Accepted for publication in *TESOL Quarterly* (Wiley Blackwell)

To What Extent Does Long-Term Foreign Language Education Improve Spoken Second Language Lexical Proficiency?

Kazuya Saito[1]

Abstract

The current study examined lexical aspects of second language (L2) speech attainment in the foreign language (FL) classroom setting (i.e., several hours of target language input per week). A total of 72 second-year university students with seven years of FL study and no experience abroad participated in the study. Their spontaneous speech was analyzed via a set of lexical measures, and then compared to that of experienced, naturalistic Japanese L2 learners of English. According to the results, their lexical proficiency was factored into three dimensions—appropriateness (global, semantic, morphosyntactic accuracy), specificity (frequency, range) and abstractness (concreteness, meaningfulness, imageability, hypernymy). Overall, extensive FL education led many participants' specificity performance to reach comparable proficiency levels to the baseline group. Approximately half of participants achieved such satisfactory proficiency in abstractness. The participants' lexical appropriateness demonstrated a great deal of individual variability, and was linked to the extent to which they had recently practiced the target language.

Key words: *Foreign language learning, Second language speech, Vocabulary use, Lexical Proficiency, Lexical sophistication*

[1] This study was supported by the Birkbeck College Additional Research Fund. I am grateful to the Journal Editor, Charlene Polio, and anonymous TQ reviewers for their helpful input and feedback on earlier versions of this manuscript, and to Kim van Poeteren, Kokoro Muramoto, Takumi Uchihara, Shungo Suzuki, and Masaki Eguchi who helped with the data collection and analyses.

In the field of second language acquisition (SLA), scholars have begun to examine the role of experience in the development of second language (L2) oral abilities in the context of foreign language (FL) classrooms, where students' L2 use is restricted to only several hours of primarily language-focused lessons per week (Muñoz, 2014). Whereas the existing literature has begun to look at the long-term effectiveness of FL education on L2 pronunciation (e.g., Nagle, 2017; Saito & Hanzawa, 2016; Simon, & D'Hulster, 2012), surprisingly little attention has been directed towards the lexical aspects of L2 speech learning and attainment in FL settings. Drawing on the recent development of the computational modeling of L2 vocabulary use (Crossley, Salsbury, & McNamara, 2015; Kyle & Crossley, 2015; Salsbury, Crossley, & McNamara, 2011), the current study examined the extent to which a total of 72 college-level students could improve the lexical appropriateness (global, semantic and morphological accuracy) and sophistication (frequency, range, abstractness) of their L2 speech after seven years of FL education in Japan without any experience studying abroad.

Background

Second Language Speech Learning in Foreign Language Classrooms

In today's globalized society, developing adequate L2 oral proficiency has become increasingly crucial to achieve academic- and business-related goals by interacting with individuals from various linguistic and cultural backgrounds. While immersion/study abroad is commonly conceived as the optimal way to improve such L2 skills, a great number of L2 leaners study their target languages in FL settings. As Muñoz (2008) pointed out, however, SLA in FL settings is limited in terms of both quantity and quality. For instance, students in FL settings typically receive only a few hours of L2 input per week. The nature of this instruction could be still language-focused rather than meaning-oriented, and without many opportunities to practice the target language beyond classroom contexts.

To remedy this, scholars have extensively worked on examining how to boost the efficacy of FL pedagogy through task-based language teaching (Skehan, 2014), content and language integrated instruction (Coyle, Hood, & Marsh, 2010) and computer-assisted language learning (Plonsky & Ziegler, 2016). In Japanese English-as-a-Foreign-Language classrooms (the main focus of the study), for example, practitioners, researchers and politicians have worked together on enabling the Japanese public to accomplish "the basic and practical communication abilities"—an educational goal set by the Ministry of Education's Action Plan 2003. To promote students' communicative use of language, a number of changes have been made thus far to the existing curriculums, including: increasing the number of oral communication classes at both junior and senior high schools; having college entrance examinations put equal emphasis on assessing all four skills (reading, writing, speaking, listening); and revising teacher training programs to equip pre- and in-service teachers with more adequate, advanced L2 English proficiency (MEXT, 2014).

Under such FL conditions, obtaining comprehensible and intelligible (rather than nativelike) speech is considered as a more feasible, realistic and appropriate goal for many L2 learners (Trofimovich & Isaacs, 2012). However, it is still a crucial research initiative to scrutinize the extent to which L2 learners can ultimately improve their oral proficiency in FL settings, and what kinds of FL experiences are instrumental to leading to such successful L2 speech learning. Examining the pedagogical *potential* and *limits* of FL will in turn help promote

not only a critical assessment of the existing FL system, but also a constructive discussion on the conceptualization/development of future educational reforms.

To date, a number of empirical investigations have been conducted to examine the process and product of L2 development in FL settings (Jaekel, Schurig, Florian, & Ritter, 2017; Llanes & Muñoz, 2013; Muñoz, 2006, 2014; Nagle, 2017; Saito, 2018; Saito & Hanzawa, 2016; Riney & Flege, 1998; Ojima, Matsuba-Kurita, Nakamura, Hoshino, & Hagiwara, 2011; Simon & D'Hulster, 2012). According to their findings, L2 learners' speech is likely tied to the length of FL experience, especially when they are involved with a sufficiently wide variety of instructional treatment (Saito & Hanzawa, 2016 for oral communication and pronunciation training); seek ample opportunities to practice the target language outside classrooms (Muñoz, 2014 for watching TV, reading books and writing emails; Saito & Hanzawa, 2016 for conversational activities with international students); and have high-level awareness of phonological accuracy (Cebrian, 2006 for college students with English Philology majors).

In the field of instructed SLA, scholars have also explicated the role of experience by conducting quasi-experimental studies with a pre- and post-test design to further scrutinize how to best draw L2 learners' attention to form in various classroom settings. These effective techniques include explicit phonetic instruction (Kissling, 2013), corrective feedback (Lee & Lyster, 2016), controlled and communicative tasks (Lord, 2008) and self assessment (de Saint Léger, 2009). With these focus-on-form techniques, teachers can help L2 learners make the most of their L2 experience even if it is limited in quantity under classroom conditions (Derwing & Munro, 2005). In terms of timing of instruction (immediate vs. delayed effects), there has been some indication that L2 learners tend to show instructional gains very quickly, especially when the nature of instruction is explicit and language-focused. Comparably, the sustainability of such intentional L2 learning has remained highly controversial (Norris & Ortega, 2000).

Different from naturalistic SLA, however, the degree of L2 learners' final attainment after years of FL education seems to be unrelated to their initial age of learning (Jaekel et al., 2017; Muñoz, 2006, 2014; Ojima et al., 2011). This is arguably because most of FL learning is explicit in nature, and may be more advantageous for older (rather than younger) learners who can make the most of their cognitive maturity, their already-developed L1 literacy and their accumulative classroom learning experiences (Muñoz, 2008). In such acquisition-limited environments, the amount of meaningful L2 input is not sufficient enough for triggering incidental and implicit learning, which is subject to age effects in naturalistic SLA (DeKeyser & Larson-Hall, 2005).

