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

Running Head:
ROLE OF LOR IN LATE SLA

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

Acknowledgement

This study was partially funded by the Grant-in-Aid for Scientific Research (No. 26770202) from the Ministry of Education in Japan. I am grateful to Tracey Derwing, Pavel Trofimovich, and *Language Learning* reviewers for providing constructive comments on earlier versions of this paper. I gratefully acknowledge George Smith and Ze Shan Yao who helped data collection, and all the volunteer participants for the project. Address correspondence to: Kazuya Saito, Department of Applied Linguistics and Communication, Birkbeck, University of London, 30 Russell Square, London WC1B 5DT, UK. Email: k.saito@bbk.ac.uk

Abstract

The aim of this study is to evaluate the effects of language experience—operationalized as length of residence—on late second language learners. Data collected from 65 participants consists of three groups of Japanese learners of English (n = 13 Short-, Mid-, Long-LOR Groups, respectively) and two groups of baseline speakers (n = 13 for inexperienced Japanese speakers and native English speakers, respectively). The global quality of their spontaneous speech production was initially judged by 10 native speaking raters of English based on 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 L2 comprehensibility as a result of the continuous development of good prosody, optimal speech rate, and proper lexicogrammar usage, while a great amount of L2 experience was required to enhance accentedness, which entailed refined segmental accuracy, vocabulary richness, and grammatical complexity. These results, in turn, suggest L2 learners continue to improve L2 oral proficiency over an extensive period of L2 immersion (e.g., 6 years of LOR), and they 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.

Key words: Experience effects, L2 oral ability, Late bilingualism, Comprehensibility, Accentedness, Pronunciation, Fluency, Lexicon, Grammar

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

Adult second language (L2) speech learning is a multifaceted phenomenon influenced by many affecting 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 fields 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 native and non-native speakers. A usage-based account of SLA, for instance, posits that the more frequently L2 learners encounter specific linguistic features in the input, the stronger connections they will establish between these features and the various contexts in which they occur. This frequency-driven mapping of form and meaning enables L2 learners to ultimately attain automatic processing of L2 as formulaic units in response to any relevant situational and linguistic cues (N. Ellis, 2006).

What remains controversial in the existing SLA theories, however, concerns whether, to what degree, and how such experience effects *continue* to help late L2 learners improve their proficiency beyond the early phase of L2 development and ultimately approximate nativelike performance in the long run. To investigate the role of experience in interlanguage development (i.e., the restructuring and enhancement of the L2 system) and ultimate attainment (i.e., the plateaued and asymptotic L2 performance), previous speech research has extensively studied 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. Below, I will first review two competing theoretical positions—the Critical Period Hypothesis (e.g., DeKeyser & Larson-Hall, 2005) versus the Speech Learning Model (e.g., Flege, 2009)—and their different predictions as to the effect of experience, operationalized as LOR, on the initial, mid 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 late Japanese learners of English with various proficiency levels (e.g., beginner to advanced).

Background

Critical Period Hypothesis (CPH)

Certain researchers have claimed that late bilinguals who start learning L2 after puberty have little access to an assumed automatic learning mechanism by which to acquire language through mere exposure due to the passing of a critical period for implicit language learning. The strong version of the CPH states that there is no L2 learning after puberty regardless of increased experience (e.g., Patkowski, 1990; Scovel, 2000). The proponents of other versions of CPH assume that late SLA benefits from L2 input and interaction 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 leaners practice L2 in a similar manner to learning other general cognitive skills such as mathematics and computer programming (DeKeyser, 2007). Previous research has indeed noted 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 (i.e., 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 production 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, 1991; Flege, Bohn, & Jang, 1997; Trofimovich & Baker, 2006) but not with those who have passed the rate of learning advantage (LOR > 6 months) (Flege, 1988; Flege, Munro, & Fox, 1994; Munro, 1993). According to Munro and Derwing's (2008) longitudinal research on adult ESL learners' vowel acquisition, it was suggested that most L2 speech learning takes place within the first three to four 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 exceptional learners with high language aptitude. For example, Granena and Long (2013) found that late learners' near-nativelike pronunciation abilities were exclusively limited to those with good sound-symbol correspondence and grammatical inferencing abilities regardless of their LOR profiles (see also DeKeyser, 2000).

Speech Learning Model (SLM)

Other researchers have emphasized that late L2 learners continue to have the capacity for L1 acquisition active even after puberty and may apply it to post-pubertal SLA (Best & Tyler, 2007; Bialystok, 1997; Birdsong, 2006; Hopp & Schmid, 2013)[1]. Specifically, 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, the extensive amount of L2 input and interaction is to lead both early and late L2 learners to achieve near-nativelike 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 somewhat limited quality and quantity of L2 input (Muñoz & Llanes, 2014). For example, they may have many opportunities to receive input from native speakers of various dialects while also receiving input from non-native speakers from diverse L1 backgrounds. Some immigrants may also choose to only use L1 at home and work within the same language background community (see 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 simulate social and psychological environments that early bilinguals generally benefit from such as 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 improvement, provided the main language of communication is L2 (e.g., university-level international students) and not L1 (e.g., immigrant workers). Flege (2009) found that the frequency of L1 and L2 use strongly predicted the extent to which certain L2 learners can actually make the most use of 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 that L2 learners never lose the plasticity for language learning with their L2 abilities consistently susceptible to their dominant language use. This subsequently suggests that even late L2 oral proficiency development can be

[1] These key researchers listed here share the view that adult and early bilingualism draw on the same language acquisition system, but they do not necessarily alienate 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.

characterized as a gradual, constant, and extensive process, as L1 acquisition is established in a sufficiently robust way by the end of 5 to 10 years (Werker & Tees, 1999).

Motivation for the Current Study

As reviewed above, the two competing theories—the CPH versus the SLM—offer different predictions for experience effects on late SLA. For 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). For 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 to 10 years). Examining the role of experience in the initial, mid and final state of late L2 oral proficiency development is thus crucial in SLA theory building. Such research will shed light on one of the most fundamental questions regarding the underlying mechanism for late SLA: How long does it take for late L2 learners to reach ultimate attainment and how far can they improve 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 for assessing late L2 speaking proficiency. That is, most of the aforementioned LOR research has exclusively depended on highly controlled measures such as word and sentence repetition tasks whereby participants simply repeat audio and written prompts without generating any free production. Although these measures enable researchers to elicit and analyze global (e.g., foreign accentedness) or specific (e.g., segmentals, suprasegmentals) pronunciation features under highly controlled conditions, they also contradictorily allow participants to carefully monitor their correct linguistic forms by drawing on their explicit knowledge (R. Ellis, 2005); L2 performance of this kind is arguably claimed to merely reflect on "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, 2007). To measure late L2 learners' more natural oral proficiency, many SLA researchers have emphasized the importance of eliciting L2 speech production at a spontaneous— rather than controlled— level by encouraging them to pay equal attention to the phonological, lexical, and grammatical aspects of language to convey their intended message in the most efficient and effective way (e.g., Spada & Tomita, 2010). Such spontaneous speech production is claimed to mirror the present state of L2 learners' segmental, prosodic, temporal, lexical and grammatical competence (Trofimovich & Baker, 2006; Schmid & Hopp, 2013). For a more detailed discussion on pros and cons for various controlled and spontaneous measures, see Piske et al. (2001).

