	L2	L1	L2
Mean syllable duration (ms)	215 (25)	286 (68)	189 (26)	294 (49)
Mean silent pause duration				
Within ASU (ms)	552 (110)	711 (205)	635 (132)	739 (158)
Between ASU (ms)	650 (162)	820 (276)	677 (168)	893 (238)
Number of				
Silent pauses/second	0.375 (0.098)	0.485 (0.115)	0.296 (0.100)	0.483 (0.103)
Filled pauses/second	0.180 (0.109)	0.265 (0.159)	0.237 (0.110)	0.338 (0.162)
Repetitions/second	0.063 (0.050)	0.060 (0.049)	0.017 (0.019)	0.049 (0.042)
Corrections/second	0.036 (0.026)	0.059 (0.030)	0.044 (0.017)	0.071 (0.031)

Note: ASU, analysis of speech units.

between-subjects variable and language (L1 or L2) as the within-subjects variable. Table 2 shows the significant main effects and interactions. For all measures of fluency, a significant main effect of language was found, showing that, overall, participants were less fluent in their L2 (Dutch) than in their L1 (either Turkish or English). As can be seen from the eta squared column (η^2) in Table 2, all of these effects can be considered to be large (Cohen, 1988). The largest effect was observed for mean syllable duration ($\eta^2 = 0.79$) and the smallest (yet still "large") effect for number of repetitions ($\eta^2 = 0.15$). For the measure number of repetitions per second, there was also a main effect of language group. However, a significant interaction was also obtained, which means that the effect of language group was dependent on whether L1 or L2 was spoken. Finally, a significant interaction was also found for mean syllable duration and for number of silent pauses per second.

For the measures that showed significant interactions, follow-up t tests were carried out comparing Turkish and English speakers in the L1 and in the L2. It turned out that each interaction could be explained by the fact that the language groups showed large differences between groups in the L1s (as evidenced by Cohen ds > 0.8), but no differences in the L2. For syllable duration, English syllables in the L1 were, on average, longer than Turkish syllables in the L1, t $(49) = -4.02$, $p < .001$; $d = 1.15$, but the speakers did not significantly differ in their L2 ($t < 1$). For silent pauses per second, it was found that Turkish native speakers produced fewer pauses in their L1, $t\,(49) = -2.85$, $p = .006$; $d = 0.81$, but

Table 2. *Results of mixed between-within analyses of variance with language group (English versus Turkish) as the between-participants variable and language (first vs. second language) as the within-participants variable*

	Language Group		Language		Interaction	
df (1, 49)	*F*	η^2	*F*	η^2	*F*	η^2
Mean syllable duration (ms)	2.67		223.09*	0.79	11.54*	0.04
Mean silent pause duration						
Within ASU (ms)	2.96		62.37*	0.62	3.12	
Between ASU (ms)	1.09		81.36*	0.54	1.22	
Number of						
Silent pauses/second	2.29		147.92*	0.71	10.23*	0.05
Filled pauses/second	3.79		32.64*	0.40	0.03	
Repetitions/second	8.88*	0.15	12.30*	0.15	19.49*	0.24
Corrections/second	2.61		61.14*	0.55	0.01	

Note: ASU, analysis of speech units.
*$p < .05$.

again, no differences were found between the language groups in their L2 ($t < 1$). Similarly, with respect to repetitions, Turkish speakers produced fewer repetitions per second in their L1 than the English native speakers, $t\,(49) = -4.91$, $p < .001$; $d = 1.40$, but no differences were found between the Turkish and English speakers in their L2 ($t < 1$).

Predicting L2 fluency

Before running the regression analyses to investigate the predictive strength of the L2 fluency measures over the L1 measures, simple correlations between all fluency measures were computed. Table 3 shows the Pearson correlations between the fluency measures in the L1 and in the L2. The upper half (above the diagonal) shows the correlations between the fluency measures in the L1, and the bottom half shows the correlations in the L2. The diagonal in Table 3 shows the relation, for each measure of fluency, between L1 and L2.