Though revealing, the long-term effectiveness of FL education for high-level L2 speech attainment still remains unclear, as the aforementioned studies contain several methodological limitations. First, most have focused only on the phonological aspects of L2 speech (Cebrian, 2006; Nagle, 2017; Riney & Flege, 1998; Saito, 2018; Saito & Hanzawa, 2016; Simon & D'Hulster, 2012); very few have highlighted other aspects of L2 speech attainment, including vocabulary (cf., Llanes & Muñoz, 2013). Second, many of the FL studies conducted have included participants with advanced L2 proficiency who had studied abroad; this has prevented scholars from expounding the effect of FL experience *alone* on L2 speech learning (Muñoz, 2014; Riney & Flege, 1998). Third, few studies have compared the attained proficiency of FL students with any comparison against/with baseline groups. As a result, we have yet to answer to what degree these FL learners could be considered as successful (or unsuccessful) against any realistic/appropriate benchmarks, such as experienced, naturalistic and advanced L2 learners. The current study was designed to respond to all these concerns.

Modeling Spoken L2 Lexical Proficiency

According to Crossley's computational framework of L2 vocabulary use (e.g., Crossley et al., 2015; Kyle & Crossley, 2015; Salsbury et al., 2011), the lexical characteristics of speech can be conceptualized from two different perspectives: (a) appropriateness and (b) sophistication. The first dimension, appropriateness, concerns how accurately L2 learners can access semantic and morphosyntactic properties of words in context (see Crossley et al., 2015 for collocational accuracy; Foster & Wigglesworth, 2016 for weighted lexicogrammar accuracy; Saito, Webb, Trofimovich & Isaacs, 2016 for overall comprehensibility). The other dimension, sophistication, relates to how L2 learners use more sophisticated, infrequent, specific and abstract words (see Crossley, Salsbury, & McNamara, 2009 for hypernymy; Kyle & Crossley, 2015 for frequency and range; Salsbury et al., 2011 for concreteness, meaningfulness and imageability).

To date, many empirical studies have shown that lexical appropriateness and sophistication factors can explain a large amount of the variance (60-70%) in trained raters' assessments of spoken L2 proficiency in ACTFL (Crossley et al., 2011, 2015), TOEFL (Kyle & Crossley, 2015) and the Test for English Majors in China (Lu, 2012). Assuming that L2 learners' proficiency develops on the continuum of "beginner" to "advanced" (as operationalized in general proficiency tests), the results suggest that the two major dimensions—appropriateness, sophistication—interact to reflect lexical aspects of L2 speech learning over time to a great degree. According to previous longitudinal investigations, L2 learners' vocabulary use indeed quickly becomes more accurate (Llanes & Muñoz, 2013; Mora & Valls-Ferrer, 2012) and more diverse and complex (Tavakoli, 2018) within a short period of immersion (2-3 months of study abroad). Focusing on international students in ESL classrooms over one academic year, Crossley et al.'s seminal studies demonstrated that the participants' L2 lexical proficiency gradually grew by using less concrete, meaningful and imageable words (Salsbury et al., 2011) and accessing a wider range of hypernymy levels (Crossley et al., 2009).

Motivation for the Current Study

The current study took a step forward by using Crossley and colleagues' lexical proficiency framework to analyze the vocabulary use development of 72 second-year FL students who had studied L2 English for seven years in FL settings without any experience abroad. In particular, the analyses highlighted a total of 10 lexical measures which tapped into appropriateness (global, semantic and morphosyntactic accuracy) and sophistication (frequency, range, concreteness, meaningfulness, imageability, hypernymy). Furthermore, L2 learners' ultimate attainment in FL classrooms was linked to a range of experience factors (language-focused vs. oral communication vs. content-based classes at junior high school, high school and university).

Two methodological innovations of this study were noteworthy here. First, to control for the influence of pronunciation and fluency factors, all the speech samples were transcribed and subjectively analyzed by trained raters (Crossley et al., 2015) as well as objectively via the Tool for the Automated Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015). Second, to provide a realistic assessment of the participants' L2 performance (the extent to which L2 learners' lexical proficiency could be considered as "successful"), they were evaluated in relation to a group of advanced Japanese learners of English in Canada. These learners moved to Canada after receiving six-to-nine years of EFL education in Japan, and had been using L2 English as their primary language of communication at work and family on a daily basis for more than 10 years. This decision was made in conjunction with Ortega's (2013) claim that any

aspect of L2 learners' proficiency should be compared with other groups of L2 learners rather than with native speakers. In the case of the current study, experienced Japanese learners of English were chosen as their performance was hypothesized to indicate the upper limit of late SLA (i.e., advanced L2 proficiency), which could serve as a realistic role model for learners hoping to achieve comprehensible and intelligible (rather than nativelike) oral proficiency skills. The following research questions were thus formulated:

1. To what extent can Japanese FL students with seven years of FL experience ultimately improve their L2 vocabulary use after seven years of FL education relative to advanced naturalistic L2 learners' performance?
2. How was the participants' L2 lexical appropriateness and sophistication performance related to the quantity and quality of previous and current FL experience?

Method

Participants

FL students. Given that the primary objective of the study was to examine the impact of long-term FL education on vocabulary aspects of L2 oral proficiency, a series of pilot studies were conducted to identify the most ideal participants to this end. In Muñoz's (2014) similar research on ultimate attainment in FL settings, for example, their participants had approximately 10 years of FL learning (similar to naturalistic SLA research: e.g., Abrahamsson, 2012). As observed in Muñoz (2014) and according to the preliminary results of our pilot studies, however, the extensive approach (FL > 10 years) could inevitably include a number of students who have studied abroad, which will in turn conflate the objective of the current study. At the university where the data was collected, most students have started learning English from Grade 7; some interested students can apply for and join study abroad programs from the second term of Year 2. In order to test the final quality of L2 speech learning resulting solely from FL education in Japan, a decision was made to collect the data from second year students when they had finished their "first" term.

Originally, a total of 75 foreign language learners (age: 19-20 years) were recruited from a range of social sciences and humanities programs (e.g., business, economics, marketing, psychology) at a large private university in Tokyo, Japan. Since three participants did not complete the data collection for various reasons, a total of 72 students (36 males, 36 females) participated in the study. The participants' FL backgrounds were carefully checked to ensure that they met all the following conditions. First, their L1 was Japanese, and both of their parents were L1 Japanese speakers. Second, the participants started learning English from secondary school (i.e., Grade 7 without any FL education in preschool and elementary school), allowing us to examine the effects seven years' worth of FL instruction on their L2 vocabulary use. Third, the participants had never studied abroad except for short family trips (less than 10 days). According to their general English proficiency test scores (measured via TOEIC), their proficiency ranged from 300 to 915, indicating that they ranged between the A2 (Basic User) and B2 (Independent User) levels.