Recently, Derwing and Munro (2013) conducted longitudinal research to investigate how late L2 learners' spontaneous speech production changed at three different points across seven 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 non-native speakers in English progressively refined their overall comprehensibility (i.e., ease of understanding)— rather than accentedness (i.e., linguistic nativelikeness)— as a function of additional exposure and interaction. Conversely, those without such communicative intensions exhibited little improvement in comprehensibility and accentedness over time. The study provided a broad developmental pattern in late SLA that motivated regular L2 users to continue to improve their oral proficiency by selectively attending to certain linguistic features related to "successful L2 communication" instead of "the mastery of L2" (for definitions of comprehensibility and accentedness, see also Derwing & Munro, 2009).

Whereas Derwing and Munro's (2013) research has provided empirical support to the significant role of additional experience in late L2 oral learning, it also brought to light several important questions that future studies need to further investigate such as the way to define and analyze L2 oral proficiency. On the one hand, the precursor study's findings were based on global language ratings of accentedness and comprehensibility. On the other hand, L2 speaking proficiency has been traditionally characterized as a composite (rather than monolithic) phenomenon constituting various linguistic domains spanning pronunciation, fluency, vocabulary and grammar (De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012). The conceptualization of L2 oral ability has 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., Crosslely, Salsubury, & McNamara, 2014; Isaacs & Trofimovich, 2012) or based on human raters' holistic scores at a broader level (e.g., Saito, Trofimovich, & Isaacs, in press; Pinget, Bosker, Quené, & de Jong, 2014; Derwing, Rossiter, Munro, & Thomson, 2004). Despite such methodological variability among the existing literature, however, there is a consensus that L2 learners improve pronunciation, fluency, vocabulary and grammar at different learning rates at the initial stage of L2 oral learning (e.g., beginner to intermediate proficiency) than at the later stage of L2 oral learning (e.g., intermediate to advanced proficiency) (Isaacs & Trofimovich, 2012; Saito et al., in press).

To better understand the LOR effects on L2 ability learning, the primary aim of the current study is to replicate the findings of Derwing and Munro (2013) while focusing on a different group of L2 learners and type of analysis. 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 (from the onset to upper limit of late SLA) (see below for details). Second, the Japanese learners' spontaneous speech production was not only assessed for overall accentedness and comprehensibility (Derwing & Munro, 2009) but also submitted to pronunciation, fluency, vocabulary and grammar analyses. As such, the second aim of the study is to examine the differential effects of Japanese learners' LOR profiles on the 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) quality of the spontaneous production data.

LOR can only truly be considered as a rough index for the amount of input and interaction, especially for those who stay in their own communities and exclusively carry out the same routine conversations over the course of a day (Piske, MacKay, & Flege, 2001). Even when the quantity (i.e., LOR) is the same, there can be tremendous variation in the quality of the experience such as the amount of exposure to aural and written input in the media (Jia & Aaronson, 2003). There is no reason to doubt that adult L2 learners may greatly differ in terms of their language aptitude (DeKeyser, 2000) and levels of education (Derwing & Munro, 2009). However, the primary goal of the study is 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 measured through LOR profiles (among many other affecting factors) can significantly explain the level of variance in late L2 oral proficiency development when the analysis focuses only on those with as ample opportunities and motivation to use L2 as their main language of communication. Similar to the precursor research (i.e., Derwing & Munro, 2013) as well as many previous L2 speech studies on this topic (e.g., Flege et al., 1997; Trofimovich & Baker, 2006), the current study was designed to quantitatively examine the

relationship between language experience (LOR = 8 months to 13 years) and late SLA, but qualitative analysis on the nature of input and interaction was not further pursued.

Method

Participants

Japanese learners. Given that LOR effects can be clearly observed among immigrants that use L2 on a daily basis (Flege & Liu, 2001), caution was exercised to highlight, in particular, late learners without many opportunities to use L1 Japanese. Such participants were recruited through posts on social network websites for Japanese immigrants in Montreal where their Japanese community is relatively limited (e.g., 0.06% for immigrant population in Quebec) (Statistics Canada, 2008). Based on the results of the individual interview, 39 Japanese learners of English (M_{age} = 31.4 years ranging 21 to 43 years) were carefully selected as suitable participants for the study based on the following criteria.

First, since the theoretical debates between 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), the 39 Japanese learners of English with more than eight months of LOR were carefully chosen as ideal participants for the study. Second, all of the participants arrived in Canada after 19 years of age (M_{AOA} = 27.3 years ranging from 19 to 39 years) after receiving 6 to 10 years of English education, typically through the grammar translation method in Japan. Importantly, their age of arrival was not significantly related to their LOR (r = -.229, p = .160). Third, all of the participants reported very frequent use of L2 (their main language of communication at home and/or work was English).[2] Whereas most of the participants with LOR around one to two years were enrolled in private language institutes to improve their oral proficiency in English for various academic and career goals (they had invested time and money to study abroad in Canada), those with LOR of more than few years were either graduate students at English-speaking universities or full-time workers who dealt mainly with English-speaking customers.

These participants were equally divided into three LOR groups which were hypothesized to represent the initial, mid and final states of late SLA, respectively: (a) Short LOR (8 months to 1 year) for beginner learners who partially or fully complete the initial quick improvement over the first six months of LOR (Munro & Derwing, 2008); (b) Mid LOR (1 to 5 year) for intermediate learners who likely demonstrate a small and limited size of improvement due to the lack of the rate of learning advantage (Munro, 1993); and (c) Long LOR (5 to 13 years) for

[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 on participants' self-reports (typically based on a 6-point scale), as operationalized in many previous L2 speech studies (e.g., Flege et al., 1995). Notably, the subjective nature of these methods without any longitudinal observation has been considerably criticized (e.g., Flege, 2009). Acknowledging the inherent difficulty in quantifying how much L2 learners usually receive L2 input and interact with other native and non-native speakers in L2, I argue that dichotomous units (e.g., "*Which language do you mainly use, English or Japanese*?") rather than a continuum 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 at private and business settings. This is because the former type of question can more directly tap late bilinguals' most frequently-used language without taking into account their potentially different definitions of what it means by the upper (e.g., "*frequent use of L1/L2*") and lower (e.g., "*infrequent use of L1/L2*") end of the continuum. See Ranta and Meckelborg (2009) for innovative methodological options on this topic, such as blogging.

advanced learners who have reached their stage of ultimate attainment (Patkowski, 1990). Results of a one-way between groups ANOVA demonstrated no significant differences between the three groups in terms of mean age of arrival, $F(2, 36) = .285$, $p = .754$.