For these correlations, Turkish and English speakers are collapsed into one group. As can be gleaned from Table 3, with respect to the intercollinearity of these measures within L1, most correlations are low (under $r = .3$) to moderate (under $r = .5$; see Cohen's, 1988, rules of thumb of effect sizes for correlations). The correlation between the two pause duration measures turned out to be strong ($r = .71$). One should also note the direction of the relations. Most are positive, indicating that the more fluent one speaks in the L1 with respect to one aspect, the more fluent one also tends to speak with respect to another aspect. The relation between silent pause duration within ASU and number of repetitions, however, is negative, suggesting that there may be a weak trade-off between these two

Table 3. *Pearson correlations between fluency measures within first language (upper half) and within second language (bottom half), and with the relation between first and second languages on the diagonal (in bold)*

	1	2	3	4	5	6	7
1. Mean syllable duration (ms)	**0.37***	0.24	0.18	−0.03	0.11	0.29*	−0.20
2. Mean silent pause duration within ASU (ms)	0.31*	**0.65***	0.71*	0.12	0.10	−0.28*	0.15
3. Mean silent pause duration between ASU (ms)	0.17	0.79*	**0.76***	0.46*	−0.22	−0.20	−0.01
4. Number of silent pauses/second	0.28*	0.40*	0.36*	**0.62***	−0.25	0.22	−0.02
5. Number of filled pauses/second	0.15	−0.37*	−0.48*	−0.25	**0.73***	0.35*	0.43*
6. Number of repetitions/second	0.11	−0.12	−0.30*	0.07	0.53*	**0.60***	0.35*
7. Number of corrections/second	0.10	−0.01	−0.15	0.19	0.45*	0.42*	**0.68***

Note: ASU, analysis of speech units.
*$p < .05$.

measures: Speakers who tend to exhibit many repetitions in their L1 tend to show relatively short silent pauses.

Turning to the correlations between the measures in the L2, shown in the bottom half of the table, two relations can be considered strong: Similarly to the relations in the L1, there is a strong relation between the two L2 measures of pause durations. The relation between number of filled pauses and number of repetitions can be considered strong. Three correlations are negative: between filled pauses and both measures of duration of silent pauses and between number of repetitions and duration of pauses between ASU. This suggests that, in the L2, there may be a trade-off between these measures of fluency, in that L2 speakers who tend to use many filled pauses and repetitions will use, in general, shorter pauses. At the same time, the frequency measures such as number of filled pauses, number of repetitions, and number of repairs cluster together in the sense that L2 speakers who tend to use many filled pauses will also exhibit many repetitions and repairs.

Turning to the diagonal in Table 3, one can see that all correlations between the L1 and the L2 fluency measures are statistically significant, ranging from $r = .37$ for syllable duration to $r = .76$ for silent pause duration between ASU. These Pearson correlations already give an indication of the strength between the L1 and L2 fluency measures overall. To answer our third research question, however, a regression analysis that takes language group into account (English vs. Turkish) is needed in order to test whether the measures of L2 fluency should be corrected *in the same way* for both language groups.

In the previous section it was found that Turkish and English native speakers differ in their L1 fluency behavior. English speakers in English produce longer syllables on average and use more silent pauses and more repetitions than Turkish speakers do in Turkish. Therefore, in this section, it was tested whether the linear regression models that predict L2 measures from L1 measures (see the diagonal in Table 3) were improved by adding language group as a (dummy) predictor variable. Furthermore, it was tested whether adding the interaction between language group (as dummy variable) and L1 fluency measure significantly improved the model. In other words, it was tested whether the slope of the regression line could be assumed to be the same for the English and Turkish native speakers. By testing whether the slope could be assumed to be the same for the two language groups, it was tested whether the predictive strength of the L1 measures over the L2 measures could be assumed to be the same. Finally, the residuals of the best fitting (and most parsimonious) model for each measure of fluency were saved and used as corrected measures of L2 fluency in subsequent analyses.

The interaction between language group and fluency measures in the L1 never proved to be significant ($ps > .05$, R^2 increase $< .04$). Therefore, even though large differences were found for some measures of L1 fluency between the two language groups, the relation between L1 and L2 fluency measures is not different for the native speakers of English and the native speakers of Turkish. At the same time, due to overall differences in the two languages, adding a different intercept for the Turkish native speakers significantly improved the model for some measures. Table 4 shows the standardized regression weights and corresponding standard errors for the intercepts, slopes, and adjusted intercepts for Turkish native speakers (if the model including a separate intercept proved to be better), as well as total R^2.

