Title:
Developing Second Language Oral Ability in Foreign Language Classrooms: The Role of the Length and Focus of Instruction and Individual Differences

Running Head:
L2 SPEECH IN FOREIGN LANGUAGE CLASSROOMS

Authors:
Kazuya Saito, Birkbeck, University of London
Keiko Hanzawa, Waseda University

Corresponding Author:
Kazuya Saito
Birkbeck, University of London
The Department of Applied Linguistics and Communication
30 Russell Square, London WC1B 5DT, UK.
Email: k.saito@bbk.ac.uk

Acknowledgement

Abstract

The current study aimed to examine how instruction can impact the global, segmental, prosodic and temporal qualities of second language (L2) oral ability in foreign language (FL) settings (i.e., a few hours of target language input per week). Spontaneous speech was elicited via a timed picture description task from 56 Japanese freshman college students who had studied English through FL instruction from Grades 7 to 12 without any experience abroad. The tokens were rated for global accentedness and then submitted to segmental, prosodic and temporal analyses. According to statistical analyses, (a) the participants' oral performance widely varied in relation to the length and focus of FL instruction, the frequency of their conversations in the L2, and aptitude; and (b) their diverse proficiency levels were particularly predicted by the amount of extra FL activities inside (i.e., pronunciation training) and outside (i.e., cram school) of high school (but not junior high) classrooms. The results in turn suggest that whereas extensive FL instruction (> 875hr) itself does make some difference in L2 oral ability development, its pedagogical potential can be increased by how students optimize their most immediate FL experience beyond the regular syllabus.

Key words. Foreign language learning, Classroom SLA, Second language pronunciation, Segmentals, Suprasegmentals, Fluency

Developing Second Language Oral Ability in Foreign Language Classrooms

Over the past 40 years, second language acquisition (SLA) researchers have paid much attention to how second language (L2) learners can develop their oral abilities according to a set of affecting variables, such as the quantity (Flege, 2009) and quality (Jia & Aaronson, 2003) of input and interaction, age of acquisition (Abrahamsson & Hyltenstam, 2009), language aptitude (Granena, 2013), motivation (Moyer, 1999), and willingness to communicate in the L2 (Derwing & Munro, 2013). Recently, certain researchers (e.g., Llanes & Muñoz, 2014; Muñoz & Llanes, 2014) have begun to examine the extent to which the findings from naturalistic settings, where L2 learners are intensively exposed to a target language on a daily basis, are generalizable to foreign language (FL) learning conditions, typically characterized by a few hours of L2 input per week. The current study was designed to examine (a) the extent to which adolescent Japanese learners of English[1] could enhance the global, segmental, prosodic, and temporal quality of their spontaneous speech solely after receiving six years of instruction (Grade 7-12) within FL classrooms; and (b) how the level of proficiency achieved was related to the length and focus of instruction that students had received as well as their language aptitude and motivation profiles.

Background

Predictors for Successful L2 Speech Learning

One of the most well-researched topics in SLA is the role of individual differences in determining adolescent and adult L2 learners' speaking ability in naturalistic settings (for a review, see Piske, MacKay, & Flege, 2001). For instance, it has been shown that late L2 learners' varied oral proficiency is likely determined by how much L2 input they receive, and how they use the L2 to interact with other native and non-native speakers. Although late L2 learners can choose to mainly use their first language within the same ethnic community even in an L2 speaking country, those who report high L2 use tend to demonstrate continued improvement in the segmental (e.g., Saito & Brajot, 2013) and suprasegmental (e.g., Trofimovich & Baker, 2006) aspects of pronunciation as well as listening comprehension skills and proper grammar usage (Flege & Liu, 2001) after an extensive amount of L2 immersion. These studies have provided some evidence that late learners' L2 oral proficiency development can be characterized as a gradual, constant, and extensive process (Flege, 2009).

Much research attention has also been directed towards the role of language aptitude in explaining the variability in adolescent and adult L2 speech learning. Language aptitude has been conceptualized as a subset of linguistic abilities, including phonetic coding (i.e., connecting sounds to relevant symbols), grammatical sensitivity (i.e., recognizing the functions of linguistic parts in sentences), and memory capacity (i.e., rote memorization). It has been measured by several tests including the Modern Language Aptitude Test (Carroll & Sapon, 1959), the Pimsleur Language Aptitude Battery (Pimsleur, 1966), and the LLAMA Language Aptitude Tests (Meara, 2005). L2 language aptitude is essentially innate and independent of other cognitive characteristics of individual learners, such as attitude, motivation, and beliefs about the target language (Skehan, 2002). According to Ortega's (2009) review, language aptitude scores likely explain 16-36% ($r = .4-.6$) of variance in proficiency levels, final course grades, and teachers' ratings under instructed conditions. In addition, language aptitude has been found to

---

[1] In line with the existing L2 speech literature (e.g., Flege, 2009), these students were defined as *late* learners who began learning an L2 after puberty (e.g., 12 years), and were adolescents or young adults at the time of the investigation.

predict the extent to which late learners can continue to enhance their L2 general performance and ultimately simulate nativelike performance (Granena, 2013).

Finally, another index for late SLA is related to professional and integrative motivation. For example, certain L2 learners with a high degree of motivation to speak an L2 will likely demonstrate advanced L2 oral ability (e.g., Moyer, 1999). Recently, Derwing and Munro (2013) conducted longitudinal research to investigate how late immigrants' English oral ability changed at three different time points across seven years of residence in Canada. The results showed that while those with high willingness to communicate with other interlocutors in English progressively refined their L2 oral ability as a function of additional immersion, those without such communicative intentions exhibited little improvement over time. L2 learners' concern for pronunciation accuracy has also been shown to contribute to learners' reduced foreign accentedness (Flege, Munro, & MacKay, 1995).

In contrast, there exists some empirical evidence showing that motivation is unrelated to L2 performance (Oyama, 1976; Thompson, 1991).This discrepancy in the literature suggests that defining what constitutes L2 learners' motivation in L2 speech learning is difficult, since the nature of motivation widely varies in relation to learners' goals, intentions and self-images as well as their learning contexts (Dörnyei, 1994). Future research of this kind needs to precisely examine which aspects of learners' motivation relate to linguistic development, especially by elaborating valid methods for quantifying motivation (Piske et al., 2001).

Role of Foreign Language Education

Recently, what has attracted an increasing amount of attention in the field of L2 education research is whether and to what degree the findings on L2 speech learning resulting from immersion under naturalistic conditions can be generalized to FL instructional settings. According to Larson-Hall (2008), the FL instruction involves "minimal input conditions" and typically constitutes "no more than four hours of instruction per week" (p. 36). As Muñoz pointed out (2008), the following conditions typically characterize FL learning: (a) Instruction is limited to 2-4 sessions of approximately 50 minutes per week; (b) exposure to the target language during these class periods may be limited in source (mainly the teacher), quantity (not all teachers use the target language as the language of communication in the classroom) and quality (there is large variability in teachers' oral fluency and general proficiency); (c) the target language is not the language of communication between peers; and (d) the target language is not spoken outside the classroom (p. 578-579).

Although it is clear that attaining nativelike L2 performance solely based on such limited L2 experience is extremely difficult or even unrealistic (Derwing & Munro, 2009; Levis, 2005), what remains unclear is the extent to which instruction alone can be facilitative of L2 oral ability development in FL classroom. Investigating the efficacy of FL instruction is thus of critical pedagogical and theoretical relevance to late SLA for several reasons. First, many adult L2 learners still learn their L2 in such FL classrooms as preparation for future career- and academic-related goals. There has been some suggestion that what and how they learn from FL instruction as pre-departure training strongly predicts the degree of positive change in the linguistic and affective dimensions of their SLA processes after they actually start working and studying abroad (e.g., DeKeyser, 2007). Second, although child SLA is generally facilitated by an extensive exposure to the target language (e.g., Flege, 2009), recent longitudinal investigations have provided some evidence that late SLA greatly benefits from formal instruction (sometimes even more than study-abroad), especially when it comes to short-term improvement (e.g., Llanes

& Muñoz, 2013; Muñoz & Llanes, 2014). These studies in turn suggest that further research on the facilitative role of FL instruction in late SLA is an important initiative.