English and Japanese baseline. As a baseline for nativelike production in 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 that the Japanese learners did. All participants were native speakers of north-eastern Canadian and American English with a mean age of 21.3 years. As a baseline 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.

Detailed information about all of the participants is summarized in Table 1. Preliminary analysis regarding the effects of LOR on their English /r/ production was reported in Saito and Brajot (2013). In the current study, the overall linguistic qualities of the same dataset were re-analyzed from not only segmental but also global, prosodic, temporal, lexical and grammatical perspectives.

INSERT TABLE 1 HERE

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 the last three pictures (Pictures A, B, C, see below) for the final analysis; and (c) had only five 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 a certain length of spontaneous (rather than controlled) speech production 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.

These key words were carefully chosen to elicit segmentals 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, their performance was assumed to reflect the present state of their phonological abilities without careful avoidance of 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 participants' clear understanding of the procedures. The participants first described four

pictures randomly presented as distracters, and then described the remaining three pictures (A, B, C) randomly presented for the main analysis. In total, the participants generated 390 tokens (65 Japanese and baseline talkers × 3 pictures). Approximately 10 seconds of the beginning of a picture description (M = 8.3 sec ranging from 4.0 to 11.5 sec) was extracted for each participant. Since three pictures were described, there was an average of 24.0 seconds (ranging 14.6 from 32.7 sec) of free speech samples for the comprehensibility and accentedness judgment for each participant. The length of the entire samples for every participant can be considered as suitable compared to similar L2 speech research (e.g., Derwing & Munro, 2013 for 30 sec; Hopp & Schmid, 2013 for 10-20 sec).

Global Analyses

Inexperienced raters. To judge the global qualities (accentedness and comprehensibility) of the spontaneous speech samples, 10 native English undergraduate students were recruited at an English-speaking university in Vancouver (M_{age} = 23.2 years). As operationalized in the previous L2 speech research (Derwing & Munro, 2009), the judgement of accentedness and comprehensibility by definition refers to *naïve* raters' intuitive impression about L2 speech production. Thus, efforts were made to carefully select the raters based on their lack of familiarity and contact with Japanese learners of English. According to the language background questionnaire, their average self-reported familiarity with Japanese-accented English was 1.3 (SD = 0.5) on a 6-point scale (*6 = Very much*, *1 = Not at all*). Furthermore, all of them were business and psychology major students without much linguistics-related experience.

Accentedness and comprehensibility measures. First, the raters received a brief explanation on the definitions of accentedness (i.e., different patterns of speech sounds compared to their native language) (Isaacs & Trofimovich, 2012) and comprehensibility (i.e., the degree of ease or difficulty in listeners' understanding of L2 speech) (Derwing & Munro, 2009) (for the scripts of the explanation, see Appendix A) from a trained research assistant. The raters then familiarized themselves with the picture prompts and key words, and practiced the judgement 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 reflects on listeners' initial intuitions and impressions about L2 speech. They were reminded to use the entire scale to assess the comprehensive proficiency range of the dataset (e.g., "*nativelike*" for English baseline; "*near-nativelike*" for advanced Japanese learners with extensive length of residence; "*beginner*" for inexperienced Japanese learners who had just arrived in Canada). The entire session took approximately 1.5 hours.

Pronunciation, Fluency, Vocabulary, and Grammar Analyses

In terms of specific areas of language (i.e., pronunciation, fluency, vocabulary, and grammar), recent L2 speech research has corroborated human raters' intuitive judgments of various aspects of spontaneous speech production, such as segmentals (Piske et al., 2001), temporal fluency (Pinget et al., 2014; Derwing et al., 2004) and lexical accuracy, variation and richness (Crossley, Salsbury, & Mcnamara, 2014). According to Saito et al.'s (in press) validation study, the raters' intuitions about pronunciation, fluency, vocabulary, and grammar judgements were predictive of the actual linguistic properties of L2 speech production at a fine-grained level (summarized in Table 2).

INSERT TABLE 2 HERE

Following this vein of L2 assessment research, sub-domains of L2 speaking proficiency were defined and assessed based on the eight rater-based categories spanning the linguistic dimensions of pronunciation (segmentals, word stress, intonation), fluency (speech rate), vocabulary (appropriateness, richness) and grammar (accuracy, complexity)[3].

Experienced raters. To conduct linguistic analyses on the phonological, lexical and grammatical qualities of the spontaneous speech samples, five native English speakers were recruited as experienced raters (M_{age} = 29.4 years: 2 males, 3 females). All of them were graduate students in Applied Linguistics in an English-speaking university in Montreal. They reported teaching experience in various ESL and EFL settings ($M_{\text{years of teaching}}$ = 4.0 years ranging from 1 to 10 years), and previous training experience specific to pronunciation teaching (a semester-long course on applied phonetics and pronunciation teaching). Their 6-point familiarity with Japanese accented English varied, with a mean of 3.4 ranging from 1 to 5 (*6 = Very much*, *1 = Not at all*).

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 in a single WAV file. The raters listened to and evaluated each speech stimuli 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 via MATLAB. 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 of the tokens. 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". The slider was initially placed in the middle of each scale, and the raters were told that even a small movement of the slider will represent 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 A).

Transcript-based measures. To ensure raters' vocabulary and grammar analyses without being distracted by the pronunciation accuracy and fluency, I followed the methodology

[3] Such rater-based categories can be further reduced into 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 sub-domains of L2 speaking proficiency at a macro (i.e., rater-based categories) rather than a micro (i.e., actual linguistic properties) level. This is because L2 speech production in the study was conceptualized and analyzed based on *minimum* but still *perceptible* units to human raters. For further empirical research and discussion on more abstract (rather than broad) constructs of L2 oral proficiency, see Saito et al. (in press), de Jong et al. (2012), and Isaacs and Trofimovich (2012).

used in Crossley et al. (2014), which asked raters to read written transcripts of the speech samples. To this end, all speech tokens were first transcribed and verified by a trained research assistant, and then edited and cleaned up via modifying pronunciation-specific errors, such as those related to given target words (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, non-basic grammar).

