LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

Experience Effects on the Development of Late Second Language Learners' Oral Proficiency

Kazuya Saito

Birkbeck, University of London

The aim of this study was to evaluate the effects of second language (L2) experience–operationalized as length of residence (LOR) in Canada—on late Japanese learners of English. Data collected from 65 participants consisted of three groups of learners (short-, mid-, and long-LOR groups) and two baseline groups of native Japanese and native English speakers, with 13 participants in each group. The global quality of the participants' spontaneous speech production was initially judged by 10 native-speaking English raters for accentedness (linguistic nativelikeness) and comprehensibility (ease of understanding) and then submitted to segmental, prosodic, temporal, lexical, and grammatical analyses. According to the results, LOR was generally predictive of improved comprehensibility through its association with adequate and varied prosody, optimal speech rate, and proper lexicogrammar usage. In contrast, contributions of LOR to accentedness remained unclear, with less accented speech linked to refined segmental accuracy, vocabulary richness, and grammatical complexity. These findings suggest that learners continue to improve in their L2 oral proficiency over an extensive period of L2 immersion (e.g., 6 years of LOR), and they likely do so by paying selective attention to certain linguistic domains closely linked to comprehensibility—but not necessarily relevant to accentedness—for the purpose of successful L2 communication.

Keywords experience; late bilingualism; L2 oral proficiency; comprehensibility; accentedness; pronunciation

This study was funded by the Grant-in-Aid for Scientific Research in Japan (No. 26770202). I am grateful to Tracey Derwing and anonymous *Language Learning* reviewers for their helpful input and feedback on the content of this manuscript and to George Smith and Ze Shan Yao who helped with data collection and analyses. I also acknowledge many insightful comments from the journal editor, Pavel Trofimovich.

Correspondence concerning this article should be addressed to Kazuya Saito, Birkbeck, University of London, Department of Applied Linguistics and Communication, 30 Russell Square, London, United Kingdom WC1B 5DT. E-mail: k.saito@bbk.ac.uk

DOI: 10.1111/lang.12120

Introduction

Adult second language (L2) speech learning is a multifaceted phenomenon influenced by many factors such as first language (L1) influences (e.g., Best & Tyler, 2007), cognitive aging (e.g., Birdsong, 2005), attitude and aptitude (e.g., Ioup, Boustagi, El Tigi, & Moselle, 1994), motivation (e.g., Moyer, 1999), level of education (e.g., Derwing & Munro, 2009), and ethnic identity (e.g., Gatbonton, Trofimovich, & Segalowitz, 2011). There is a general consensus among researchers in the field of second language acquisition (SLA) that L2 learners' developing system is enhanced as they increase their amount of relevant L2 experience through intensive exposure to the L2. Usage-based approaches to SLA, for instance, posit that the more frequently L2 learners encounter specific linguistic features in the input, the stronger connections they establish between these features and the various contexts in which these features occur. This frequency-driven mapping of form and meaning enables L2 learners to ultimately attain automatic processing of the L2, often in terms of formulaic patterns, in response to any relevant situational and linguistic cues (N. Ellis, 2006).

What remains controversial in current SLA theories, however, concerns whether, to what degree, and how such experience effects *continue* to help late L2 learners (i.e., learners exposed to the L2 in adulthood) improve their proficiency beyond the early phase of L2 development and approximate nativelike performance in the long run. To investigate the role of experience in interlanguage development (i.e., restructuring and enhancement of the L2 system) and ultimate attainment (i.e., plateaued and asymptotic L2 performance), previous speech research has extensively focused on L2 learners' length of residence (LOR), with an assumption that longer residence in an L2-speaking country may entail a larger amount of input and interaction, compared to shorter periods of residence. Below, I will first review two competing theoretical positions—the Critical Period Hypothesis (CPH; e.g., DeKeyser & Larson-Hall, 2005) and the Speech Learning Model (SLM; e.g., Flege, 2009)—and their different predictions as to the effect of experience, operationalized as LOR, on the initial, middle, and final stages of late L2 oral proficiency development. Next, I will present the results of the current study, which examined the predictive power of LOR for the global, segmental, prosodic, temporal, lexical, and grammatical qualities of L2 speech by late Japanese learners of English from various proficiency levels (e.g., beginner to advanced).

Background

CPH

Several researchers working within the framework of the CPH have claimed that late bilinguals (i.e., those who start learning L2 after puberty) have little access to an assumed automatic learning mechanism by which to acquire language through mere exposure and that reduced access to this learning mechanism is due to the passing of a critical period for implicit language learning (often claimed to take place around puberty). The strong version of the CPH states that there is no L2 learning after puberty regardless of increased language experience (e.g., Patkowski, 1990; Scovel, 2000). The proponents of other versions of the CPH assume that benefits from L2 input and interaction in late SLA only occur during the early phase of L2 immersion (e.g., Abrahamsson & Hyltenstam, 2009; DeKeyser, 2000; Granena & Long, 2013). According to this position, post-critical-period SLA relies on explicit and conscious strategies, suggesting that adult L2 learners practice their L2 in a manner similar to learning other general cognitive skills, such as mathematics and computer programming (DeKeyser, 2007).

Previous research has indeed shown a general tendency for adult L2 learners to demonstrate quick improvement over the first few months of LOR, followed by a leveling off, despite additional linguistic input, which roughly corresponds to the power law of practice (for a review, see DeKeyser & Larson-Hall, 2005). It has been found that experienced learners (LOR > 5 years) tend to show less foreign accent in their L2 speech when their performance is compared with inexperienced learners who are still in the early phase of L2 oral proficiency development (LOR < 6 months; Flege & Fletcher, 1992; Flege, Bohn, & Jang, 1997; Trofimovich & Baker, 2006) but not with those who are beyond the initial stages of learning (LOR > 6 months; Flege, 1988; Flege, Munro, & Fox, 1994; Munro, 1993). According to Munro and Derwing's (2008) longitudinal research on adult English as a second language (ESL) learners' vowel acquisition, most L2 speech learning takes place within the first 3 to 4 months of LOR. In this regard, the CPH assumes that any continued improvement in L2 proficiency after the initial stage of learning is unrelated to further exposure and is instead attributed to individual differences, such as high language aptitude of exceptional learners. For example, Granena and Long (2013) found that late learners' near-native pronunciation abilities were exclusively limited to those with good sound–symbol correspondence skills and grammatical inferencing abilities regardless of learners' LOR profiles (see also DeKeyser, 2000).

SLM

Other researchers working within alternative theoretical frameworks, such as Flege's (2009) SLM, have emphasized that late L2 learners continue to have the capacity for language acquisition active even after puberty and may apply it to postpubertal SLA (Best & Tyler, 2007; Bialystok, 1997; Birdsong, 2006; Hopp & Schmid, 2013).[1] Specifically, as Flege (2003) pointed out, "the capacity to accurately perceive the phonetic properties of L2 speech sounds and to establish new categories based on those properties remains intact across the life span" (p. 345). Thus, extensive amounts of L2 input and interaction can lead both early and late L2 learners to achieve near-native L2 performance (Flege, 2009). Compared to early learners who tend to receive a substantial amount of native-speaker input from their caregivers and peers, late learners are typically exposed to L2 input of limited quality and quantity (Muñoz & Llanes, 2014). For example, late learners may have many opportunities to receive input from native speakers of various dialects while also receiving input from nonnative speakers from diverse L1 backgrounds. Some immigrants may also choose to use only their L1 at home and work, especially if they live in a community sharing the same language (Jia & Aaronson, 2003). Therefore, this theoretical position assumes that late L2 learners may continue to show improvement in relation to their additional L2 experience, as long as they take advantage of the social and psychological environments that early bilinguals generally benefit from, in terms of frequent use of L2 on a daily basis (Bialystok, 1997).

Flege and Liu (2001) showed that additional L2 input, through increased LOR, significantly correlates with continued L2 speech learning, provided that the main language of communication is the L2 (e.g., for university-level international students) and not the L1 (e.g., for immigrant workers). Flege (2009) found that frequency of L1 and L2 use strongly predicted the extent to which certain L2 learners could make the most of their relevant experience and thus improve the ultimate quality of their L2 performance after years of LOR (see also Flege, Frieda, & Nozawa, 1997). These studies lend some evidence to the conclusion that L2 learners never lose the capacity for language learning, such that their L2 is continuously influenced by the use of their dominant language. This then suggests that even late L2 oral proficiency development can be characterized as a gradual, constant, and extensive process, similar to the processes involved in L1 acquisition by children in the first several years of their lives (Werker & Tees, 1999).