From Table 4, and more specifically from the slope column in Table 4, one can see that all measures of L2 fluency can, to a certain extent, be predicted on the basis of the L1 fluency measure alone. The higher this slope, the less residual variance remains for L2-specific variance. Another finding is that the success in predicting the L2 measure from L1 behavior and language group ranges from 21% (for mean syllable duration) to 57% (for mean length of pauses between ASU).

Relating corrected and uncorrected measures to vocabulary scores

From the linear regression models predicting L2 fluency from L1 fluency (and language group), the residuals (the corrected measures of L2 fluency) were saved.

Table 5 presents the total R^2 of the vocabulary scores as predicted by the different measures of fluency. The first column shows that the L1 measures are never significantly related to our indicator of L2 proficiency (i.e., to L2 vocabulary knowledge).[1] The second column shows that most of the uncorrected L2 measures are significantly related to the vocabulary scores, except for the two measures of silent pause durations. There is quite some variability in the amount of variance explained: between 9% by number of filled pauses and 30% by syllable duration. From Table 5 one can also see that for all the frequency measures (number of silent pauses, filled pauses, repetitions, and repairs) there are no big changes between the uncorrected L2 measures (second column) and the corrected L2 measures (third

Table 4. *Standardized regression weights (standard errors) of best fitting regression models predicting second language fluency measures from first language fluency measures and from language group as the dummy variable (0 = English, 1 = Turkish)*

	Intercept	Slope	Language Group Adjust.	Total R^2
Mean syllable duration (ms)	−0.29	0.52*	0.63*	.21
	(0.19)	(0.15)	(0.29)	
Mean silent pause duration				
Within ASU (ms)	0.00	0.65*		.42
	(0.11)	(0.11)		
Between ASU (ms)	0.00	0.76*		.57
	(0.09)	(0.09)		
Number of				
Silent pauses/second	−0.25	0.72*	0.52*	.44
	(0.15)	(0.12)	(0.23)	
Filled pauses/second	0.00	0.73*		.53
	(0.10)	(0.10)		
Repetitions/second	−0.32	0.79*	0.68*	.43
	(0.16)	(0.13)	(0.26)	
Corrections/second	0.00	0.68*		.46
	(0.10)	(0.11)		

Note: ASU, analysis of speech units.
*$p < .05$.

column) in explaining vocabulary scores. For syllable duration, the difference is larger: the explained variance increases from 30% to 41%.

Note that the original uncorrected scores are mathematically the same as the L2-corrected scores plus (a constant multiplying) the L1 measures. In other words, the difference between these two types of scores is that, in the uncorrected measure, L1 fluency behavior is incorporated. From the first column in Table 5, one can see that the L1 score by itself is never a significant predictor. For the four frequency measures of fluency we can therefore conclude that although the uncorrected scores and the corrected scores predict the vocabulary scores about equally well, it must be the L2-specific behavior within the uncorrected scores that is actually the explanatory factor that is related to the L2 vocabulary score. Adding information about L1 fluency behavior (by measuring the uncorrected score), does not lead to added explained variance.

As mentioned before, for the measure syllable duration, correcting the L2 scores for L1 behavior leads to an increase in explaining vocabulary scores. To test whether this is a significant increase, we compared a linear model including both L1 and L2 measures to the linear model including only the L2 uncorrected measure. The more complex model (see the fourth column in Table 5) uses L1 scores plus corrected scores to explain vocabulary scores. One can see this in Table 5: the R^2 from the L1 measure (first column) plus the R^2 of the corrected measure (third column) always equals the total R^2 of the model, including both

Table 5. *Total R^2 of L2 vocabulary scores predicted by measures of fluency*

Total R^2	L1 Measure	L2 Measure	Corrected Measure	L1 and L2 Measure	Difference Measure (L1–L2)
Mean syllable duration (ms)	.00	.30*	.41*	.42*	.41*
Mean silent pause duration					
Within ASU (ms)	.00	.01	.03	.03	.04
Between ASU (ms)	.02	.02	.00	.02	.00
Number of					
Silent pauses/second	.03	.17*	.16*	.19*	.11*
Filled pauses/second	.01	.09*	.12*	.13*	.10*
Repetitions/second	.03	.13*	.12*	.14*	.08*
Corrections/second	.01	.10*	.11*	.12*	.09*

Note: L2, second language; L1, first language; ASU, analysis of speech units.
*$p < .05$.