Experienced Japanese learners. Ten experienced Japanese learners of English were recruited as a baseline group in Calgary, Canada. All of them were considered as late bilinguals, as they received nine years of FL education in secondary and postsecondary schools in Japan and then moved to Canada during their 20's ($M_{age\ of\ arrival}$ = 21.4 years; *Range* = 18 to 26 years). The length of their stay in Canada varied between 10 and 23 years ($M_{length\ of\ residence}$ = 14.0 years),

indicating that their performance likely reached ultimate attainment—a standard typically used in age-related SLA research (DeKeyser & Larson-Hall, 2005). According to the results of individual interviews, all of them reported using L2 English as their main language communication at work and home, and estimated their use of L2 English as very often ("6" on a 6-point scale, where *1 = very infrequent, 6 = very frequent*).

Speech Materials

As widely used in earlier L2 speech research (e.g., Derwing & Munro, 1997) and recently adopted for L2 vocabulary research (e.g., Authors, 2016b), the participants' spontaneous speech was elicited via a picture narration task. After taking one minute to familiarize themselves with the contents of an eight-frame cartoon story, they proceeded to describing the sequence of the events (e.g., two strangers bumping into each other and accidentally switching their identical-looking suitcases). Their narratives were recorded in a sound-proof booth (for the FL participants) or in a quiet room (for the experienced Japanese learners) using a high-quality digital audio recorder (set to a 44.1 kHz sampling rate with 16-bit quantization).

The length of the samples widely varied (*M* = 3 min 15 sec: *Range* = 1 min 56 sec to 5 min 11 sec) and could be considered comparable to other similar L2 vocabulary studies (e.g., Crossley et al., 2015; Lu, 2012 for 3 min). Since the focus of the study lay in the analyses of the vocabulary (rather than pronunciation and fluency) dimensions of L2 speech, the speech samples were transcribed and cleaned by removing filled pauses (e.g., "ah, eh, oh, um") and fixing obvious mispronunciation problems based on contextual information (e.g., "road" mispronounced as "load"). The average number of words of the resulting files was 120.6 words (*Range* = 61 to 268 words). According to the results of Grubbs' tests, no outliers were found in the dataset in terms of the length and number of words of 72 narratives.

Appropriateness Analyses

Traditionally, L2 learners' abilities to choose semantically appropriate words with accurate morphosyntax markers have been examined by tallying the number of erroneous instances in a certain sentence unit (e.g., Yuan & Ellis, 2003 for per clause). Certain scholars (e.g., Foster & Wigglesworth, 2016) have casted doubts on such dichotomous ways of accuracy judgements (correct or incorrect). They have rather emphasized the importance of using human raters' subjective judgements of lexical errors based on their supposedly different amount of impact on communicative adequacy. Taking into account different levels of error gravity, for example, linguistically trained raters have previously assessed lexical appropriateness using four different rubrics (*1 = entirely accurate; 2 = minor errors; 3 = serious errors; 4 = very serious errors hindering meaning conveyance*) (Foster & Wigglesworth, 2016) or using a 6-point scale (*1 = minimum accuracy, 6 = maximum accuracy*) (Crossley et al., 2015).

Following this line of thought, linguistically trained raters participated in the current study in order to assess three different constructs of lexical appropriateness via a computerized procedure. These measures included (a) global accuracy (i.e., overall ease of understanding) (Authors, 2016b), (b) semantic accuracy (i.e., the selection of appropriate words in context) (Authors, 2017), and (c) morphosyntactic accuracy (i.e., the accurate use of tense, aspects, agreement, plurality and word order) (Ruivivar & Collins, 2017). All the transcripts were displayed in a randomized order via a tailor-made, MATLAB-based software. While reading each transcript, the raters made their judgements on global, semantic and morphosyntactic

accuracy by using a moving slider, which automatically recorded their scores on a 1000-point scale (*0 = nontargetlike, 1000 = targetlike*). For onscreen labels, see Appendix.

Raters. A total of five raters (2 males, 3 females) were recruited at a university in Montreal, Canada, for the appropriateness analyses. All of them were graduate students in Applied Linguistics with a great deal of experience with taking numerous linguistics courses and conducting various kinds of L2 speech analyses. They reported high familiarity with Japanese-accented speech (M = 5.2 on a 6-point scale; *1 = not familiar at all, 6 = highly familiar*) due to their previous teaching experience and/or exposure to foreign-accented L2 speech in ESL/EFL classrooms ($M_{\textit{years of ESL/EFL teaching experience}}$ = 3.3 years: *Range* = 1 to 11 years).

Procedure. As with previous studies (e.g., Saito, Trofimovich, & Isaacs, 2017; Ruivivar & Collins, 2017), the raters first received detailed explanation on the purpose of the project (examining lexical profiles of Japanese learners' L2 English speech); on the definitions for comprehensibility, lexical appropriateness and morphosyntactic accuracy; and on the interpretations of the endpoints (what it meant by the most left and right-hand sides of each rating continuum). The training scripts were detailed in Appendix.

Next, the raters practiced the procedure with three transcripts which were not included in the main dataset. For each transcript, the raters explained their decisions and the trained researcher provided feedback to ensure that the raters fully understood the three different scales. Finally, they proceeded with the judgements of 82 transcripts (72 FL students + 10 experienced learners). These transcripts were presented in a randomized order via the MATLAB-based software; and the raters were explicitly told about the purpose of the assessment (i.e., they were to rate the lexical accuracy of speech produced by Japanese learners of English with varied proficiency levels. The entire session took approximately 50 minutes.

Inter- and Intra-Rater Reliability. The five raters' inter-rater agreement was checked and considered relatively strong via Cronbach alpha analyses for global accuracy (α = .95), semantic accuracy (α = .88) and morphosyntactic accuracy (α = .90). After completing the rating task, the raters were interviewed on the extent to which (a) they understood the rated categories (*1 = "I did not understand at all", 9 = "I understand this concept well"*) and (b) could comfortably and easily use them when rating (*1 = "very difficult", 9 = "very easy and comfortable"*). Their self-reports ranged between 8 and 9, demonstrating high-levels of understanding and confidence to use the three measures throughout their appropriateness judgements.

Sophistication Analyses

Following Kyle and Crossley's (2015) proposed concept of lexical sophistication, three different aspects of vocabulary use were automatically measured spanning frequency, range and abstractness via TAALES (Kyle & Crossley, 2015).

Frequency. Word frequency was calculated by dividing the total sum of frequency scores (in reference to the Corpus of Contemporary American English) by the number of all the words with frequency scores. Higher word frequency scores indicate that L2 learners likely include less frequent words in their texts, indicating more advanced L2 proficiency levels (Crossley et al., 2009). Following Kyle and Crossley's recommendation, the logarithmically transformed scores were used for the frequency analyses in order to control for Zipfian effects common in word frequency lists (i.e., the first 3,000 to 4,000 word families tend to be recycled intensively).