The written transcripts were presented via the MATLAB software 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 A).

Training and rating sessions. The entire training and rating sessions took place for three 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.

Day 1. A trained research assistant provided the raters with thorough instructions on the eight pronunciation, vocabulary and grammar measures (for training scripts, see Appendix A). To check the accuracy and reliability of their linguistic analyses, the raters first judged phonological, lexical and grammatical qualities of a total of 40 non-native speakers' English speech samples, and then their audio- and transcript-based ratings were compared to the actual linguistic properties of the tokens. According to the results reported in Saito et al. (in press), not only did these raters judge the major components of L2 spontaneous speech production in an accurate manner (see Table 2) but also the inter-rater agreement among the raters' judgment scores was high as well (Cronbach alpha > .900). The entire session took approximately three hours with a 10-minute intermission at the halfway point.

Day 2. They first received recapped instructions for the four audio-based measures. Then, they familiarized themselves with picture prompts and key words of Pictures A, B, C, and practiced the procedure by rating three audio samples (not included in the main dataset). 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 dataset of 65 audio samples. The entire session took approximately 1.5 hours.

Day 3. After receiving recapped instructions on the four transcript-based measures and feedback on their practice ratings of the same three tokens (not included in the main dataset), they rated the 65 written samples. The entire session took approximately 0.5 hours.

Results

Global Analyses

Inter-rater agreement. Cronbach's alpha was calculated to check inter-rater agreement among the 10 raters' global judgement scores of 156 samples (52 Japanese talkers × 3 pictures) excluding 13 native speakers whom they invariably provided highest rating scores. In line with previous L2 comprehensibility research (e.g., Derwing & Munro, 2009), the results also found relatively high alpha levels for accentedness ($\alpha = .954$) and comprehensibility ($\alpha = .969$). Mean comprehensibility and accent ratings for each talker (3 picture descriptions) were computed by

pooling the data over all listeners. All participants' accentedness and comprehensibility scores were summarized in Table 3.

INSERT TABLE 3 HERE

Baseline comparisons. The first aim of the analysis was to investigate the interlanguage characteristics of the 39 Japanese learners (LOR = 8 months to 13 years) relative to inexperienced Japanese learners (LOR < 1 month), and the extent to which their performance simulated nativelike production. To this end, these Japanese learners' accentedness and comprehensibility scores were combined as one group (i.e., Japanese Learners), and analyzed compared to those of Japanese and English Baseline groups. The learners' accentedness and comprehensibility judgement scores were separately submitted to one-way ANOVAs with one between-group factor (3 groups [Japanese Learners, Japanese Baselines, English Baselines]). The results demonstrated significant main effects for accentedness F (2, 62) = 144.02, $p < .001$, and comprehensibility, F (2, 62) = 85.48, $p < .001$. According to the results of Bonferroni probability adjustments for multiple comparisons, accentedness and comprehensibility scores were significantly different between the three groups ($p < .001$). This suggests that the Japanese learners beyond approximately one year of LOR demonstrate better English proficiency than inexperienced Japanese learners with little experience overseas, but their performance should be considered as near-native rather than a nativelike level.

LOR effects. The second aim of the analysis was to examine whether and to what degree the 39 Japanese learners' performance differed as a function of LOR after the initial spike in improvement. In this regard, the subsequent analyses were conducted on the comprehensibility and accentedness scores of the 39 Japanese learners (LOR = 8 months to 13 years of LOR). A one-way ANOVA with one between-group factor (3 Groups [Short-, Mid-, Long-LOR]) was conducted on the learners' accentedness and comprehensibility scores as the dependent variable. The results found a significant main effect of Group for accentedness, $F(2, 36) = 4.939$, $p = .013$, and comprehensibility, $F(2, 36) = 7.258$, $p = .002$. According to the results of Bonferroni multiple comparisons, despite no significant difference between Mid- and Long-LOR ($p > .900$), both of the groups outperformed Short-LOR in terms of accentedness and comprehensibility:

- Accentedness: Long > Short ($p = .031$), Mid > Short ($p = .022$)
- Comprehensibility: Long > Short ($p = .007$), Mid > Short ($p = .005$)

Given that group analysis (i.e., ANOVA) is subject to the influence of the categorization specified in the study (Short-, Mid- and Long-LOR), the relationship between LOR and L2 performance was also investigated using a set of simple regression analyses on the learners' proficiency as the dependent variable and their LOR profiles as the independent variable. The results showed that LOR was significantly related to the Japanese learners' comprehensibility scores, $R = .362$, $F(1, 37) = 5.595$, $p = .023$, and thus moderately accounted for 13.1% of variance in their comprehensibility scores. Yet, LOR was not a significant predictor of their accentedness scores, $R = .267$, $F(1, 37) = 2.832$, $p = .101$.

Since LOR was a significant predictor for L2 comprehensibility development, the model needs to be scrutinized to find whether the relationship between LOR and proficiency remains equally linear or conforms to any abrupt discontinuity (i.e., the beginning of ultimate attainment) at certain points in the LOR continuum (8 months to 13 years). A piecewise regression analysis

was conducted on the learners' comprehensibility scores as the dependent variable and their LOR profiles as the independent variable. This analysis allowed us to further examine whether the inclusion of any breakpoint could improve their fitting levels to the dataset (for details of calculation method, see Appendix B) and has regularly been used in the previous SLA literature (e.g., Birdsong & Molis, 2001). According to the results, a sloping segment followed by a horizontal line was found to be the best fitting model for the LOR-comprehensibility function, $F(3, 35) = 5.984$, $p = .003$. The model demonstrated an optimal breakpoint of 3.04 years of LOR, after which the nature of the regression line changed from linear to horizontal ($p > .05$) (see Figure 1).

Taken together, the results suggest that Japanese learners continue to improve their comprehensibility scores, especially for the first three years of LOR, and then reach a point of ultimate attainment. In contrast, the effects of LOR on accentedness reduction remain unclear especially after the participants enjoy the rate of learning advantage within a LOR of one year. Last, their accentedness/comprehensibility performance was substantially different from the native English speaker baseline.