Finally, it is important to note that previous L2 education research (e.g., Spada & Tomita, 2010) has yet to reveal the pure effects of instruction on late SLA, especially from a longitudinal perspective. In response to some researchers' skepticism of any significant impact of formal instruction on acquisition (e.g., Krashen, 2013), many quasi-experimental intervention studies have been conducted in FL classrooms with pre- and post-test designs. The results have revealed that certain kinds of L2 instruction, especially those combining form and meaning in a complementary manner (i.e., focus on form), can facilitate learners' linguistic performance not only at a controlled but also at a spontaneous level (Norris & Ortega, 2000; Spada & Tomita, 2010). However, as Norris and Ortega (2000) pointed out, these findings need to be interpreted with caution, since most of these instructed SLA studies have corroborated the pedagogical value of a brief amount of instruction. For example, the mean length of instruction of the 41 studies featured in Spada and Tomita (2010) was approximately three hours, ranging from 20 minutes to nine hours. In addition, these studies have documented learners' improvement on only a few targets due to the brevity of instructional treatment. Very few studies have ever examined how much L2 learners can continue to improve in various linguistic areas of L2 oral ability after an extensive amount of instruction over a prolonged period of time.

Although there still exists much room for research on the role of instruction in late SLA, some researchers have argued that studying a FL in a classroom setting may be unrelated to, in particular, the development of L2 oral ability. For example, FL classrooms have been criticized as "a *restricted setting*" (*emphasis original*, Best & Tyler, 2007, p. 19), as most lack the opportunity for systematic and abundant conversational experience with native speakers, which is thought to be the most important source of L2 speech learning (see also Long, 2007). This is because learners in many FL settings, such as Japanese students in English-as-a-Foreign-Language (EFL) classrooms (the focus of the study), learn the L2 for not only communication-oriented, but also examination-driven purposes (Kozaki & Ross, 2008). Furthermore, language-focused methods (e.g., grammar-translation teaching) are likely considered the most preferred way of L2 instruction (Yashima, Zenuk-Nishide, & Shimizu, 2004). Importantly, Doughty (2003) emphasized that instruction without much contextualized use of language simply leads to "the accumulation of metalinguistic knowledge about language" (p. 271); thus what L2 learners learn through FL education may not necessarily be integrated into their interlanguage system (Jiang, 2007). Indeed, students who receive focus-on-form instruction often fail to generalize what they have learned to communicative contexts outside of classrooms (Gatbonton & Segalowitz, 2005), and their improvement pattern for one grammatical target (e.g., progressives) is reported to be short-lived, especially when they move on to different structures (e.g., simple past) (Lightbown, 1983).

Recently, however, the quality of formal instruction has dramatically changed, especially thanks to the increasingly important role of English as an international language of communication in many parts of the world: many L2 learners of English have begun to consider the development of oral proficiency skills as a path to have successful social interaction in English in the context of a global society. For example, many secondary-level school students in Japanese FL classrooms are reported to have different kinds of intrinsic, instrumental and integrative types of motivation to study English as an L2 (e.g., entrance examinations, cultural exploration, and future business) (Kimura, Nakata, & Okumura, 2001). Since 2003, the Ministry of Education, Culture, Sports, Science and Technology in Japan has introduced a range of

educational reforms (i.e., an Action Plan to Cultivate "Japanese with English Abilities") to emphasize the importance of using English as a practical tool for communication. As Yashima et al. (2004) pointed out, the FL context in Japan (and possibly in other east Asian countries) is multidimensional rather than monolithic, such that FL students have begun to study English to pursue not only "a short-term realistic goal related to examinations and grades" but also "a somewhat vague long-term objective related to using English for international/intercultural communication" (p. 121). In light of L2 learners' changing perceptions of the goal of L2 instruction, and the growing awareness of the value of oral communication and conversation activities in many FL classrooms, it is high time to reconsider the pedagogical potential and limitations of FL instruction.

Motivation for Current Study

Though few in number, some attempts have been made to investigate the effectiveness of extensive FL instruction in late SLA. One such example is Muñoz (2006), who conducted a longitudinal investigation on how Spanish-Catalan Bilingual learners of English improved in a range of linguistic abilities. In this project, four groups of learners who started learning English at different ages (i.e., 8, 10, 14, 18+ years) were compared at three different points of time after receiving the same amount of instruction (i.e., 2, 4, 8 years). The results showed that their speaking performance, elicited via a range of oral tasks (e.g., picture narratives), increased over time regardless of their starting age profiles. Muñoz (2006) found similar outcomes in the domains of listening comprehension skills, fluency, and vocabulary. Additionally, adolescent and adult learners actually tended to demonstrate earlier and more substantial gains from the same amount of L2 instruction as child learners. Adolescent and adult learners are able to do so by making the best of their cognitive maturity (e.g., advanced logical and deductive reasoning, memory and processing capacities), literacy knowledge (e.g., larger L1 vocabulary size, greater phonological and morphological awareness) and accumulated experience at school (e.g., familiarity to learning the L2 under minimal input conditions) (see also Muñoz & Singleton, 2011).

The current study aimed to further examine how and to what extent FL instruction alone enables adolescent and adult learners to improve their L2 oral abilities. Specifically, we focused on analyzing the global (foreign accentedness), segmental (consonant and vowel errors), prosodic (word stress, intonation), and temporal (speech rate) quality of the spontaneous speech of 56 Japanese learners of English who had just finished six years of FL education in Japan—from Grade 7 to 12 (12-17 years old)—without any experience abroad. Their performance was compared to that of 10 experienced late Japanese learners in Canada (+20 years of L2 immersion) who were assumed to represent the final state of SLA. Subsequently, we examined under what conditions such FL efficacy can be increased according to the length and focus of FL instruction that learners received, as well as their motivation and language aptitude profiles. Accordingly, two research questions were formulated as follows:

1. To what extent can an extensive amount of FL instruction (6 years) impact the development of adolescent L2 learners' oral ability?
2. Which variables—length and focus of instruction, frequent L2 conversation, aptitude, motivation—predict the outcomes of late SLA in FL classrooms?

Method

Participants

FL students. In total, 56 freshman students at a university in Japan voluntarily participated in the study (age range: 18-19 years). Data collection was administered within one

month after the students had entered the college. Students were recruited via a flyer which specified the necessary conditions for participating in the project:

- Participants must be native speakers of Japanese (they must have received language input only in Japanese from their native speaking Japanese parents from birth).
- They had started learning English from secondary school (the fact that these students had not received any English lessons at elementary and/or private language school led us to assume that they had *zero* knowledge of English at the beginning of Grade 7).[2]
- They had never traveled in English speaking countries for more than one month. None had any prior study-abroad experience (this factor ensured that they had studied English only through FL instruction).

Participating students were majoring in either business and marketing, or international relations and liberal studies. According to their language background questionnaire, they received only a few hours of English lessons per week during junior high school and high school. As reported in previous research (e.g., Yashima et al., 2004), the content of the FL syllabus in Japanese English education is typically two-fold. Whereas teachers and learners focus on memorizing vocabulary and idiomatic expressions, practicing sentence translations and engaging in intensive and extensive reading as a main and short-term goal, they gradually start paying attention to oral communication and conversation activities as a secondary, long-term goal. For details of the length and type of FL instruction that the participants received, see Table 3.

Experienced Japanese learners. To establish a baseline for the current study (i.e., the upper limit of late L2 learners' oral performance), rather than using native speakers of English, the decision was made to recruit highly experienced Japanese learners of English who had already reached ultimate attainment due to their extensive amount of L2 experience. Many SLA scholars (e.g., Cook, 2002; Ortega, 2009) have emphasized that any L2 phenomenon needs to be examined within non-native speakers themselves rather than in relation to a native speaker model, given that few non-native speakers can actually achieve perfect nativelike performance, especially when they start learning the L2 after the age of 12 (Abrahamsson & Hytemstam, 2009), and that non-native accents are a normal aspect of L2 speech production (Derwing & Munro, 2009; Flege et al., 1995).

In line with age-related SLA research standards (e.g., DeKeyser, 2013), 10 late experienced Japanese learners at the point of ultimate attainment were carefully recruited in Vancouver based on the quantity and quality of their extensive L2 experience. They had all arrived in Canada after the age of 18 ($M_{\text{age of arrival}}$ = 24.1 years), resided there for more than 20 years ($M_{\text{length of residence}}$ = 24.7 years), and reported that their main language of communication either at home or work was English. According to their language background questionnaire, they indeed demonstrated highly frequent use of English (M = 5.7 from 1 = *Very infrequent* to 6 = *Very frequent*).[3] Their performance was thus judged to well reflect the end state of late SLA, the

---

[2]Although some learners in other FL contexts may report some contact with the L2, especially through informal activities such as watching TV (Muñoz, 2014), it is extremely rare to find any TV programs solely in English without dubbing nor subtitles on any public TV stations in Japan. Although there is some possibility that certain Japanese parents may be interested in sending their kids to private language institutes or/and exposing them to English TV programs via DVDs, any participants with such early English education backgrounds were carefully controlled for and eliminated from the current analysis.