Range. Word range was calculated by dividing the total sum of range scores by the number of words in the texts with range scores. The range index corresponds to how widely certain words are used in different types of documents with the assumption that more proficient L2 learners can access more specific words which have been narrowly used and observed in certain contexts and genres.

Abstractness. Another crucial lexical sophistication factor concerns the psycholinguistic properties of words related to the notion of abstractness. In the current study, the abstractness aspects of L2 vocabulary use were analyzed in terms of native speakers' subjective judgements (concreteness, meaningfulness, imageability) and corpus-based measures (hypernymic relation). All of these lexical dimensions have been found to associate with L2 vocabulary learning, as more proficient L2 learners tend to use less familiar, concrete and imageable words (Salsbury et al., 2009) with less superordinate terms (Crossley et al., 2009).

- **Concreteness, and Meaningfulness and Imageability**: In TAALES, native speakers' subjective ratings were collected for perceived concreteness, meaningfulness and imageability of approximately 4,000 words based on the MRC psycholinguistics database (Coltheart, 1981). Each transcript was submitted to TAALES to calculate average concreteness, meaningfulness and imageability scores.
- **Hypernymymic Relation**: Lexical abstractness has also been objectively conceptualized as semantic hierarchy where a word is located (i.e., hypernymy). For example, "color" is more abstract than "green" because the former is a superordinate term for the latter. In TAALES, hypernymy scores were calculated for each transcript by dividing the total number of superordinate terms by the total number of words.

Surveying Foreign Language Experience

As often seen in previous FL research (e.g., Saito & Hanzawa, 2016; Muñoz, 2014), all the participants' English learning experience was surveyed via individual interview sessions targeting the following themes:

Length of Instruction. Since length of instruction has been found to be a major factor influencing successful FL learning (Muñoz, 2006, 2014), the participants were asked how many hours of English classes per week they had attended in junior high school (Grade 7-9), senior high school (Grade 10-12) and university (one year).

Focus of Instruction. Given that oral communication has been considered a crucial component of L2 English education in secondary school in Japan, the participants self-reported the presence of oral communication classes during junior (Grade 7-9) and senior high school (Grade 10-12). Their answers were coded binarily: 0 (no) or 1 (yes).[2] As for the type of instruction received during university, all the participants reported taking a certain number of language-focused English lessons (where their performance was assessed based on the accurate and fluent use of L2 English) but also given choices to enroll in a range of content-based classes (where their performance was assessed based on the breadth and depth of their understanding of subjects in concern, such as economics, marketing and psychology). Accordingly, the participants reported how many hours of language- and content-based classes they had taken per week.

[2] Following our precursor research (Saito & Hanzawa, 2016), the presence of oral communication classes was coded "yes" when they were offered as a stand-alone module (i.e., at least one session per week). To ensure that participants correctly understood our intension, their relevant experience profiles were carefully surveyed through an individual interview on an interactive mode.

Amount of Extracurricular L2 Use: Assumedly, the participants greatly differed in terms of the amount of extracurricular practice with L2 English outside classrooms—another crucial influencing factor (Muñoz, 2014; Saito & Hanzawa, 2016). To this end, the participants reported how many hours they had spent on extracurricular activities per week. In particular, their answers were separately analyzed for how much they had studied L2 English during prep school in junior high school (as preparation for high school entrance exams) and senior high school (as preparation for college entrance exams); and how much they engaged in conversation activities with native and non-native speakers during their first year at university.

Results

Construct Validity of Lexical Variables

The first objective of the statistical analyses was to examine the factors underlying the nine vocabulary measures which were hypothesized to tap into appropriateness (global, semantic, morphosyntactic accuracy) and sophistication (frequency, range, concreteness, meaningfulness, imageability, hypernymy) aspects of L2 lexical proficiency. To this end, all the participants' performance scores were submitted to a factor analysis with the minimum Kaiser criterion eigenvalue set to 1.0. Given that some of the nine measures were conceptually inter-related especially within each lexical dimension (e.g., global, semantic and morphosyntactic accuracy for "appropriateness"), the decision was made to use the oblique rotation method (i.e., Promax) rather than the orthogonal rotation method (e.g., Varimax). The factorability of the entire dataset was confirmed via two tests: the Bartlett's test of sphericity ($\chi^2 = 495.87$, $p < .001$) and the Kaiser-Meyer-Olkin measure of sampling adequacy (.735). As summarized in Table 1, a "three-factor" solution was identified, accounting for 78.5% of the total variance in the nine lexical variables.

Table 1
Summary of a Three-Factor Solution Based on a Factor Analysis of the Nine Lexical Appropriateness and Sophistication Measures

	Factor 1	Factor 2	Factor 3
A. Appropriateness			
Global accuracy	.079	-.029	**.807**
Semantic accuracy	-.216	-.018	**.804**
Morphosyntactic accuracy	-.150	-.004	**.829**
B. Sophistication			
Frequency	.334	**.741**	.207
Range	-.262	**.877**	-.266
Concreteness	**.906**	.001	-.126
Meaningfulness	**.906**	.069	-.112
Imageability	**.848**	.176	.001
Hypernymy	**.755**	-.326	-.092

Note. All loadings > .5 were highlighted in bold.

Whereas all the appropriateness measures were clearly clustered into the one single factor (Factor 2), the sophistication measures were clearly divided into two subcomponent factors. The results here suggest that the corpus-based frequency and range measures (Factor 3) were methodologically and thematically different from all of the abstractness-related measures (i.e., concreteness, meaningfulness, imageability and hypernymy) (Factor 1).

In sum, a total of the nine lexical measures adopted in the current study were assumed to tap into three different aspects of L2 learners' lexical proficiency: using not only appropriate L2 vocabulary (global, semantic and morphosyntactic accuracy), but also less frequent and more content-specific words (frequency, range) entailing more abstract concepts (concreteness, meaningfulness, imageability, hypernymy). To reduce type one error, the resulting factor scores (Appropriateness, Specificity, Abstractness) were used for the rest of the statistical analyses.

FL Students vs. Experienced Learners

The next objective of the statistical analyses was to assess the pedagogical potential and limitations of seven years of FL education by comparing the FL students' vocabulary performance with experienced Japanese learners. Results of the participants' factor scores (Appropriateness, Specificity, Abstractness) were descriptively summarized in Table 2. In the current study, a unique statistical analysis was performed to count how many FL participants could be considered as "successful" in the sense that their performance was statistically "comparable" to the experienced Japanese learners (the benchmark in the current study). In many naturalistic L2 ultimate attainment studies (e.g., Abrahamson, 2012), certain L2 learners could be regarded as having reached "near-nativelike" proficiency if their performance fell within two SDs of the baseline group's average scores (i.e., native speaker controls). Using the same methodology as near-nativelike proficiency analysis, we counted the number of FL students whose factor scores (Appropriateness, Specificity, Abstractness) fell within a two-SD range of the experienced Japanese learner group for all measures.