INSERT FIGURE 1

Pronunciation, Fluency, Vocabulary and Grammar Analyses

Inter-rater agreement. For the five raters' audio-based measures, the five raters demonstrated high inter-rater agreement in terms of their Cronbach's alpha value for segmentals ($\alpha = .845$), word stress ($\alpha = .821$), intonation ($\alpha = .806$) and speech rate ($\alpha = .824$). With respect to their transcript-based measures, the experienced raters again demonstrated high agreement when it came to vocabulary (Cronbach $\alpha = .756$ for appropriateness; $\alpha = .789$ for richness) and grammar ($\alpha = .745$ for accuracy; $\alpha = .757$ for complexity). Since their inter-rater reliability was still sufficiently satisfactory relative to L2 research standards ($\alpha = .700$-800) (Larson-Hall, 2010), their audio and transcript ratings were averaged across five raters in order to derive a single mean score per speaker for each rated category. The descriptive results of their vocabulary and grammar scores are presented in Table 3.

Baseline comparisons. First, the participants were categorized into three groups ($n = 13$ for Japanese Baselines; $n = 13$ for English Baselines; $n = 39$ for Japanese Learners). Then, their pronunciation, fluency, vocabulary, and grammar scores were submitted to a set of one-way ANOVAs with one between-group factor (Japanese Learners, Japanese Baselines, English Baselines), respectively. The results demonstrated significant main effects of Group for segmentals, $F(2, 62) = 143.24$, word stress, $F(2, 62) = 168.71$, intonation, $F(2, 62) = 117.56$, speech rate, $F(2, 62) = 66.50$, lexical appropriateness, $F(2, 62) = 25.655$, lexical richness, $F(2, 62) = 25.655$,lexical richness, $F(2, 62) = 21.711$, grammatical accuracy, $F(2, 62) = 37.289$, and grammatical complexity, $F(2, 62) = 25.357$ ($p < .001$). Mean contrasts revealed that Japanese learners' pronunciation and fluency performance was significantly different from that of Japanese and English Baselines ($p < .05$), suggesting that the Japanese learners make some significant gains in all linguistic aspects of L2 oral proficiency after eight months of L2 immersion but still fail 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 a set of one-way ANOVAs with one between-group factor (Short-, Mid-, Long-LOR), respectively. The results revealed significant main effects of Group for word stress, $F(2, 36) = 11.848$, p

< .001, intonation, $F(2, 36) = 8.075$, $p = .001$; speech rate, $F(2, 36) = 7.909$, $p = .001$; lexical appropriateness, $F(2, 36) = 4.577$, $p = .017$; grammatical accuracy, $F(2, 36) = 11.772$, $p < .001$; and grammatical complexity, $F(2, 36) = 3.653$, $p = .036$. Yet, significant effects were not detected for segmentals, $F(2, 36) = 2.638$, $p = .085$, and lexical richness, $F(2, 36) = 1.458$, $p = .246$. Mean comparisons yielded significant differences as follows:

- 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 ($p = .016$)
- Grammatical accuracy: Long > Short ($p < .001$), Mid > Short ($p = .020$)

A set of simple linear regression analyses were conducted to examine the extent to which LOR (8 months to 13 years) was predictive of the segmental, prosodic, fluency, lexical and grammatical qualities of the 39 Japanese learners' L2 speech production development. The results showed that LOR significantly predicted their suprasegmental and lexicogrammar accuracy, explaining 28.8% of variance in word stress, $R = .537$, $F(1, 37) = 14.977$, $p < .001$; 29.4% in intonation, $R = .542$, $F(1, 37) = 15.414$, $p < .001$; 32.7% in speech rate, $R = .572$, , $F(1, 37) = 17.973$, $p < .001$; 14.0% of variance in lexical appropriateness, $R = .374$, $F(1, 37) = 15.414$, $p = .019$; and 23.9% of variance in grammatical accuracy, $R = .489$, $F(1, 37) = 6.027$, $p = .002$. Conversely, the LOR-proficiency function failed to reach statistical significance in segmentals, $R = .299$, $F(1, 37) = 3.631$, $p = .065$, lexical richness, $R = .199$, $F(1, 37) = 1.532$, $p = .224$, and grammatical complexity, $R = .199$, $F(1, 37) = 4.034$, $p = .052$.

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 were 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 showed that the regression was improved with the modified model described first as a sloping segment and then horizontal in terms of (a) word stress, $F(3, 35) = 9.887$, $p < .001$, (b) intonation, $F(3, 35) = 6.993$, $p = .001$, (c) speech rate, $F(3, 35) = 6.481$, $p = .002$, and (d) grammatical accuracy, $F(3, 35) = 12.511$, $p < .001$. The extensive amount of LOR was suggested as a break point for word stress (5.62 years), intonation (5.99 years), and speech rate (5.50 years), respectively; the mid LOR for grammatical accuracy (3.04 years). For lexical appropriateness, however, the LOR-proficiency regression was best described as two horizontal segments at different levels, $F(2, 36) = 10.753$, $p < .001$, with a breaking point of LOR = 1.31 years (see Figure 2).

INSERT FIGURE 2

To summarize, the statistical results indicate that (a) Japanese learners continue to improve lexicogrammatical accuracy within the first three years of LOR and suprasegmetal performance within the first five years of L2 experience, but (b) they may need much more L2 experience (LOR > 5 years) to demonstrate any significant improvement in their segmental and grammatical complexity performance.

Linguistic Correlates of Comprehensibility and Accentedness

The final analysis focused on illustrating the relationship between pronunciation, vocabulary and grammar qualities, on the one hand, and the overall comprehensibility and accentedness of L2 speech production, on the other hand. To examine the phonological influences on the global judgements, a series of partial correlation analyses were first conducted to see how the four pronunciation scores (segmentals, word stress, intonation, speech rate) of all participants were associated with comprehensibility and accentedess judgements, respectively, when their lexicogrammar scores were factored out. As shown in Table 3, all of the pronunciation categories were significantly correlated with comprehensibility and accentedness at a $p < .01$ level (Bonferroni corrected). According to the Fisher r-to-z transformation, no statistical difference in the strength of correlation coefficients between the four pronunciation categories and comprehensibility was found at a $p < .008$ level (Bonferroni corrected). Yet, the results showed that segmentals and word stress were more strongly tied with accentedness than speech rate ($p < .001$). This in turn suggests that, while all pronunciation elements are equally related to comprehensibility, the talkers' pronunciation of words (segmental accuracy with correct word stress) is more strongly predictive of accentedness than their delivery of sentences with good intonation and optimal speech rate.

INSERT TABLE 3

Next, a series of partial correlation analyses were conducted to examine the differential lexical and grammatical contributions to comprehensibility and accentedness when the influence of pronunciation categories were controlled. It was found that the proper (rather than sophisticated) vocabulary and grammar usage was significantly related to comprehensibility, but none of the lexicogrammar variables predicted accentedness (see Table 4). According to the results of the Fisher r-to-z transformation, no significant differences were found in the correlation coefficients between the four lexicogrammar variables and comprehensibility/accentedness, respectively ($p > .008$).

