Title:

The Role of Age of Acquisition in Late Second Language Oral Proficiency Attainment

Running Head:

AGE EFFECTS ON LATE SLA

Corresponding Author:

Kazuya Saito

Birkbeck, University of London

The Department of Applied Linguistics and Communication

30 Russell Square, London WC1B 5DT, UK.

Email: k.saito@bbk.ac.uk

Abstract

The current project examined whether and to what degree age of acquisition (AOA), defined as the first intensive exposure to a second language (L2) environment, can be predictive of the end state of post-pubertal L2 oral proficiency attainment. Data were collected from 88 experienced Japanese learners of English and two groups of 20 baseline speakers (inexperienced Japanese speakers and native English speakers). The global quality of their spontaneous speech production was first 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, AOA was negatively correlated with the accentedness and comprehensibility components of L2 speech production, owing to relatively strong age effects on segmental and prosodic attainment. Yet, significant age effects were not observed in the case of fluency and lexicogrammar attainment. The results in turn suggest that AOA plays a key role in determining the extent to which learners can attain advanced-level L2 oral abilities via improving the phonological domain of language (correct consonant and vowel pronunciation, adequate and varied prosody); and that the temporal and lexicogrammatical domains of language (optimal speech rate, the proper vocabulary and grammar usage) may be enhanced with increased L2 experience, regardless of age.

Key words: Age, L2 oral ability, Late bilingualism; Comprehensibility; Foreign accentedness; Pronunciation; Fluency; Lexicon; Grammar

.

The Role of Age of Acquisition in Late Second Language Oral Proficiency Attainment

Whereas late second language (L2) learners tend to demonstrate a great amount of improvement in relation to increased L2 experience, especially around the early phase of second language acquisition (SLA) processes (i.e., rate of learning), many researchers have extensively examined the extent to which they can continue to enhance learners' oral ability (i.e., ultimate attainment) in a way that could lead to *near* nativelike proficiency. On the one hand, few bilinguals demonstrate perfect proficiency in all aspects of language like monolinguals do (Abrahamsson & Hyltenstam, 2009; Flege, Yeni-Komshian, & Liu, 1999). On the other hand, some learners are able to attain high-level L2 performance, and the incidence of successful SLA is influenced by several factors, such as the linguistic distance between first language (L1) and L2 structures (Best & Tyler, 2007; Flege, 2003), aptitude (DeKeyser, Alfi-Shabtay, & Ravid, 2010; Granena & Long, 2013), the quality and quantity of L2 input (Flege & Liu, 2001; Jia, & Aaronson, 2003), cognitive aging (Birdsong, 2005, 2006; Hakuta, Bialystok, & Wiley 2003), motivation (Dörnyei & Kubanyiova, 2014; Derwing & Munro, 2013), level of education (Derwing & Munro, 2009; Spada & Tomita, 2010), and ethnic identity (Gatbonton, Trofimovich, & Segalowitz, 2011; Pavlenko, & Blackledge, 2004).

Among these factors, previous L2 speech research has paid by far the most attention to examining learners' age of acquisition (AOA), defined as the first intensive exposure to input and interaction in an L2 speaking environment.[1] There has been a great deal of empirical evidence which has shown that 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 (e.g., AOA < 16 years)

(e.g., Abrahamsson, 2012; Abrahamsson & Hyltenstam, 2008, 2009; DeKeyser et al., 2010; Flege, Munro, & MacKay, 1995; Flege et al., 1999; Granena & Long, 2013; Hopp & Schmid, 2013; Johnson & Newport, 1989; Munro & Mann, 2005; Patkowski, 1880, 1990). However, 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., AOA > 16 years) (e.g., Birdsong, 2005 vs. DeKeyser & Larson-Hall, 2005).

In what follows, I will first provide an overview on two competing theoretical explanations for age effects on late bilingualism (i.e., Critical Period Hypothesis vs. Cognitive Aging Hypothesis). Subsequently, I will review how recent studies have examined the interlanguage development of L2 oral ability from pronunciation, fluency, vocabulary, and grammar research perspectives. Last, I will present the results of the current study, which examined in depth the role of AOA in predicting the global, segmental, prosodic, temporal, lexical, and grammatical qualities of L2 oral proficiency attainment by 88 experienced Japanese late arrival (>16 years old) learners.

Background

Critical Period Hypothesis (CPH)

Several researchers have claimed that age effects are found only for early bilinguals but not for late bilinguals, due to the fundamental and qualitative differences between the two SLA processes (e.g., Abrahamsson, 2012; DeKeyser & Larson-Hall, 2005; Granena & Long, 2013; Johnson & Newport, 1989; Paradis, 2009; Patkowski, 1990; Scovel, 2000). From birth, early learners progressively lose access to an assumed language-specific cognition system procedurally represented in the brain, a system used to pick up the L2 through mere exposure to natural input in an automatic and incidental manner. This results in strong age effects on the final

quality of early bilingualism. Following the maturational accounts for L1 acquisition, Abrahamsson (2012) argued that the gradual loss of cerebral plasticity for language acquisition is correlated with the neurologically-determined myelination processes of cortical neurons (Pulvermüller & Schumann, 1994). After passing such a critical and optimal period for implicit language acquisition, late SLA processes do not always benefit from simply being exposed to L2 input. Since the influence of aging effects apply only to implicit and automatic language learning, late learners' ultimate attainment patterns are not associated with their AOA profiles (e.g., Abrahamsson, 2012; Abrahamsson & Hyltenstam, 2008, 2009; DeKeyser, 2000; DeKeyser et al., 2010; Johnson & Newport, 1989; Granena & Long, 2013; Patkowski, 1980, 1990).

During post critical period SLA, late learners likely draw on explicit (rather than implicit) strategies via declarative memory to learn the L2 in a manner similar to the intentional and effortful learning of other general cognitive skills, such as mathematics and computer programming (Abramahsson, 2012; DeKeyser et al., 2010). As is the case with developing domain general cognition (e.g., Anderson, 1993), it has been shown that late learners' L2 speech learning is characterized by the power law of learning—a quick improvement over the first few months of length of residence (LOR) in the L2 environment, followed by a levelling-off, despite additional practice and environmental input (for a review, see DeKeyser & Larson-Hall, 2005). For example, experienced learners tend to note superior L2 speech ability when their performance is compared to beginning learners (LOR < 1 year) (Flege & Fletcher, 1991; Flege, Bohn, & Jang, 1997), but not when compared to intermediate learners (LOR > 1 year) (Flege, Munro, & Fox, 1994; Larson-Hall, 2006).

In terms of their ultimate attainment, the upper limit and incidence of near-nativelike performance in late SLA is not linked with AOA but is instead subject to individual learner

differences such as exceptional learners with high language learning aptitude. DeKeyser (2000) found that near-native performance on oral grammaticality judgment tests by late Hungarian learners of English significantly correlated with their high analytical aptitude (see also Abrahamsson & Hyltenstam, 2008; DeKeyser et al., 2010). In L2 pronunciation attainment, Granena and Long (2013) examined how late Spanish-Chinese bilinguals' speaking skills (i.e., reading aloud a paragraph) were related to various domains of their language learning aptitude measured via the Llama Language Aptitude Test (Meara, 2005). The results demonstrated a strong link between their foreign accentedness scores and certain aspects of the aptitude test such as sound-symbol correspondence (i.e., connecting sounds with relevant symbols) and grammatical inferencing (i.e., discovering grammar rules in an unknown language).

Cognitive Aging Hypothesis (CAH)

Other researchers have claimed that the ultimate L2 attainment of both early and late bilinguals can be susceptible to age effects throughout the lifespan without any cut-off point, suggesting that the language learning capacity used in successful L1 speech acquisition remains active even after puberty and can be applied to late SLA (Best & Tyler, 2007; Bialystok, 1997; Birdsong, 2005, 2006; Flege, 2003; Hakuta et al., 2003; Hopp & Schmid, 2013).[2] According to this theoretical position, one underlying cause for more salient foreign accents in older rather than younger learners can be environmental, as opposed to maturational, in nature. That is, some late immigrants may be exposed to somewhat limited L2 input by choosing to exclusively use the L1 at home and work within the same language background community, although early learners tend to receive a substantial amount of native speaker input from their caregivers and peers (Jia & Aaronson, 2003). This theoretical position, therefore, suggests that late learners continue to learn new sounds as long as they can meet similar socio-psychological conditions

that early learners likely benefit from (Bialystok, 1997), such as high-frequent use of the L2 (Flege & Liu, 2001) and high willingness to communicate in the L2 (Derwing & Munro, 2013).