Table 2
Summary of Lexical Factor Scores (Appropriateness, Specificity, Abstractness)

	Lexical dimension	*M*	*SD*	*Min*	*Max*
FL students (*n* = 72)	Appropriateness	-0.239	0.795	-1.784	1.450
	Specificity	0.721	1.005	-2.055	2.021
	Abstractness	0.148	0.965	-1.736	2.744
Experienced learners (*n* = 10)	Appropriateness	1.661	0.720	0.256	2.599
	Specificity	-0.519	0.872	-2.397	0.939
	Abstractness	-1.068	0.530	-1.659	0.096

The results indicated the following hierarchy in terms of the degree of L2 vocabulary achievement after seven years of FL education in Japan: Specificity > Abstractness > Appropriateness. First, a majority of the participants' specificity scores (*n* = 62 out of 72 Japanese learners, 86.1%) could be considered comparable to the experienced Japanese learners. Second, approximately half of them reached the proficiency standard of the experienced Japanese learners in terms of abstractness (*n* = 31, 43.0%). Finally, the ratio of such successful FL students was relatively low in appropriateness (*n* = 24, 33.3%).[3]

[3] Although using the 2-SD threshold has been widely used to identify near-nativelike and advanced L2 proficiency learners in previous SLA literature (e.g., Abrahamsson, 2012), one reviewer pointed out the importance of reporting the number/proportion of FL learners whose performance reached "one" SD of the experienced Japanese group's mean scores. Using the 1-SD threshold as a stricter benchmark, the ratio of successful FL students decreased for Appropriateness (*n* = 6, 8.3%), Specificity (*n* = 44, 66.1%) and Abstractness (*n* = 19, 26.3%). Again, the following pattern of learning success was confirmed: Specificity > Abstractness > Appropriateness.

Relationships between Length, Focus and Timing of FL Experience and L2 Achievement

The third objective of the statistical analyses was to examine the extent to which the FL students' lexical appropriateness, specificity and abstractness achievement could be related to the length, focus and timing of their L2 English experiences over the course of seven years of FL education. As summarized in Table 4, the FL students' previous and recent L2 experience was diverse in terms of both quantity and quality. During secondary school (Grades 7-12), the FL students appeared to spend a wide range of time practicing L2 English inside (262-612 hours for junior high school; 525-787 hours for senior high school) and outside (0-840 hours for junior high school, 0-1500 hours for senior high school) of classrooms. Importantly, about half of the students reported receiving oral communication classes, while the other half did not.

Similarly, Table 3 shows a great deal of individual variability among the students' FL experience during their time at university. Whereas all the participants took language-focused classes (90-270 hours), some made efforts to further enrich their FL experience by taking content-based lessons (*M* = 51.5 hours; *Range* = 0-200 hours) and seeking conversation activities with other native and non-native speakers (*M* = 51.4 hours; *Range* = 0-660 hours).

Table 3
Summary of Participants' Past and Recent FL Experience

A. Length of instruction	*Mean*	*SD*	*Range*	
			Min	*Max*
Total hours of instruction at junior high school (3 years)	284.3	52.4	262	612
Total hours of instruction at high school (3 years)	618.5	30.5	525	787
Total hours of instruction at university (1 year)	192.5	104.8	90	420

B. Focus of instruction	*Mean*	*SD*	*Range*	
			Min	*Max*
Oral communication at junior high school (3 years)	Yes (*n* = 26), No (*n* = 46)			
Oral communication at high school (3 years)	Yes (*n* = 38), No (*n* = 34)			
Total hours of language-focused instruction at university (1 year)	141.2	47.9	90	270
Total hours of content-based instruction at university (1 year)	51.5	66.2	0	200

C. Extracurricular activities	*Mean*	*SD*	*Range*	
			Min	*Max*
Total hours of prep schools at junior high school (3 years)	152.3	180.2	0	840
Total hours of prep schools at high school (3 years)	254.4	271.3	0	1500
Total hours of conversation activities during university (1 year)	51.4	108.5	0	660

Lastly, a set of correlation and multiple regression analyses were performed to elucidate which experience variables were relatively important for determining the FL students' success in L2 vocabulary achievement. To avoid problems with multicollinearity, the decision was made to reduce the 10 experience variables to common underlying factors via a factor analysis. The factorability of the data was confirmed according to the Bartlett's test of sphericity ($\chi^2 = 622.898$, $p < .001$) and the Kaiser-Meyer-Olkin measure of sampling adequacy (.451). A Promax rotation was performed, yielding four factors which accounted for 70.8% of variance.

Table 4
Summary of a Four-Factor Solution Based on a Factor Analysis of the 10 Experience Variables

	Factor 1 (Recent Classroom Experience)	Factor 2 (Past Classroom Experience)	Factor 3 (Past Extra Experience)	Factor 4 (Recent Extra Experience)
A. Length of instruction				
Junior high school	.038	**.817**	-.199	.095
High school	.072	**.832**	.034	-.108
University	**.990**	.048	-.033	.040
B. Focus of instruction				
Oral communication at junior high school	-.118	-.008	**.824**	-.087
Oral communication at high school	-.306	-.138	**.532**	-.375
Language-focused instruction at university	**.867**	.178	.097	.042
Content-based instruction at university	**.909**	-.056	-.120	.034
C. Extracurricular activities				
Prep schools at junior high school	.111	-.084	**.617**	.219
Prep schools at high school	-.077	-.279	-.012	**.784**
Conversation activities during university	.183	.366	.154	**.694**

Note. All loadings > .5 were highlighted in bold.

As summarized in Table 4, Factor 1 was labeled as "Recent Classroom Experience," as it featured the total number of form- and content-oriented lessons FL students received during university. Factor 2 was labeled as "Past Classroom Experience," covering both the amount of FL instruction that the FL students had received in secondary school (Grades 7-12). However, it was difficult to interpret how the other experience variables were clustered onto Factors 3 and 4. Factor 3 covered what the students did beyond the language-focused activities inside (oral communication) and outside (prep school) FL classrooms especially during junior high school (Grades 7-9). Given that only half of the participants reported their experience in oral communication during junior and senior high school (see Table 3), this could be considered as one form of "extra" FL experience. In this regard, Factor 3 was labeled as "Past Extra Experience." Factor 4 featured what the participants did outside FL classrooms during high school (prep school) and university (conversation activities). Given that this factor corresponded to the participants' relatively recent experience outside FL classrooms (high school and university), it was labeled as "Recent Extra Experience." However, the labels for Factors 3 and 4 should be considered tentative, a limitation that I will revisit in the Limitation section.

To illustrate a general picture of the experience and achievement link, a set of Pearson correlation analyses were first performed to examine the associations between the participants' lexical factor scores (Appropriateness, Specificity, Abstractness) and experience factor scores (Past Classroom, Past Extra, Recent Classroom, Recent Extra). In line with Plonsky's (Plonsky & Oswald, 2014) field-specific benchmark, correlations were interpreted in terms of both significance and strength. As shown in Table 5, the participants' lexical appropriateness performance was significantly and strongly tied to the extent to which they had practiced inside classrooms at the university ($r = .417$, $p < .001$) (Bonferroni corrected). Marginal and weak correlations were found between lexical appropriateness and recent extra experience ($r = .417$, $p = .043$) and lexical specificity and past classroom ($r = .210$, $p = .056$).