[3] In contrast, their use of French (the other official language in Canada) and Japanese (their first language) was limited (M = 1.1, 2.4, respectively).

result of much experience and practice, and was considered *near-nativelike* in performance (Abrahamsson & Hyltenstam, 2009).

Speaking Task

Traditionally, L2 speech has been elicited via controlled speech tasks, such as paragraph and sentence readings (i.e., repeating audio and written prompts) (Piske et al., 2001). However, since adult L2 learners can carefully monitor the linguistic forms they use (Jiang, 2007), such highly controlled performance has been criticized for eliciting "language-like behavior" rather than "actual L2 proficiency" (Abrahamsson & Hyltenstam, 2009, p. 254). To tap into the present state of L2 learners' interlanguage representations, many SLA researchers have emphasized the importance of adopting spontaneous speech tasks, during which L2 learners are induced to pay equal attention to the phonological domain as well as the temporal, lexical, grammatical, and discoursal domains of language to convey their communicative intentions (Spada & Tomita, 2010) under time pressure (R. Ellis, 2005).

Similar to previous L2 pronunciation studies (e.g., Derwing & Munro, 2009; Trofimovich & Isaacs, 2012), participants' spontaneous speech was elicited via a timed picture description task. As conceptualized and validated in Saito, Trofimovich and Isaacs (in press-a), the task adopted in the study was carefully designed to elicit a certain length of spontaneous speech data without excessive hesitations and dysfluencies from the participants, who had a wide range of L2 proficiency. First, instead of using a series of thematically-linked images (e.g., Derwing & Munro, 2009), speakers described seven separate pictures, with three keywords printed as hints. Second, to control for speakers' lack of familiarity with the task, the first four pictures were used for practice and the last three were targeted for analyses. Third, to minimize the amount of conscious speech monitoring (see R. Ellis, 2005), speakers were given only a very small amount of planning time (i.e., only 5s) before describing each picture.

The three target pictures depicted a table left out in a driveway in heavy rain (keywords: rain, table, driveway), three men playing rock music with one singing a song and the other two playing guitars (keywords: three guys, guitar, rock music), and a long stretch of road under a cloudy blue sky (keywords: blue sky, road, cloud). The keywords were intentionally chosen to push Japanese learners to use problematic segmental and syllable structure features and show their pronunciation abilities. For instance, Japanese speakers have been reported to neutralize the English /r/-/l/ contrast ("rain, rock, brew, crowd" vs. "lane, lock, blue, cloud") and to insert epenthetic vowels between consecutive consonants (/dəraɪvə/ for "drive," /θəri/ for "three," /səkaɪ/ for "sky") and after word-final consonants (/teɪbələ/ for "table," /myuzɪkə/ for "music") in borrowed words (i.e., Katakana).

All speech recordings were carried out individually with both the FL students and experienced Japanese learners in university labs using a digital Marantz PMD 660 audio recorder (44.1 kHz sampling rate with 16-bit quantization). To ensure that all speakers understood the procedure, the researcher (a native speaker of Japanese) delivered all instructions in Japanese. The participants then described the seven pictures, using the first four pictures as a practice. The remaining three pictures (A, B, C, in that order) were used for the main analysis. In total, the speakers generated 168 picture descriptions (3 pictures by 56 Japanese and 10 English speakers).

On average, about 5-10s from the beginning of each description was extracted for each speaker. Three picture descriptions (Pictures A, B, C) for each speaker were combined and stored in a single audio file, resulting in a total mean length of 25s for the three picture descriptions combined (18.5-40.3 s). Compared to the 15-30s samples used for rating in similar pronunciation studies (e.g., Derwing & Munro, 2009), the entire duration of these samples was

considered to be sufficient for eliciting elicit listeners' impressionistic ratings of speech. In total, 66 speech samples were created from 56 FL students and 10 experienced Japanese learners.

Global Analyses

The global quality of L2 speech was assessed by native speaking raters on the continuum of *foreign accentedness*. The accentedness index refers to how different an L2 speaker's accent sounds from that of the native-speaker community (e.g., Derwing & Munro, 2009) and is measured via naïve listeners' intuitions without relying on training and background (e.g., Flege et al., 1995; Muñoz & Llanes, 2014). This measure has been extensively used in the previous L2 speech literature (e.g., Flege et al., 1995) and is reported to reflect various aspects of language, including pronunciation, fluency, vocabulary and grammar (Trofimovich & Isaacs, 2012).

Raters. Following the definition of naïve listeners in Isaacs & Thomson (2013), five native speakers of English (2 males, 3 females) were recruited at an English-speaking university in Montreal. The raters were born and raised in English-speaking homes in Canada ($n = 3$ in Montreal, 2 in Toronto). All of the raters ($M_{age} = 21.6$ years) were undergraduate students with non-linguistic backgrounds (e.g., business, psychology) and reported no previous teaching experience in SL/FL classrooms. They reported relatively low familiarity with Japanese-accented English ($M = 1.8$ from 1 = *Not at all* to 6 = *Very much*). None of the raters reported any hearing problems.

Procedure. The test was run offline using a custom software, Z-Lab (Yao, Saito, Trofimovich, & Isaacs, 2013), developed using commercial software package (MATLAB 8.1, The MathWorks Inc., Natick, MA, 2013). The raters used a free moving slider on a computer screen to assess the foreign accentedness of the speech samples. If the slider was placed at the leftmost end of the continuum, labeled with a frowning face (indicating very negative), it was recorded as "0"; if it was placed at the rightmost end of the continuum, labeled with a smiley face (indicating very positive), it was recorded as "1000." First, the raters received a brief explanation of the construct of foreign accentedness from a trained research assistant (for training scripts and onscreen labels, see the Appendix). After a practice run, wherein the raters rated three speech samples (not included in the main dataset), they listened to 66 speech samples in a randomized order. To tap into the initial intuitions and impressions of foreign accented speech, each sample was played only once for the raters' judgment. The listening test was designed such that the raters were allowed to make foreign accentedness judgments only after listening to the entire sample. The raters were always reminded that the entire speech samples well represented a wide range of Japanese learners of English with various proficiency levels (not only FL students but also experienced Japanese learners), and were thus encouraged to use the entire scale as much as possible. The whole session took approximately one hour.

Inter-rater reliability. Similar to previous research (e.g., Derwing & Munro, 2009), the five inexperienced raters showed high reliability values among their accentedness ratings, (Cronbach's alpha = .94). Thus, mean rating scores were calculated by pooling over the five inexperienced raters, and then given to each token produced by the participants.

Pronunciation and Fluency Analyses

In the study, L2 oral ability was defined not only as a broad concept of global foreign accentedness, but also as a specific phonological phenomenon, spanning segmentals, prosody, and fluency (Trofimovich & Isaacs, 2012). Such subdomains of L2 speech have typically been measured via objective instruments of acoustic analyses (Piske et al., 2001). Since these measures are designed to analyze the segmental and temporal features of L2 speech when the phonetic contexts of the target sounds (e.g., following and preceding vowels, speech and

articulation rate) are strictly controlled, it remains unclear whether they can be appropriately applied to more uncontrolled and conversational speech samples.

In this regard, recent L2 pronunciation studies have also used extensively trained raters' subjective judgments of the pronunciation and fluency aspects of L2 speech. For example, previous research has examined segmentals (Piske, Flege, MacKay, & Meador, 2011), prosody (Field, 2005), and temporal fluency (Bosker, Pinget, Quené, Sanders, & de Jong, 2013; Derwing, Rossiter, Munro, & Thomson, 2004). In these studies, native speaking raters (usually with much pedagogical and linguistic experience) directly assess specific aspects of L2 speech embedded in extemporaneous speech after receiving explicit training on the target sounds being evaluated. The human judgement method has been found to be highly trustworthy, because these raters can selectively attend to the targetlikeness of segmentals, prosody and fluency by drawing on their own intuitions without being distracted by other non-nativelike use of language (e.g., vocabulary and grammar errors). Thus, the experienced human rating method was adopted in this study, whereby linguistically trained raters assessed pronunciation and fluency aspects of L2 speech using four measures (segmentals, word stress, intonation, speech rate) developed in the extensive L2 speech research (e.g., Derwing et al., 2004), and validated in a previous project (Saito, Trofimovich, & Isaacs, in press-b).