Motivation for the Current Study

As reviewed above, the two competing frameworks—the CPH versus the SLM—offer different predictions for experience effects in late SLA. According to the CPH, a significant relationship between LOR and L2 proficiency should be observed only during the initial stage of learning (e.g., LOR < 6 months). Within the SLM, LOR can be predictive of the development of late L2 learners' oral ability given an extensive period of L2 immersion (e.g., LOR = 5–10 years). Examining the role of experience in the initial, middle, and final stages of late L2 oral proficiency development is thus crucial for SLA theory building. Such research will shed light on one of the most fundamental questions regarding the underlying mechanisms of late SLA: How long does it take for late L2 learners to reach ultimate attainment and how extensively can they improve their L2 oral ability?

One possible reason for the mixed findings among the previous studies to date could be attributed to the construct validity of the outcome measures and types of analyses used in assessing late L2 oral ability. That is, most of the aforementioned LOR research has exclusively relied on highly controlled measures, such as word and sentence repetition tasks, whereby participants repeat audio and written prompts without generating any free production. Although these measures enable researchers to elicit and analyze global (e.g., accentedness) or specific (e.g., segmental or suprasegmental accuracy) pronunciation features under highly controlled conditions, these measures also allow participants to carefully monitor their linguistic output by drawing on their explicit knowledge (R. Ellis, 2005); L2 performance of this kind arguably reflects "language-like behavior" rather than "actual L2 proficiency" (Abrahamsson & Hyltenstam, 2009, p. 254).

In fact, it has been widely reported that late L2 learners' performance significantly differs under formal (e.g., word and sentence reading) versus communicative (e.g., picture narratives) task conditions (Major, 2008). To measure late L2 learners' oral proficiency more naturalistically, many SLA researchers have emphasized the importance of eliciting L2 speech production at a spontaneous—rather than controlled—level by encouraging speakers to pay equal attention to the phonological, lexical, and grammatical aspects of their language to convey intended meanings in the most efficient and effective way (e.g., Spada & Tomita, 2010). Such spontaneous speech production is claimed to reflect the present state of L2 learners' segmental, prosodic, temporal, lexical, and grammatical competence (Hopp & Schmid, 2013; for a more detailed discussion of various controlled and spontaneous measures, see Piske, MacKay, & Flege, 2001).

Derwing and Munro (2013) conducted longitudinal research investigating how late L2 learners' spontaneous speech production changed at three different points across 7 years of their residence in Canada (LOR = 0, 2, 7 years). Following the predictions of the SLM, the results demonstrated that the immigrants with high willingness to communicate with native and nonnative speakers in English progressively refined their overall comprehensibility (ease of understanding)—rather than accentedness (linguistic nativelikeness)—as a function of additional exposure and interaction (for further definitions of comprehensibility and accentedness, see Derwing & Munro, 2009). Conversely, those without such communicative intentions exhibited little improvement in comprehensibility or accentedness over time. This study revealed a broad developmental pattern in late SLA, showing that L2 users can continue to improve their oral proficiency by selectively attending to certain linguistic features related to successful L2 communication rather than to L2 mastery.

Whereas Derwing and Munro's (2013) research has provided empirical support for the significant role of additional experience in late L2 learning, it has also brought to light several important questions that future studies need to investigate, including how to define and analyze L2 oral proficiency. On the one hand, Derwing and Munro's findings were based on global language ratings of accentedness and comprehensibility. On the other hand, L2 oral ability has been traditionally characterized as a composite (rather than monolithic) phenomenon involving various linguistic domains spanning pronunciation, fluency, vocabulary, and grammar (de Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012). Conceptualizations of L2 oral ability have substantially varied between previous studies, especially depending on whether L2 speech production is analyzed via a set of objective instruments at a fine-grained level (e.g., Crossley, Salsubury, & Mcnamara, 2014; Isaacs & Trofimovich, 2012) or based on human raters' holistic scores at a broader level (e.g., Saito, Trofimovich, & Isaacs, in press-a, in press-b; Pinget, Bosker, Quené, & de Jong, 2014; Derwing, Rossiter, Munro, & Thomson, 2004). Despite such methodological variability, however, there is a general consensus that L2 learners improve pronunciation, fluency, vocabulary, and grammar at different learning rates in initial (e.g., beginner to intermediate proficiency), compared to later (e.g., intermediate to advanced proficiency), stages of L2 oral learning (Isaacs & Trofimovich, 2012; Saito et al., in press-a).

To better understand LOR effects on L2 learning, the primary aim of the current study was to replicate the findings of Derwing and Munro (2013) while focusing on a different group of L2 learners and type of analys is. First, a cross-sectional approach was adopted to investigate spontaneous speech production of Japanese learners of English who had a sufficiently broad range of LOR profiles (see below). Second, the learners' spontaneous speech production was assessed not only for overall accentedness and comprehensibility (Derwing & Munro, 2009) but also submitted to pronunciation, fluency, vocabulary, and grammar analyses. Therefore, the second aim of the study was to examine differential effects of Japanese learners' LOR profiles on various qualities of the learners' spontaneous L2 production, including its global (accentedness, comprehensibility), segmental (consonants and vowels), prosodic (word stress and intonation), temporal (speech rate), lexical (appropriate and rich vocabulary), and grammatical (accurate and complex grammar) characteristics.

Obviously, LOR can only be considered as a rough index of the amount of language input and interaction, especially for those who reside and work in predominantly L1-based communities or engage in similar conversations over the course of a day (Piske et al., 2001). Even when the quantity of L2 input and interaction is the same, as measured through LOR, there can still be tremendous variation in the quality of experience, for example, in terms of the amount of exposure to aural and written input in the media (Jia & Aaronson, 2003). There is also no reason to doubt that adult L2 learners likely greatly differ in terms of their language aptitude (DeKeyser, 2000) and levels of education (Derwing & Munro, 2009). However, the primary goal of this study was not to test the validity of LOR as a direct measure of actual L2 experience (cf. Ranta & Meckelborg, 2013) but instead to examine the extent to which language experience, as measured through LOR profiles (among many other factors), can significantly explain variation in late L2 oral proficiency development when the analysis focuses only on learners with ample opportunities and motivation to use the L2 as their main language of communication. Following prior research (e.g., Derwing & Munro, 2013; Flege et al., 1997; Trofimovich & Baker, 2006), the current study was designed to quantitatively examine the relationship between language experience (LOR = 8 months–13 years) and late SLA, but qualitative analyses of the nature of input and interaction were not pursued.

Method

Participants

Japanese Learners

Given that LOR effects can be clearly observed among immigrants that use their L2 on a daily basis (Flege & Liu, 2001), it was important to select late learners without many opportunities to use L1 Japanese. Such participants were recruited through posts on social network Web sites for Japanese immigrants in Montreal where the Japanese community is small (e.g., .06% out of all immigrant population in Quebec, according to Statistics Canada, 2008). Based on the results of individual interviews, 39 Japanese learners of English ($M_{\text{age}} = 31.4$; 21–43 years) were carefully selected as participants following three criteria. First, because theoretical debates targeting the CPH and SLM concern whether and to what degree increased L2 experience can facilitate late SLA beyond the early phase of L2 immersion (LOR > 6 months), all 39 Japanese learners of English had more than 8 months of LOR in Canada. Second, all participants arrived in Canada after 19 years of age ($M_{\text{AOA}} = 27.3$; 19–39 years), and all had completed 6 to 10 years of English education, typically through a grammar translation method in Japan. Importantly, their age of arrival was not significantly related to their LOR ($r = -.23$, $p = .16$). Third, all participants reported very frequent use of L2, such that their main language of communication at home and/or work was English.[2] Whereas most of the participants with LOR around 1 to 2 years were enrolled in private language institutes to improve their oral proficiency in English for various academic and career goals (i.e., they had invested time and money to study abroad in Canada), those with LORs of more than a few years were either graduate students at English-speaking universities or full-time workers who dealt mainly with English-speaking customers.

The participants were divided into three equal LOR groups that were hypothesized to represent the initial, middle, and final states of late SLA: (a) Short LOR (8 months–1 year) for beginner learners who have either partially or fully completed an initial learning stage over the first 6 months of LOR (Munro & Derwing, 2008); (b) Mid LOR (1–5 years) for intermediate learners who have likely demonstrated limited improvement due to the lack of the rate of learning advantage (Munro, 1993); and (c) Long LOR (5–13 years) for advanced learners who have likely reached the stage of ultimate attainment (Patkowski, 1990). Results of a one-way between-groups analysis of variance (ANOVA) demonstrated no significant differences between the three groups in terms of age of arrival, $F(2, 36) = .29, p = .75$.