INSERT TABLE 4

Thus, the results indicate that (a) comprehensibility is equally related to the segmental, prosodic, temporal, lexical and grammatical qualities of L2 speech production, and (b) accentedness is mainly accounted for by pronunciation (particularly segmentals and word stress) rather than lexical and grammar variables.

Discussion

Whereas the role of experience in interlanguage development still continues to be a source of theoretical debate (DeKeyser & Larson-Hall, 2005 vs. Flege, 2009), Derwing and Munro (2013) have recently revealed an important longitudinal finding that late L2 learners continuously enhanced their comprehensibility (but not minimised accent) when their L2 oral ability was assessed at the time of two and seven years of their residence in Canada, respectively. This cross-sectional study tested the generalizability of the findings in the context of highly motivated Japanese learners of English with varying LOR profiles beyond the early phase of late SLA (i.e., 8 months to 13 years).

If we look at the Japanese learners' L2 proficiency based on their foreign accentedness scores, as has been the case with most previous LOR research (e.g., Flege, 1988), their substantial improvement seemed to be limited to the first year of L2 experience, beyond which

their performance was subject to individual differences regardless of increased LOR. The results appeared 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). By taking into account multiple linguistic dimensions of L2 speech production, however, the results further revealed that 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 cross-sectional data showed that LOR was a significant predictor for the improved comprehensibility but not for the reduced accentedness of Japanese learners' L2 speech production, especially within approximately three years of their residence in Canada. In terms of the pronunciation and lexicogrammar analyses, the results also indicated that the extensive amount of LOR seemed to facilitate the development of certain linguistic features affecting comprehensibility such as word stress, intonation, speech rate, and lexicogrammatical accuracy. However, the role of experience remained unclear for the development of the other linguistic features strongly related to accentedness such as segmentals, vocabulary richness, and grammatical complexity. Furthermore, the results of the piecewise regression analyses suggest that their performance reaches a plateau in naturalistic settings as a function of a different amount of LOR according to pronunciation (5-6 years), grammar (3-4 years), and vocabulary (1 year). The results of the linguistic scrutiny on the dataset do not concur with the strong version of the CPH, which hypothesizes the presence of a LOR-proficiency link exclusively at the initial stage of late SLA (LOR < 6 months). Rather, these results echo the SLM, which assumes that L2 learners may continue to show their improved oral ability later in life over an extended period of L2 immersion (LOR = 5-6 years).

At the same time, however, the role of experience in late L2 oral proficiency development, unlike early bilingualism and L1 acquisition, can also be characterized as multidimensional in nature as a result of learners' overall goals (accentedness vs. comprehensibility) and their relevant linguistic processing (i.e., functional vs. sophisticated language use). First and foremost, the results presented here largely indicated that experience effects are evident in the linguistic correlates of comprehensibility but not those of accentedness. As L2 learners gain experience in L2, they may prioritize the development of good prosody, optimal speech rate, proper vocabulary, and grammar usage (over segmental accuracy and the sophisticated use of vocabulary richness and grammar) for the purpose of achieving successful communication in their private, business, and academic settings. The findings can be well accounted for by the interactionist view, which states what learners essentially aim to achieve during their social interactions with their interlocutors—comprehensibility rather than accentedness—serves as a crucial variable, especially for late SLA processes (Gass & Mackey, 2006).

The Interaction Hypothesis (e.g., Long, 1996) claims that language learning takes place precisely when comprehensibility is retrieved during L2 conversational exchanges between native and non-native speakers (or non-native and non-native speakers). Namely, communication breakdowns are more likely when L2 learners interact with other native and non-native speakers, and this results in intuitive or conscious efforts to correct impaired linguistic accuracy by seeking successful comprehensibility. 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 they would not otherwise be able to decipher alone

(Dunn & Lantolf, 1998). This whole conversational move is hypothesized to be facilitative of the rate and ultimate attainment of 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), the theory indicates that L2 learners selectively strive to notice and practice linguistic domains closely linked to comprehensibility (but not necessarily relevant to accentedness) (e.g., Mackey, Gass, & McDonough, 2000). Drawing on the interactionist perspective on language learning, the findings in the current study concur 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 non-native speakers in an L2 community (Flege & Liu, 2001), and their L2 speaking proficiency is assessed based on their performance on comprehensibility-related linguistic features (i.e., suprasegmentals, vocabulary appropriateness, grammatical accuracy: Isaacs & Trofimovich, 2012).

Limitations

While the findings reported here provide useful insights into the relationship between experience and late SLA, several crucial limitations need to be acknowledged. First, the results based on such a small dataset (n = 39 Japanese learners) needs to be interpreted with caution, because some of the findings could be simply due to artifacts of statistical analyses. For example, a breakpoint in the regression models (the end of the LOR-proficiency link) in the study may have corresponded to information density rather than a point of L2 ultimate attainment. Since there were considerable differences between the groups in their range of LORs, correlations in the lower end of the LOR range (e.g., 0.8-1 years for Short LOR) are more likely to be found than those in the higher end (e.g., 5-13 years for Long LOR). While the current study successfully replicated Derwing and Munro (2013), the generalizability of the findings on L2 comprehensibility development still need to be tested, especially with larger sample sizes.

Second, any suggestions, especially regarding learners' lexical and grammatical competence, should only be considered as tentatively due to crucial 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 full comprehension of the lexical and grammatical influences on comprehensibility and accentedness following norms in L2 vocabulary and grammar research (e.g., Lu, 2012 for 3 minutes). Another limitation could be that the nature of the task (describing a picture cartoon) did not elicit a sufficiently wide range of infrequent lexical items. Given that all participants may have used similar frequent lexical items, the predictive power of lexical indices did not reach statistical significance, especially after L2 learners had reached minimum lexical thresholds to successfully complete the task (see Crowther, Trofimovich, Isaacs, & Saito, forthcoming).

Third, sub-dimensions of L2 speaking proficiency were measured only based on human raters' pronunciation, fluency, vocabulary and grammar judgements. They can be further scruited via objective 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 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, some may argue that the Japanese learners actually used more domain general learning and explicit strategies (rather than incidental and automatic learning processes) to improve their L2 oral ability over the extensive period of their immersion in Canada. That is, the current study has yet to provide a conclusive answer in regards to the detailed nature of the underlying system that late L2 learners primarily draw on to improve their oral ability (the presence or absence of the implicit learning mechanism even during post-pubertal SLA).[4] Of further interest would be a study that highlights another crucial affecting factor for late SLA—learners' age of acquisition (AOA) (i.e., the first intensive exposure to input and interaction in an L2 speaking environment). Although AOA is a relatively strong predictor of the end state of SLA (i.e., 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 country before puberty, it has remained highly controversial whether, to what degree, and how such age effects can be germane to late bilinguals whose immersion in the L2 country starts after puberty (e.g., Birdsong, 2005 vs. DeKeyser & Larson-Hall, 2005).