Raters. Different from the global analyses, following the definition of experienced raters by Isaacs and Thomson (2013), five native speaking raters (3 males, 2 females) were recruited based on their linguistic and pedagogical experience. They were born and raised in English-speaking homes in Canada (3 from Montreal, 2 from Ontario). All of them were graduate students in the Department of English at a university in Montreal. All of them had received training in phonetics and phonology, and reported a sufficient amount of teaching experience in SL/FL settings ($M = 4.0$ years from 2 to 6 years). They reported relatively high familiarity with Japanese-accented English ($M = 4.4$ from $1 = $ *Not at all* to $6 = $ *Very much*). None of the raters reported any hearing problems.

Segmental, prosodic and temporal measures. The raters listened to 66 samples played in a randomized order via Z-Lab (Yao et al., 2013). For each audio sample, they used the same moving slider (1000-point scale: $1 = $ *non-targetlike*, $1000 = $ *targetlike*) to evaluate four segmental, prosodic, and temporal aspects of L2 speech at the same time: (a) segmental errors (substitution, omission, or insertion of individual consonants or vowels); (b) word stress errors (misplaced or missing primary stress); (c) intonation (appropriate, varied versus incorrect and monotonous use of pitch); and (d) speech rate (speed of utterance delivery).

Procedure. The sessions took place in a quiet room on two different days, with the first day for training (about 3 hours), and the second day devoted to evaluating the audio files of the current dataset (about 3 hours). For training scripts and onscreen labels for the audio- and transcript-based measures, see the Appendix.

Training phase. The five raters in the current study first received thorough instructions from a trained research assistant on the eight different domains of pronunciation and fluency. The definitions and training transcripts were elaborated from previous research focusing on segmentals (Piske et al., 2011), word stress (Field, 2005), intonation (Hahn, 2004), and speech rate (Derwing et al., 2004). They then proceeded to practice the judgment procedure using the dataset of Trofimovich and Isaacs (2012), which consisted of a total of 40 non-native speakers' picture narratives.

As separately reported in detail in Saito et al. (in pree-b), the validity of their pronunciation and fluency judgments were examined from various angles. In terms of the

accuracy of their ability to judge specific phonological features in L2 speech, the raters' pronunciation and fluency judgement scores were compared with the corresponding linguistic properties that Trofimovich and Isaacs (2012) measured via a range of objective instruments (e.g., acoustic analyses). The results identified significant correlations between the pronunciation and fluency ratings and the relevant linguistic dimensions, which were briefly summarized in Table 1.

<div align="center">TABLE 1</div>

In addition, not only did the raters demonstrate relatively high inter-rater agreement calculated by Cronbach's alpha for segmentals (.93), word stress (.93), intonation (.91) and speech rate (.94), but they also reported a high level of understanding of each category on a 9-point scale (1 = I did not understand this concept at all; 9 = I understand this concept well) for segmentals ($M = 8.9$), word stress ($M = 8.7$), intonation ($M = 8.7$), and speech rate ($M = 8.9$).

Rating phase. On the second day, they first recapped the main points of Day 1 and made preparations for the rating procedure in the main rating sessions. After receiving a review of the instructions on the four pronunciation and fluency categories and familiarizing themselves with the picture prompts and key words for the current dataset, the five raters practiced rating five practice samples (i.e., picture descriptions of Japanese learners not included in the main analysis). For each sample, the raters explained their decisions and received feedback on their accuracy based on their understanding of the categories. Subsequently, the raters proceeded to perform audio-based judgments of 66 audio files.

Inter-rater reliability. Similar to the training phase, high inter-rater agreement was found among the five experienced raters' linguistic judgment in terms of pronunciation (Cronbach's $\alpha_{segmentals} = .97$: $\alpha_{word\ stress} = .95$; $\alpha_{intonation} = .94$) and fluency ($\alpha_{speech\ rate} = .95$). The raters' scores were therefore considered sufficiently consistent and were averaged across five experienced raters to derive a single score per rated category for each speaker.

Interrelationships between linguistic scores. Simple correlation analyses were performed to investigate the degree of independence between the audio ratings (see Table 2). A Fisher r-to-z transformation was also conducted to check the different strength of the correlation coefficients ($p = .008$, Bonferroni corrected). For the audio-based measures, the raters' segmental scores were more strongly related to their word stress scores ($r = .96$) than speech rate scores ($r = .76$, $p < .001$). The speech rate scores were more closely related to intonation ($r = .92$) than segmentals ($r = .76$, $p < .001$) or word stress ($r = .80$, $p = .006$). The results suggested that the four rater-based linguistic categories were considered to tap into three domains of L2 phonological proficiency—correct word pronunciation (segmentals, word stress), prosody (word stress, intonation), and rhythmic fluency (intonation, speech rate).

<div align="center">TABLE 2</div>

Questionnaire Instruments

The FL students filled out a questionnaire which consisted of a set of items regarding the length and focus of FL instruction they had received in junior high school and high school as well as the frequency of L2 conversation and their motivation for learning English at the time of the project (see Table 3). Acknowledging that the construct validity of self-reports remains controversial because some students may have difficulty remembering (Piske et al., 2001), the participants were guided to report their previous FL learning experience during interactive interviews with the researcher, similar to what was done in Muñoz (2014). The items included for the final analysis were grouped into four sub-categories:

(1) **Length of instruction**. Although previous FL studies do not agree on the significance of age of initial learning on acquisition (Larson-Hall, 2008 vs. Muñoz, 2006), length of instruction has been found to be a significant predictor of FL success (Muñoz, 2008, 2014). Following FL research standards, the length of instruction was measured by asking participants to retrospectively self-report the total number of hours of FL instruction inside (e.g., English language arts lessons) and outside (e.g., cram schools) the classroom in junior high school and high school.

(2) **Focus of instruction**. It has been documented that Japanese EFL classrooms have begun to increase the amount of speaking activities and the number of native speaking teachers, despite their continuing strong emphasis on exam preparation through grammar translation teaching methods (e.g., Kozaki & Ross, 2011). In addition, certain L2 education researchers have emphasized the key role of pronunciation training as a part of oral communication classes in order to enhance the perceived comprehensibility of students' speech (Derwing & Munro, 2009). In our study, the presence of not only oral communication classes (taught by native and non-native teachers) but also any pronunciation training during junior high school and high school was surveyed through the questionnaire.

(3) **Frequency of L2 conversations**. Given the significant role of frequent L2 use through conversation with native and non-native speakers in late SLA in naturalistic (Flege, 2009) and classroom (Muñoz, 2014) settings, we also examined if this variable facilitated L2 oral ability development under FL conditions. As in previous research (Flege, 2009), the participants were asked to self-report the total number of minutes of conversation with native and non-native interlocutors per week at the time of the project. Unlike the oral communication classes, which provided teacher-centered speaking activities, this factor was included to reveal to what degree the participating students made an effort to find other native and non-native speakers of English in Japan, and actually interact with them in English in a meaningful manner.

(4) **Motivation**. The L2 motivation questionnaire was carefully tailored to the Japanese EFL context, where FL students likely have "dual orientations for studying English" with an equal focus on test preparation and intercultural communication (Yashima et al., 2004, p. 121). FL students were asked to rate the amount of their integrative (e.g., expanding cultural knowledge and perspectives, making English-speaking friends) and instrumental (e.g., studying and working abroad) motivation for learning English on a 6 point scale (1 = *Disagree*, 6 = *Agree*).

Language Aptitude

The FL students' language aptitude was measured by the LLAMA test (Meara, 2005). Building on the Modern Language Aptitude Test (Carroll & Sapon, 1959), this test consists of four subtests focusing on vocabulary learning, grammatical inference, sound-symbol correspondence, and sound recognition. The entire testing session took approximately 30 minutes. Similar to previous research on the relationship between LLAMA test scores and naturalistic SLA (Granena, 2013), the participants' language aptitude was calculated using a composite score derived from their individual performance on each sub-test (recorded from 0 to 100).

<div align="center">

Results

</div>

Individual Differences among FL Students

The first aim of the statistical analysis was to provide an overview of the individual differences among the 56 FL students. Since seven participants did not complete all of the items on the questionnaire for various individual reasons, the descriptive results reported here were based on the questionnaires of 49 students (see Table 3). Over six years of secondary school education in Japan, the participants received an average of 932.1 hours of FL instruction (range: 875-1662) and 365.5 hours of extra FL activities, such as assignments and cram school instruction (range: 0-1155). At the time of the project, the students reported very limited opportunities to speak in the L2 with native speakers and non-native speakers outside of the classroom, spending approximately only five minutes per week. Based on the above, it can be said that the learning environment of the participants in this study concurs with the pre-existing definition of FL classrooms (Larson-Hall, 2008; Muñoz, 2008).

