Accepted for publicaion in *Studies in Second Language Acquisition* (Cambridge Univesity Press)

APTITUDE, EXPERIENCE AND SECOND LANGUAGE PRONUNCIATION PROFICIENCY DEVELOPMENT IN CLASSROOM SETTINGS: A LONGITUDINAL STUDY

Kazuya Saito, Yui Suzukida, and Hui Sun
Birkbeck, University of London

The current study longitudinally examined the influence of aptitude on second language (L2) pronunciation development when 40 first-year Japanese university students engaged in practice activities inside and outside English-as-a-Foreign-Language classrooms over one academic year. Spontaneous speech samples were elicited at the beginning, middle and end points of the project, analyzed for global, segmental, syllabic, prosodic and temporal aspects of L2 pronunciation, and linked to their aptitude and experience profiles. Results indicated that the participants generally enhanced the global comprehensibility of their speech (via reducing vowel insertion errors in complex syllables) as a function of increased classroom experience during their first semester, and explicit learning aptitude (associative memory, phonemic coding) appeared to help certain learners further enhance their pronunciation proficiency through the development of fluency and prosody. In the second semester, incidental learning ability (sound sequence recognition) was shown to be a significant predictor of the extent to which certain learners continued to improve and ultimately attain advanced-level L2 comprehensibility, largely thanks to improved segmental accuracy.

Adult second language acquisition (SLA) is a multifaceted phenomenon whose process and product are greatly affected not only by factors related to experience (e.g., how second language [L2] learners have practiced the target language), but also by those which are learner-internal (e.g., to what extent they are cognitively and socially adept at L2 pronunciation learning). Adopting a longitudinal approach (i.e., where learners engaged in practice activities inside and outside English-as-a-Foreign-Language (EFL) classrooms over one academic year), the current study examined how 40 college-level Japanese students with various aptitude scores (in terms of sound sequence recognition, phonemic coding, and associative memory) could improve the global

We are grateful to Peter Skehan, the journal associate editor, Andrea Révész, and two anonymous *SSLA* reviewers for their constructive feedback on an earlier version of the manuscript. We also acknowledge Keiko Hanzawa, Takumi Uchihara, George Smith and Ze Shan Yao for their help for data collection and analyses. The project was funded by the Grant-in-Aid for Scientific Research in Japan (No. 26770202).

(comprehensibility), segmental (consonant/vowel errors), syllabic (schwa vowel insertion), prosodic (wrong/missing stress) and temporal (breakdown, speed, fluency) dimensions of their L2 pronunciation.

BACKGROUND

Experience, Individual Differences and Second Language Acquisition

In the field of SLA, many scholars would at least agree with the fundamental view that much improvement takes place as long as L2 learners make consistent efforts to practice the target language. Such experience effects (i.e., more practice is better) could be strong especially when L2 learners have started learning the target language under new environments (see Derwing, Thomson, & Munro, 2006 for longitudinal evidence). However, these experience factors alone may not fully explain the extent to which L2 learners can promptly, substantially and continuously improve their proficiency, as they engage in more practice opportunities with different types of native and non-native interlocutors in various social settings. The quality of SLA beyond the early stage of L2 learning could be susceptible to a great deal of individual variability and is arguably linked to certain learner-intrinsic (rather than -extrinsic) factors, such as professional motivation (Moyer, 2014), affect (Gkonou, Daubney, & Dewaele, 2017), and aptitude (Granena & Long, 2013).

Whereas the relevant discussion on experience and individual differences has been primarily concerned with naturalistic SLA, a growing number of researchers have begun to explore the generalizability of the research findings to foreign language (FL) classroom contexts, where access to the L2 is restricted to several hours of instruction per week—the focus of the current study. Under such limited input conditions, L2 learners' improvement patterns could be strongly tied to different FL learning backgrounds in terms of length and type of instruction (Spada & Tomita, 2010). According to previous investigations, adult L2 learners in various FL classrooms can develop a range of L2 skills as a function of increased input and practice inside and outside classrooms (Zhang & Lu, 2013 for lexical knowledge; Larson-Hall, 2008 for grammar knowledge; Muñoz, 2006 for listening proficiency). Furthermore, certain learners appear to be capable of achieving high-level L2 performance after receiving an extensive amount of FL instruction (6-10 years) (Muñoz, 2014). To account for the incidence of successful foreign language learning, especially at the relatively later stages of classroom SLA, a great deal of research attention has been directed towards one learner-internal factor in particular—aptitude.