English and Japanese Baseline Groups

To provide a baseline measure of nativelike production for the global, phonological, temporal, lexical, and grammatical analyses, 13 native English undergraduate students at an English-speaking university in Montreal completed the same oral task. These participants were native speakers of northeastern Canadian and American English with a mean age of 21.3 years. To provide a baseline measure for Japanese learners with little L2 experience, 13 native speakers of Japanese who had just arrived in Canada (LOR < 1 month) were recruited at private language schools in downtown Montreal to complete the oral task. Their mean age was 28.5 years.[3] Background information about the participants is summarized in Table 1.

Speaking Task

In line with previous L2 speech research (e.g., Derwing & Munro, 2013; Hopp & Schmid, 2013; Isaacs & Trofimovich, 2012), spontaneous speech was elicited via a picture description task. Due to the relatively demanding nature of the task (Derwing et al., 2004), especially for beginner L2 learners (e.g., LOR < 1 year), it was modified such that participants (a) described seven different pictures using three key words below each picture (instead of a series of pictures in a sequence without any hints), (b) used the first four pictures as practice to get used to the task procedure and then completed the last three pictures (Pictures A, B, C, see below) for the final analysis; and (c) had only 5 seconds of planning time before describing each picture. This task allowed Japanese learners of English with a wide range of oral proficiency levels to generate spontaneous (rather than controlled) speech without too many long filled and unfilled pauses. Pictures A, B, and C 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 two others playing guitars (keywords: three guys, guitar, rock music), and a long stretch of road under a cloudy, blue sky (keywords: blue sky, road, cloud), respectively. The key words were carefully chosen to elicit segments and syllable structures especially difficult for Japanese learners of English. For example, Japanese learners have been reported to neutralize the English /r/-/l/ contrast (*rain*, *rock*, *brew*, *crowd* vs. *lane*, *lock*, *blue*, *cloud*) and substitute borrowed words by inserting 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 this way, the learners' performance was assumed to reflect the current state of their pronunciation abilities without opportunities for them to avoid using these difficult pronunciation features.

All speakers were recorded in a quiet office using a Roland-05 audio recorder, set at 44.1 kHz sampling rate and 16-bit quantization, and a unidirectional condenser microphone. All instructions were delivered in Japanese by the researcher (a native speaker of Japanese) to ensure the speakers' clear understanding of the procedures. The speakers first described four pictures

Table 1 Characteristics of participant groups

	Japanese baseline ($n = 13$)		Short LOR ($n = 13$)		Mid LOR ($n = 13$)		Long LOR ($n = 13$)		English baseline ($n = 13$)	
	M (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)	Range
Testing age (years)	28.2 (5.7)	21–40	28.9 (7.3)	21–40	30.5 (3.6)	25–36	34.9 (4.1)	28–43	21.3	20–28
AOA (years)			28.1 (0.1)	19–35	27.6 (1.2)	21–34	26.3 (2.5)	21–38		
LOR (years)			0.9 (0.1)	0.7–1.0	2.9 (1.2)	1.2–4.7	8.6 (2.5)	5–13		
Gender	11 females, 2 males		10 females, 3 males		10 females, 3 males		10 females, 3 males		6 females, 7 males	

Note. AOA = age of acquisition, LOR = length of residence.

randomly presented as distracters and then described the remaining three pictures (A, B, C) randomly presented for the main analysis. In total, the speakers generated 390 tokens (65 Japanese and baseline speakers × 3 pictures). Approximately 10 seconds from the beginning of each picture description ($M = 8.3$; 4.0–11.5 seconds) was extracted for each speaker. Because each speaker described three pictures, there was on average 24.0 seconds of extemporaneous speech (14.6–32.7 seconds) per speaker available for comprehensibility and accentedness judgment. This sample length can be considered suitable, compared to similar L2 speech research (e.g., 30 seconds in Derwing & Munro, 2013; 10–20 seconds in Hopp & Schmid, 2013).

Global Analyses

To judge the global qualities (accentedness, comprehensibility) of the spontaneous speech samples, 10 native English undergraduate students were recruited at an English-speaking university in Vancouver ($M_{\mathrm{age}} = 23.2$ years). As operationalized in previous L2 speech research (Derwing & Munro, 2009), the judgment of accentedness and comprehensibility refers to naïve raters' intuitive impression about L2 speech. Thus, efforts were made to carefully select the raters based on their lack of familiarity and contact with Japanese learners of English. The raters' average self-reported familiarity with Japanese-accented English was 1.3 (1–2) on a 6-point scale (1 = *not at all*, 6 = *very much*). Furthermore, all raters were business and psychology major students without much linguistics-related experience.

First, the raters received a brief explanation of accentedness, which refers to different patterns of speech sounds compared to the raters' L1 (Isaacs & Trofimovich, 2012), and comprehensibility, which refers to the degree of ease or difficulty in listeners' understanding of L2 speech (Derwing & Munro, 2009), from a trained research assistant (for rater-training scripts, see Appendix S1 in the Supporting Information online). The raters then familiarized themselves with the picture prompts and keywords and practiced the judgment procedure by assessing the global qualities of five speech samples (not included in the main dataset) based on a 9-point scale for accentedness (1 = *little accent*, 9 = *heavily accented*) and comprehensibility (1 = *easy to understand*, 9 = *hard to understand*). Last, the raters evaluated all speech samples presented in a randomized order via the *Praat* software (Boersma & Weenink, 2012). Each picture description was played only once on the assumption that accentedness and comprehensibility reflect listeners' initial intuitions and impressions about L2 speech. They were reminded to use the entire scale to assess the proficiency

Table 2 Summary of linguistic predictors for raters' pronunciation, fluency, vocabulary, and grammar judgment in Saito et al. (in press-b)

Rated measure	Linguistic predictor
A. Phonology	
Segmentals	Vowel and consonant errors
Word stress	Word stress errors
Intonation	Intonation errors
Speech rate	Mean length of run, frequency of unfilled pauses, articulation rate
B. Lexicogrammar	
Lexical appropriateness	Lexical errors
Lexical richness	Type frequency, token frequency
Grammatical accuracy	Grammatical errors
Grammatical complexity	Subordinate clause ratio

range of the data set consisting of English baseline and Japanese learners with varied LOR profiles. The entire session took approximately 1 hour 30 minutes.

Pronunciation, Fluency, Vocabulary, and Grammar Analyses

In terms of specific areas of language (i.e., pronunciation, fluency, vocabulary, grammar), recent L2 speech research has relied extensively on human raters' intuitive judgments of various aspects of spontaneous speech production, such as segmentals (Piske et al., 2001); temporal fluency (Derwing et al., 2004; Pinget et al., 2014); and lexical accuracy, variation, and richness (Crossley et al., 2014). According to Saito et al.'s (in press-b) validation study, raters' intuitions about pronunciation, fluency, vocabulary, and grammar judgments were predictive of the actual linguistic properties of L2 speech analyzed at a fine-grained level (summarized in Table 2). Following this L2 assessment research, subdomains of L2 oral proficiency were defined and assessed based on eight rater-based categories spanning the linguistic dimensions of pronunciation (segmentals, word stress, intonation), fluency (speech rate), vocabulary (appropriateness, richness), and grammar (accuracy, complexity).[4]

To conduct linguistic analyses of the phonological, lexical, and grammatical qualities of the spontaneous speech samples, five native English speakers (2 males, 3 females) were recruited as experienced raters ($M_{age} = 29.4$ years). All of them were graduate students in applied linguistics at an English-speaking university in Montreal. They reported experience teaching English in various settings ($M = 4.0$; 1–10 years) and previous training specific to pronunciation teaching (a semester-long course on applied phonetics and pronunciation teaching). Their mean rated familiarity with Japanese accented English (1 = *not at all*, 6 = *very much*) was 3.4 (1–5).

Audio-Based Measures

To provide sufficient phonological information for the pronunciation analyses, the three picture descriptions (Pictures A, B, C) from each participant were combined and stored as a single wav file. The raters listened to and evaluated each speech sample using the following pronunciation and fluency categories: (a) segmentals (substitution, omission, or insertion of individual consonant and vowel sounds); (b) word stress (misplaced or missing primary stress); (c) intonation (appropriate, varied use of pitch moves); and (d) speech rate (speed of utterance delivery).