With respect to the focus of FL instruction, about a half of the students reported that their syllabus included oral communication which was primarily taught by either Japanese English teachers or native speaking teachers. Despite the importance of pronunciation instruction in L2 speech learning as noted by many experts (e.g., Derwing & Munro, 2009), only a small portion of the students reported receiving pronunciation-focused training ($n = 5$ for junior high school, $n = 13$ for high school). The results of the LLAMA test indicated that the students had very diverse language aptitude profiles (35-78 out of 100 points). Finally, although the students were equally motivated to learn English to study abroad in the near future, the levels of their professional (job-related) and integrative (expanding cultural perspectives) motivation varied greatly.

TABLE 3 HERE

Effects of FL Instruction

The second aim of the statistical analysis was to closely examine and compare the L2 oral ability of Japanese FL students and experienced Japanese immigrants in Canada. As summarized in Table 4, the students' speaking performance was positively evaluated, with mean linguistic scores of 500 out of 1000 in all of the global and phonological domains. Since none of the participants were rated as "zero," the results indicated that six years of FL instruction did make some tangible impact on the Japanese students' pronunciation and fluency abilities as well as their overall foreign accentedness. At the same time, their performance was subject to a great deal of individual variability (i.e., their linguistic scores widely ranged from 100 to 800).

TABLE 4 HERE

A set of independent-samples t-tests showed that their performance was significantly different from that of the experienced Japanese learners with large effects for foreign accentedness ($t = -10.69$, $p < .001$, $d = 3.48$), segmentals ($t = -10.53$, $p < .001$, $d = 3.62$), word stress ($t = -9.97$, $p < .001$, $d = 3.25$), intonation ($t = -7.79$, $p < .001$, $d = 3.03$), and speech rate ($t = -7.16$, $p < .001$, $d = 3.17$). In addition, we also examined how many FL students could reach the range of these experienced Japanese learners' performance. Following the research literature on nativelikeness (see DeKeyser, 2013), we calculated the means and standard deviations (SD) of the baseline group for each speech measure, and then counted how many FL students' oral performance fell within two SDs of the baseline mean values. Out of the 56 FL students, very few reached this nativelike performance for accentedness ($n = 4$), segmentals ($n = 2$), word stress ($n = 3$), and intonation ($n = 7$). Furthermore, none of them showed such high proficiency in terms of speech rate.

Predictors for Successful FL Learning

The third aim of the statistical analysis was to identify which variables—length and focus of instruction, L2 conversation, language aptitude, motivation—influenced the individual

differences among the FL students' oral ability. To this end, we report here whether the 16 variables were significantly related to the FL students' global and phonological aspects of L2 speech using Spearman rho correlation analyses, and how these variables differentially interact to predict the students' oral ability using factor and regression analyses.

Correlation analyses. As seen in Table 3, some items on the questionnaire had very large standard deviations (e.g., Q2, Q4, Q11, and Q12). Thus, a set of Spearman rho correlation analyses (appropriate for nonparametric data) was conducted to check for the presence of any significant link between the 16 questionnaire variables and 5 proficiency scores. According to the results (Table 5), four significant predictors were identified including: (a) the total amount of instruction outside of high school (for accentedness, segmentals, word stress); (b) pronunciation training in high school (for segmentals); (c) the frequency of conversations with non-native speakers at the time of the project (for all measures); and (d) aptitude (for segmentals, word stress, speech rate).

TABLE 5

Factor and regression analyses. While the correlation analyses found a general pattern that the FL students' oral ability significantly varied according to the aforementioned affecting variables, it was important to further pursue the relative predictive power of these variables for the impact of FL instruction on L2 speech learning. To avoid multicollinearity problems, we first examined the set of 16 predictors in Table 3 to see if it could be reduced by combining the predictors into factors. The raw questionnaire scores were submitted to a Principal Component Analysis (PCA) with Varimax rotation and the Kaiser criterion eigenvalue set at 1. The factorability of the entire dataset was examined and validated via two tests: Bartlett's test of sphericity ($\chi^2 = 295.35$, $p < .001$) and the Kaiser-Meyer-Olkin measure of sampling adequacy (.361).

As summarized in Table 6, the PCA revealed six factors accounting for 69.1% of the total variance in the original dataset. The resulting six PCA factors were then used as predictor variables in separate stepwise multiple regression analyses to examine their contribution to the global, segmental, prosodic, and temporal qualities of the FL students' oral ability, respectively.

TABLE 6 HERE

To determine the appropriateness of conducting a set of multiple regression analyses with a relatively small sample size ($N = 49$), several necessary conditions were carefully checked. First, as explained above, the 16 predictors originally included in the questionnaire were reduced to 6 predictors by way of PCA. Second, the normality of each dependent variable (the global, segmental, prosodic, and temporal scores) was confirmed by Kolmogorov-Smirnov tests ($p > .05$). Finally, it was determined that the power to find a medium effect size in a multiple regression with 49 participants was .58, which has been considered a minimum requirement ($> .50$) in the field of second language acquisition research (Larson-Hall, 2010).

According to the results of the multiple regression analyses, Factor 1 significantly explained variance in foreign accentedness (14.7%), segmentals (16.3%), and word stress (16.4%); the other factors did not reach statistical significance as predictors for the FL students' L2 oral ability (see Table 7). Factor 1 consisted of three variables (length of FL instruction outside of the classroom during high school, pronunciation training in high school, conversation with non-native speakers at the time of the study), and was labeled "recent and extra FL experience." This is because the variables clustered in this factor concerned the degree to which the students maximized their FL experience beyond the regular syllabus at school via cram

schools, pronunciation training and conversation with non-native speakers, especially in the latter part of FL education (Grades 10-12).

TABLE 6 HERE

Linguistic Correlates of Foreign Accentedness

The final aim of the statistical analysis was to examine how the global construct of L2 oral ability (foreign accentedness) was related to the four linguistic categories (segmentals, word stress, intonation, speech rate) which tapped into three domains of L2 phonological proficiency—correct word pronunciation (segmentals, word stress), prosody (word stress, intonation), and rhythmic fluency (intonation, speech rate). The results of the simple correlation analyses ($p = .003$, Bonferroni corrected) showed that foreign accentedness was significantly correlated with all linguistic categories (segmentals, word stress, intonation, speech rate), and its relationship with correct pronunciation of words (segmentals, word stress) was particularly strong ($r > .70$).

TABLE 7 HERE

Discussion

In light of the growing body of empirical evidence showing that older and more cognitively mature learners can achieve greater gains at a faster rate compared to younger learners when the amount of L2 input and interaction is extremely limited in foreign language settings (e.g., Muñoz & Singleton, 2011), the main purpose of the current study was to further scrutinize the complex mechanisms underlying the facilitative role of FL instruction in late L2 oral ability learning. To this end, we analyzed the global, phonological, and temporal qualities of the spontaneous speech of Japanese freshman college students with a history of FL instruction from Grades 7 to 12, and no experience abroad.

Our first research question asked to what extent extensive FL instruction impacted the development of adolescent learners' oral abilities. Compared to when they started learning English (Grade 7, with no knowledge of the target language), the students demonstrated intermediate level linguistic scores (300-500 out of 1000 points) for their L2 speaking performance in terms of foreign accentedness as well as pronunciation and fluency abilities at the time of the project (when they had completed six years of FL learning). Their performance as a whole ($n = 56$) was significantly different from a baseline group of experienced Japanese learners in Canada who had reached the final state of naturalistic SLA after 20 years of L2 immersion. Very few of our participants reached the range of the baseline group's performance solely based on FL instruction. In response to the debate on the role of FL instruction (which is likely decontextualized in nature and void of ample opportunities for conversation) in late SLA (Norris & Ortega, 2000 vs. Spada & Tomita, 2010), these results provided some indication regarding its potentials (i.e., some positive change in all domains of learners' linguistic competence) and limitations (i.e., much room for improvement compared to ultimate attainment in naturalistic settings).