Importantly, this position also assumes that the final quality of late SLA is closely related to AOA, as evidenced in L1 acquisition and early bilingualism. This is because AOA acts as a barometer for a subset of learner intrinsic variables affecting SLA, such as the degree of L1 and L2 development. Late SLA builds on the common linguistic space, where the L1 has been fully developed, resulting in inevitable foreign accent development (e.g., Best & Tyler, 2007; Flege et al., 1995). The mutual interaction between L1-L2 categories in turn indicates not only that late learners who arrive in the L2 environment during earlier adulthood attain better L2 proficiency after a larger amount of L2 practice (Baker, Trofimovich, Flege, Mack, & Halter, 2008; Yeni-Komshian, Flege, & Liu, 2000), but also that the intensive and constant use of the L2 alters their L1 performance (e.g., Bialystok & Miller, 1999; Hopp & Schmid, 2013). Another crucial variable is the age-related decline in many human cognitive functions, such as working memory, executive control, speech sound processing, or inhibition of task-irrelevant information (Hakuta et al., 2003). Birdsong (2005, 2006) ascribed the notion of cognitive aging to the biologically (but not maturationally) defined aging process in the brain system, such as decreases in brain volume and nigrostriatal dopamine (starting at age 20 years). According to Birdsong (2006), the dopamine system is believed to promote "defossilization, an undoing of automatized nontargetlike linguistic performance" (p. 32) while preventing L2 learners from drawing on their L1-related strategies during L2 processes.

Several studies lend some evidence to the significant role of AOA in late bilinguals' attainment of various linguistic abilities. Birdsong and Molis (2001) showed that AOA (> 16 years) was predictive of the oral grammaticality judgement scores of Spanish learners of English

(see also Bialystok & Miller, 1999; Hakuta et al., 2003). In the context of 33 late Hungarian learners of English, Hellman (2011) showed that the ratio of their nativelike lexical attainment (vocabulary size and depth of word knowledge) was negatively correlated with AOA (> 16 years). Finally, Flege, Birdsong, Bialystok, Mack, Sung, and Tsukada (2006) examined 36 late Korean-English bilinguals' sentence production (AOA > 20 years) and found a significant correlation between their accent ratings and AOA. According to the results, AOA accounted for 30.3% of variance in the late learners' pronunciation performance (r = -.55).

In sum, the Critical Period Hypothesis and Cognitive Aging Hypothesis present sharply contrastive beliefs as to the predictive power of age for late L2 ultimate attainment. The CPH predicts "discontinuity in the AOA-proficiency", due to a fundamental and qualitative change in learning potential after the mid teens (DeKeyser & Larson-Hall, p. 97). The CAH suggests "a linear monotonic decline of learning over the [AOA] spectrum, with age effects continuing past the point at which maturation has ceased" (Birdsong, 2005, p. 115). Whereas previous findings have been mixed (e.g., Granena & Long, 2013 vs. Flege et al., 2006), the disagreement between the positions is probably due to the complex nature of the assessment methodologies and the developmental patterns inherent in adult L2 speech production. Below, I will review a wide range of measures that relevant studies have adopted to analyze L2 speech production, and the way adult L2 learners can promote the development of pronunciation, fluency, vocabulary, and grammar as they increase their amount of L2 experience in naturalistic settings.

Assessing and Developing L2 Speech Production

One important factor in researching late learners' oral proficiency concerns whether their performance is elicited at a controlled or spontaneous speech level. Many AOA researchers have exclusively drawn on controlled speech tasks, such as paragraph and sentence readings, whereby

participants chorally repeat audio and written prompts (see Piske, MacKay, & Flege, 2001 for review). Researchers may prefer these controlled tasks in order to highlight certain features in L2 speech production that are of particular interest to them. Yet, these tasks also allow adult L2 learners to focus specifically on carefully monitoring their correct linguistic forms, drawing on their explicit knowledge (Jian, 2007). Given that L2 learners generally demonstrate better proficiency under formal rather than free production conditions (Major, 2007), such highly controlled L2 performance is claimed to merely mirror "language-like behavior" rather than "actual L2 proficiency" (Abrahamsson & Hyltenstam, 2009, p. 254). In this regard, many SLA researchers have emphasized the importance of tapping the present state of L2 learners' oral competence by adopting spontaneous speech tasks (e.g., picture narratives) which push L2 learners to pay equal attention to not only the phonological but also the temporal, lexical, grammatical, and discoursal domains of language to convey their communicative intentions in an effective and efficient manner (Spada & Tomita, 2010) under time pressure (Ellis, 2005).

Recently, L2 speech research has begun to analyze how late learners can improve their oral proficiency, especially at a spontaneous speech level, by adopting a range of linguistic measures in L2 pronunciation, fluency, vocabulary and grammar research. Derwing and Munro (2013) conducted longitudinal research to probe how late Slavic and Chinese immigrants in Canada could enhance the global qualities of their L2 speech production by comparing their performance at three different points of time (LOR = 0, 2, 7 years). The results showed that their overall comprehensibility (ease of understanding) gradually improved from the onset to the endpoint of the data collection, especially among the Slavic learners, who generally reported positive attitudes towards interacting with native and non-native speakers in the L2. Additionally,

the global foreign accentedness (linguistic nativelikeness) of all participants remained unaltered over time (see also Derwing & Munro, 2009; Derwing, Rossiter, Munro, & Thomson, 2004).

Using a cross-sectional research design and late Japanese learners of English with short, mid and long LOR profiles (0 to 10 years), Saito (in press) not only replicated Derwing and Munro's (2013) research findings (i.e., experience effects for comprehensibility rather than accentedness), but also found indications of certain developmental patterns. Specifically, the learners improved specific elements of their L2 oral ability at different learning rates in relation to increased LOR. According to the results of simple and piecewise regression analyses, much learning appeared to take place at the initial and mid stages of SLA (LOR = 1 to 3 years) in terms of their proper lexicogrammar usage. Continuous development seemed to be observed over an extensive period of time (LOR = 5 to 6 years) in terms of the prosodic (word stress, intonation) and temporal (speech rate) domains of L2 speech production (see also Trofimovich & Baker, 2006). However, the amount of improvement in their sophisticated use of language (segmental accuracy, vocabulary richness, grammatical complexity) was relatively limited.

Taken together, recent speech studies (Saito, in press; Derwing & Munro, 2013) suggest that L2 learners tend to selectively pay attention to improving the functional use of language (i.e., speech rate, adequate and varied prosody, proper lexicogrammar usage) with a view of achieving successful communication in the L2 (i.e., comprehensibility) throughout the *initial* and *mid* states of late SLA. It is important to note that these previous studies were designed to examine how adult L2 learners develop certain aspects of their interlanguage system (especially those crucial for comprehensibility) in relation to an increasing amount of experience (LOR = 0 to 6 years). However, they have yet to answer whether and to what degree these learners can *ultimately* improve all domains of L2 speech production (influencing not only comprehensibility but also

accentedness), and thereafter reach near nativelike proficiency at the *end* state of late SLA. In particular, it still remains unclear how late learners' potentially varied proficiency is related (or unrelated) to their varied ages of arrival in early, mid and late adulthood.

Many reviews of the theoretical and methodological standards in age-related SLA research (e.g., Birdsong, 2005; DeKeyser & Larson-Hall, 2005) emphasize that the AOA-proficiency function needs to be examined in the context of experienced L2 learners with plateaued and asymptotic proficiency. Such learners must have gone through an extensive amount of practice and immersion in the L2 on a daily basis, and manifest positive integrative and/or instrumental motivation towards using the target language (see also Derwing & Munro, 2013).

To further advance this line of L2 speech research, unlike the precursor study, which sought to highlight interlanguage development (LOR = 0 to 6 years), the current investigation is exclusively concerned with the end state of L2 speech production attainment (LOR = 6 to 42 years), focusing on a relatively large number of late bilinguals (88 experienced Japanese learners of English). The study aimed to examine the extent to which learners' age of acquisition (AOA = 16 to 40) can be predictive of L2 oral proficiency in conjunction with different learning goals (comprehensibility, accentedness) and various linguistic domains (segmentals, suprasegmentals, fluency, vocabulary, grammar).

Method

Talkers

Experienced Japanese learners. The project was widely advertised on regional community websites and local newspapers in both Montreal and Vancouver, Canada, where the number of Japanese immigrants is relatively low (e.g., 0.06% in Quebec and 1% in British

Columbia) (Statistics Canada, 2008). Originally, 108 Japanese learners of English were identified as late bilinguals who had already reached their plateau (i.e., little room for further L2 development) in line with two necessary conditions in the previous literature: (a) age of arrival in Canada beyond 16 years; and (b) six years of LOR (for similar definitions of late bilinguals, see; Birdsong & Molis, 2001; Johnson & Newport, 1989).