Table 5
Summary of Correlations between Lexical Achievement and Experience Profiles

	Appropriateness		Specificity		Abstractness	
	r	*p*	*r*	*p*	*r*	*p*
Past Classroom Experience	.163	.170	.210	.056	.100	.401
Past Extra Experience	-.004	.974	-.129	.280	-.095	.425
Recent Classroom Experience	.414	<.001*	.024	.842	-.133	.264
Recent Extra Experience	.239	.043	.179	.133	-.100	.405

Note. * indicates $p < .012$ (Bonferroni corrected)

To further examine the relative weights of the experience factors for L2 vocabulary achievement, the FL students' lexical factor scores were submitted to stepwise regression analyses with the four experience factor scores—Past Classroom, Past Extra, Recent Classroom, Recent Extra—as predictor variables. According to the results, the participants' recent classroom experience significantly accounted for 17.1% of variance in their lexical appropriateness performance, $F(1, 70) = 14.458$, $p < .001$. However, the regression models did not reach statistical significance for the participants' specificity and abstractness performance at a $p < .05$ level.

Discussion

To date, there has been an increasing amount of evidence that older and more cognitively mature learners likely improve their L2 proficiency to a greater extent than younger learners in the FL educational context, which consists of only several hours of explicit L2 learning per week (e.g., Muñoz, 2006, 2008, 2014). In the context of 72 second-year university students who had practiced L2 English exclusively under such FL conditions for seven years without any study abroad experience (6 years of secondary school + 1 year of university), the current study aimed to elucidate the long-term effectiveness of FL education on vocabulary aspects of L2 speech attainment in relation to the length, focus and timing of relevant L2 experience. Whereas the nine measures were adopted in line with Crossley's framework of L2 vocabulary use (Crossley et al., 2015; Kyle & Crossley, 2015), the results of the factor analyses showed that they were found to tap into three different dimensions of the participants' L2 lexical proficiency—appropriateness (global, semantic, morphosyntactic accuracy), specificity (frequency, range) and abstractness (concreteness, meaningfulness, imageability, hypernymy).

In response to RQ1, which inquired about the pedagogical potentials and limits of long-term L2 vocabulary learning in FL settings (950-2966 hours), the participants' lexical appropriateness, specificity and abstractness performance was compared to that of experienced Japanese learners of English in Canada. In this study, the experienced Japanese learners (rather than native speakers of English) were chosen as a baseline group, as their performance could serve as a realistic goal/endpoint of attainment for the FL students in the current study. Overall, extensive FL education (7 years) led certain participants to reach comparable levels of lexical proficiency with the degree of achievement being uniquely related to different dimensions of L2 vocabulary use (Specificity > Abstractness > Appropriateness). According to the results, (a) a majority of the participants' performance in specificity was considered to be successful, as they fell within two SDs of the mean ratings of the baseline group (86.1%); (b) approximately half of FL students achieved satisfactory levels of lexical proficiency in abstractness (43.0%); and (c) only a handful of the FL students actually demonstrated such successful performance in appropriateness (33.3%).

In terms of RQ2, which inquired about the experience-related predictors for successful L2 speech learning in FL settings, the results of the correlation and multiple regression analyses indicated that the FL students' L2 achievement was generally related to a wide range of recent and previous experience factors. The findings here concur with a number of previous FL studies which have evidenced strong experience effects on classroom SLA (e.g., Muñoz, 2006, 2014; Ojima et al., 2011; Riney & Flege, 1998). More specifically, the current study revealed that lexical appropriateness (global, semantic and morphosyntactic accuracy) of the FL students' performance were moderately tied to the extent to which they had recently practiced L2 English inside language-focused and content-based classes during university (explaining 17.1% of variance).

The findings are in line with previous research on the role of FL education in L2 pronunciation learning (Authors, 2016a), confirming that L2 learners' attainment of in FL classrooms could be related to relatively recent (rather than past) FL experience. The argument on the optimal timing of receiving L2 input (i.e., more recent is better) echoes the usage-based theoretical account of SLA. According to this position, L2 learners' form-meaning mappings can be best reinforced, depending on how much (quantity), in what way (quality) and how recently (immediacy) they practice the target language. Once certain L2 forms are entrenched as

constructions, L2 learners can access them quickly and accurately when they next encounter similar conversational, learning and linguistic contexts where they may be used (Ellis, 2006).

At the same time, it is noteworthy that weak associations were also found between the specificity scores of the FL students' performance and their past L2 use (the number of instruction hours) during secondary school. These results hint that FL learning could be somewhat influenced by what L2 learners practice in the early stages of their L2 English education. The findings can be explained from two different angles. First, factors related to age of L2 learning may still account for a small portion of variance in individual differences in SLA even in FL classrooms (as argued by Larson-Hall, 2008). Second, as shown earlier, many of the FL students' achievement was considered more successful in lexical specificity (86.1%) than in lexical appropriateness (33.3%) at least in the context of the current study. This in turn suggests that the acquisition of relatively easy features (specificity) could be strongly predicted by how much L2 learners practiced the target language at the onset of their FL learning, and by extension, how quickly they could reach satisfactory levels of performance within a relatively short period of time.

Conclusion, Implications and Future Directions

The current study provided two overall conclusions on the long-term effectiveness of FL education for the lexical aspects of L2 speech learning by comparing the performance of students with 7 years of FL experience with that of experienced, naturalistic L2 English learners. On the one hand, extensive FL education (7 years) allowed the vocabulary usage of many students to become adequately appropriate and sophisticated (using less frequent, more narrowly-ranged, and more abstract words in a contextually appropriate manner). On the other hand, the degree of achievement resulting from extensive FL instruction could be uniquely influenced by different dimensions of L2 vocabulary use. According to the results, the ratio of FL learners varied in the following order: Specificity (86.1%) > Abstractness (43.0%) > Appropriateness (33.3%). In particular, the acquisition of the relatively difficult features (semantic and morphosyntactic appropriateness) could be related not only to how frequently, but also to how recently learners have practiced the target language.