L1 and L2 (except for rounding error). When testing whether the L1 plus the L2 differed from the L2 uncorrected model, there was only a significant change in R^2 for the measure syllable duration, $F\ (1) = 9.87$, $p = .003$. Because the difference in explained variance cannot be due to the nonsignificant addition of L1 itself (additional $R^2 = 0.004$), we can conclude that correcting the L2 scores of syllable duration for L1 behavior leads to significant better predictions of vocabulary scores. For none of the other fluency variables were significant differences found.

DISCUSSION

Think back to Oscar and Mark, the two speakers of Dutch as L2 introduced at the beginning of this paper. Compared to other L2 speakers with similar proficiency, Mark seems to use many filled pauses in his L2, whereas Oscar uses only a few. Similarly, compared to other speakers, Mark uses many and Oscar only a few filled pauses in their L1. To measure L2-specific fluency, which pertains to L2-specific difficulties in speech planning, should L2 measures be adjusted for L1 behavior? In other words, should the measure of L2-specific filled pauses for Oscar and Mark be about the same, after correcting for L1 filled pause behavior?

This study was aimed to test whether L1 fluency behavior should be taken into account to gauge L2-specific measures of L2 fluency. This goal originates from Segalowitz's (2010) proposal to "partial out sources of variability that are not related specifically to the disfluencies in L2 but that characterize a person's general performance in the given testing conditions" (p. 40). To test whether such corrected scores do form better L2-specific measures of fluency, L1 and L2 speech data of 51 L2 speakers of Dutch were gathered. A score on a productive L2 vocabulary task was used as an approximation of L2 proficiency.

It was found that all fluency measures could, to a certain extent, be predicted on the basis of L1 fluency behavior (our first research question). The amount of

explained variance of the L2 measures from the L1 measures and the L1 itself ranged from 21% for the measures of *speed fluency* (syllable duration) to as much as 57% for a measure of *breakdown fluency* (mean length of pauses between ASU). The stronger the relation between the L1 measure and the L2 measure, the more this measure can be seen as reflecting a speaker's general performance given the test conditions because the L2 fluency behavior can, to a large extent, be predicted on the basis of the L1 behavior alone. Derwing et al. (2009) also related measures of L1 fluency to measures of L2 fluency; much like the current study, they found high correlations. Note, however, that there are two main differences between this study and Derwing et al. (2009) with respect to relating L1 to L2. In their study, besides number of pauses per second, global measures of fluency were used, such as speech rate and pruned syllables per second, which incorporate speed and breakdown aspects of fluency. In the current study, we opted for separate measures for separate aspects of fluency. Another difference is that Derwing et al. (2009) used the same tasks in the L1 as in the consecutive measurement moments in the L2. Therefore, they may have found some effects of repetition, which can interfere with fluency behavior. In the current study, we chose to use maximally similar tasks in the two languages, without any repetitions.

To answer our second research question, corrected and uncorrected L2 fluency scores were compared in the extent to which they could predict L2 vocabulary scores. For most measures of L2 fluency, the uncorrected scores and the corrected L2 fluency scores were equally related to the vocabulary score. For mean syllable duration, however, the L2 measure from which the L1 variance of syllable duration had been partialed out was a significantly better predictor of L2 vocabulary knowledge than was the uncorrected L2 syllable duration measure. The amount of explained variance in vocabulary knowledge increased from 30% to 41% when the L1 measure was partialed out. For the other measures of fluency, no differences in explained variance were found.

To summarize, in answer to our second research question, we can conclude that for the fluency measure syllable duration, a corrected score is more strongly related to a measure of L2 proficiency than is the original uncorrected L2 measure. For other measures of fluency, the uncorrected and corrected measures predict L2 proficiency equally well, but it should be noted that for these measures it is L2-specific variance that is in common with L2 vocabulary knowledge and that the L1 measure (incorporated in the original uncorrected score) does not add explained variance.