Foreign Language Aptitude

Foreign language aptitude refers to a set of perceptual and cognitive abilities which are assumed to help L2 learners acquire a target language in an effective and efficient fashion. Different from other individual difference variables in SLA (e.g., motivation), aptitude is considered to be a relatively stable trait which is unlikely to change with varying amounts of L2 learning experience. An abundance of aptitude research was launched in the 1950's, when Carroll and Sapon (1959)

proposed their influential aptitude framework relevant to successful FL learning. This initial framework included such concepts as associative memory (remembering new word form-meaning pairings) and phonemic coding ability (analyzing and remembering unfamiliar sounds), and led to the development of an aptitude test still widely used today—the Modern Language Aptitude Test (MLAT). According to earlier validation studies (e.g., Carroll, 1965), MLAT scores were found to be correlated with L2 learners' test performance and grades in various FL classrooms, which were, at the time, rife with the use of explicit learning strategies to acquire the target language across a limited number of short-term, language-focused classes.

Although the original concept of aptitude was strongly tied to the initial stage of form-oriented FL learning, more recent studies have expounded what dimensions of aptitude relate to L2 learners' ultimate attainment after years of varied L2 learning experience in both classroom and naturalistic settings. By using a composite test battery consisting of 11 cognitive tasks (i.e., Hi-LAB), Linck et al. (2013) investigated the aptitude profiles of 450+ personnel from the U.S. government and military who had studied various foreign languages for a long period of time (> 10 years). According to the results, those with high-level reading and listening performance according to the Defense Language Proficiency Tests had similarly high explicit language learning ability (associative memory), working memory (phonological short term memory) and implicit learning ability (serial reaction time).

A growing number of empirical studies have found that LLAMA test scores—developed by Meara (2005) and used in the current study—can successfully predict various aspects of L2 lexicogrammar development. Extending the MLAT, the LLAMA test incorporates not only key elements for form-focused FL learning (associative memory, phonemic coding, grammatical inferencing), but also a new measure for the ability to recognize sound sequence patterns in spoken language in a relatively incidental fashion (sound sequence recognition)—a crucial component for successful L1 acquisition (Saffran, 2014), that is also linked to L2 grammatical attainment (Granena, 2013). Under FL learning conditions (the context of the study), explicit learning aptitude (associative memory, phonemic coding, grammatical inferencing) has been found to successfully predict the extent to which L2 learners benefit from their teachers' explicit instruction (Yalçin & Spada, 2016) and corrective feedback (Yilmaz & Granena, 2016).

In his synthesis of research on aptitude and SLA, Skehan (2016) argued that different constructs of aptitude may be uniquely related to different putative stages of SLA: analyzing incoming input → automatizing partially acquired knowledge → attaining advanced-level use of language. In this view, those dimensions of aptitude related to auditory processing, such as phonemic coding, enhance the phonological buffer component of working memory. This in turn allows L2 learners to hold more information regarding unfamiliar sounds (i.e., input processing), and makes it available for more detailed and refined linguistic analyses (i.e., noticing). Thus, L2 learners with such auditory processing abilities (e.g., phonemic coding) can better focus on improving the accurate use of their L2 speech thanks to their enhanced ability to analyze input, especially when they are explicitly asked to do so under classroom conditions (Yilmaz & Koylu, 2016). Certain talented L2 learners can also rely on associative memory to establish stronger form-meaning mappings by quickly combining relevant linguistic forms with what they signify. They can remember, maintain and control even vast amounts of declarative knowledge (e.g., what to say and how to convey it) during actual use of language (i.e., proceduralization). After much practice in various conversational settings, learners with good memory can become more accurate *and* fluent (i.e., automatization) in their language use (Schneiderman & Desmarais, 1988).[1]

The attainment of advanced-level L2 proficiency requires L2 learners to further improve their representational systems and processing abilities in the target language. To this end, L2 learners may need not only explicit and intentional but also implicit and incidental learning aptitude (e.g., serial reaction time, sound sequence recognition). Learners with such high aptitude scores can make the most of their L2 experience by not only processing/analyzing incoming pronunciation, vocabulary and grammar units of language during the explicit learning process, but also by attending to more abstract and intuitive combinations of sound and word sequences without awareness (i.e., lexicalization). In fact, the latter type of aptitude (incidental and implicit one) has been found to facilitate the mastery of nativelike L2 use among experienced L2 learners in naturalistic settings (Abrahamsson & Hyltenstam, 2008; Granena, 2013).