The two competing theories (CPH vs. SLM) equally assume that very few late learners can attain nativelike performance, however, owing to two essentially different causes: (a) the lack of access to the human language learning system after the passing of the critical period (DeKeyser & Larson-Hall, 2005) and/or (b) the age-related decline in many human cognitive functions (starting at age 20 years) such as working memory, executive control, speech sound 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 AOA can be a significant predictor for not only early but also late bilinguals' ultimate attainment because they use the same learning mechanism for L1 and L2 acquisition. As many SLA researchers (e.g., DeKeyser & Larson-Hall, 2005) point out, however, future research of this kind needs to include an ample number of participants who have already reached the upper limit of SLA after an extensive amount of L2 immersion (LOR > 6 years), as opposed to the current investigation including only 13 advanced L2 learners with such extensive L2 experience.

Conclusion

Building on the recent L2 speech research (Derwing & Munro, 2013), this cross-sectional study re-examined the extent to which late L2 learners' oral proficiency can be enhanced in relation to their additional amount of experience, operationalized as LOR. According to the results, LOR was generally predictive of improved L2 comprehensibility as a result of the continuous development of good prosody, optimal speech rate, and proper lexicogrammar usage, but it required a great amount of L2 experience to enhance accentedness, which entails refined segmental accuracy, vocabulary richness, and grammatical complexity. These results successfully replicated Derwing and Munro's (2013) longitudinal dataset, suggesting an ability to learn new language is maintained over a life span, and experience effects on late L2 learners'

[4] To my knowledge, no single SLA research (including the current study) has ever 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.

oral proficiency development and attainment simulate a gradual, constant, and extensive learning processes over an extensive period of L2 immersion (e.g., 6 years of LOR). Noteworthy is that L2 learners improve their oral proficiency 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.

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. 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. 109127). New York: Oxford.University Press.

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*.

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (forthcoming). Does task affect second language comprehensibility?

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: 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: Mouton de Gruyter.

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

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

Gatbonton, E., Trofimovich, P., & Segalowitz, N. (2011). Ethnic group affiliation and patterns of development of a phonological variable. *The 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.

Hakuta, K., Bialystok, E., &Wiley, E. (2003). Critical evidence: A test of the critical-period hypothesis for second-language acquisition. *Psychological Science, 14,* 31–38.

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*. 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.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Review, 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: 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.

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. (2012). 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. *The Canadian Modern Language Review, 69*, 1-33.

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). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*.

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 Statistics Canada: 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.

Table 1. *Characteristics of Four Japanese Groups and One English Group*

	Group									
	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

Table 2. *Summary of Linguistic Predictors for Human Raters' Pronunciation, Fluency, Vocabulary and Grammar Judgement in Saito et al. (in press)*

Rater judgement measures	Linguistic predictors
A. Phonology	
Segmentals	No. of vowel and consonant errors
Word stress	No. of word stress errors
Intonation	No. of intonation errors
Speech rate	Mean length of run; no. of unfilled pauses; articulation rate
B. Lexicogrammar	
Lexical appropriateness	No. of lexical errors
Lexical richness	Type frequency, token frequency
Grammatical accuracy	No. of grammatical errors
Grammatical complexity	Subordinate clause ratio

Table 3. *Means and Standard Deviations for Rated Global, Phonological, Temporal, and Lexicogrammatical Qualities of Japanese Learners and Japanese and English Baseline' Picture Descriptions*

	Low LOR (*n* = 13)	Mid LOR (*n* = 10)	Long LOR (*n* = 10)	Japanese Baseline (*n* = 10)	English Baseline (*n* = 10)
A. Global ratings (9 points)					
Accentedness	7.3 (0.6)	6.0 (1.4)	5.9 (1.6)	7.8 (0.6)	1.1 (0.1)
Comprehensibility	5.3 (1.0)	3.9 (1.2)	3.8 (1.2)	6.0 (1.0)	1.0 (0.1)
B. Audio ratings (1000 points)					
Segmentals	310 (94)	406 (131)	429 (181)	272 (106)	994 (7)
Word stress	363 (72)	461 (75)	542 (124)	354 (81)	986 (27)
Intonation	268 (90)	348 (104)	453 (150)	256 (92)	885 (58)
Speech rate	408 (152)	541 (156)	645 (148)	294 (154)	977 (27)
C. Transcript ratings (1000 points)					
Lexical appropriateness	675 (127)	758 (123)	801 (58)	629 (103)	912 (49)
Lexical richness	410 (220)	402 (178)	517 (167)	241 (87)	700 (209)
Grammatical accuracy	378 (139)	551 (182)	667 (134)	337 (132)	892 (111)
Grammatical complexity	277 (144)	296 (152)	427 (166)	186 (73)	640 (234)

Note. 9 point scale (*1 = Little accent, easy to understand, 9 = Heavily accented, hard to understand*); 1000 point scale (*1 = Non-targetlike production, 1000 = Targetlike production*)

Table 4. *Partial Correlations Between Four Rated Pronunciation Variables and Comprehensibility and Accentedness, with the Influence of Four Lexicogrammar Variables Controlled*