An unsolved problematic issue with the corrected scores used in this study is that a researcher or language tester will need to sample both L1 and L2 data to obtain the corrected scores, but these corrected scores will be dependent on the particular sample of speakers. The corrected scores are calculated on the basis of the regression predicting L2 scores from L1 scores *for a particular group* of speakers. Future research should ideally test whether the regression slopes found in the present study replicate (for more L2 speakers, at different proficiency levels, for more L1s, and for more L2s). Only if in this future research the regression slopes predicting L2 behavior and L1 behavior turn out to be quite stable, one could utilize "standardized" corrections of L2 measures, calculated from the L2 measures and the L1 measures. To circumvent calculating regression slopes for particular

samples of L2 speakers and therefore calculating corrected scores differently for each sample of speakers, one could also use difference scores as corrected scores (e.g., L1 syllable duration minus L2 syllable duration).

In Table 5, the last column represents the results of such additional analyses using difference scores instead of corrected scores. Similar to the corrected measures, five of the difference measures significantly predicted vocabulary scores, and the amount of variance explained is about the same, albeit somewhat lower for four of these measures when using difference scores rather than corrected measures (compare the last column of Table 5 with the third column). These results suggest that correcting for L1 fluency behavior by using residuals might lead to more precise measures of L2-specific fluency than simply taking the difference scores. The question remains, however, whether the slopes found in the present study generalize to populations with different L1s and L2s, and with different L2 proficiency levels.

Another result from the analyses predicting L2 proficiency from fluency measures is that for duration of pauses no significant relation could be found. One should be cautious, however, in concluding that these fluency measures are not related to L2 proficiency at all. In this study, proficiency was approximated by a measure of vocabulary knowledge alone. In a recent large-scale study (De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012b), however, multiple measures of L2 proficiency were used: vocabulary knowledge, grammar knowledge, pronunciation skills, lexical retrieval speed, and on-line sentence building speed. In that study (with $N = 179$) it was found that all measures of fluency were related to measures of proficiency. Similar to the present study, it was found that syllable duration was the most strongly related to measures of proficiency ($R^2 = 0.50$) and that duration of pauses was only minimally related to measures of L2 proficiency ($R^2 = 0.05$).

Our third research question asked whether the results obtained were generalizable for typologically different L1s. To answer this question, we examined the differences between the native speakers of Turkish and English found in our study. All of the differences between these language groups were in the L1 and not in the L2. Furthermore, the nature of the relation between the L1 and L2 fluency measures was the same for the two language groups. We therefore conclude that, in our sample, we could not find any evidence against generalizing the conclusions from our first and second research questions.

However, we did find differences between Turkish and English native speakers' measures of fluency in their L1. Although these differences are of no importance to the research questions of this paper, in what follows, we will give tentative explanations for the cross-linguistic differences we found. We can explain the difference between L1 Turkish and L1 English syllable durations if we consider the (frequent) syllable structures of the two languages. In Turkish, consonant–consonant–vowel–consonant and consonant–vowel–consonant–consonant are extremely rare, whereas in English two or more consecutive consonants within a syllable are more frequent. The difference between syllable duration for the English and Turkish native speakers can now readily be explained: because Turkish syllables tend to be shorter in terms of number of phonemes, it follows that the mean syllable duration as measured in milliseconds is likewise shorter than that of the English speakers.

We also found L1 differences in number of silent pauses: English speakers paused more often than Turkish speakers did. We can explain this difference when we assume that in speech planning processes, for each language, the lexical word is the unit of encoding (e.g., Dell, 1986; Levelt, 1989; Levelt et al. 1999). Turkish is an agglutinative language, where words are concatenations of morphemes, which together with the root can combine into long words (Lewis, 2001). For instance, *from our bookcases*, three words in English, would translate into a single word with seven syllables in Turkish: *kitaplıklarımızdán* (example taken from Kabak & Vogel, 2001). When words are longer, as in Turkish, speakers have fewer opportunities to pause. This might explain why English speakers pause more often than Turkish speakers do.