Based on the findings, and considering the results of previous FL literature, some tentative suggestions can be made to help L2 learners make the most of the FL education to achieve comprehensible L2 lexical proficiency in a most effective and efficient manner. According to the current study, whereas the vocabulary use of many L2 learners was sufficiently diverse and sophisticated after 7 years of instruction, there seems to be much room for improvement in the appropriateness dimension of L2 lexical proficiency in particular. Many L2 vocabulary scholars have recommended that students should focus on mastering the form, meaning and use of the first 3000-4000 word families as a priority, as these are thought to cover most of the lexical items used in oral communication (Adolphs & Schmitt, 2001). Instructed SLA research has convincingly shown that such form/meaning mapping processes may be optimized when students are guided to attend to accuracy (typically through explicit instruction and corrective feedback) during meaning-oriented activities (for a review, see Ortega, 2009). Although policy makers, researchers and practitioners have extensively debated the incorporation of such communicative focus-on-form practices as a part of early FL education, its benefits remain highly controversial (Muñoz, 2008 vs. Larson-Hall, 2008). It is important to remember, however, that the current study adds that what directly relates to successful L2 learning in FL classrooms could be their most recent experience; in this regard, the primary focus

of any discussion on improved FL education systems should be concerned with the nature of pedagogy in the later stages of FL learning in particular (e.g., high school, university).

To close, four primary limitations of the current study need to be addressed with a view of designing more robust L2 vocabulary development studies of this kind. First and foremost, the study entirely drew on cross-sectional data, which would stop us from making any discussion as to the causality of the associations between experience and acquisition. As a remedy, efforts were made to recruit participants with unique backgrounds (i.e., studying L2 English exclusively in classroom settings since Grade 7); as such, we aimed to estimate how seven years of EFL experience could impact on vocabulary aspects of L2 learners' speech attainment in the long run. To further pursue the predictive power of experience-related factors for L2 speech learning, however, future studies should adopt a "longitudinal" design wherein scholars can track participants' development patterns for a certain period of time in relation to the quantity and quality of instruction (cf. Saito, 2018; Saito & Hanzawa, 2017).

Second, all the findings need to be interpreted with caution, as they derived from the very specific population (72 FL students in Japan and 10 experienced Japanese residents in Canada) using only one single, highly structured task: describing an eight-frame picture cartoon. As the results of the factor analyses pointed out, the participants' L2 lexical proficiency could be clustered into three subgroups of appropriateness, specificity and abstractness. However, it needs to be emphasized that the three-factor solution is "specific" to the context of the current study; and that the generalizability of the framework should be further tested with larger dataset with L2 learners with various L1 backgrounds and diverse proficiency levels (see the lexical correlates of "sophistication," Kyle & Crossley, 2015).

For future research, multiple task formats should also be adopted. According to the previous task-based language assessment literature (e.g., Skehan, 2014), it has been shown that L2 learners are induced to pay more attention to accuracy aspects of language when tasks entail tight structures with clear storylines to convey, compared to when tasks are more freely-structured, personal, informal and familiar (e.g., oral interview). In the latter tasks with a lot of freedom in structure and content, L2 learners may prioritize conceptualization (what to say) over formulation (how to say), supposedly demonstrating more increased complexity and sophistication in their lexicogrammar use (e.g., Foster & Skehan, 1996). Such free-speaking tasks should be adopted to elicit and assess L2 learners' lexical sophistication knowledge in particular (using more specific and abstract words).

Third, although the study supported the role of FL experience in elucidating the lexical aspects of L2 speech learning, such experience factors explained at most 20% of the variance. Although the experience questionnaire was adopted based on the previous FL literature (Authors, 2016a; Larson-Hall, 2008; Munoz, 2014), the results here indicate that this methodology needs more elaboration, validation and refinement in order to better capture precisely what FL students have generally experienced inside and outside classrooms at different time points. Indeed, the loadings of the experience factors were unclear/problematic, particularly for Factors 3 and 4 (Past and Recent Extra Experience). To this end, future studies may need to adopt not only a cross-sectional, but also a longitudinal design to track the complex portraits of FL experience over time (e.g., Saito, 2018; Saito et al., 2018; Riney & Flege, 1998).

The small-to-medium predictive power of the experience factors also suggest that other learner factors beyond experience (e.g., individual differences) could be related to successful foreign language learning. Indeed, Ortega's (2009, p. 146) individual differences framework has suggested that L2 learners' processing of received input differs— especially in classroom

settings—according to three main learner *internal* factors: cognition (how they internalize language in the brain), conation (how they undertake language-related actions of their own will) and affect (how they feel while using language). In the field of SLA, it has been shown that L2 learners' global proficiency is generally corelated with aptitude (Skehan, 2014), conation (Boo, Dörnyei. & Ryan, 2015) and emotion (Gkonou, Daubney, & Dewaele, 2017); to our knowledge, however, such topics have never been examined as potential predictors for L2 vocabulary learning in classroom settings.

Finally, as acknowledged earlier, the current study focused on a very unique group of L2 learners who had learned the target language only through FL education. Therefore, the findings need to be replicated with different groups of L2 learners. As an intriguing direction of future L2 vocabulary research, it would be interesting to probe how L2 learners' appropriateness and sophistication changes over an extensive period of immersion in naturalistic settings (length of residence = 1-10 years), and the extent to which their lexical proficiency can ultimately approximate nativelike performance according to different ages of arrival. Examining such experience and age effects on L2 lexical attainment would provide us with a full-fledged picture of the role of experience and age in L2 speech development and ultimate attainment, which has been exhaustively examined in the field of L2 phonology (but not L2 vocabulary).

References

Abrahamsson, N. (2012). Age of onset and nativelike L2 ultimate attainment of morphosyntactic and phonetic intuition. *Studies in Second Language Acquisition, 34*, 187–214. DOI: 10.1017/S0272263112000022

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics, 24*, 425–438. DOI: 10.1093/applin/24.4.425

Baker, A. (2014). Exploring Teachers' knowledge of second language pronunciation techniques: Teacher cognitions, observed classroom practices, and student perceptions. *TESOL Quarterly*, *48*, 136–163. DOI: 10.1002/tesq.99

Boo, Z., Dörnyei. Z., & Ryan, S. (2015). L2 motivation research 2005–2014: Understanding a publication surge and a changing landscape. *System, 55*, 147–157. DOI: 10.1016/j.system.2015.10.006

Cebrian, J. (2006). Experience and the use of non-native duration in L2 vowel categorization. *Journal of Phonetics, 34*, 372–387. DOI: 10.1016/j.wocn.2005.08.003

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, *33*, 497-505.

Coyle, D., Hood, P., & Marsh, D. (2010). *Content and language integrated learning*. Cambridge: Cambridge University Press.

Crossley, S. A., Kyle, K., & Salsbury, T. (2016), A usage-based investigation of L2 lexical acquisition: The role of input and output. *The Modern Language Journal 100*(3), pp. 702–715. DOI: 10.1111/modl.12344

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning, 59*, 307–34. DOI: 10.1111/j.1467-9922.2009.00508.x

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly, 45* (1), 182–193. DOI: 10.5054/tq.2010.244019

Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics, 36*, 570–590. DOI: 10.1093/applin/amt056

DeKeyser, R., & Larson-Hall, J. (2005). What does the critical period really mean? In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 88–108). New York, NY: Oxford University Press.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility. *Studies in second language acquisition*, *19*, 1-16.