The speech samples were delivered in a randomized order 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 listened to each sample (with an option to repeat until they felt satisfied) and then used a free-moving slider on a computer screen to assess the four phonological categories in the samples. If the slider was placed at the leftmost end of the continuum, labeled with a frowning face (indicating the negative endpoint), the rating was recorded as 0; if it was placed at the rightmost end of the continuum, labeled with a smiley face (indicating the positive endpoint), it was recorded as 1,000. The slider was initially placed in the middle of each scale, and the raters were told that even a small movement of the slider represents a significant difference in the rating. Except for the frowning and smiley faces and accompanying brief verbal descriptions for the endpoints of each pronunciation category, the scale included no numerical labels or marked intervals (for onscreen labels, see Appendix S1 in the Supporting Information online).

Transcript-Based Measures

To ensure that the raters were not distracted by pronunciation accuracy and fluency when evaluating vocabulary and grammar, they followed the procedure used by Crossley et al. (2014), which asked raters to read written transcripts of the speech samples. All speech samples were first transcribed and verified by a trained research assistant and then edited to remove pronunciation-specific errors, such as those related to the keywords (e.g., *rock music* pronounced as *lock music*, *table* spoken as *devil*), obvious mispronunciations based on contextual information of the pictures (*outside* pronounced as *ought side* was transcribed as “outside,” *lonely* pronounced as *lawn Lee* was transcribed as “lonely”), and orthographic markings of pausing (e.g., uh, um, oh, ehh). The final transcripts were assessed for the following lexical and grammatical dimensions: (a) lexical appropriateness (accuracy of vocabulary), (b) lexical richness (varied and sophisticated use of vocabulary), (c) grammatical accuracy (errors in word order, grammar endings, agreement), and (d) grammatical complexity (use of sophisticated, nonbasic grammar). The written transcripts were presented via the Z-Lab software (Yao et al., 2013) in a random order. The raters read the three transcripts for Pictures A, B, and C, with the descriptions displayed on screen in the same order, and assessed their lexical and grammatical content with similar free moving sliders (for onscreen labels, see Appendix S1 online).

Training and Rating Sessions

The training and rating sessions took place on 3 consecutive days, with the first day used for the training, the second day for the pronunciation and fluency analyses, and the third day for the vocabulary and grammar analyses. During the initial session, a trained research assistant provided the raters with thorough instructions on the eight pronunciation, vocabulary, and grammar measures (for training scripts, see Appendix S1). To check the accuracy and reliability of their linguistic analyses, during this session, the raters first judged phonological, lexical, and grammatical qualities of 40 L2 speech samples (unrelated to the current project) and then their audio- and transcript-based ratings were compared to the actual linguistic properties of those samples. According to the results reported in Saito et al. (in press-b), not only did these raters judge the major components of spontaneous L2 speech in an accurate manner (see Table 2), they also showed high inter-rater agreement (Cronbach’s $\alpha > .90$). The entire session took approximately 3 hours to complete, with a 10-minute break at the halfway point.

During the next session, the raters first received a recap of instructions for the four audio-based measures. Then, they familiarized themselves with the picture prompts and keywords for Pictures A, B, and C and practiced the procedure by rating three audio samples (not included in the main data set). For each practice sample, they were asked why they made their decisions and then received feedback to ensure that the rated categories were understood and applied appropriately. Finally, the raters proceeded to rate the main data set of 65 audio samples. The entire session took approximately 1 hour 30 minutes. During the final session, after reviewing instructions for the four transcript-based measures and receiving feedback on their practice ratings of the same three samples (not included in the main data set), they rated 65 written samples. The session took approximately 30 minutes.

Results

Global Analyses

Inter-Rater Agreement

Cronbach's α was calculated to check inter-rater agreement for the 10 raters' global judgment scores of 156 samples (52 Japanese speakers × 3 pictures), excluding 13 native speakers for whom they invariably provided the highest scores. In line with previous L2 comprehensibility research (e.g., Derwing & Munro, 2009), these analyses revealed relatively high consistency levels for accentedness ($\alpha = .95$) and comprehensibility ($\alpha = .97$). Therefore, mean comprehensibility and accent ratings for each speaker (across the three picture descriptions) were computed by pooling the data over all raters. The speakers' accentedness and comprehensibility scores are summarized in Table 3.

Baseline Comparisons

The first aim of the global analyses was to investigate the interlanguage characteristics of the 39 Japanese learners (LOR = 8 months–13 years) relative to inexperienced Japanese speakers (LOR < 1 month), examining the extent to which their performance corresponded to nativelike production. To this end, the 39 Japanese learners' accentedness and comprehensibility scores were treated as part of one group (Japanese learners) and were then compared to those of Japanese and English baseline groups. The accentedness and comprehensibility judgment scores for the three groups of speakers were submitted separately to univariate ANOVAs. The results demonstrated significant main effects for accentedness, $F(2, 62) = 144.02, p < .001, \eta_p{}^2 = .82$, and comprehensibility, $F(2, 62) = 85.48, p < .001, \eta_p{}^2 = .70$. According to Bonferroni-adjusted multiple comparisons, accentedness and comprehensibility scores were significantly different between the three groups ($p < .001$). This suggests that the Japanese learners beyond approximately 1 year of LOR demonstrated better English proficiency (in terms of accentedness and comprehensibility), compared to the inexperienced Japanese learners with little experience overseas, but that their performance should be considered substantially different from nativelike.

LOR Effects

The second aim of the global analyses was to examine whether and to what degree the 39 Japanese learners' performance differed as a function of LOR

Table 3 Means (standard deviations in parentheses) for rated global, phonological, temporal, and lexicogrammatical properties in the speakers' L2 picture descriptions

	Japanese baseline ($n = 13$)	Low LOR ($n = 13$)	Mid LOR ($n = 13$)	Long LOR ($n = 13$)	English baseline ($n = 13$)
A. Global ratings (9 points)					
Accentedness	7.8 (0.6)	7.3 (0.6)	6.0 (1.4)	5.9 (1.6)	1.1 (0.1)
Comprehensibility	6.0 (1.0)	5.3 (1.0)	3.9 (1.2)	3.8 (1.2)	1.0 (0.1)
B. Audio ratings (1000 points)					
Segmentals	272 (106)	310 (94)	406 (131)	429 (181)	994 (7)
Word stress	354 (81)	363 (72)	461 (75)	542 (124)	986 (27)
Intonation	256 (92)	268 (90)	348 (104)	453 (150)	885 (58)
Speech rate	294 (154)	408 (152)	541 (156)	645 (148)	977 (27)
C. Transcript ratings (1000 points)					
Lexical appropriateness	629 (103)	675 (127)	758 (123)	801 (58)	912 (49)
Lexical richness	241 (87)	410 (220)	402 (178)	517 (167)	700 (209)
Grammatical accuracy	337 (132)	378 (139)	551 (182)	667 (134)	892 (111)
Grammatical complexity	186 (73)	277 (144)	296 (152)	427 (166)	640 (234)

Note. 9-point scale (1 = *little accent, easy to understand*, 9 = *heavily accented, hard to understand*); 1,000-point scale (1 = *non-targetlike production*, 1,000 = *targetlike production*).

after the initial spike in improvement. To address this goal, all subsequent analyses were conducted on the comprehensibility and accentedness scores of the 39 Japanese learners (LOR = 8 months–13 years). Univariate ANOVAs comparing the Short-, Mid-, and Long-LOR groups' accentedness and comprehensibility scores revealed a significant main effect for accentedness, $F(2, 36) = 4.94, p = .013, \eta_p^2 = .22$, and comprehensibility, $F(2, 36) = 7.26$, $p = .002, \eta_p^2 = .29$. According to Bonferroni-adjusted multiple comparisons, despite no significant difference between Mid- and Long-LOR groups ($p > .90$), both of these groups outperformed the Short-LOR group in terms of accentedness (Long > Short, $p = .031$; Mid > Short, $p = .022$) and comprehensibility (Long > Short, $p = .007$; Mid > Short, $p = .005$).

Given that group analyses (i.e., through ANOVAs) are subject to the influence of distinct categories specified in the study (short vs. mid vs. long LOR), the relationship between LOR and L2 performance was also investigated using a set of simple regression analyses, with the learners' accentedness and comprehensibility as the dependent variables and their LOR profiles as the independent variable. The results showed that LOR was significantly related to the Japanese learners' comprehensibility, $R = .36$, $F(1, 37) = 5.60, p = .023$, accounting for 13.1% of variance in their comprehensibility scores. Yet LOR was not a significant predictor of their accentedness, $R = .27$, $F(1, 37) = 2.83$, $p = .10$.