To narrow down the sample to only those who used the L2 on a daily basis with ample opportunities for practice, 88 participants (13 males and 75 females) were carefully selected based on their language background questionnaires and interview data during the testing session according to the following criteria: (a) their self-reported use of English was above "4" (on a 6 point scale: 1. Very infrequent – 6. Very frequent) (M = 5.4); and (b) their primary language of communication either at home or work was English.[3] At the time of testing, the mean age of the 88 participants was 45.9 years (ranging from 30 to 70 years); their mean age of arrival in Canada was 26.1 years (ranging from 16 to 40 years); and their mean length of residence was 17.8 years (ranging from 6 to 42 years). They reported six to nine years of English learning experience (typically through grammar translation methods) in secondary school settings in Japan prior to their arrival in Canada.[4] Although most of the participants had little knowledge of French, 12 participants (8 from Montreal, 4 from Vancouver) reported having limited exposure to French.

Japanese and English baseline. To establish baseline speech data for the experienced Japanese learners, two groups of native speakers of Japanese and English were recruited. For the Japanese Baseline, 10 native speakers of Japanese (2 males and 8 females) who had just arrived in Canada with little L2 experience (LOR < 1 month), were recruited at private language schools in downtown Montreal (M_{age} = 17.9 years). For the English Baseline, 10 native speakers of north-eastern Canadian and American English (5 males, 5 females) who were undergraduate

students at an English-speaking university in Montreal were recruited (M_{age} = 25.1 years). Preliminary analysis regarding the effects of age on their English /r/ production was reported in Saito (2013). In the current study, the overall linguistic qualities of the same dataset were re-analyzed from not only segmental, but also prosodic, temporal, lexical and grammatical perspectives.

Speaking Task

For the sake of easy comparison, the same speaking task in Author's (Saito, in press) precursor study (i.e., timed picture description) was used to elicit the participants' spontaneous production. Following L2 research standards, spontaneous production was defined as free speech that L2 learners produced in order to convey their intended message (Spada & Tomita, 2010) under communicative pressure, without much room for conscious monitoring (Ellis, 2005). Picture narrative tasks, wherein participants describe one particular drawing (e.g., Munro & Mann, 2005) or several pictures in a sequence (e.g., Derwing & Munro, 2013)[5], are some of the most commonly used tasks in L2 pronunciation research. However, these tasks have been identified as relatively demanding compared to other spontaneous speaking tasks, such as monologues and interactive interviews (Derwing et al., 2004). In the present study, in order to elicit a certain length of spontaneous (rather than controlled) speech production without too many long filled and unfilled pauses, a picture narrative task was slightly modified in the following manner: (a) The participants described seven different pictures using three key words below each picture as hints; (b) the first four pictures were used as practice for participants to get used to the task procedure; (c) the last three pictures (Pictures A, B, C, see below) were used for the final analysis; and (d) five seconds of planning time was given before participants started describing each picture under some communicative pressure.

Pictures A, B, and C respectively 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). One critique of any spontaneous speech task of this kind is that learners can avoid difficult pronunciation features through careful word choice and syntactic errors more salient to native speakers' accentedness judgement (Munro & Mann, 2005). Special attention was given, therefore, to selecting keywords which would elicit segmental, prosodic and syllabic structures especially difficult for Japanese learner 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 (i.e., Katakana) by inserting epenthesis vowels between consecutive consonants (/dəraıvə/ for "drive," /θəri/ for "three," /səkaı/ for "sky") and after word-final consonants (/teıbələ/ for "table," /myuzıkə/ for "music").

All speech recordings were carried out individually in quiet rooms in university labs, community centers, or participants' homes in Montreal or Vancouver, using a digital Roland-05 audio recorder (44.1 kHz sampling rate with 16-bit quantization). All instructions were delivered in Japanese by the researcher (a native speaker of Japanese) to ensure that all speakers understood the procedure. The speakers first described four pictures randomly presented as distracters, and then described the remaining three pictures randomly for the main analysis. In total, the participants generated 324 tokens (108 Japanese and baseline talkers × 3 pictures). Approximately 10 seconds of the beginning of each picture description (M = 8.5 sec ranging from 4.0 to 12.5 sec) were extracted for each participant. Since three pictures were described, an average of 25.7 seconds (ranging 14.4 from 34.0 sec) of free speech samples were generated by

each participant for subsequent global, pronunciation, vocabulary, and grammar judgments. The length of the entire sample for each participant can be considered suitable compared to similar L2 speech research (e.g., Derwing & Munro, 2013 for 30 sec; Hopp & Schmid, 2013 for 10-20 sec).

Global Analyses

Raters. To judge the global qualities (accentedness and comprehensibility) of the spontaneous speech samples, 10 native speakers of English (5 males, 5 females) were recruited at an English-speaking university in Vancouver. As operationalized in previous L2 speech research (e.g., Derwing & Munro, 2009), the judgement of accentedness and comprehensibility by definition refers to *naïve* raters' intuitive impressions about L2 speech production, without relying on any pre-existing descriptors nor background knowledge. Thus, efforts were made to carefully select raters based on a lack of familiarity and contact with Japanese learners of English.

All of the raters in the study were undergraduate students with a mean age of 28.3 years. They majored in non-linguistic disciplines (e.g., business, psychology) and reported little familiarity and contact with Japanese-accented English (M = 1.3 from 1 = *Not at all* to 6 = *Very much*). According to the definition laid out in Isaacs and Thomson (2013), these raters can be considered as inexperienced. All raters reported having normal hearing.

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). After familiarizing themselves with the picture prompts and key words, they practiced the judgement procedure in a quiet room by evaluating five preliminary speech samples (not included in the

main analysis) presented via the speech analysis software, *Praat* (Boersma & Weenink, 2012) based on a 9-point scale for accentedness (1 = *little accent*, 9 = *heavily accented*) and comprehensibility (1 = *easy to understand*, 9 = *hard to understand*). Afterwards, they randomly heard and rated each of the 324 picture descriptions in a randomized order. Each picture description was played only once on the assumption that accentedness and comprehensibility corresponds to listeners' initial intuitions and impressions about L2 speech. They were explicitly told that the dataset represented a range of ability levels, from nativelike speakers to beginners, and were asked to use the entire scale. Since the entire session took approximately three hours, they took a 10 minute break halfway through to avoid listener fatigue.

Pronunciation, Fluency, Vocabulary and Grammar Analyses

In the literature, L2 speaking proficiency has been not only conceptualized based on global language ratings of accentedness and comprehensibility (e.g., Derwing & Munro, 2009), but also characterized as a composite phenomenon constituting various linguistic domains, spanning pronunciation, fluency, vocabulary and grammar (De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012). Whereas these subdomains of L2 speech have been traditionally analyzed via a set of objective instruments at a fine-grained level (e.g., Isaacs & Trofimovich, 2012), 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 (Bosker, Pinget, Quené, Sanders, de Jong, 2013; Derwing, Rossiter, Munro, & Thomson, 2004) and lexical accuracy, density, diversity and sophistication (Crossley, Salsbury, & McNamara, 2014).

Following this latter line of L2 assessment research, the current study adopted the human rater method, whereby experienced raters with linguistic and pedagogical backgrounds analyzed

specific areas of language (i.e., pronunciation, fluency, vocabulary, and grammar) in conjunction with the eight categories developed and validated in Saito, Trofimovich and Isaacs (in press-a). These categories included the linguistic dimensions of pronunciation (segmentals, word stress, intonation), fluency (speech rate), vocabulary (appropriateness, richness) and grammar (accuracy, complexity)[6].

Raters. Unlike accentedness and comprehensibility, which allows for the use of inexperienced raters' intuitive judgements, raters for the pronunciation, fluency, vocabulary and grammar analyses were expected to have a great deal of relevant experience and knowledge to make reliable, accurate, and consistent judgements of the multiple linguistic aspects of L2 speech production (Saito et al., in press-a). In this regard, five experienced raters (2 males, 3 females), who were graduate students in Applied Linguistics at an English-speaking university in Montreal, were carefully selected (Isaacs & Thomson, 2013). Their mean age was 29.4 years and reported (a) previous teaching experience in various ESL and EFL settings (M = 4.0 years of teaching); (b) previous training experience specific to pronunciation, fluency, vocabulary and grammar analyses and teaching; and (c) varied familiarity with Japanese accented English (M = 3.4 from 1 = *Not at all* to 6 = *Very much*).

Audio measures. The three picture descriptions were combined and stored in a single WAV file for each talker in order to provide the raters with sufficient phonological information for their judgements. The raters listened to each sample (with an option to repeat until they felt satisfied) delivered via the MATLAB software, and then used a free moving slider on a computer screen to assess the four phonological and temporal categories of the tokens: (a) segmentals (substitution, omission, or insertion of individual consonants or vowels); (b) word stress

(misplaced or missing primary stress); (c) intonation (appropriate, varied versus incorrect and monotonous use of pitch); and (d) speech rate (speed of utterance delivery).