To date, the existing literature comprises of a wealth of cross-sectional investigations on the role of aptitude in L2 *lexicogrammar* learning (see Li, 2016 for a meta-analysis). However, comparatively less is known about influence aptitude has on the development of L2 *pronunciation* proficiency, particularly from a longitudinal perspective. The main objective of the current study is to further the existing aptitude research agenda by moving the past investigation of L2 lexicogrammar learning, and exploring aptitude's link with L2 pronunciation learning.

Second Language Pronunciation Proficiency

Traditionally, second language pronunciation proficiency has been conceptualized and analyzed through a range of global, segmental, syllabic, prosodic and temporal measures. Given that achieving nativelike L2 pronunciation is difficult even for early bilinguals relative to other domains of language such as vocabulary and grammar (Granena & Long, 2013), many researchers have emphasized the importance of assessing the global quality of adult L2 learners' speech in terms of "comprehensibility" (i.e., to what degree L2 learners' pronunciation is easy to understand) rather than "accentedness" (i.e., the extent to which L2 speech sounds like native speaker baseline) (Derwing & Munro, 1997). With respect to specific constructs of L2 pronunciation, many researchers (e.g., Trofimovich & Isaacs, 2012) have illustrated: (a) how L2 learners can accurately pronounce new consonantal and vocalic sounds (i.e., segmentals) in both simple (e.g., Consonant-Vowel [CV]) and complex (e.g., CVC, CCVCC) syllable structures with the correct placement of stress (i.e., prosody); and (b) how such pronunciation forms can be delivered without many pauses and repetitions with optimal tempo (i.e., fluency).

Derwing, Munro and Thomson's (2008) longitudinal investigation of adult L2 English learners in Canada showed that the participants significantly enhanced the overall comprehensibility (but not accentedness) of their speech over the course of two years of immersion when their main language communication was L2 rather than L1. There is ample cross-sectional evidence that a great amount of L2 experience (e.g., more than 10 years of residence) may be needed for the acquisition of more refined segmental (e.g., Schmid, Gilbers, & Nota, 2014) and prosodic (e.g., Trofimovich & Baker, 2006) accuracy, as well as the development of more nativelike fluency (e.g., Lahmann, Steinkrauss, & Schmid, 2017).

To explore the extent to which L2 learners can ultimately improve their pronunciation proficiency in FL settings, our precursor research (Saito, 2017) elucidated the linguistic and learning profiles of 50 Japanese L2 learners of English with extensive FL backgrounds. Whereas the participants' overall pronunciation forms were generally considered to be intelligible, certain

learners with greater explicit learning aptitude demonstrated more advanced-level L2 oral ability, owing particularly to their better command of difficult phonological features. According to the results of the correlation analyses, phonemic coding ability was moderately correlated with the segmental and prosodic qualities of L2 speech, and associative memory was weakly associated with the temporal qualities of L2 speech. However, their relatively incidental learning ability (i.e., sound sequence recognition) was unrelated to any aspects of their speech performance (for similar results in naturalistic SLA, see Granena & Long, 2013).

Whereas the results here provide some evidence to support the interaction between different types of aptitude (phonemic coding vs. associative memory) and speech (pronunciation accuracy vs. fluency) (Skehan, 2016), such findings need to be interpreted with caution due to the cross-sectional nature of the dataset. Although L2 learners with high aptitude scores demonstrated relatively greater accurate and fluent L2 pronunciation performance at the time of the project, the results could be explained not only by differences in their aptitude levels, but also by differences in the nature of the FL input (in and outside the classroom) they had obtained throughout their FL learning experience—another crucial affecting factor for the outcomes of L2 learning in FL classrooms. As a remedy, multiple data collection points are needed in order to track not only L2 learners' developmental patterns, but also their experience within specific timescales; such longitudinal designs would shed light on the full-fledged picture of aptitude effects in SLA while controlling for relevant experience variables. The current study is designed to respond to these concerns.