Similarly, it may be that English speakers more often have the chance to repeat a word because words are shorter. Because stopping midword is nonpreferred (Levelt, 1983), we can then also explain why English speakers repeat themselves more often than Turkish speakers do. In terms of number of syllables as well as in terms of duration, Turkish words in our data were on average longer than English words. In terms of duration, Turkish words were on average 438 ms ($SD = 54$), whereas English words' duration averaged 293 ms ($SD = 40$); in terms of syllables, Turkish words averaged 2.3 syllables ($SD = 0.20$), and English words were on average 1.36 syllables ($SD = 0.09$) in length. Furthermore, pausing midword seldom occurred, either in the English or in the Turkish speech data, corroborating the hypothesis that a lexical word is the unit of speech planning.

Research into cross-linguistic differences in pausing behavior is scarce. One exception is the work by Riazantseva (2001). She compared pausing phenomena between Russian and English and found that Russian speakers used longer silent pauses than English speakers did. She related this finding to cultural differences between Russian speakers and American English speakers. Although we explain the differences between Turkish and English in our data as typological differences, cultural differences may likewise play a role.

CONCLUSIONS

On the basis of the current findings, we conclude that research into L2 speaking will benefit from utilizing corrected measures of L2 fluency by sampling both L1 and L2 speech. For instance, for relating L2-specific linguistic processing to utterance fluency, at least for the measure syllable duration, adjusting for L1 behavior yields more precise measures. For diagnostic language tests, it may be profitable to have L1 behavior as a baseline. For instance, it would be futile for an L2 speaker to strive for using very few filled pauses in his L2 when he tends to be an "uhm"-er in his L1. One may even hypothesize that for learners to increase their L2 fluency, they need to change their speaking style in any language, including in their L1. For criterion-referenced tests, when criteria with respect to fluency have been predefined, it does not make sense to use corrected measures of L2 fluency because this would mean that each speaker's L1 behavior would serve as his or her own criterion. At the same time, the present study has shown that at least for syllable duration, basing conclusions about a learner's L2 proficiency on the uncorrected measure is basing this conclusion partly on the learner's personal

speaking style, which is unrelated to his L2 proficiency. Whether or not this is fair is beyond the scope of this paper.

Another issue beyond the scope of this paper concerns the relation between L2 corrected measures of utterance fluency and measures of perceived fluency. Perhaps it is the case that in real life interactions, listeners can distinguish between disfluencies that are related to developing L2 proficiency and disfluencies (hesitations) that are due to an individual's personal speaking style. If that is the case, raters in L2 tests may already base their judgments on corrected fluency. Although we believe that it is unlikely that listeners (thus raters) are able to distinguish between these causes of disfluencies, further research is needed to solve this matter.

Although many language tests ask raters, who may or may not take personal speaking style into account when rating on L2 fluency, to judge L2 fluency, automated language tests use objective utterance measures as diagnostic measures of speaking proficiency. On the basis of our findings, and those of De Jong et al. (2012b), we would argue that in automated tests and in research investigating L2-specific processing on the basis of utterance fluency, duration of pauses should play a modest role in estimating L2-specific cognitive fluency and that (a corrected measure of) syllable duration should play a stronger role.

APPENDIX A

The appendix gives a short description of the Dutch tasks and the mirroring English and Turkish tasks. The Turkish tasks were translated from the English tasks. Descriptions of the tasks performed by the speakers in their L1 (English or Turkish) are the following:

Task 1 *simple, informal, descriptive:* The participant speaks with a friend and describes the type of apartment he is looking for.

Task 2 *simple, formal, descriptive:* The participant, who has just witnessed a crime/accident occur on the street, describes what happened to a police officer.

Task 3 *simple, informal, persuasive:* The participant advises his/her brother on how to choose between quitting his current job to work full-time on developing a new career and remaining at his current job while studying part-time for his new career.

Task 4 *simple, formal, persuasive:* The participant is present at a neighborhood meeting in which an official has just proposed building a new casino at a location near a school. The participant speaks up, suggesting another location that would be more acceptable.

Task 5 *complex, informal, descriptive:* The participant tells a friend about a piece in the newspaper about home sales in rural versus suburban areas.

Task 6 *complex, formal, descriptive:* The participant is the principal of a high school and calls a new science teacher to tell him about the courses he will be teaching.