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, labelled with a smiling face (indicating very positive), it was recorded as “1000”. The slider was initially placed in the middle of each scale. The raters were told that even a small movement of the slider may represent a fairly large difference in the rating score. Except for the frowning and smiling faces and accompanying brief verbal descriptions for the endpoints of each pronunciation and fluency category, the scale included no numerical labels or marked intervals of any kind (see the Appendix for examples of the onscreen labels).

Transcript measures. The three picture descriptions were transcribed and stored in a single text file for each talker ($M_{\text{no. of words}} = 42.5$ words per talker). The raters read the written files to conduct the lexicogrammatical analysis without being distracted by pronunciation errors (Crossley et al., 2014; Patkowski, 1980). To this end, all speech tokens were first transcribed by a trained research assistant, and then 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) (Lu, 2012).

The final written transcripts were presented via the MATLAB software in a random order. The raters read three short paragraphs within each transcript displayed on the screen in the same order as they were heard, and used similar free moving sliders to assess four lexicogrammatical

categories: (a) lexical appropriateness (accuracy of vocabulary); 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).

Training and rating sessions. The entire pronunciation, fluency, vocabulary and grammar analyses sessions took place over three days, with Day 1 for the training phase (2 hours), Day 2 for the audio rating phase (2 hours), and Day 3 for the transcript rating phase (1 hour).

Day 1. The raters first received thorough instructions from a trained research assistant on the eight different elements of pronunciation, fluency, and lexicogrammar (see Appendix for training scripts). They then proceeded to practice the judgement procedure in a quiet room by evaluating a total of 40 non-native speakers' picture narratives, first for audio-based measures (segmentals, word stress, intonation, speech rate), and, second for transcript-based measures (lexical appropriateness, lexical richness, grammatical accuracy, grammatical complexity).

As reported in Saito et al. (in press-a) and summarized in Table 1, these raters' ratings were significantly correlated with key linguistic properties of the tokens, which as a result confirmed the accuracy and reliability of the participating human raters' abilities to analyze the phonological, temporal, lexical and grammatical qualities of L2 speech production.

TABLE 1 HERE

Day 2. After receiving recapped instructions on the four pronunciation and fluency categories and familiarizing themselves with the picture prompts and key words for the current dataset, the raters first practiced rating five audio picture descriptions of Japanese learners (not

included in the main analysis). For each practice sample, the raters explained their decisions and received feedback on the accuracy of their understanding of the categories. Subsequently, the raters proceeded to rate the main dataset of 108 audio samples with a 10-minute intermission halfway through.

Day 3. The raters first received recapped instruction on the four lexicogrammar categories and feedback on their practice ratings of the same three tokens (not included in the main dataset). Subsequently, they rated the 108 transcript samples.

Inter-rater Reliability

Cronbach's alpha was calculated to check inter-rater agreement among the 10 inexperienced raters' global scores of 324 samples (108 talkers × 3 picture descriptions) and the five experienced raters' audio and transcript ratings of 108 speech samples (3 picture descriptions combined per talker). In line with previous L2 comprehensibility research (e.g., Derwing & Munro, 2009), the results found relatively high alpha levels for accentedness ($\alpha = .97$) and comprehensibility ($\alpha = .95$). In terms of the pronunciation analyses, the raters' judgements were overall consistent, demonstrating high reliability indexes (Cronbach's alpha) for segmentals ($\alpha = .90$), word stress ($\alpha = .87$), intonation ($\alpha = .82$), and speech rate ($\alpha = .88$). The raters showed slightly less agreement for their analyses of lexical appropriateness ($\alpha = .75$), lexical richness ($\alpha = .84$), grammatical accuracy ($\alpha = .80$), and grammatical complexity ($\alpha = .77$). The reliability indexes were overall acceptable, exceeding the benchmark value of .70-.80 in L2 research (Larson-Hall, 2010). By averaging across all listeners' ratings, one mean score was computed for each speaker according to global (accentedness, comprehensibility), phonological (segmentals, word stress, intonation), temporal (speech rate), lexical (appropriateness, richness) and grammatical (accuracy, complexity) categories, respectively.

Interrelationships between Linguistic Scores

To investigate the degree of independence between audio and transcript ratings, a set of simple correlation analyses was performed, respectively (see Table 2). Different strength of correlation coefficients were also checked using the Fisher r-to-z transformation (p = .008, Bonferroni corrected). For audio-based measures, segmental scores were more closely related to prosodic scores (r = .94 for word stress; .84 for intonation) than fluency scores (r = .75 for speech rate), $p < .001$. Prosodic scores were similarly correlated to fluency scores (r = .84 for word stress; r = .83) ($p > .08$). For transcript-based measures, relatively strong correlations were found between appropriateness and accuracy (r = .71) as well as richness and complexity (r = .84). To summarize, as conceptualized and validated in the previous study (Saito et al., in press-a), the eight rater-based linguistic categories were considered to tap into four domains of L2 speaking proficiency—pronunciation (segmentals, word stress, intonation), fluency (word stress, intonation, speech rate), the proper (appropriateness, accuracy) and sophisticated (richness, complexity) usage of lexicogrammar.

TABLE 2 HERE

Results

Linguistic Characteristics of L2 Oral Proficiency Attainment

The first aim of the statistical analysis was to investigate the global, phonological, temporal, and lexicogrammatical qualities of experienced Japanese learners' oral proficiency attainment relative to the performance of two baseline groups of Japanese (LOR < 1 month) and English native speakers.

Global analyses. Participants' accentedness and comprehensibility scores are summarized in Table 3. These were used as dependent variables and submitted to a two-way ANOVA with Group ("experienced learners," Japanese baseline," "English baseline") as a between-subject factor and Domain (accentedness, comprehensibility) as a within-subject factor. The results found a significant Group × Domain interaction effect, $F(2, 105) = 18.860$, $p < 0.001$. According to Bonferroni multiple comparisons, the experienced Japanese learners' comprehensibility scores were rated higher than their accentedness scores. The experienced Japanese learners also significantly outperformed the Japanese baseline ($p < .001$), but performed more poorly compared to the English baseline ($p < .001$) in terms of both sets of scores.

Pronunciation, fluency, vocabulary and grammar analyses. Participants' linguistic scores are summarized in Table 3. A two-way ANOVA was conducted using participants' audio (segmentals, word stress, intonation, speech rate) and transcript (lexical appropriateness and richness, grammatical accuracy and complexity) rating scores as the dependent variables. The results yielded a significant interaction effect for Group and Domain for the audio measures, $F(6, 315) = 4.785$, $p < .001$, and the transcript measures, $F(6, 315) = 5.209$, $p < .001$. Bonferroni pairwise comparisons showed that, for all linguistic domains (pronunciation, fluency, vocabulary, grammar), the experienced Japanese learners showed better performance than the Japanese baseline ($p < .001$), but still differed significantly from the English baseline ($p < .001$).

TABLE 3 HERE

Age Effects on Attained L2 Oral Proficiency

The second aim of the statistical analysis was to examine whether and to what degree the 88 experienced Japanese learners' AOA was predictive of their attained oral ability via a set of simple and partial correlation analyses.

Global analyses. To check the normality of the dataset for subsequent correlation analyses, participants' accentedness and comprehensibility scores were submitted to Grubb's tests, identifying no outliers in both domains ($p > .05$). The simple correlation between the global rating score and AOA was significant for accentedness, r (87) = .346, p = .001, and comprehensibility, r (87) = .429, $p < .001$. A scatterplot for the AOA-proficiency relationship is presented in Figure 1.

FIGURE 1 HERE

As many researchers have pointed out, AOA effects are likely confounded with learners' LOR (i.e., the earlier they arrive in an L2 country, the longer they stay) (e.g., Flege et al., 1995); the two variables were indeed significantly correlated in the current study, r (87) = -.315, p = .003. To this end, partial correlation analyses were conducted to examine the relationship between AOA, and accentedness and comprehensibility, when the other confounding variable (i.e., LOR) was controlled. With LOR factored out, the AOA-proficiency relationship still remained significant for accentedness, r (85) = .315, p = .003, and comprehensibility, r (85) = .412, $p < .001$.[7]

Pronunciation, fluency, vocabulary and grammar analyses. According to Grubb's tests, one outlier was found in the context of lexical appropriateness ($z = 3.88$, $p > .05$); this participant's score was eliminated for the relevant analyses. As for the pronunciation and fluency

analyses, results of simple correlation between the four audio rating scores and AOA was significant for segmentals, r (87) = -.299, p = .005, word stress, r (87) = -.309, p = .003, and intonation, r (87) = -.235, p = .027. Yet, it did not reach statistical significance for speech rate, r (87) = -.175, p = .102. As for the vocabulary and grammar analyses, AOA was not significantly correlated with any of the transcript rating scores, such as lexical appropriateness, r (86) = -.160, p = .137, lexical richness, r (87) = .122, p = .257, grammatical accuracy, r (87) = -.051, p = .635, and grammatical complexity, r (87) = .084, p = .436. Scatterplots for the relationship between AOA and linguistic proficiency are presented in Figures 2 and 3.