As for our second research question, which examined the variables predicting successful late SLA in FL classrooms, it is important to emphasize here that these FL students' oral ability varied greatly, and that some of the FL students reached the proficiency range of experienced Japanese learners. Why did certain FL students show such high-level oral ability? In line with previous research, the results of the correlation analyses demonstrated that their L2 oral ability levels were significantly related to the length of instruction (Muñoz, 2006), pronunciation training (Saito, 2012), the current frequency of L2 conversation opportunities (Muñoz, 2014), and language aptitude (Ortega, 2009).

Interestingly, the results of multiple regression analyses further revealed how these predictors interacted to determine the FL students' widely diverse speaking performance (global foreign accentedness, segmentals, word stress), which was particularly explained by a composite factor consisting of three variables related to "recent and extra FL experience". That is, to make the best of FL instruction under restricted input conditions, what is important seems to be (a) how much the students practiced English outside of classrooms during high school; (b) whether they received pronunciation training during their high school oral communication classes; and (c) how often they used the L2 in oral communication, especially with non-native speakers, at the time of the project. To summarize, whereas six years of FL instruction itself (> 875hr) led to tangible gains in all linguistic domains of L2 speech, regardless of students' various language aptitude and motivation profiles, certain students with greater amounts of *extra* FL activities tended to demonstrate better pronunciation abilities, and to speak with less perceived foreign accentedness.

Our results here concur with those of Muñoz's (2014) FL study which found a range of extracurricular practice activities (e.g., watching TV, writing letters/emails, reading books, conversations with native and non-native speakers) to be significant predictors of the fluency aspects of students' oral performance. To date, the advantages of pronunciation training (e.g., Saito, 2012) and social interaction (e.g., Flege, 2009) for adolescent and adult SLA has been extensively documented in the previous literature. However, it is uncertain why the amount of practice in cram schools outside of high school (range: 612-1332hr) was strongly related to successful FL learning in this study. Given that the chief goal of cram schools is to prepare Japanese high school students for entrance exams, it is reasonable to assume that these exams reflect what is studied in cram schools. The content of entrance exams consists mainly of reading (50% for School of International Liberal Studies; 100% for School of Commerce) and listening (30% for School of International Liberal Studies) comprehension questions; the exams are essentially designed to measure students' abilities to comprehend (but not necessarily produce) written and oral texts within time limits.

The content of the exams mentioned above leads us to speculate about two broad patterns regarding the nature of FL activities that many FL students—at least our participating students—typically experience during high school. At first, students may initially start with decontextualized activities, such as rote vocabulary memorization and discrete grammar exercises. Yet, they may ultimately be pushed towards a great deal of comprehension practice, such as intensive and extensive reading and listening activities, especially in cram schools, where they invest extra money and time with the view of attaining high scores in the entrance exams. According to comprehension-based teaching proposals, exposing L2 learners to large amounts of oral and written input may be one of the most beneficial ways to lead to successful learning not only in comprehension, but also in the production, especially for beginner L2 learners with emerging L2 knowledge (e.g., Asher, 1969 for Total Physical Response; Krashen, 2013 for The Natural Approach; VanPatten, 2004 for Processing Instruction).

Taken together, the results of this study suggest that certain L2 students can attain relatively advanced oral proficiency under FL conditions, especially when they have extra opportunities to improve not only their production performance via pronunciation training and conversation with non-native speakers, but also their comprehension skills via a great deal of reading/listening practice beyond the regular FL syllabus. At the same time, it is also important to remember that most of such successful FL learners substantially failed to reach the upper limits of naturalistic SLA—the high-level L2 speaking performance represented by the

experienced Japanese learners in Canada. Thus, the FL-only approach may not always be ideal for adequately proficient L2 learners, because their speaking performance levels-off somewhat after extensive amounts of FL instruction (e.g., Trofimovich, Lightbown, Halter, & Song, 2009). At this point, it is suggested that students need to be pushed to engage in intensive exposure to L2 input and interaction, especially via study-abroad programs, and further refine the accuracy and fluency of their output abilities (DeKeyser, 2007).

The last factor for discussion relates to the optimal timing of receiving L2 input. It is important to reiterate that the above-mentioned significant predictors for successful FL learning included what the participating students had done in Grades 10 to 12 (high school)—not in Grades 7 to 9 (junior high school).Our results are consistent with previous research which shows that successful L2 learning in FL settings can be linked to the amount of L2 input received (i.e., how much students studied the target language inside and outside of FL classrooms) (Munoz, 2014). In addition, our study contributes to the field by demonstrating that such FL efficacy may be strongly related to the type (what kinds of pedagogical activities students were involved with) and timing (how recently students received and experienced such instructional treatment) of L2 input. Although much discussion has been directed towards the quality and quantity of L2 input in late SLA, few studies have explored how the L2 experiences that learners have at different points of time affect SLA processes (e.g., Flege, 2009). This is possibly because it is methodologically difficult to measure and define L2 experience by keeping track of the amount, type, and timing of the target language exposure of certain L2 learners via longitudinal research designs (cf. Ranta & Meckelborg, 2013).

Usage-based theoretical accounts of SLA have emphasized that humans learn language as "optimal word processors" (N. Ellis, 2006, p. 8). As such, L2 learners are adaptively sensitive to not only how often (i.e., frequency), but also how recently (i.e., immediacy) certain linguistic items are used in particular discourse situations (i.e., contexts). As experience with the L2 increases, therefore, learners can attain increasingly robust associative representations by which to quickly and accurately predict and use the most relevant linguistic constructions in response to any linguistic and contextual cues. Extending this line of thought, the results of the study suggest that it is not only *how much* and *in what way*, but also *when* FL students practice the target language that relates to successful FL learning. Whereas some researchers have debated the role of early English education in FL contexts, arguably because it provides FL learners with a larger amount of instruction and practice (Larson-Hall, 2008 vs. Muñoz, 2006), our findings add that the nature and timing of FL instruction needs to be taken into account for the purposes of designing optimal FL syllabi.

Conclusion and Future Directions

To our knowledge, the current study was one of the first attempts to provide a comprehensive picture of the impact of FL instruction on L2 oral ability learning under restricted input conditions. In the context of Japanese EFL students, the results provide three broad findings: (a) the participants' oral performance widely varied in relation to the length and focus of FL instruction, the frequency of their conversations in the L2, and aptitude; (b) their diverse proficiency was particularly predicted by the amount of extra FL activities inside (i.e., pronunciation training) and outside (i.e., cram school) of high school (but not junior high) classrooms; and (c) very few reached the proficiency range of the baseline group's near-nativelike performance solely based on FL instruction. The results in turn suggest that whereas extensive FL instruction (> 875hr) itself does make some difference in L2 oral ability learning,

its pedagogical potential can be increased by how students optimize their most immediate FL experience beyond the regular syllabus.

Given the exploratory nature of the project, several directions need to be addressed for future FL studies of this kind. First, it is crucial to acknowledge that the sample size of the study was relatively small, as evidenced by the small-to-medium power (cf. Larson-Hall, 2010). The findings are based on participants with widely varying proficiency levels and heterogeneous FL profiles, and thus should be considered as tentative. Therefore, the results of the study need to be replicated with different methodologies in the context of a larger number of FL learners with various L1/L2 backgrounds. For example, although the current study exclusively concerned foreign accentedness, some L2 speech researchers (e.g., Derwing & Munro, 2009) have argued that L2 learners' oral ability should be measured based on ease of understanding (i.e., comprehensibility) and speech intelligibility, given that even heavily accented speech can be highly comprehensible and intelligible to interlocutors (Levis, 2005). Furthermore, given that the language aptitude variable (measured via the LLAMA test) was identified as a significant predictor of the effectiveness of FL instruction, it would be intriguing to test the generalizability of the findings together with other major language aptitude tests used in the field of SLA (see Ortega, 2009). Future FL studies also need to further examine other affecting variables not included in the current investigation, such as language and cognitive skills (de Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012).

Second, it is worth noting that the results of the study were exclusively based on spontaneous speech elicited from the timed picture description task. The generalizability of the findings should be tested with different task modalities, because native speakers tend to perceive the same L2 learners' oral abilities in a significantly different manner according to how their speech is elicited. Derwing et al. (2004) showed that L2 learners' comprehensibility and fluency scores were rated more positively in monologue- and dialogue-based tasks than in a picture-narrative task (see also Crowther, Trofimovich, Isaacs, & Saito, 2015). According to the task-based SLA literature, in tasks requiring online planning, L2 learners tend to use more appropriate lexical items with correct grammar (Yuan & R. Ellis, 2003), and appear to produce more complex but less speech in tasks requiring some form of decision and subjective opinions (Skehan, 2009).