Derwing, T., & Munro, M. (2005). Second language accent and pronunciation teaching: A research–based approach. *TESOL Quarterly*, *39*, 379–397. DOI: 10.2307/3588486

de Saint Léger, D. (2009). Self-assessment of speaking skills and participation in a foreign language class. *Foreign Language Annals*, *42*, 158–178. DOI: 10.1111/j.1944-9720.2009.01013.x

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics. 27*, 1–24. DOI: 10.1093/applin/ami038

Foster, R., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*, 299–323.

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, *36*, 98-116. DOI: 10.1017/S0267190515000082

Gkonou, C., Daubney, M., & Dewaele, J. M. (2017). *New Insights into Language Anxiety: Theory, Research and Educational Implications*. Multilingual Matters.

Jaekel, N., Schurig, M., Florian, M., & Ritter, M. (2017). From early starters to late finishers? A longitudinal study of early foreign language learning in school. *Language Learning*, *67*, 631-664. DOI: 10.1111/lang.12242

Kissling, E. M. (2013). Teaching pronunciation: Is explicit phonetics instruction beneficial for FL learners? *The Modern Language Journal*, *97*, 720–744. DOI: 10.1111/j.1540-4781.2013.12029.x

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*, 757-786. DOI: 10.1002/tesq.194

Larson-Hall, J. (2008). Weighing the benefits of studying a foreign language at a younger starting age in a minimal input situation. *Second Language Research, 24*, 35–63. DOI: 10.1177/0267658307082981

Lee, A. H., & Lyster, R. (2016). Effects of different types of corrective feedback on receptive skills in a second language: A speech perception training study. *Language Learning*, *66*, 809-833. DOI: 10.1111/lang.12167

Llanes, À., & Muñoz, C. (2013). Age effects in a study abroad context: Children and adults studying abroad and at home. *Language Learning*, *63*, 63-90.

Lord, G. (2008). Podcasting communities and second language pronunciation. *Foreign Language Annals, 41*, 374-389. DOI: 10.1111/j.1944-9720.2008.tb03297.x

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Review, 96*, 190–208. DOI: 10.1111/j.1540-4781.2011.01232_1.x

MEXT. (2014). Results of survey in current implementation of plans of establishing schools in 2013 academic year. Retrieved from http://www.mext.go.jp/b_menu/houdou/26/02/1344114.htm

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. DOI: 10.1002/tesq.34

Muñoz, C. (2006). The effects of age on foreign language learning: The BAF Project. In C. Muñoz (Ed.), *Age and the rate of foreign language learning* (pp. 1-40). Clevedon: Multilingual Matters.

Muñoz, C. (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Applied Linguistics, 24*, 578–596. DOI: 10.1093/applin/amm056

Muñoz, C. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Applied Linguistics, 35*, 463–482. DOI: 10.1093/applin/amu024

Nagle, C. (2017). A longitudinal study of voice onset time development in L2 Spanish stops. *Applied Linguistics*. DOI: 10.1093/applin/amx011

Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50,* 417–528. DOI: .1111/0023-8333.00136

Ojima, S., Matsuba-Kurita, H., Nakamura, N., Hoshino, T., & Hagiwara, H. (2011). Age and amount of exposure to a foreign language during childhood: Behavioral and ERP data on

the semantic comprehension of spoken English by Japanese children. *Neuroscience research*, *70*, 197-205. DOI: 10.1016/j.neures.2011.01.018

Ortega, L. (2009). *Understanding second language acquisition*. London: Hodder Education.

Ortega, L. (2013). SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn. *Language Learning*, *63*(s1), 1-24. DOI: 10.1111/j.1467-9922.2012.00735.x

Plonsky, L., & Oswald, F. L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878-912. DOI: 10.1111/lang.12079|

Plonsky, L., & Ziegler, N. (2016). The CALL–SLA interface: Insights from a second-order synthesis. *Language learning & Technology, 20*, 17–37

Riney, T. J., & Flege, J. E. (1998). Changes over time in global foreign accent and liquid identifiability and accuracy. *Studies in Second Language Acquisition*, *20*, 213-243.

Ruivivar, J., & Collins, L. (2017). The effects of foreign accent on perceptions of nonstandard Grammar: A pilot study. *TESOL Quarterly*. DOI: 10.1002/tesq.374

Saito, K., Suzukida, Y., & Sun, H. (2018). Aptitude, experience and second language pronunciation proficiency development in classroom settings: A longitudinal study. *Studies in Second Language Acquisition*. DOI: 10.1017/S0272263117000432

Saito, K. (2018). Individual differences in second language speech learning in classroom settings: Roles of awareness in the longitudinal development of Japanese learners' English /r/ pronunciation. *Second Language Research*. DOI: 10.1177/0267658318768342

Saito, K., & Hanzawa, K. (2016). Developing second language oral ability in foreign language classrooms: The role of the length and focus of instruction and individual differences. *Applied Psycholinguistics, 37*, 813-840.

Saito, K., & Hanzawa, K. (2017). The role of input in second language oral ability development in foreign language classrooms: A longitudinal study. *Language Teaching Research*.

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics, 38,* 439-462.

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016a). Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness and sense relations. *Studies in Second Language Acquisition, 37*, 677-701.

Salsbury, T., Crossley, S. A, & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research, 27*, 343–360. DOI: 10.1177/0267658310395851

Simon, E., & D'Hulster, T. (2012). The effect of experience on the acquisition of a non-native vowel contrast. *Language Sciences*, *34*, 269-283. DOI: 10.1016/j.langsci.2011.10.002

Skehan, P. (2014). The context for researching a processing perspective on task performance. In P. Skehan (Ed.), *Processing Perspectives on Task Performance* (pp. 1-26). The Netherlands: John Benjamins.

Tavakoli, P. (2018). L2 development in an intensive Study Abroad EAP context. *System*, *72*, 62-74. DOI: 10.1016/j.system.2017.10.009

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*, 905–916. DOI: 10.1017/S1366728912000168

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics, 24*, 1–27. DOI: 10.1093/applin/24.1.1

APPEDIX

Summary of Training Scripts and Onscreen Labels for Each Rated Appropriateness Dimension

Global accuracy (comprehensibility)	This dimension refers to how much effort it takes to understand what someone is trying to convey. If you can understand (what the picture story is all about) with ease, then the speaker is highly comprehensible. However, if you struggle and must read very carefully, or in fact cannot understand what is being said at all, then a speaker has low comprehensibility.
Difficult to understand ⟷ Easy to understand	
Semantic accuracy	This dimension refers to the semantic appropriateness of the vocabulary words used by the speaker. If the speaker uses incorrect or inappropriate words in context, including words from the speaker's native language, lexical accuracy is low. On the other hand, lexical accuracy is high if the speaker has all the lexical items required to accomplish the speaking task and does so using semantically precise lexical expressions.
Many inappropriate words ⟷ Consistently appropriate	
Morphosyntactic accuracy	This dimension refers to the number of errors that the speaker makes, including errors in morphological ending (e.g., tense, aspects, agreement, plurality) and word order.
Poor grammar ⟷ Excellent grammar	