FIGURE 2 HERE

FIGURE 3 HERE

Partial correlation analyses were also performed to illustrate the impact of AOA on the learners' pronunciation, fluency, and lexicogrammar performance by separating any other experience-related factors (i.e., LOR effects) from the age function. After the variable of the learners' LOR profiles were removed, the AOA-proficiency link remained significant for segmentals, r (85) = -.240, p = .025, and word stress, r (85) = .260, p = .015, but became marginal for intonation, r (85) = -.210, p = .050, and non-significant for speech rate, r (85) = -.184, p = .087. As for lexicogrammar, the partial correlation analyses still failed to find any significant power of AOA for lexical appropriateness, r (84) = -.143, p = .188, lexical richness, r

(85) = .182, p = .091, grammatical accuracy, r (85) = -.044, p = .684, and grammatical complexity, r (85) = .113, p = .299.

Discussion

In the context of late Japanese-English bilinguals (AOA > 16 years), the current study aimed to examine whether and to what degree age of acquisition can predict their post-pubertal L2 oral proficiency attainment after years of input and interaction with native and non-native speakers through extensive residence in the L2 environment (LOR > 6 years). Overall, the results provide three broad findings. First, the experienced Japanese learners' L2 oral ability demonstrated significantly better global, phonological, temporal, lexical, and grammatical qualities compared to that of inexperienced Japanese learners (LOR < 1 month), although their performance was substantially different from that of native speakers of English. Second, AOA was significantly predictive of the late learners' global L2 oral ability (accentedness and comprehensibility), arguably owing to relatively strong age effects on segmental and prosodic attainment. Third, AOA did not relate to the temporal and lexicogrammatical domains of attained L2 speech production.

By and large, these results do not provide the necessary support for the predictions of the strong version of the CPH, which explicitly hypothesizes the absence of age effects on any linguistic areas of late bilingualism (due to the close of a critical period). Rather, the data can be well explained in support of the predictions of the CAH, which assumes the existence of language-specific cognition across the lifespan. That is, both young and adult L2 learners alike successfully and continuously enhance their L2 oral ability, given ample opportunities and high motivation to use the L2 (Flege & Liu, 2001); and the final state quality of L2 learners' near-nativelike performance is equally subject to age effects before and after puberty (Birdsong &

Molis, 2001). For the latter point, the correlation coefficients on the AOA-proficiency relationship among the late learners in the study (r = .346 for accentedness, r = .429 for comprehensibility) are somewhat comparable to those of early learners (AOA < 16 years) in the previous literature (e.g., r = .360-.560 for Granena & Long, 2013).

Noteworthy, however, is that the age factor differentially, not monolithically, predicted late L2 speech production attainment. The pronunciation, fluency, vocabulary and grammar analyses demonstrated that the predictive power of AOA was strong, especially in the phonological (segmental and prosodic) domain of language rather than the temporal and lexicogrammatical domains. Such complex results lead us to consider several possible accounts for the multifaceted nature of age effects on late SLA. One relevant discussion involves the recently proposed process-oriented model for L2 speech production development (e.g., Isaacs & Trofimovich, 2012). According to the model, native speakers draw on different realms of linguistic information (pronunciation, fluency, vocabulary, and grammar) when they perceive the L2 oral proficiency of beginner, intermediate, and advanced level learners. For example, Derwing and Munro showed that whereas good prosody (intonation) was invariably related to native speakers' evaluation of all groups of learners, they likely prioritized temporal over segmental information to make proficiency judgments for inexperienced learners (Derwing & Munro, 1997) and vice versa for experienced learners (Munro & Derwing, 1995) (see also Saito, Trofimovich, & Isaacs, in press-b). Similarly, Isaacs and Trofimovich (2012) found that prosodic qualities (word stress) equally predicted beginner, intermediate and advanced levels of L2 comprehensibility. In contrast, temporal qualities (mean length of run) only distinguished between beginner and intermediate level learners.

The relative weights of the linguistic influences on native speakers' assessment patterns shed some light on how L2 learners enhance their rate and attainment of L2 speech production as they increase their L2 experience at the initial, mid and final phases of SLA. That is, the continuous development of optimal speech rate, proper lexicogrammar, and good prosody is characteristic of the initial to mid phases of L2 speech learning, and refined segmental and prosodic performance is representative of the mid to final phase of L2 speech learning (Saito, in press). Situated within this developmental framework, the results of the experienced Japanese learners (LOR > 6 years) suggest that AOA can be a good predictor of segmental and prosodic attainment, arguably because it indicates to what degree these experienced learners can attain advanced levels of L2 oral proficiency via improving the phonological domain of language at the later stage of L2 speech learning. Conversely, given that the development of fluency and lexicogrammar is a crucial part of the initial to mid (but not final) stages of L2 speech learning, obtaining optimal speech rate and proper lexicogrammar usage can be achieved by virtue of being extensively exposed to L2 input despite different timings of AOA (Saito, in press; Trofimovich & Baker, 2006)

Such differential effects of age on pronunciation, fluency, vocabulary and grammar attainment well reflects on the continuum of easy, moderate, and difficult linguistic features which has been suggested by the extensive nativelikeness research in SLA. Whereas even late L2 learners likely attain some aspects of nativelike vocabulary performance (e.g., Hellman, 2011 for vocabulary size), the attainment of such high proficiency tends to occur very infrequently in grammar (Flege et al., 1999) and entail an extensive amount of L2 experience in speech and articulation rate, rhythm, and the number of pauses (Munro & Derwing, 2014). Furthermore, the incidence of nativelikness itself is found to be extremely rare compared to segmental

(Abrahamsson, 2012) and prosodic accuracy (Trofimovich & Baker, 2006), and vocabulary richness and grammatical complexity (Abrahamsson & Hyltenstam, 2009). As shown in the current study, the Japanese learners likely reached their upper limits of the proper lexicogrammar usage and optimal fluency as long as they had an adequate amount of L2 experience through at least six year of L2 immersion; but it may require not only the extensive LOR but also the early AOA profiles for Japanese learners to further enhance the prosodic and segmental qualities of L2 speech production to attain advanced, sophisticated, and near-nativelike proficiency.

Limitations

Given the exploratory nature of the project, several methodological limitations need to be acknowledged with a view of future replication studies. First and foremost, the findings, especially those regarding the role of AOA in lexicogrammar attainment, should be interpreted with caution due to an obvious methodological problem inherent in the study's instruments. That is, only 30 seconds of the participants' spontaneous speech production was used for the raters' transcript-based judgements. Although the length of speech samples (30 sec) can be considered to have provided sufficient phonological information for judgement (e.g., Derwing & Munro, 1997; Hopp & Schmid, 2013; Isaacs & Trofimovich, 2012), it may have failed to provide sufficient written data for even trained raters to analyze the detailed relationship between AOA, and vocabulary and grammar performance.

For example, although our dataset constituted an average of approximately 50 words per talker, the robust analyses of certain lexical measures (e.g., lexical richness) may require more than 100 words (Koizumi & In'nami, 2012). Previous research indeed has used longer speech samples for vocabulary and grammar analyses (3 min for Lu, 2012; 5 min for Yuan & Ellis, 2003). In addition, the four categories (lexical appropriateness and richness, grammatical

accuracy and complexity) in the study may not be comprehensive enough to capture the numerous layers of participants' vocabulary and grammar performance (e.g., see Lu, 2012 for different correlation coefficients between 20 lexical richness measures and L2 oral proficiency).