Task 7 *complex, informal, persuasive:* After watching a movie about global warming, the participant discusses the issue with a friend and tries to convince him that more solar/wind energy production is the best solution.

Task 8 *complex, formal, persuasive:* The participant, who is the manager of a nursing home, addresses the board of directors and discusses the advantages and disadvantages of building more facilities.

Descriptions of the tasks performed by the speakers in their L2 (Dutch) are the following:

Task 1 *simple, informal, descriptive:* The participant speaks on the phone to a friend, describing the new apartment of friends who have recently moved.
Task 2 *simple, formal, descriptive:* The participant, who witnessed a road accident some time ago, is in a courtroom, describing the accident to the judge.
Task 3 *simple, informal, persuasive:* The participant advises his/her sister on how to choose between (or combine) childcare, further education, and working.
Task 4 *simple, formal, persuasive:* The participant is present at a neighborhood meeting in which an official has just proposed to build a school playground across the street from the school itself. The participant takes the floor and argues against the planned location of the playground.
Task 5 *complex, informal, descriptive:* The participant tells a friend about the development of unemployment rates among women and men over the last 10 years.
Task 6 *complex, formal, descriptive:* The participant works at the employment office of a hospital and tells a candidate for an open nursing position what the main tasks of the vacant position are.
Task 7 *complex, informal, persuasive:* The participant discusses the pros and cons of three means of transportation (public transportation, bicycles, and automobiles) in solving the problem of traffic congestion.
Task 8 *complex, formal, persuasive:* The participant, who is the manager of a supermarket, addresses a neighborhood meeting and argues for one of three alternative plans for building a parking garage.

ACKNOWLEDGMENTS

Part of this research was funded by the Netherlands Organisation for Scientific Research by NWO Grant 254-70-030 (to J.H.H. and R.S.) and by a grant from Pearson Language Testing (to N.H.D.J.). We thank Margarita Steinel for finding Turkish participants and for working together on this project. We thank our research assistants Cem Keskin and Erica Bouma for all the meticulous transcriptions, as well as Veysel Yüce and Iske Bakker for running the experiments, and Canan Gonencay for translating the English tasks to Turkish. We also thank Anne-France Pinget and Hans Rutger Bosker for scoring the vocabulary tasks and for valuable discussions, and we thank Jelle Goeman for statistical advice.

NOTE

1. For the linear models predicting L2 vocabulary from L1 fluency measures, we used adjusted L1 measures if the adjusted intercepts for Turkish speakers turned out to be significant predictors for the measures of L2 fluency (i.e., for syllable duration, number of silent pauses per second, and number of repetitions per second). The adjustment

involved adding the mean difference between the two language groups for the Turkish speakers, leading to the same means for both groups.

REFERENCES

Beglar, D., & Hunt, H. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing, 16*, 131–162.

Chambers, F. (1997). What do we mean by fluency? *System, 25*, 535–544.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

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

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

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2012b). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics.* Advance online publication. doi:10.1017/S0142716412000069

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*, 283–321.

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

Fillmore, C. J. (1979). On fluency. In C. J. Fillmore, D. Kempler, & W. S. Y. Wang (Eds.), *Individual differences in language ability and language behaviour* (pp. 85–102). New York: Academic Press.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2001). 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). Amsterdam: John Benjamins.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. New York: Academic Press.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics, 30*, 461–473.

Howell, P., & Au-Yeung, J. (2002). The EXPLAN theory of fluency control applied to the diagnosis of stuttering. In E. Fava (Ed.), *Current issues in linguistic theory series: Pathology and therapy of speech disorders* (pp. 75–94). Amsterdam: John Benjamins.

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

Kabak, B., & Vogel, I. (2001). The phonological word and stress assignment in Turkish. *Phonology, 18*, 315–360.

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

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*, 33–51.

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

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14*, 41–104.

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

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*, 1–37.

Lewis, G. (2001). *Turkish grammar*. Oxford: Oxford University Press.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Erlbaum.

Ramsay, R. W. (1968). Speech patterns and personality. *Language and Speech, 11*, 54–63.

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

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

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

Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*. Unpublished doctoral dissertation, University of California, Berkeley.

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

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). Amsterdam: John Benjamins.

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

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency. *Studies in Second Language Acquisition, 27*, 567–595.