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

Note. *α < .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*

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

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

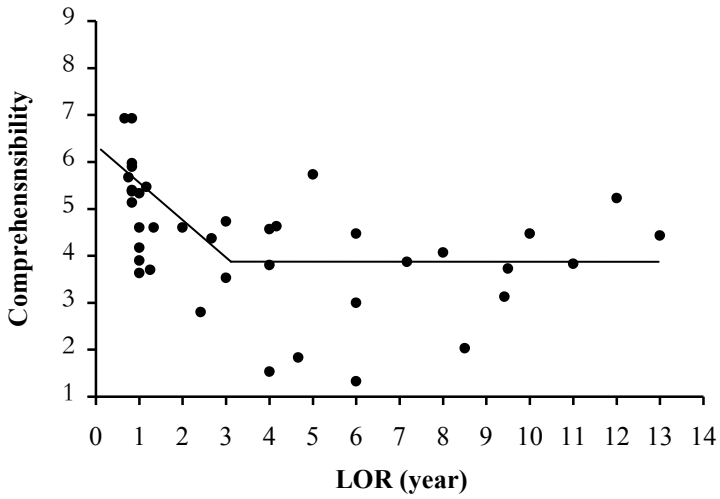
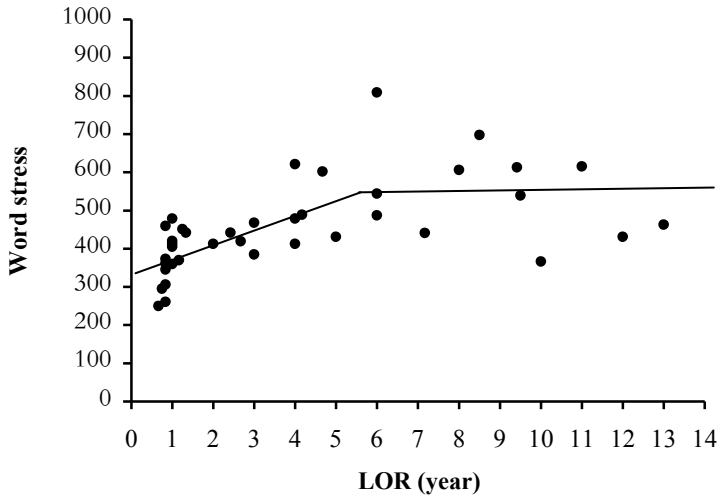
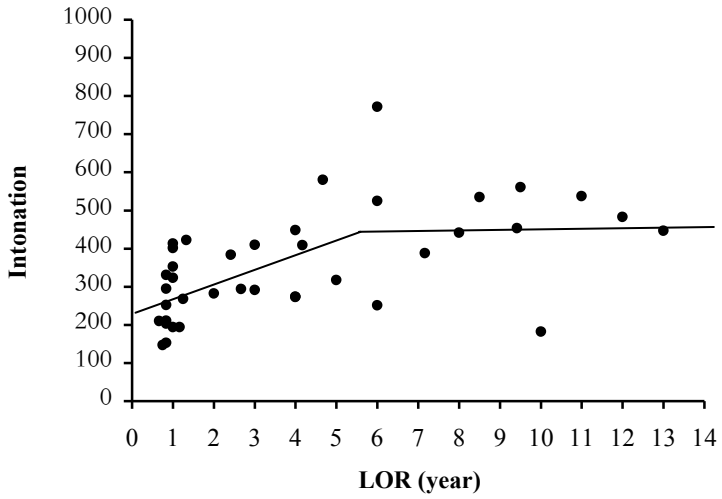
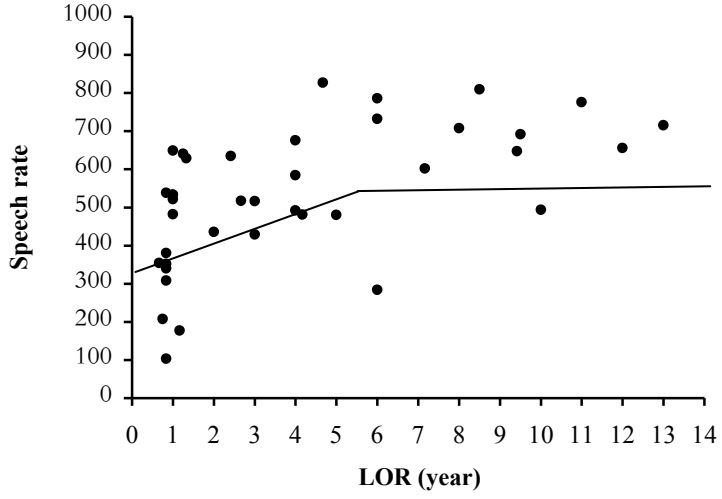


Figure 1. Comprehensibility scores (1 = *easy to understand*, 9 = *hard to understand*) and LOR profiles of 39 Japanese learners (8 months < x < 13 years)


Word stress
1000
900
800
700
600
500
400
300
200
100
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
LOR (year)


Intonation
1000
900
800
700
600
500
400
300
200
100
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
LOR (year)


Speech rate
1000
900
800
700
600
500
400
300
200
100
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
LOR (year)

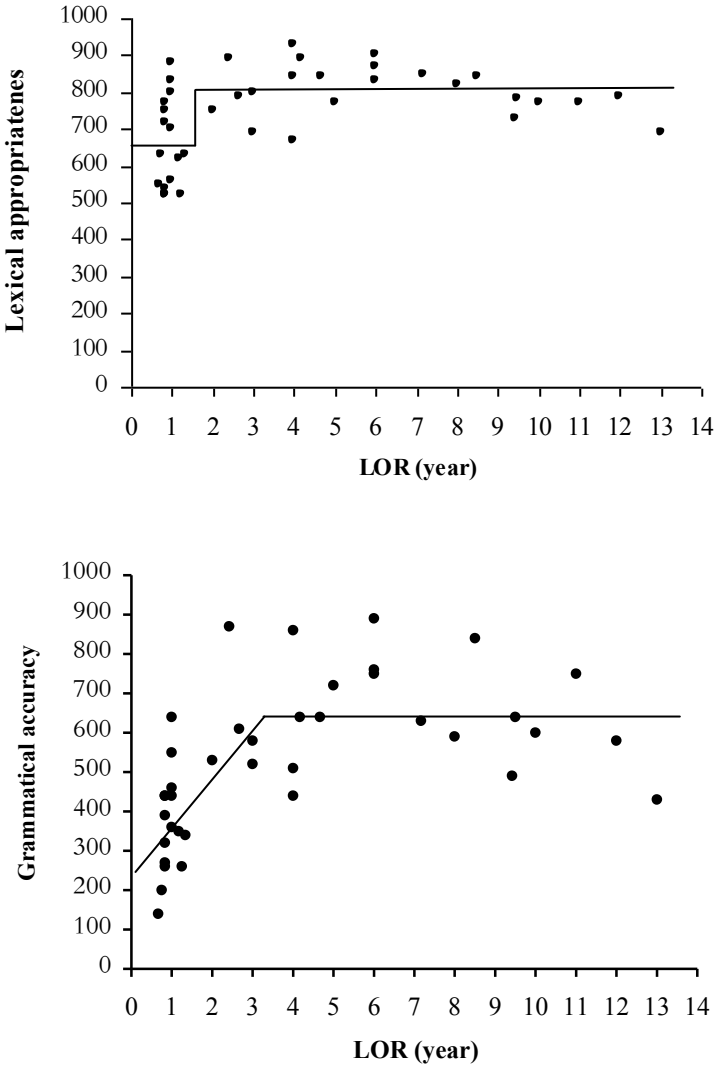


Figure 2. Pronunciation, fluency, vocabulary, and grammar scores (0 = *non-targetlike*, 1000 = *targetlike*) and LOR profiles of 39 Japanese learners (8 months to 13 years)