Another important issue concerns the type of speaking task. It is crucial to reiterate that the tentative suggestions on AOA effects in the study were solely based on the timed picture description task; the nature of the task (i.e., describing each picture with three key words and five seconds of pre-planning time) may have failed in eliciting a sufficiently wide range of various lexical items. One could argue that the predictive power of the lexicogrammar factors did not reach statistical significance in any contexts, probably because all participants were allowed to simply use similar kinds of frequent and familiar lexical items. In fact, Crowther, Trofimovich, Isaacs and Saito (in press) showed that native speaking raters tended to attend to pronunciation *and* lexicogrammar factors only when L2 speech was elicited via a relatively difficult speaking test (i.e., the TOEFL iBT integrated task). In contrast, they relied exclusively on the pronunciation factor when L2 speech was elicited based on a relatively easy task (i.e., the IELTS long-turn task) (see also Derwing et al., 2004 for similar task effects and L2 speaking proficiency). The generalizability of the results (especially related to lexicogrammatical attainment) in the study need to be tested within the context of various speaking tasks, especially more argumentative, formal and complex ones whereby L2 learners are induced to demonstrate a more varied and sophisticated use of L2 vocabulary (see Hulstijn, Schoonen, de Jong, Steinel, & Florijn, 2012).

Finally, future research is warranted to scrutinize the direct causes of the AOA-proficiency correlation in late bilingualism such as cross-linguistic influence (Hopp & Schmid, 2013) and/or the cognitive aging factor (Birdsong, 2005, 2006). One potential way to address

this is to investigate whether learners' AOA is related to not only the quality of L2, but also L1 speech production. If we accept the view that extensive L2 use negatively affects L1 performance, it seems reasonable to assume that earlier AOA profiles can equally predict not only better L2 performance, but also more L1 attrition. In terms of the influence of the cognitive aging factor on the age function in late SLA, participants' levels of cognitive and neurobiological development can be first measured via instruments previously used and validated in the cognitive psychology literature (e.g., Simon task: Bialystok, Craik, Klein, & Viswanathan, 2004). Subsequently, it would be intriguing to examine how the various aging conditions of early and late arrivals can be differentially related to their attained L2 proficiency. Such future research will in turn highlight the intricate relationship between learners' AOA, the magnitude of intrinsic L1/L2 interaction, the state of neurological and cognitive development, and various linguistic elements of L2 speech production.

Conclusion

The current study was designed to investigate the predictive power of AOA for the global, phonological, temporal, lexical and grammatical qualities of post-pubertal L2 speech production attainment by late experienced Japanese learners. According to the results of the global analyses, AOA was negatively correlated with accentedness and comprehensibility in L2 speech production, suggesting that the aging factor remains pertinent to not only early but also late SLA throughout the age spectrum (Birdsong, 2005, 2006). It is important to reiterate here that one potential reason for the significant age function in the study can be attributed to the fact that participants who had many opportunities to process input and interaction in the L2 with native and non-native speakers for many years (LOR > 6 years) were carefully chosen (Derwing &

Munro, 2013), and their performance was measured at a spontaneous (rather than controlled) speech level (Hopp & Schmid, 2013).

Additionally, the results of the pronunciation, fluency, vocabulary and grammar analyses revealed that such post-pubertal age effects were particularly strong in the case of the segmental and prosodic attainment, which is a crucial linguistic characteristic of advanced level L2 oral proficiency (Isaacs & Trofimovich, 2012; Saito et al,, in press-b). In contrast, AOA played a negligible role in temporal and lexicogrammar attainment, probably because most L2 learners have already passed the certain threshold needed for successful communication as a function of LOR instead of AOA profile (Saito, in press; Trofimovich & Baker, 2006).

Extending previous L2 speech studies of this kind (Saito, in press; Derwing & Munro, 2013), the current study leads to three tentative conclusions about the underlying mechanism for late SLA. First, regular and motivated L2 users are assumed to make steady improvement in the temporal and lexicogrammatical domains of language (optimal fluency, good prosody, appropriate vocabulary, accurate grammar) over an extensive period of stay in the L2 environment ($0 < LOR < 5$ years) for the purpose of successful L2 social interaction (Saito, in press; Derwing & Munro, 2013). In the long run, their age of acquisition seems to be an important index for determining the extent to which they can attain advanced-level L2 oral proficiency, specifically via improving the phonological domain of language (correct consonant and vowel pronunciation, good prosody) (Saito, 2013; Trofimovich & Baker, 2006).

The first two conclusions motivate us to propose the last conclusion: that even adult L2 learners may draw on qualitatively and fundamentally similar language learning mechanisms used for early SLA and L1 acquisition with a lifelong gradual negative change in their L2 attainment with increasing age (Best & Tyler, 2007; Bialystok, 1997; Birdsong, 2005, 2006;

Flege, 2003; Hakuta et al., 2003; Hopp & Schmid, 2013). To obtain a better understanding of the plasticity for language learning in late SLA (i.e., similarities and dissimilarities in L1 and L2 acquisition), we call for more research that highlights age effects on both the L1 and L2 performance of early and late bilinguals. Such research should adopt more comprehensive measures of not only pronunciation and fluency, but also lexicogrammar performance in the context of a range of speaking tasks requiring different lexicogrammatical thresholds (e.g., TOEFL iBT, IELTS). Since this study was based exclusively on Japanese learners of English, the generalizability of the results can be tested in conjunction with late bilinguals with different L1-L2 backgrounds.

References

Abrahamsson, N. (2012). Age of onset and nativelike L2 ultimate attainment of morphosyntactic and phonetic intuition. *Studies in Second Language Acquisition, 34*, 187-214.

Abrahamsson, N. & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition, 30*, 481–509.

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

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Baker, W., Trofimovich, P., Flege, J. E., Mack, M., & Halter, R. (2008). Child-adult differences in second-language phonological learning: The role of cross-language similarity. *Language and Speech*, *51,* 316-341.

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.

Bialystok, E., Craik, F. I., Klein, R., & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: Evidence from the Simon task. *Psychology and aging, 19*, 290-303.

Bialystok, E., & Miller, B. (1999). The problem of age in second-language acquisition: Influences from language, structure, and task. *Bilingualism: Language and cognition*, *2*, 127-145.

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.

Birdsong, D. (2006). Age and second language acquisition and processing: A selective overview. *Language Learning*, *56*, 9-49.

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

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

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. (in press). Does 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., Alfi-Shabta, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics, 31*, 413-438.

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.

Dörnyei, Z., & Kubanyiova, M. (2014). *Motivating learners, motivating teachers: Building vision in the language classroom*. Cambridge: Cambridge University Press.

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. (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., Birdsong, D., Bialystok, E., Mack, M., Sung, H. & Tsukada, K. (2006). Degree of foreign accent in English sentences produced by Korean children and adults. *Journal of Phonetics, 34*, 153-175.

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. & 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.

Flege, J., Yeni-Komshian, G., & Liu, S. (1999). Age constraints on second language acquisition. *Journal of Memory & Language, 41*, 78–104.

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.

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

Granena, G., & 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.

Hellman, A. B. (2011). Vocabulary size and depth of word knowledge in adult-onset second language acquisition. *International Journal of Applied Linguistics*, *21*, 162-182.

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.

Hulstijn, J.H., Schoonen, R., De Jong, N.H., Steinel, M.P, & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of

speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing, 29*, 203-221.

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

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

Johnson, J. & Newport E (1989). Critical period effects in second language learning: the influence of maturational state on the acquisition of ESL. *Cognitive Psychology, 21,* 60-99.

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System, 40*, 554-564.

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science, 10*, 110-120.

Larson-Hall, J. (2006). What does more time buy you? Another look at the effects of long-term residence on production accuracy of English /r/ and /l/ by Japanese speakers. *Language and Speech, 49*, 521-548.

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

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. *The Modern Language Review, 96*, 190-208.

Mackey, A. (Ed.) (2007). *Conversational interaction in SLA: A collection of empirical studies*. New York: Oxford University Press.

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

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.

Meara, P. (2005). *LLAMA Language Aptitude Tests*. Swansea: Lognostics.

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

Munro, M., & Derwing, T. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*, 73-97.

Munro, M., & Derwing, T. (2014, March). *How different can you get? Ten-year learning trajectories for two L2 groups.* Paper presented at the annual conference of the American Association for Applied Linguistics, Portland, OR.

Patkowski, M. (1980). The sensitive period for the acquisition of syntax in a second language. *Language Learning, 30*, 449-472.

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

Pavlenko, A., & Blackledge, A. (Eds.). (2004). *Negotiation of identities in multilingual contexts*. Clevedon: Multilingual Matters.

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

Pulvermüller, F., & Schumann, J. H. (1994). Neurobiological mechanisms of language acquisition. *Language Learning*, *44*, 681–734.

Saito, K. (2013). Age effects on late bilingualism: The production development of /r/ by high-proficiency Japanese learners of English. *Journal of Memory and Language, 69*, 546-562.

Saito, K. (in press). Experience effects on the development of late second language learners' oral proficiency. *Language Learning, 65*.

Saito, K., Trofimovich, P., & Isaacs, T. (in press-a). Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics, 35*.

Saito, K., Trofimovich, P., & Isaacs, T. (in press-b). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics, 35*.