Because LOR emerged as a significant predictor of L2 comprehensibility, the model needed to be examined further to find whether the relationship between LOR and comprehensibility remained equally linear or demonstrated any abrupt discontinuity (i.e., corresponding to the beginning of ultimate attainment) at certain points in the LOR continuum (8 months–13 years). Therefore, a piecewise regression analysis was conducted next, using the learners' comprehensibility scores as the dependent variable and their LOR profiles as the independent variable. This analysis, which allows for examining whether the inclusion of any breakpoint can improve the fit of the regression line to the data (for details of the calculation method, see Appendix S2 in the Supporting Information online), has been used regularly in previous SLA literature (e.g., Birdsong & Molis, 2001). A sloping segment followed by a horizontal line was found to be the best fitting model for the LOR-comprehensibility function, $F(3, 35) = 5.98, p = .003$. As shown in Figure 1, the model revealed an optimal breakpoint at 3.04 years of LOR, after which the nature of the regression line changed from sloped to horizontal ($p > .05$).

To summarize, the results of global analyses suggest that the Japanese learners continued to improve in their comprehensibility, especially during the

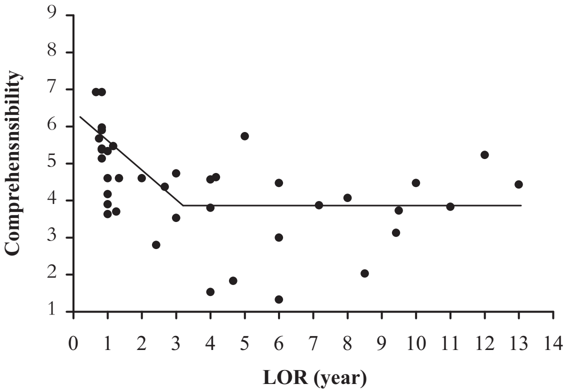

Figure 1 Comprehensibility scores (1 = *easy to understand*, 9 = *hard to understand*) and LOR profiles for 39 Japanese learners (8 months–13 years).

first 3 years of their LOR, and then reached a point of ultimate attainment and then plateaued. In contrast, the effects of LOR on nativelikeness (as measured through accentedness rating) remained unclear, particularly after the learners benefitted from the rate of learning advantage within the initial year of their LOR. Last, the learners' accentedness and comprehensibility performance was substantially different from the native English speaker baseline.

Pronunciation, Fluency, Vocabulary, and Grammar Analyses

Inter-Rater Agreement

For audio-based measures, the five raters demonstrated high inter-rater agreement for segmentals (Cronbach's $\alpha = .85$), word stress ($\alpha = .82$), intonation ($\alpha = .81$), and speech rate ($\alpha = .82$). With respect to transcript-based measures, the five experienced raters again demonstrated high agreement when it came to vocabulary ($\alpha = .76$ for appropriateness; $\alpha = .79$ for richness) and grammar ($\alpha = .75$ for accuracy; $\alpha = .76$ for complexity). Because inter-rater consistency was sufficiently high relative to L2 research standards ($\alpha > .70$, Larson-Hall, 2010), audio- and transcript-based ratings were averaged across the five raters in order to derive a single mean score per speaker for each rated category. Descriptive statistics for the rated scores are presented in Table 3.

Baseline Comparisons

First, the speakers were again categorized into three groups: Japanese learners ($n = 39$), Japanese baseline speakers ($n = 13$), and English baseline speakers ($n = 13$). Then, their pronunciation, fluency, vocabulary, and grammar scores were submitted to univariate ANOVAs that revealed significant main effects for segmentals, $F(2, 62) = 143.24$, $p < .001$, $\eta_p{}^2 = .82$; word stress, $F(2, 62) = 168.71$, $p < .001$, $\eta_p{}^2 = .85$; intonation, $F(2, 62) = 117.56$, $p < .001$, $\eta_p{}^2 = .79$; speech rate, $F(2, 62) = 66.50$, $p < .001$, $\eta_p{}^2 = .68$; lexical appropriateness, $F(2, 62) = 25.66$, $p < .001$, $\eta_p{}^2 = .45$; lexical richness, $F(2, 62) = 21.71$, $p < .001$, $\eta_p{}^2 = .41$; grammatical accuracy, $F(2, 62) = 37.29$, $p < .001$, $\eta_p{}^2 = .55$; and grammatical complexity, $F(2, 62) = 25.36$, $p < .001$, $\eta_p{}^2 = .45$. Bonferroni-adjusted comparisons revealed that the Japanese learners' performance was significantly different from that of Japanese and English baselines speakers ($p < .05$), suggesting that the learners made significant gains in all linguistic aspects of L2 oral performance after 8 months of L2 immersion but failed to reach nativelike levels.

LOR Effects

To examine the role of LOR in the Japanese learners' (LOR > 8 months) pronunciation, fluency, vocabulary, and grammar performance, their scores were submitted to univariate ANOVAs comparing the Short-, Mid-, and Long-LOR groups. The results revealed significant main effects for word stress, $F(2, 36) = 11.85$, $p < .001$, $\eta_p{}^2 = .40$; intonation, $F(2, 36) = 8.08$, $p < .001$, $\eta_p{}^2 = .31$; speech rate, $F(2, 36) = 7.91$, $p < .001$, $\eta_p{}^2 = .31$; lexical appropriateness, $F(2, 36) = 4.58$, $p = .017$, $\eta_p{}^2 = .20$; grammatical accuracy, $F(2, 36) = 11.77$, $p < .001$, $\eta_p{}^2 = .39$; and grammatical complexity, $F(2, 36) = 3.65$, $p = .036$, $\eta_p{}^2 = .08$. However, significant effects were not detected for segmentals, $F(2, 36) = 2.64$, $p = .09$, and lexical richness, $F(2, 36) = 1.46$, $p = .25$. Bonferroni-adjusted comparisons yielded significant differences for word stress (Long > Short, $p < .001$; Mid > Short, $p = .035$); intonation (Long > Short, $p = .001$); speech rate (Long > Short, $p = .001$); lexical appropriateness (Long > Short, $p = .016$); and grammatical accuracy (Long > Short, $p < .001$; Mid > Short, $p = .020$).

A set of simple linear regression analyses was conducted next to examine the extent to which LOR (8 months–13 years) was predictive of the segmental, prosodic, fluency, lexical, and grammatical qualities of the 39 Japanese learners' L2 speech. The results showed that LOR significantly predicted the learners' suprasegmental and lexicogrammar accuracy, explaining 28.8% of variance in word stress, $R = .54$, $F(1, 37) = 14.98$, $p < .001$; 29.4% of variance in intonation, $R = .54$, $F(1, 37) = 15.41$, $p < .001$; 32.7% of variance in speech rate, $R = .57$, $F(1, 37) = 17.97$, $p < .001$; 14.0% of variance in lexical appropriateness, $R = .37$, $F(1, 37) = 15.41$, $p = .019$; and 23.9% of variance in grammatical accuracy, $R = .49$, $F(1, 37) = 6.03$, $p = .002$. However, the

LOR-proficiency function failed to reach statistical significance for segmentals, $R = .30$, $F(1, 37) = 3.63$, $p = .07$; lexical richness, $R = .20$, $F(1, 37) = 1.53$, $p = .22$; and grammatical complexity, $R = .20$, $F(1, 37) = 4.03$, $p = .052$.

To investigate the existence of any breakpoints in the LOR-proficiency function in the significant linear models identified above, a set of piecewise regression analyses was conducted on Japanese learners' suprasegmental (word stress, intonation, speech rate) and lexicogrammar accuracy (lexical appropriateness, grammatical accuracy) scores as the dependent variables while their LOR profiles served as the independent variable. The results (depicted in Figure 2) showed that the regression fit was improved, with the modified model described as a sloping segment followed by a horizontal segment for (a) word stress, $F(3, 35) = 9.89$, $p < .001$; (b) intonation, $F(3, 35) = 6.99$, $p = .001$; (c) speech rate, $F(3, 35) = 6.48$, $p = .002$; and (d) grammatical accuracy, $F(3, 35) = 12.51$, $p < .001$. Relatively extensive amounts of LOR were identified before a breakpoint for word stress (5.62 years), intonation (5.99 years), and speech rate (5.50 years), while an LOR of 3.04 years was identified before a breakpoint for grammatical accuracy. For lexical appropriateness, however, the LOR-proficiency regression was best described as two horizontal segments at different levels, $F(2, 36) = 10.75$, $p < .001$, with a breakpoint of 1.31 years of LOR.

To summarize, the results indicated (a) that the Japanese learners continued to improve in lexicogrammar accuracy within the first 3 years of LOR and in suprasegmental performance within the first 5 years of L2 experience, but (b) that they may have needed much more L2 experience (LOR > 5 years) to demonstrate any significant improvement in their segmental accuracy and grammatical complexity.