Another promising direction for future research is exploring the impact of listener characteristics. The Japanese students' overall proficiency (i.e., foreign accentedness) was judged by native speakers of English who reported little familiarity with Japanese-accented English. Future research may recruit various raters, both native and non-native speakers (Derwing & Munro, 2013) with and without familiarity with foreign accented speech in the target language (Winke, Gass, & Myford, 2013), and with varying degrees of linguistic and pedagogical experience (Saito, Trofimovich, Isaacs, & Webb, in press).

References

Abrahamsson, N. & Hyltenstam, K. (2009). Age of acquisition and nativelikeness in a second language-listener perception *vs.* linguistic scrutiny. *Language Learning, 59*, 249-306.

Asher, J. J. (1969). The Total Physical Response Approach to second language learning. *The Modern Language Journal*, *53*, 3-17.

Best, C., & Tyler, M. (2007). Nonnative and second-language speech perception. In O. Bohn, & M. Munro (Eds.), *Language experience in second language speech learning: In honour of James Emil Flege* (pp. 13-34). Amsterdam: John Benjamins.

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

Carroll, J. B., & Sapon, S. M. (1959). Modern language aptitude test.

Cook, V. (Ed.). (2002). *Portraits of the L2 user* (Vol. 1). Multilingual Matters.

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility? *Modern Language Journal, 99*, 80-95.

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

DeKeyser, R. (Ed.). (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge, UK: Cambridge University Press.

DeKeyser, R. M. (2013). Age effects in second language learning: Stepping stones toward better understanding. *Language Learning*, *63*, 52-67.

Derwing, T. M. & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42*, 476-490.

Derwing, T. M., Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A seven-year study. *Language Learning*, *63*, 163-185.

Derwing, T.M., Rossiter, M.J., Munro, M.J. & Thomson, R.I. (2004). L2 fluency: Judgments on different tasks. *Language Learning, 54,* 655-679.

Dörnyei, Z. (1994). Motivation and motivating in the foreign language classroom. *The Modern Language Journal*, *78*, 273-284.

Doughty, C. (2003). Instructed SLA: Constraints, compensation, and enhancement. In M. Long & C. Doughty (Eds.), *Handbook of second language acquisition* (pp. 257-310). Malden, MA: Blackwell.

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics. 27*, 1-24.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a Second Language: A psychometric study. *Studies in Second Language Acquisition, 27*, 141-172.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly, 39*(3), 399-423.

Flege, J. E. (2009). Give input a chance! In T. Piske & M. Young-Scholten (Eds.), *Input matters in SLA* (pp.175-190). Clevedon: Multilingual Matters.

Flege, J. & Liu, S. (2001). The effect of experience on adults' acquisition of a second language. *Studies in Second Language Acquisition, 23*, 527-552.

Flege, J., Munro, M, & MacKay, I. R. A. (1995). Factors affecting degree of perceived foreign accent in a second language. *Journal of the Acoustical Society of America, 97*, 3125-3134.

Gatbonton, E. & Segalowitz, N. (2005). Rethinking the communicative approach: A focus on accuracy and fluency. *Canadian Modern Language Review, 61*, 325-353.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers*, *36*, 193-202.

Granena, G. (2013). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning, 63*, 665-703.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*, 135-159.

Jia, G. & Aaronson, D. (2003). A longitudinal study of Chinese children and adolescents learning English in the United States. *Applied Psycholinguistics*, *24*, 131-161.

Jiang, N. (2007). Selective integration of linguistic knowledge in adult second language acquisition. *Language Learning, 57*, 1-33.

Kimura, Y., Nakata, Y., & Okumura, T. (2001). Language learning motivation of EFL learners in Japan-A cross-sectional analysis of various learning milieus. *JALT Journal*, *23*, 47-68.

Kozaki, Y., & Ross, S. J. (2011). Contextual dynamics in foreign language learning motivation. *Language Learning*, *61*, 1328-1354.

Krashen, S. D. (2013). The effect of direct instruction on pronunciation: Only evident when conditions for monitor use are met? *GIST Education and Learning Research Journal*, *7*, 271-275.

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.

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

Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*, 367-377.

Lightbown, P.M. (1983). Acquiring English L2 in Quebec classrooms. In S.W. Felix & H. Wode (Eds.), *Language development at the crossroads* (pp. 101-120). Tübingen: Gunter Narr.

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.

Long, M. H. (2007). *Problems in SLA*. Mahwah, NJ: Lawrence Erlbaum.

Lyster, R. (2007). *Learning and teaching languages through content: A counterbalanced approach*. Amsterdam/Philadelphia: John Benjamins.

Meara, P. (2005). LLAMA language aptitude tests: The manual. *Swansea: Lognostics*.

Moyer, A. (1999). Ultimate attainment in L2 phonology: The critical factors of age motivation and instruction. *Studies in Second Language Acquisition*, *21*, 81-108.

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.

Muñoz, C. & Llanes, À. (2014). Study abroad and changes in degree of foreign accent in children and adults. *The Modern Language Journal*, *98*, 432-449.

Muñoz, C., & Singleton, D. (2011). A critical review of age-related research on L2 ultimate attainment. *Language Teaching*, *44*, 1-35.

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

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

Oyama, S. (1976). A sensitive period for the acquisition of a nonnative phonological system. *Journal of Psycholinguistic Research*, *5*, 261-283.

Pimsleur, P. (1966). *Pimsleur language aptitude battery*. Harcourt, Brace & World.

Piske, T., MacKay, I., & Flege, J. (2001). Factors affecting degree of foreign accents in an L2: A review. *Journal of Phonetics*, *29*, 191-215.

Piske, T., Flege, J., MacKay, & Meador, D. (2011). Investigating native and non-native vowels produced in conversational speech. In M. Wrembel, M. Kul & Dziubalska-Kołaczyk, K. (Eds.), *Achievements and perspectives in the acquisition of second language speech: New Sounds 2010* (pp. 195-205). Switzerland: Peter Lang.

Ranta, L. & Meckelborg, A. (2013). How much exposure to English do international graduate students really get? Measuring language use in a naturalistic setting. *The Canadian Modern Language Review, 69*, 1-33.

Saito, K. (2012). Effects of instruction on L2 pronunciation development: A synthesis of 15 quasi-experimental intervention studies. *TESOL Quarterly*, 842-854.

Saito, K., & Brajot, F. (2013). Scrutinizing the role of length of residence and age of acquisition in the interlanguage pronunciation development of English /r/ by late Japanese bilinguals. *Bilingualism: Language and Cognition, 16*, 847-863.

Saito, K., Trofimovich, P., & Isaacs, T. (in press-a). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*.

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

Saito, K., Trofimovich, P., Isaacs, T., & Webb, S. (in press). Re-examining phonological and lexical correlates of second Language comprehensibility: The role of rater experience. In T. Isaacs & P. Trofimovich (Eds.). *Interfaces in second language pronunciation assessment: Interdisciplinary perspectives*. Bristol, UK: Multilingual Matters.

Skehan, P. (2002). Theorising and updating aptitude. *Individual Differences and Instructed Language Learning*, *2*, 69-94.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*, 510-532.

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning, 60,* 263-308.

Thompson, I. (1991). Foreign accents revisited: The English pronunciation of Russian immigrants. *Language Learning*, *41*, 177-204.

Trofimovich, P., & Baker, W. (2006). Learning second-language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28,* 1-30.

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*, 905-916.

Trofimovich, P., Lightbown, P. M., Halter, R., & Song, H. (2009). Comprehension-based practice: The development of L2 pronunciation in a listening and reading program. *Studies in Second Language Acquisition, 31*, 609-639.

VanPatten, B. (2004). *Processing instruction: Theory, research, and commentary*. Mahwah, NJ: Erlbaum.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*, 231-252.

Yashima, T., Zenuk-Nishide, L., & Shimizu, K. (2004). The influence of attitudes and affect on willingness to communicate and second language communication. *Language Learning*, *54*, 119-152.

Yao, Z., Saito, K., Trofimvich, P., & Isaacs, T. (2013). *Z-Lab.* Retrieved August 15, 2013, from https://github.com/ZeshanYao/Z-Lab

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.

L2 SPEECH IN FOREIGN LANGUAGE CLASSROOMS

Table 1

Correlations between Pronunciation and Fluency Ratings and Coded Linguistic Variables in the Training Dataset

Rated Variable	Linguistic dimensions	Correlations
Vowel/consonant errors	Segmental errors	.64*
Word stress	Word stress errors	.72*
Intonation	Intonation errors	.54*
Rhythm	Vowel reduction	.74*
Speech rate	Articulate rate	.43*
	Mean Length of Run	.79*
	No. of filled pause	.41
	No. of unfilled pause	.49*

Note. *α < .05. The results of the correlation analyses were retrieved from Saito et al. (in press-b).