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

Stevens, G. (2006). The Age-Length-Onset problem in research on second Language Acquisition among immigrants. *Language Learning*, *56*, 671-692.

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.

Yeni-Komshian, G. H., Flege, J. E., & Liu, S. (2000). Pronunciation proficiency in the first and second languages of Korean-English bilinguals. *Bilingualism Language and Cognition*, *3*, 131-149.

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

Acknowledgement

This study was funded by the Grant-in-Aid for Scientific Research in Japan (No. 26770202). I am grateful to Pavel Trofimovich and anonymous *SSLA* reviewers for their helpful input and feedback on the content of this manuscript, and to Ze Shan Yao and George Smith who helped data analyses. I also gratefully acknowledge Midori Adachi, Yuki Matsumura, Noriko Yamane, Keiko Onishi, Yukiko Simon, and Tonarigumi for their efforts to organize the data collection for this project.

Endnote

1. The definition of AOA in the study is synonymous with age of arrival in an L2 speaking country in line with the previous age-related SLA research (e.g., Birdsong & Molis, 2001; Johnson & Newport, 1989). Although late bilinguals have typically received formal instruction before the actual date of arrival, it is highly controversial if such foreign language learning experience (e.g., a few hours per week under classroom conditions) can be considered as a part of the intensive exposure to L2 input (Best & Tyler, 2007). Different from naturalistic SLA, foreign language settings can be characterized as limited in quantity (e.g., there are few opportunities to speak L2, especially with native speakers outside of classrooms) and quality (e.g., teachers and peers have different proficiency levels) (see Muñoz, 2008).

2. The key researchers listed here share the following view: AOA can be significantly predictive of the final state of SLA across the life span because late and early bilingualism draw on the same language acquisition system. As Flege (2009) pointed out, AOA is a "macrovariable" (p. 184). In fact, the CAH researchers ascribe a wide range of different affecting variables to the relatively strong AOA effects such as environmental and experiential factors (input and interaction) (e.g., Bialystok, 1997), psycho-social factors (willingness to use and be immersed in the L2) (e.g., Derwing & Munro, 2013), degree of L1 and L2 development (e.g., Flege, 2003), reciprocal influence of the L1 and L2 (e.g., Hopp & Schmid, 2013), and declines in cognitive function associated with aging (e.g., Birdsong, 2005). In this regard, the labeling, "CAH," does not have the coherence that is typical/desirable of a hypothesis, nor the established theoretical status of the CPH. Noteworthy, however, is that despite their opinions on underlying causes of age effects, their hypotheses on the presence of the AOA-proficiency link in late SLA

stand in contrast with that of the CPH, which assumes the lack of the age function after puberty thanks to the passing of the critical period.

3. Eighteen participants who rated their frequency of English use below "3" reported their primary language communication as Japanese (n = 10) (e.g., they spoke Japanese with their family members and did not work outside) or French (n = 8) (their business involved French-speaking customers or partners were native speakers of French).

4. Among the original data pool of 108 Japanese learners, two participants reported intensive English learning experience in immersion programs in Japan. Both of them were eliminated from the final analysis, because their precise AOA profile was difficult to determine.

5. In Saito et al. (in press-a, in press-b), it was found that the first 30 sec of narratives on one picture drawing and an eight-frame cartoon provided native speaking listeners with enough linguistic information to lead to similar global, pronunciation, fluency, vocabulary and grammar judgement results.

6. 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* units, but those which were still *perceptible* to human raters.

For further examples of empirical research and discussion on more abstract (rather than broad) constructs of L2 oral proficiency, see Saito et al. (in press-a), De Jong et al. (2012), and Isaacs and Trofimovich (2012).

7. Another confounding variable for the AOA effects on late bilingualism is age at the time of testing: When participants with later AOA profiles are homogeneously older at the time of testing, it is crucial to statistically control this age-at-testing factor as a covariate because it tends to make negative and/or positive impacts on various linguistic domains of L2 performance at the time of testing (e.g., Abrahamsson, 2012; DeKeyser et al., 2010). Yet, the participants' chronological age was non-linearly related to their AOA profiles in the study, $r(87) = .091$, $p = .399$. Following Johnson's (2006) recommendation, the variable was not further analyzed in the current investigation on the AOA-proficiency link.

Table 1. *Summary of Linguistic Predictors for Human Raters' Phonological, Temporal, Lexical and Grammatical Judgement of L2 Speech in Author (Saito et al., in press-a)*

Rater judgement measures	Linguistic predictors
A. Audio measures	
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. Transcript measures	
Lexical appropriateness	No. of lexical errors
Lexical richness	Type frequency, token frequency
Grammatical accuracy	No. of grammatical errors
Grammatical complexity	Subordinate clause ratio

Table 2. *Intercorrelations between the Audio and Transcript Ratings*

A. Audio ratings			
	Word stress	Intonation	Speech rate
Segmentals	.94	.84.	.75
Word stress		.88	.84
Intonation			.83
B. Transcript ratings			
	Richness	Accuracy	Complexity
Appropriateness	.25	.71	.28
Richness		.41	.84
Accuracy			.44

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

A. Global ratings (9 points)	Japanese Learners (*n* = 88)	Japanese Baseline (*n* = 10)	English Baseline (*n* = 10)
Accentedness	5.5 (1.2)	7.7 (0.3)	1.1 (0.2)
Comprehensibility	4.3 (1.1)	6.9 (0.5)	1.2 (0.2)
B. Audio ratings (1000 points)			
Segmentals	497 (154)	267 (117)	992 (7)
Word stress	569 (123)	362 (86)	983 (30)
Intonation	491 (153)	278 (89)	865 (58)
Speech rate	616 (157)	295 (155)	978 (28)
C. Transcript ratings (1000 points)			
Lexical appropriateness	762 (100)	317 (108)	902 (48)
Lexical richness	599 (146)	300 (110)	757 (221)
Grammatical accuracy	533 (148)	300 (97)	800 (108)
Grammatical complexity	447 (155)	204 (90)	579 (112)

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

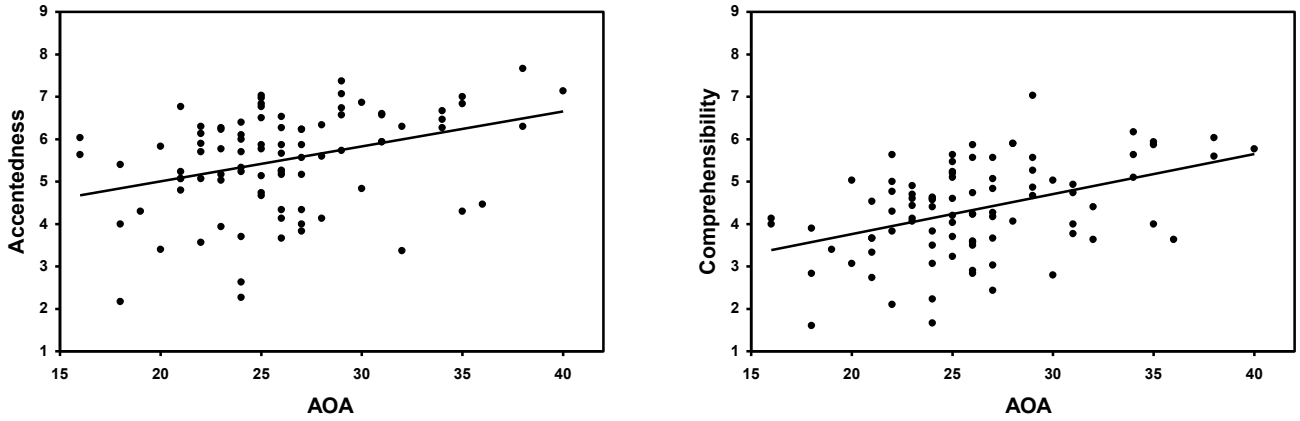

Figure 1. Global accentedness and comprehensibility scores (*1 = Little accent, easy to understand, 9 = Heavily accented, hard to understand*) plotted as a function of AOA