Linguistic Correlates of Comprehensibility and Accentedness

The final analysis focused on the relationship between the pronunciation, vocabulary, and grammar characteristics of L2 speech on the one hand, and its overall comprehensibility and accentedness on the other. To examine pronunciation influences on the global judgments of L2 speech, partial correlation analyses were first conducted to determine how the four pronunciation scores (segmentals, word stress, intonation, speech rate) of all participants were linked to comprehensibility and accentedess judgments, with lexicogrammar scores factored out. As shown in Table 4, all of the pronunciation categories were significantly correlated with comprehensibility and accentedness ($p < .01$, Bonferroni corrected). According to Fisher's r-to-z transformation, no statistical difference in the strength of correlation coefficients between the

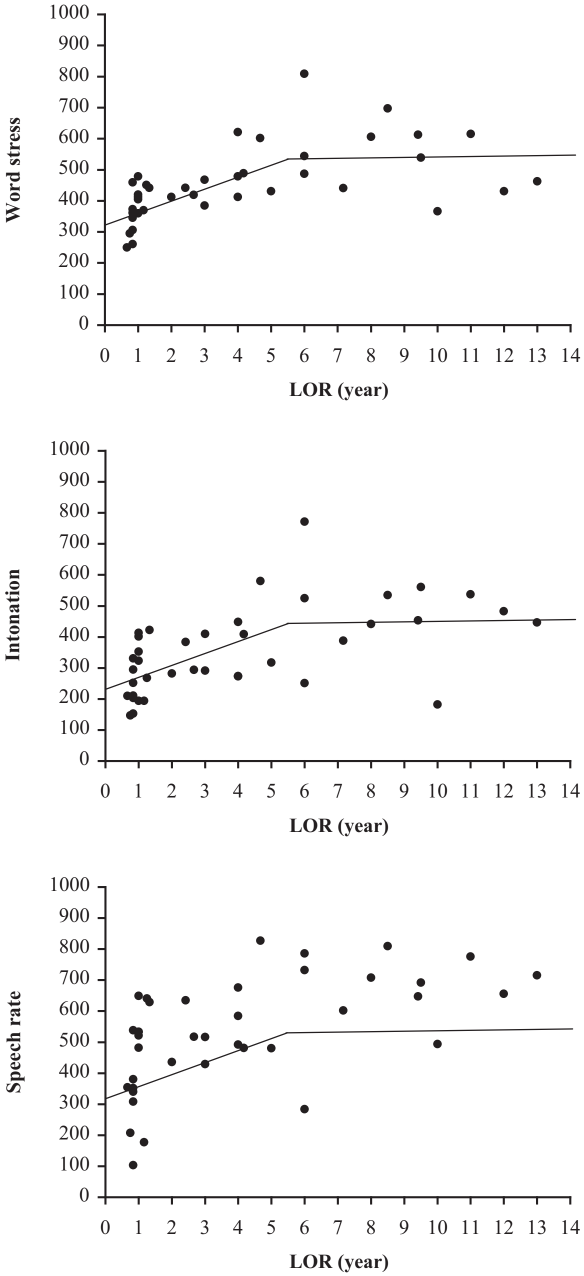

Figure 2 Pronunciation, fluency, vocabulary, and grammar scores (0 = *non-targetlike*, 1,000 = *targetlike*) and LOR profiles for 39 Japanese learners (8 months–13 years).

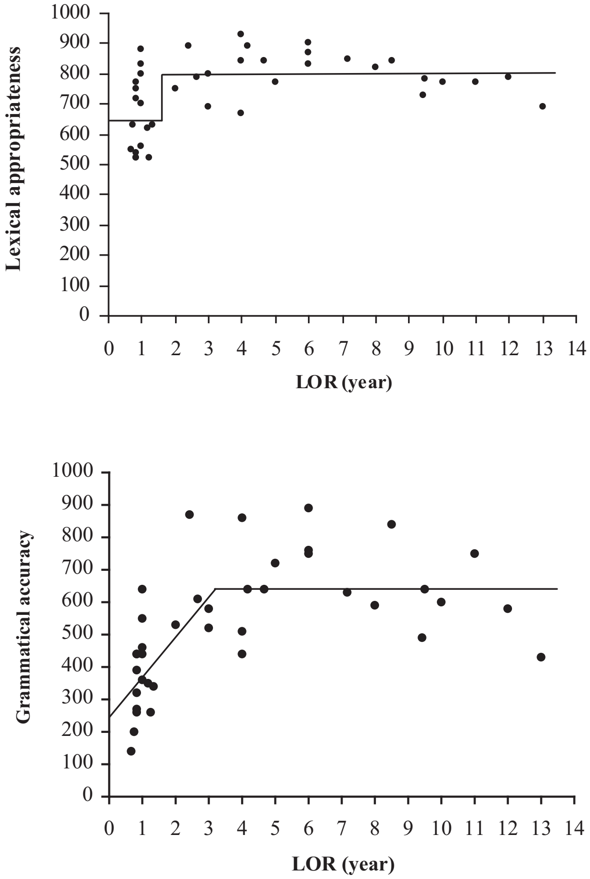

Figure 2 Continued.

Table 4 Partial correlations between four rated pronunciation variables and comprehensibility and accentedness, with the influence of four lexicogrammar variables controlled

Rated variable	Comprehensibility	Accentedness
Segmentals	.76*	.91*
Word stress	.72*	.87*
Intonation	.56*	.72*
Speech rate	.69*	.61*

Note. $^*\alpha < .01$ (Bonferroni corrected). The variables partialled out from each correlation include lexical appropriateness and richness, and grammatical accuracy and complexity.

Table 5 Partial correlations between four rated vocabulary and grammar variables and comprehensibility and accentedness, with the influence of four pronunciation variables controlled

Rated variable	Comprehensibility	Accentedness
Lexical appropriateness	.44*	.13
Lexical richness	.01	.05
Grammatical accuracy	.51*	.24
Grammatical complexity	.09	.02

Note. $^*\alpha < .01$ (Bonferroni corrected). The variables partialled out from each correlation include segmentals, word stress, intonation, and speech rate.

four pronunciation categories and comprehensibility was found ($p < .008$, Bonferroni corrected). However, segmentals and word stress were more strongly tied with accentedness than speech rate was ($p < .001$). This in turn suggests that, while all pronunciation elements were equally related to comprehensibility, the learners' pronunciation of words (segmental accuracy with correct word stress) was more strongly associated with accentedness than their delivery of sentences with good intonation and optimal speech rate.

Next, a series of partial correlation analyses was conducted to examine contributions of lexis and grammar to comprehensibility and accentedness, with the influence of pronunciation categories controlled. It was found that appropriate (rather than sophisticated) vocabulary and grammar usage was significantly related to comprehensibility but that none of the lexicogrammar variables were tied with accentedness (see Table 5). According to Fisher's r-to-z transformation, no significant differences were found in the strength of correlation coefficients between the four lexicogrammar variables and comprehensibility and accentedness ($p > .008$).

To sum up, the results indicated that (a) comprehensibility was related to segmental, prosodic, temporal, lexical, and grammatical qualities of L2 speech and (b) accentedness was mainly accounted for by pronunciation (particularly segmental and word stress accuracy) rather than lexical and grammar variables.

Discussion

The role of experience in interlanguage development continues to be a source of theoretical debate (DeKeyser & Larson-Hall, 2005, vs. Flege, 2009). Derwing and Munro (2013) recently reported an important longitudinal finding relevant to this debate. They showed that late L2 learners enhanced their comprehensibility (without minimizing accentedness) when their L2 oral ability was assessed 2 and 7 years into their residence in Canada. The current cross-sectional study tested the generalizability of this finding with respect to highly motivated Japanese learners of English with varying LOR profiles beyond the early phase of late SLA (i.e., 8 months–13 years).

If we look at the L2 proficiency of the Japanese learners in this study based on their foreign accentedness scores, as has been the case with most previous LOR research (e.g., Flege, 1988), learners' improvement seems to be limited to the first year of L2 experience, beyond which their performance is likely subject to individual differences regardless of increasing LOR. These results appear to be consistent with the CPH, which states that experience effects, if any, are concerned only with the rate of learning (DeKeyser & Larson-Hall, 2005). However, by taking into account multiple linguistic dimensions of L2 speech, the current results further revealed that the Japanese learners' LOR profiles differentially predicted global (accentedness, comprehensibility), phonological (segmentals, suprasegmentals), lexical (appropriateness, richness), and grammatical (accuracy, complexity) qualities of language.