L2 SPEECH IN FOREIGN LANGUAGE CLASSROOMS

Table 2

Inter-correlations between the Audio Ratings

A. Audio ratings			
	Word stress	Intonation	Speech rate
Segmentals	.96	.85	.76
Word stress		.88	.80
Intonation			.92

Table 3

Descriptive Statistics of Individual Differences among 49 FL Students

A. Length of FL instruction	*Mean*	*SD*	*Range*
(1) Total hours of FL instruction inside of classroom at junior high school	296.4hr	61.4	262-612.5
(2) Total hours of FL instruction outside of classroom (e.g., cramming school) at junior high school	159.6hr	192.9	0-840
(3) Total hours of FL instruction inside of classroom at high school	635.7hr	105.3	525-1312.5
(4) Total hours of FL instruction outside of classroom (e.g., cramming school) at high school	205.9hr	214.0	0-735

B. Focus of FL instruction			
(5) Oral communication class at junior high school		Yes ($n = 23$) No ($n = 26$)	
(6) Native speaking teachers in oral communication class at junior high school		Yes ($n = 16$) No ($n = 33$)	
(7) Pronunciation training in oral communication class at junior high school		Yes ($n = 5$) No ($n = 44$)	
(8) Oral communication class at high school		Yes ($n = 24$) No ($n = 25$)	
(9) Native speaking teachers in oral communication class at high school		Yes ($n = 18$) No ($n = 31$)	
(10) Pronunciation training in oral communication class at high school		Yes ($n = 13$) No ($n = 36$)	

C. Amount of L2 use	*Mean*	*SD*	*Range*
(11) Total minutes of conversation with native speakers per week at the time of the project	4.2min	14.1	0-60
(12) Total minutes of conversation with non-native speakers per week at the time of the project	6.3min	21.4	0-120

D. Language aptitude	*Mean*	*SD*	*Range*
(13) LLAMA test scores (0 to 100 points)	60.4	10.9	35-78

E. Motivation	*Mean*	*SD*	*Range*
(14) I am motivated to study English to achieve work-related goals in the future (1 = *Disagree*, 6 = *Agree*)	4.2	1.3	1-6
(15) I am motivated to study English to study abroad (1 = *Disagree*, 6 = *Agree*)	5.5	0.6	4-6
(16) I am motivated to learn English to expand my cultural perspectives (1 = *Disagree*, 6 = *Agree*)	4.8	1.2	2-6

Table 4

Means and Standard Deviations for Rated Global, Segmental, Prosodic, and Temporal Qualities of FL Students and Experienced Japanese Learners' Picture Descriptions

	FL students (*n* = 56)			Experienced learners (*n* = 10)		
A. Global ratings (1000 points)	*M*	*SD*	Range	*M*	*SD*	Range
Accentedness	287	138	80-789	806	159	503-998
B. Audio ratings (1000 points)						
Segmentals	411	119	171-834	841	118	574-988
Word stress	466	123	251-837	849	112	595-989
Intonation	439	153	105-844	850	115	629-976
Speech rate	496	187	117-795	927	44	864-990

Note. 1000 point scale (*1 = Nontargetlike production, 1000 = Targetlike production*)

Table 5

Correlations between 16 Questionnaire Variables and Five Speech Measures

	Accent	Segments	Word stress	Intonation	Speech rate
A. Total hours of FL instruction					
(1) Inside/junior high school	.01	-.02	.01	.26	.16
(2) Outside/junior high school	-.02	-.06	-.13	-.19	-.12
(3) Inside/high school	-.06	-.13	-.01	.12	.22
(4) Outside/high school	.31*	.29*	.33*	.13	.13
B. Focus of FL instruction					
(5) Oral communication class/junior high school	.01	.05	.06	-.11	-.09
(6) Native speaking teachers/junior high school	.04	.10	.07	.02	.01
(7) Pronunciation training/junior high school	-26	-.19	-.21	-20	-.16
(8) Oral communication class/high school	.07	.09	.14	-.13	-.18
(9) Native speaking teachers/high school	.05	.09	.11	-.04	-.03
(10) Pronunciation training/high school	.17	.28*	.16	.05	.07
C. Frequency of L2 conversations					
(11) Conversation with native speakers	-.08	.03	.00	-.05	-.03
(12) Conversation with non-native speakers	.44*	.37*	.40*	.53*	.50*
D. Language aptitude					
(13) LLAMA test scores (0 to 100 points)	.14	.32*	.31*	.24	.30*
E. Motivation					
(14) Work-related motivation	.01	.13	-.04	-.11	-.05
(15) Study-abroad motivation	.09	.14	.11	.10	.12
(16) Integrative motivation	.07	.26	.12	.08	.09

Note. $*\alpha < .05$

Table 6

Summary of a Six-Factor Solution Based on a Principal Component Analysis of the 16 Predictors

Factor 1 (Recent and extra FL experience)	Length of FL instruction outside of classroom at high school (.88), pronunciation training at high school (.46), conversation with non-native speakers (.41)
Factor 2 (Regular FL classroom experience)	Length of FLinstruction inside the classroom at junior high school (.82) and high school (.77)
Factor 3 (Academic motivation)	Motivation for study abroad (.71), aptitude (.55), pronunciation training at junior high school (.55)
Factor 4 (Oral communication class)	Oral communication class at junior high school (.84) and at high school (.86),Native speaking teachers in oral communication class at junior high school (.80) and high school (.77)
Factor 5 (Amount of L2 use)	L2 use with native speakers (.74) and with non-native speakers (.60), Length of FL outside of classroom at junior high school (.75)
Factor 6 (Professional and integrative motivation)	Motivation for job (.62), cultural perspectives (.84), aptitude (.54)

Note. All eigenvalues > 1

Table 7.
Significant Results of Multiple Regression Analyses Using the FL Factors as Predictors of L2 Oral Ability

Predicted variable	Predictors	R	Adjusted R^2	F	p
Accentedness	Recent and extra FL experience	.408	.147	8.779	$p = .005$
Segmentals	Recent and extra FL experience	.426	.163	9.735	$p = .003$
Word stress	Recent and extra FL experience	.428	.164	9.884	$p = .001$
Intonation			*n.a. (p > .05)*		
Speech rate			*n.a. (p > .05)*		

Table 8.

Correlation Coefficients between Foreign Accentedness and Four Rated Linguistic Categories

	Segmentals	Word stress	Intonation	Speech rate
Foreign accentedness	.74*	.77*	.65*	44.*
	($p < .001$)	($p < .001$)	($p < .001$)	($p = .001$)

Note. * demonstrates statistical significance

Appendix

Training materials and onscreen labels for pronunciation, fluency and lexicogrammar judgement

A. Pronunciation and fluency categories

Segmental errors	This refers to errors in individual sounds. For example, perhaps somebody says "*road*" "*rain*" but you hear an "l" sound instead of an "r" sound. This would be a consonant error. If you hear someone say "*fan*" "*boat*" but you hear "*fun*" "*bought*," that is a vowel error. You may also hear sounds missing from words, or extra sounds added to words. These are also consonant and vowel errors.
Word stress	When an English word has more than one syllable, one of the syllables will be a little bit louder and longer than the others. For example, if you say the word "computer", you may notice that the second syllable has more stress (comPUter). If you hear stress being placed on the wrong syllable, or you hear equal stress on all of the syllables in a word, then there are word stress errors.
Intonation	Intonation can be thought of as the melody of English. It is the natural pitch changes that occur when we speak. For example, you may notice that when you ask a question with a yes/no answer, your pitch goes up at the end of the question. If someone sounds "flat" when they speak, it is likely because their intonation is not following English intonation patterns.
Speech rate	Speech rate is simply how quickly or slowly someone speaks. Speaking very quickly can make speech harder to follow, but speaking too slowly can as well. A good speech rate should sound natural and be comfortable to listen to.

☹   ☺

1. Vowel and/or consonant errors

Frequent		**Infrequent or absent**

2. Word stress errors affecting stressed and unstressed syllables

Frequent		**Infrequent or absent**

3. Intonation (i.e., pitch variation)

Too varied or not varied enough		**Appropriate across stretches of speech**

4. Speech rate

Too slow or too fast		**Optimal**