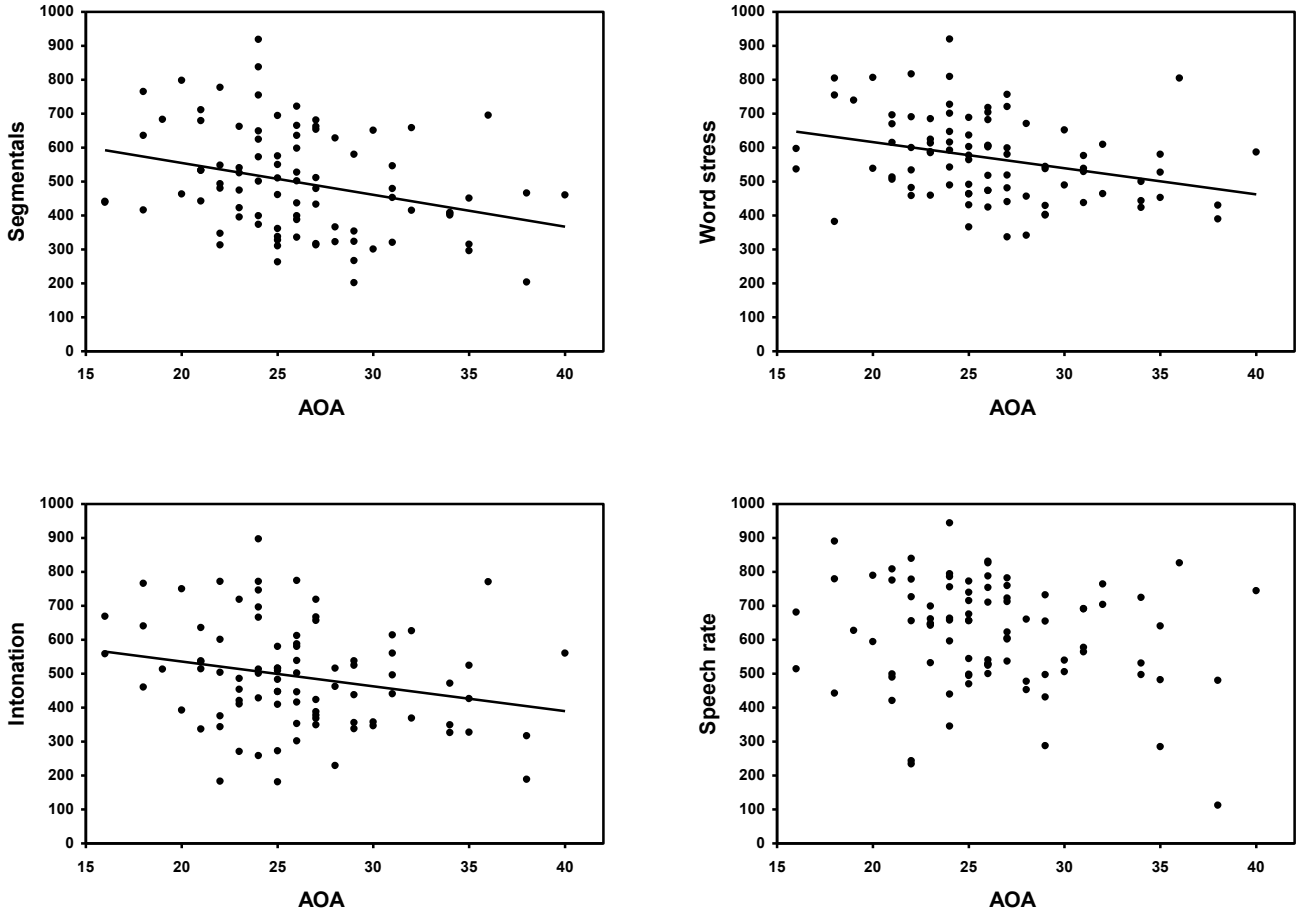

Figure 2. Pronunciation and fluency scores (*0 = Nontargetlike production, 1000 = Targetlike production*) plotted as a function of AOA

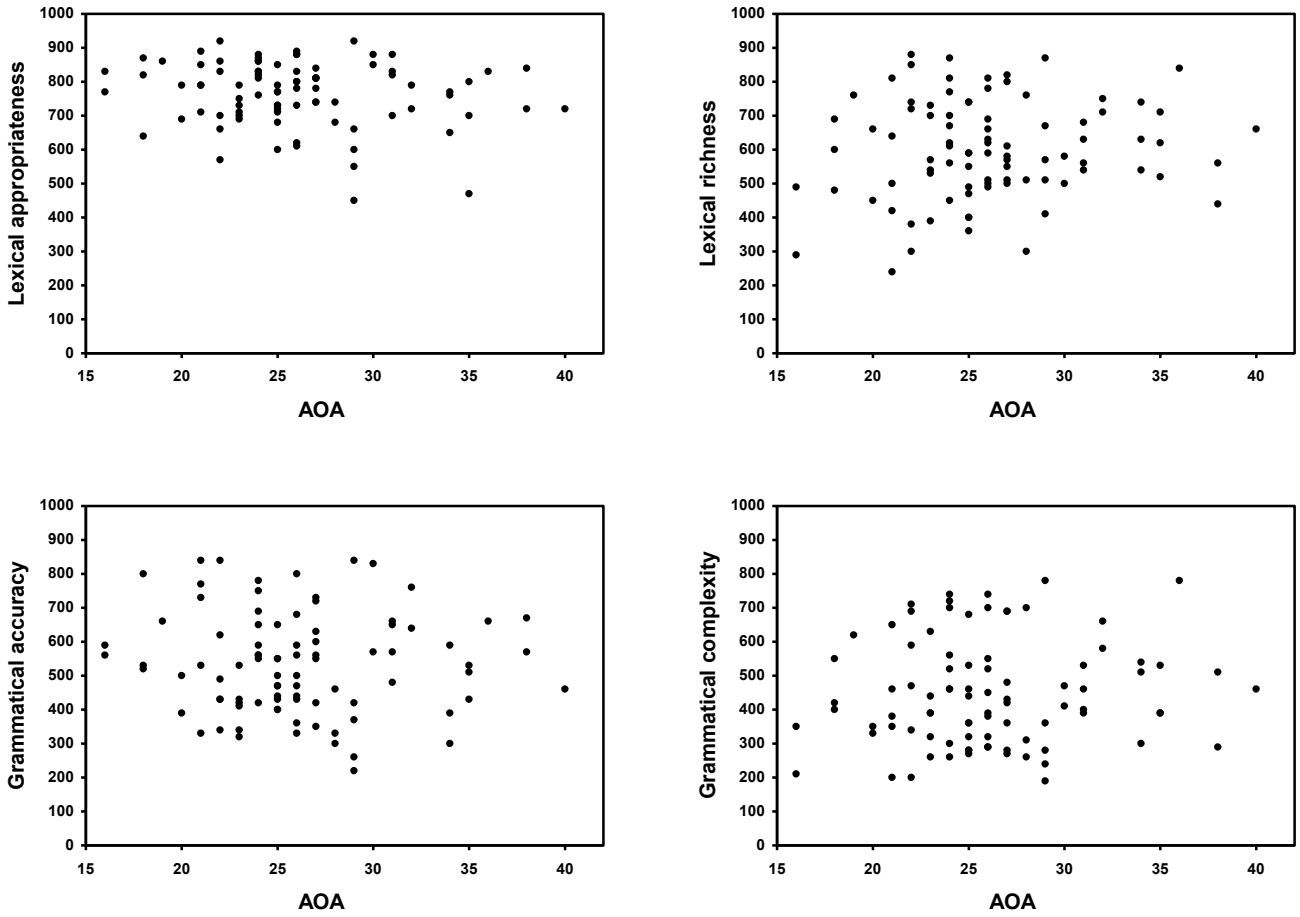

Figure 3. Vocabulary and grammar scores (*0 = Nontargetlike production, 1000 = Targetlike production*) plotted as a function of AOA

Appendix

Training materials and onscreen labels for audio- and transcript-based measures

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

1. Vowel and/or consonant errors

Frequent	**Infrequent or absent**

2. Word stress errors affecting stressed and unstressed syllables

Frequent	**Infrequent or absent**

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

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

4. Speech rate

Too slow or too fast	**Optimal**

B. Vocabulary and grammar categories	
Lexical appropriateness	This dimension refers to the appropriateness of the vocabulary words used by the speaker. If the speaker uses incorrect or inappropriate words, including words from the speaker's native language, lexical accuracy is low. On the other hand, lexical accuracy is high if the speaker has all the lexical items required to accomplish the speaking task and does so using frequently-used and/or precise lexical expressions.
Lexical richness	This dimension also refers to the vocabulary used by the speaker. What is important here, however, is how sophisticated this vocabulary is, taking into account the demands of the speaking task. If the speaker uses a few simple, unnuanced words, the speech lacks lexical richness. However, if the speaker's language is characterized by varied and sophisticated uses of English vocabulary, the speech is lexically rich.
Grammatical accuracy	This refers to the number of grammar errors that the speaker makes, including errors in word order and morphological ending.
Grammatical complexity	This dimension is about the complexity and sophistication of the speaker's grammar. If the speaker uses basic, simple or fragmented structures or sentences, grammatical complexity is low. Grammatical complexity is high if the speaker uses elaborate and sophisticated grammar structures.