CURRENT STUDY

Research Objectives

This study reports on a longitudinal investigation of the complex relationship between aptitude, experience and L2 pronunciation learning over one academic year among 40 first-year Japanese university students with six years of FL education. Spontaneous speech samples were elicited from the learners at the beginning, mid and end of the year, and analyzed for global (comprehensibility) and specific (segmentals, syllables, prosody, fluency) constructs of L2 pronunciation proficiency. Following Meara's (2005) framework, three components of aptitude potentially related to pronunciation development were featured and measured via the LLAMA test—(a) associative memory (LLAMA-B), (b) phonemic coding (LLAMA-E) and (c) sound sequence recognition (LLAMA-D). Most notably, care was taken in this study to track all participants' FL experience (inside/outside classrooms) within a specific time framework of the study—one academic year.

Predictions

In keeping with Skehan's (2016) acquisition-oriented model of aptitude reviewed earlier, auditory processing (phonemic coding) is assumed to be instrumental to input processing and noticing; associative memory to proceduralization and automatization; and sound sequence recognition to high-level L2 proficiency attainment. Accordingly, three predictions were formulated. First, L2

learners with higher phonemic coding abilities (LLAMA-E) would show more gains in the global, segmental, syllabic and prosodic accuracy aspects of L2 pronunciation proficiency, given their presumed higher ability to process and analyze the phonetic subunits of words in the L2 input. Next, those with greater associative memory would deliver their L2 speech more rapidly and smoothly (breakdown/repair fluency) by virtue of drawing on their stronger and more durable access to what they had already acquired (speed fluency). Since such explicit learning aptitude is assumed to facilitate the speed of classroom learning (Carroll, 1965), high-aptitude learners are expected to show gains especially within the first semester, when they have just entered university and begun to study L2 English in a new FL environment.

Comparatively, incidental learning is hypothesized to relate to how much (rather than quickly) L2 learners can enhance their proficiency, when they receive ample L2 input over a long time span; such outcomes of incidental learning could be gradual, but consistent and extensive even after their rate advantage (i.e., considerable/quick learning during the first semester) (Hulstijn, 2003). In this regard, certain students, who had the relatively high incidental learning aptitude—sound sequence recognition (LLAMA-D), would continue to improve their pronunciation proficiency during the first semester *and* the second semester. By maximizing their FL experience via processing input explicitly as well as implicitly, these exceptional learners were presumed to be better able to detect and integrate segmental/prosodic patterns in the L2 input into their representation systems. Owing to this, it was assumed that they could further refine their phonological accuracy and accessing efficiency in production to approach more sophisticated, nativelike levels.

METHOD

Participants

As a part of a larger project which set out to examine the speaking abilities of 100+ Japanese FL students at several universities in the Tokyo area in Japan, the data collection for the current study was conducted during the Spring (Semester 1: 15 weeks) and Fall 2013 (Semester 2: 15 weeks) semesters. A total of 40 first-year students (19 males, 21 females) enrolled in various arts and social sciences programs (e.g., business, marketing, psychology, international relations) at a large university in Tokyo were carefully recruited as participants for the study based on the following conditions. All of them were native speakers of Japanese (with both of their parents being L1 Japanese speakers). They were freshman students in their first semester at the university, and ranged in age from 18 to 19 years. They had started learning English from secondary school (Grade 7), and lacked any prior study-abroad experience. Given their length of FL learning (i.e., 6 years), the participants were considered as relatively experienced FL learners (for a similar definition, see Muñoz, 2014). At the outset of the project, the participants' general proficiency test scores (i.e., TOEIC) varied from 490 to 845 out of 990 (M = 643.6, SD = 113.2), suggesting that their CEFR bands could be considered from B1/B2 (Independent users) to C1 (Proficient users).

To survey their FL experience and record their L2 speech, an individual interview was scheduled with each participant (30 min per session) at the beginning (T1: April) and end of Semester 1 (T2: July) as well as the end of Semester 2 (T3: March). Approximately one month

after the end of the final post test (T3), the participants took the LLAMA aptitude test (40 min per session).

Although the screening criteria ensured some degree of similar demographic attributes among the participants, the participants demonstrated much individual variability in terms of their FL experience within the time framework of the project (one academic year). In our preliminary analyses (reported in another venue: Saito & Hanzawa, 2017), we surveyed the extent to which they had practiced L2 English inside and outside classrooms throughout the project (one academic year). Subsequently, we examined how their varied FL experience backgrounds influenced various aspects of their oral performance at T1, T2 and T3, measured through native raters' subjective pronunciation/lexicogrammar judgements.