With respect to global analyses, the current cross-sectional data showed that LOR was a significant predictor of improved comprehensibility but not reduced accentedness in the Japanese learners' L2 speech, especially within approximately 3 years of their residence in Canada. In terms of the pronunciation and lexicogrammar analyses, the results also indicated that an increasing LOR seemed to facilitate the development of certain linguistic features affecting comprehensibility, including word stress, intonation, speech rate, and lexicogrammar accuracy. However, the role of LOR remained unclear in the development of the linguistic features strongly related to accentedness, such as segmentals, vocabulary richness, and grammatical complexity. Furthermore, piecewise regression analyses suggested that the learners' performance reached a plateau in a naturalistic setting after different amounts of LOR for pronunciation (5–6 years), grammar (3–4 years), and vocabulary (1 year). Taken together, linguistic analyses of the current dataset do not concur with the strong version of the CPH, which hypothesizes the presence of a LOR-proficiency link exclusively in the initial stage of late SLA (LOR < 6 months). Rather, these results support the SLM, which assumes that L2 learners may continue to show improvement in oral ability later in life over an extended period of L2 immersion (LOR = 5–6 years).

At the same time, the role of experience in late L2 oral proficiency development, unlike in early bilingualism and L1 acquisition, can also be characterized as multidimensional in nature, in recognition of learners' overall learning goals (reduced accentedness vs. improved comprehensibility) and various types of linguistic processing relevant to these goals (i.e., functional vs. sophisticated language use). First and foremost, the current results indicated that experience effects were evident for linguistic correlates of comprehensibility rather than accentedness. As learners gain L2 experience, they may prioritize the development of good prosody, optimal speech rate, as well as proper vocabulary and grammar usage (over segmental accuracy and sophisticated use of vocabulary and grammar) for the purpose of achieving successful communication in their private, business, and academic settings. The findings can be well accounted for by an interactionist view, which states that comprehensibility rather than accentedness is what learners essentially aim to achieve during their interactions with interlocutors and that comprehensibility is a crucial variable, especially in late SLA (Gass & Mackey, 2006).

The Interaction Hypothesis (e.g., Long, 1996) claims that language learning takes place precisely when comprehensibility is at stake during conversational exchanges between native and nonnative speakers (or between nonnative speakers). Put differently, communication breakdowns are likely when L2 learners interact with other native and nonnative speakers, and these breakdowns result in intuitive or conscious efforts to correct impaired linguistic accuracy as learners seek successful comprehension. It is during such negotiation for meaning that learners have opportunities to receive modified input and produce self-modified output (Long, 2007) while collaborating with experts to solve problems that they would not otherwise be able to decipher alone (Dunn & Lantolf, 1998). Such conversational moves are hypothesized to be facilitative of the rate and ultimate attainment in late L2 development (Mackey & Goo, 2007).

Given that certain linguistic features affect comprehensibility and thus trigger negotiation for meaning more than others (Isaacs & Trofimovich, 2012; Pinget et al., 2014), an interactionist perspective suggests that L2 learners selectively strive to notice and practice linguistic domains closely linked to comprehensibility but not those necessarily relevant to accentedness (e.g., Mackey, Gass, & McDonough, 2000). Drawing on the interactionist perspective of language learning, the findings of the current study suggest that late bilingualism is driven by the development of L2 comprehensibility. That is, L2 learners continue to improve their oral abilities as long as they have the desire to attain increasingly smooth, accurate, and successful communicative skills throughout much of their social interaction with native and nonnative speakers in an L2 community (Flege & Liu, 2001), and as long as their L2 speaking ability is assessed based on their performance on comprehensibility-related linguistic features, such as suprasegmentals, vocabulary appropriateness, and grammatical accuracy (Isaacs & Trofimovich, 2012).

Limitations

While the findings reported here provide useful insights into the relationship between linguistic experience and late SLA, several crucial limitations need to be acknowledged. First, the results based on a small data set ($n = 39$) need to be interpreted with caution, because some of the findings could simply be artifacts of statistical analyses. For example, a breakpoint in the regression models (a discontinuity in the LOR–proficiency relationship) may have reflected information density rather than a specific point in development. Because there were considerable differences between the groups in their LOR ranges, the nature of correlations may essentially differ in the lower (e.g., .7–1 years for Short LOR) and upper (e.g., 5–13 years for Long LOR) end of the LOR range. While the current study supported the longitudinal results reported by Derwing and Munro (2013), the generalizability of the findings for L2 comprehensibility development still needs to be examined, especially with larger sample sizes and in longitudinal designs.

Second, any suggestions, especially regarding learners' lexical and grammatical performance, should only be considered as tentative due to methodological limitations. For instance, the length of speech samples (approximately 30 seconds per participant) was relatively short for conducting robust vocabulary and grammar analyses. Although 30 seconds can be sufficient for providing phonological information in line with L2 speech research standards (e.g., Derwing & Munro, 2013), longer speech samples may be needed to provide a comprehensive picture of lexical and grammatical influences on comprehensibility and accentedness, following current practice in L2 vocabulary and grammar research (e.g., 3 minutes in Lu, 2012). Another limitation is that the nature of the task (describing pictures) did not elicit a sufficiently wide range of infrequent lexical items. Given that all speakers may have used similar, frequent lexical items, the predictive power of lexical indices may not have reached statistical significance, especially if speakers had reached minimum lexical thresholds to successfully complete the task (see Crowther, Trofimovich, Isaacs, & Saito, in press). Third, in this study, L2 speaking was assessed through human raters' pronunciation, fluency, vocabulary, and grammar judgments. However, L2 speaking can be further studied via instrumental measurements, such as acoustic analyses of formant frequencies for segmentals (e.g., Saito & Brajot, 2013) and fundamental frequencies for prosody (Trofimovich & Baker, 2006); articulation rate for fluency (e.g., Derwing et al., 2004); and computational modeling of lexical breadth, depth, variation, richness, and sophistication for vocabulary (e.g., Crossley et al., 2014).

Finally, the somewhat limited sample of participants tested in this study does not allow us to completely reject the predictions of the CPH. Although the results presented here were, by and large, in agreement with the SLM, it is still possible to argue that the Japanese learners relied on domain-general learning and on explicit strategies (rather than on incidental and automatic learning processes) to improve their L2 oral ability over the varied period of their immersion in Canada. That is, research is yet to provide a conclusive answer in regards to the underlying learning mechanisms that late L2 learners draw on to improve their oral ability, for example, in terms of the presence or absence of an implicit learning mechanism during postpubertal SLA.[5] Of further interest would be a study that highlights another crucial factor for late SLA, namely, learners' age of acquisition, typically defined as the age of first intensive exposure to input and interaction in an L2-speaking environment. Although age of acquisition is a relatively strong predictor of the end state of SLA, such that the earlier L2 learners arrive, the better the quality of their ultimate attainment tends to be, especially for early bilinguals who arrive in an L2 environment before puberty, it has remained highly controversial whether, to what degree, and how such age effects can be relevant to late bilinguals whose L2 immersion starts after puberty (e.g., Birdsong, 2005, vs. DeKeyser & Larson-Hall, 2005).

Although the two competing theories (CPH vs. SLM) both assume that very few late learners can attain nativelike L2 performance, they attribute such effects to two essentially different causes. In the case of the CPH, it is the lack of access to a language learning mechanism after the passing of the critical period (DeKeyser & Larson-Hall, 2005). For the SLM, age effects are caused by an age-related decline in many human cognitive functions, such as working memory, executive control, speech processing, or inhibition of task-irrelevant information (Birdsong, 2005). Thus, these two theoretical positions provide different predictions as to the role of age of acquisition in late L2 learners' ultimate attainment. The CPH proponents assume that age of acquisition is unrelated to post-critical-period SLA because early bilingualism is essentially different from late bilingualism. In contrast, the SLM proponents may claim that age of acquisition can be a significant predictor of not only early but also late bilinguals' ultimate attainment because they use the same learning mechanism for L1 and L2 acquisition. However, as many SLA researchers point out (e.g., DeKeyser & Larson-Hall, 2005), future research targeting this issue needs to include a greater number of language users who have reached the upper limit of SLA after an extensive amount of L2 immersion (LOR > 6 years) (cf. Saito, in press), compared to the 13 advanced L2 learners with extensive L2 experience in the current study.