According to the post-hoc interviews, none of the participants received any specialized pronunciation training throughout the project—a typical phenomenon in many foreign language (Saito, 2014) and second language (Derwing et al., 2008) classrooms. In addition, the results identified small correlations between the number of classes and fluency development only during Semester 1. However, such experience effects were not found during Semester 2, suggesting that other individual difference factors beyond experience variables, such as aptitude, could have had a comparatively greater impact on successful L2 pronunciation learning.

In the current study, we revisited the dataset to elucidate how different types of aptitude (explicit vs. incidental learning aptitude) could facilitate two supposedly different phases of their pronunciation learning—Semesters 1 and 2. For the former, the associations between experience factors and L2 pronunciation development could be relatively strong; for the latter, such experience effects could be lessened. To reflect the multifaceted nature of L2 pronunciation proficiency, the global, syllabic, prosodic and temporal qualities of the participants' speech were also scrutinized by a total of seven measures developed specifically for the current project.

Aptitude Test

The LLAMA test was used to measure the participants' associative memory (LLAMA-B), phonemic coding (LLAMA-E) and sound sequence recognition (LLAMA-D), (Meara, 2005). Using visual and verbal materials adapted from a British-Columbian indigenous language and a Central-American language (rather than digits and symbols unrelated to natural language), the LLAMA test is considered to measure aptitude specific to human language learning.

To reduce the participants' awareness towards what they were doing (i.e., intentional/explicit learning) and correspond to their incidental learning aptitude, especially during the LLAMA-D session, the tests were given in the following order: LLAMA-D → LLAMA-B → LLAMA-E. Sub-tests were automatically scored out of 75 for the LLAMA-D, and 100 for LLAMA-B and -E.

LLAMA-D. This subtest assesses participants' ability to detect and memorize novel sound sequences and abstract phonetic regularities[2] in a language *without* awareness. Such phonological aptitude is believed to help L2 learners segment aural streams of speech into lexical units in order to identify words in comprehension, and automatize their lexical access in production (Speciale, Ellis, & Bywater, 2004). First, the participants listened to 10 sound strings from a computer to check if they could hear them without any difficulty. At this stage, the participants were not told about the presence/purpose of the test so as to avoid evoking the participants' intentional learning.[3]

Subsequently, the participants proceeded to the testing phase, where they listened to 30 items and then detected whether they had heard them during the initial sound-check session.

LLAMA-B. This subtest gauges the participants' ability to memorize new vocabulary items based on the association between alphabetical strings (2-7 letters) and corresponding objects (cf. the paired-associates test in MLAT). First, the participants were explicitly told about the objective of the subtest (learning vocabulary followed by recollection) and asked to memorize a relatively large amount of new vocabulary items ($N = 20$) within two minutes. In the following testing phase, the participants were asked to match randomly chosen names with the corresponding visuals (20 items).

LLAMA-E.This subtest evaluates the participants' ability to learn new sound-symbol correspondences by associating sound strings with unfamiliar alphabetical symbols (i.e., phonemic coding) (cf. the phonetic script test in MLAT). First, they were explicitly asked to learn and remember the relationship between 24 recorded syllables and their corresponding phonetic symbols within two minutes. Subsequently, their recollection memory was tested if they could correctly identify corresponding symbols after listening to a combination of two syllables (a total of 20 items).

Experience Survey

At the end of Semester 1 (T2) and Semester 2 (T3), all the students were individually interviewed to estimate how they had practiced L2 English throughout the term. As was operationalized in previous FL studies (e.g., Muñoz, 2014), they reported in a retrospective manner two experience factors directly related to successful classroom SLA—(a) the total hours of English classes they had taken per week (i.e., the length of classroom learning); and (b) the total number of hours they had weekly spent practicing English through informal conversations with other native and non-native speakers of English (i.e., extracurricular activities). For the former (L2-use inside classrooms), the students were given the option to register in not only language-focused classes (comprising reading, listening, writing and speaking activities with a primary focus on form), but also content-based classes (delivering various subject matter lessons in L2 English), which is increasingly common in many FL classrooms all over the world (e.g., Content and Language Integrated Learning). For the latter (L2 use outside classrooms), some students sought conversation opportunities especially with international students at the university, whereas others rarely used L2 English outside classrooms for conversational purposes.