Conclusion

Building on recent L2 speech research (Derwing & Munro, 2013), this cross-sectional study reexamined the extent to which late L2 learners' oral proficiency can be enhanced as a function of additional amount of experience, operationalized as LOR. LOR was generally found to be predictive of improved L2 comprehensibility as a result of continuous development of good prosody, optimal speech rate, and proper lexicogrammar usage. However, reducing accent required a great amount of L2 experience or depended on other factors (e.g., age of acquisition), with less accented speech linked to refined segmental accuracy, vocabulary richness, and grammatical complexity. These results supported earlier longitudinal findings by Derwing and Munro (2013), implying that an ability to learn a new language is maintained over a learner's life span and that experience effects on late L2 learners' oral proficiency development and attainment follow a gradual learning trajectory over an extensive period of L2 immersion (e.g., 6 years of LOR). Last but not least, L2 learners appear to improve their oral ability by paying selective attention to certain linguistic domains closely linked to comprehensibility (but not necessarily relevant to accentedness) for the purpose of successful L2 communication.

Final revised version accepted 26 July 2014

Notes

[1] The key researchers listed here share the view that adult and early bilingualism draw on the same language acquisition system, but they do not necessarily contradict the SLM in other respects; for example, see Best and Tyler's (2007) account of the relationship between language experience and restructuring patterns in SLA under the Perceptual Assimilation Model.

[2] To select Japanese immigrants who intensively used English (rather than lived within the Japanese community), the participants' main language of communication was determined via individual interviews. Some may claim that the frequency of L1/L2 use should be measured through participants' self-reports (typically based on a 6-point scale), as operationalized in previous L2 speech studies (e.g., Trofimovich & Baker, 2006). Notably, the subjective nature of these methods without any longitudinal observation has been criticized (e.g., Flege, 2009). Acknowledging the inherent difficulty in quantifying how much L2 input and interaction practice learners usually engage in with other native and nonnative speakers, I argue that

dichotomous answers (e.g., *Which language do you mainly use, English or Japanese?*) rather than a continuous scale (e.g., *How often do you use English and Japanese from 1 "very infrequent" to 6 "very frequent*"*?*) can better approximate L2 learners' general patterns of language use in private and business settings. This is because the former question can more directly tap into late bilinguals' most frequently used language without taking into account their potentially different understanding of what constitutes the upper (e.g., *frequent use of L1/L2*) and lower (e.g., *infrequent use of L1/L2*) endpoints on a continuous scale (see Ranta & Meckelborg, 2013, for innovative methodological options on this topic, such as blogging).

3 A preliminary analysis targeting the effects of LOR on these same participants' English /r/ production was reported in Saito and Brajot (2013). In the current study, the overall linguistic qualities of the same dataset were reanalyzed from not only segmental but also global, prosodic, temporal, lexical, and grammatical perspectives.

4 Such rater-based categories can be further reduced to a range of corresponding linguistic properties typically measured via computerized instruments, such as *Praat* (Boersma & Weenink, 2012) and *Coh-Metrix* (Graesser, McNamara, Louwerse, & Cai, 2004). For example, the temporal domain of L2 speech production can be divided into the number of filled and unfilled pauses; articulation rate; pruned and unpruned speech rate; and the length of words, clauses, and sentences, all of which interact to influence raters' broad intuition of "fluency" (Derwing et al., 2004). In the current study, however, I focused on the subdomains of L2 speaking proficiency at a macro level (i.e., level of rater-based categories) rather than a micro level (i.e., level of actual linguistic properties of speech). For further empirical research and discussion of more abstract (rather than broad) constructs of L2 oral proficiency, see Saito et al. (in press-a, in press-b), de Jong et al. (2012), and Isaacs and Trofimovich (2012).

5 To my knowledge, no single study (including the current research) has examined in depth the extent to which late L2 learners differentially use explicit and implicit learning strategies to improve their oral proficiency over years of L2 immersion, nor have any empirically validated methodologies been devised to measure precisely how late learners acquire L2 in an explicit and implicit manner.

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.

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

Bialystok, E. (1997). The structure of age: In search of barriers to second language acquisition. *Second Language Research*, *13*, 116–137.

Birdsong, D. (2005). Interpreting age effects in second language acquisition. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 109–127). New York: Oxford University Press.

Birdsong, D., & Molis, M. (2001). On the Evidence for maturational constraints in second language acquisition. *Journal of Memory and Language*, *44*, 235–249.

Boersma, P., & Weenink, D. (2012). Praat: Doing phonetics by computer. Retrieved from http://www.praat.org

Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2014). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*. doi: 10.1093/applin/amt056

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (in press). Does speaking task affect second language comprehensibility? *Modern Language Journal*, *99*.

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. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*, 499–533.

DeKeyser, R. (2007). Skill acquisition theory. In J. Williams & B. VanPatten (Eds.), *Theories in second language acquisition: An introduction* (pp. 97–113). Mahwah, NJ: Erlbaum.

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). Oxford, UK: Oxford University Press.

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.

Dunn, W., & Lantolf, J. (1998). Vygotsky's zone of proximal development and Krashen's i+1: Incommensurable constructs; incommensurable theories. *Language Learning*, *48*, 411–442.

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.

Flege, J. E. (1988). Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America*, *84*, 70–79.

Flege, J. (2003). Assessing constraints on second-language segmental production and perception. In A. Meyer & N. Schiller (Eds.), *Phonetics and phonology in language* *comprehension and production: Differences and similarities* (pp. 319–355). Berlin, Germany: Mouton de Gruyter.

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

Flege, J., Bohn, O.-S., & Jang, S. (1997). The effect of experience on nonnative subjects' production and perception of English vowels. *Journal of Phonetics*, *25*, 437–470.

Flege, J., & Fletcher, K. (1992). Talker and listener effects on the perception of degree of foreign accent. *Journal of the Acoustical Society of America*, *91*, 370–389.

Flege, J. E., Frieda, E. M., & Nozawa, T. (1997). Amount of native-language (L1) use affects the pronunciation of an L2. *Journal of Phonetics*, *25*, 169–186.

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., & Fox, A. (1994). Auditory and categorical effects on cross-language vowel perception. *Journal of the Acoustical Society of America*, *95*, 3623–3641.

Gatbonton, E., Trofimovich, P., & Segalowitz, N. (2011). Ethnic group affiliation and patterns of development of a phonological variable. *Modern Language Journal*, *95*, 188–204.

Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research*, *29*, 311–343.

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.

Hopp, H., & Schmid, M. (2013). Perceived foreign accent in first language attrition and second language acquisition: The impact of age of acquisition and bilingualism. *Applied Psycholinguistics*, *34*, 361–394.

Ioup, G., Boustagi, E., El Tigi, M., & Moselle, M. (1994). Reexamining the critical period hypothesis: A case study of successful adult SLA in a naturalistic environment. *Studies in Second Language Acquisition*, *16*, 73–98.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34*, 475–505.

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

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

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of language* *acquisition. Vol. 2: Second language acquisition* (pp. 413–468). New York: Academic Press.

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

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Journal*, *96*, 190–208.

Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition*, *22*, 471–497.

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407–452). Oxford, UK: Oxford University Press.

Major, R. (2008). Transfer in second language phonology: A review. In J. Hansen Edwards & M. Zampini (Eds.), *Phonology and second language acquisition* (pp. 63–94). Amsterdam: John Benjamins.

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

Munro, M. J. (1993). Productions of English vowels by native speakers of Arabic: Acoustic measurements and accentedness ratings. *Language and Speech*, *36*, 39–66.

Munro, M., & Derwing, T. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. *Language Learning*, *58*, 479–502.

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.

Patkowski, M. (1990). Age and accent in a second language: A reply to James Emil Flege. *Applied Linguistics*, *11*, 73–89.

Pinget, A., Bosker, H., Quené, H., & de Jong, N. (2014). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing*, *31*, 349–365.

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

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

Saito, K. (in press). The role of age of acquisition in late second language oral proficiency attainment. *Studies in Second Language Acquisition*, *37*.

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

Saito, K., Trofimovich, P., & Isaacs, T. (in press-a). Developing a process-oriented model for linguistic influences on comprehensibility and accentedness in second language speech production. *Applied Psycholinguistics*. doi: 10.1017/S0142716414000502

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

Scovel, T. (2000). A critical review of the critical period research. *Annual Review of Applied Linguistics*, *20*, 213–223.

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

Statistics Canada. (2008). 2006 Census of Canada topic based tabulations, ethnic origin and visible minorities tables: Ethnic origin, for population, for Canada, provinces and territories, 2006 census. (Catalogue number 97–562-XWE2006002). Retrieved June 3, 2012 from http://www12.statcan.ca/census-recensement/2006/dp-pd/hlt/97--562/index.cfm?Lang=E

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.

Werker, J. F., & Tees, R. C. (1999). Influences on infant speech processing: Toward a new synthesis. *Annual Review of Psychology*, *50*, 509–535.

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

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1: Training Materials and Onscreen Labels for Pronunciation and Lexicogrammar Judgment.
Appendix S2: Calculation Procedure for Breakpoint Analysis.