Speaking Task

Given that adult L2 learners can carefully monitor their correct use of language via explicit knowledge, especially when their performance is tested via controlled production tasks (e.g., word and sentence readings), many SLA researchers (e.g., Abrahamsson & Hyltenstam, 2008) have emphasized the importance of adopting spontaneous production tasks (e.g., picture narratives) to

examine the present state of L2 learners' oral competence. In these tasks, learners are encouraged to pay equal attention to all linguistic domains of L2 speech (pronunciation, fluency, vocabulary, grammar) to convey their communicative intentions under communicative pressure (without much planning time) (Spada & Tomita, 2010).

Accordingly, following previous L2 speech studies (e.g., Derwing & Munro, 1997), a timed picture description task was adopted in the current study. As conceptualized and piloted in Saito and Hanzawa (2016), the picture description task was designed to elicit a certain length of spontaneous speech data without excessive hesitations and dysfluencies from L2 learners, even those with low proficiency levels. In this task, participants described seven separate pictures with three keywords printed as hints for each picture; this differs from previous research which has asked participants to explain a series of thematically-linked images without any linguistic support (e.g., Derwing & Munro, 1997). To equalize participants' familiarity with the task format, the first four pictures were used for practice and the last three were targeted for analyses. To minimize any conscious speech monitoring during each picture description, the participants were given only a very small amount of planning time (i.e., 5 sec) before describing each picture.

Materials. The three target pictures depicted a table left out in a driveway in heavy rain (keywords: rain, table, driveway), three men playing rock music with one singing a song and the other two playing guitars (keywords: three guys, guitar, rock music), and a long stretch of road under a cloudy blue sky (keywords: blue sky, road, cloud). The keywords were all considered as frequent (among the first 3,000 most frequent words according to the British National Corpus), and were intentionally chosen to push Japanese learners to use challenging segmental, syllabic and prosodic features (without using any avoidance strategies). For instance, Japanese speakers have been reported to neutralize the English /r/-/l/ contrast ("rain, rock, brew, crowd" vs. "lane, lock, blue, cloud") and to insert epenthetic vowels between consecutive consonants (/dəraɪvə/ for "drive," /θəri/ for "three," /səkaɪ/ for "sky") and after word-final consonants (/teɪbələ/ for "table," /myuzɪkə/ for "music") in borrowed words (i.e., Katakana). Due to the cross-linguistic differences at the prosodic level between Japanese (mora-timed) and English (syllable-timed), L1 Japanese learners typically have difficulty in pronouncing bi-syllabic words (e.g., guiTAR, MUsic), applying incorrect word stress (GUItar, muSIC) or equal stress on each syllable (GUITAR, MUSIC).

Procedure. To keep track of the participants' pronunciation proficiency development over one academic year, their spontaneous speech was collected via the timed picture description task at three testing points (T1, T2, T3).[4] All picture descriptions were recorded by means of a Marantz PMD 660 recorder with a Shure SM 10A-CN microphone (44.1 kHz sampling rate with 16-bit quantization) in a soundproof booth at the Japanese university. For each testing session (T1, T2, T3), instructions were given orally by the researcher in Japanese in order to ensure that all speakers understood the procedures. After four pictures were presented for the participants to practice, they proceeded to describing the remaining three pictures, which were used for the final analyses.

The first 10 seconds of each of the three picture descriptions were extracted, combined and stored in a single file for each participant at T1, T2 and T3, respectively. In total, 120 speech samples were generated (40 FL students × 3 testing points). In light of previous studies, we believe that the length of the speech samples in this study (30 sec per participant at T1, T2 and T3) provided sufficient phonological information for the analyses of global comprehensibility (Derwing & Munro, 1997 for 10-15 sec), pronunciation accuracy (Isaacs & Trofimovich, 2012 for 30 sec) and fluency (Bosker, Pinget, Quené, Sanders, & De Jong, 2013 for 20 sec).

Global Analyses

Raters. In the current study, we asked *untrained* listeners (without much linguistic and pedagogical experience) to make intuitive judgements of overall L2 oral ability on the continuum of comprehensibility, without any detailed descriptors or training (the raters received only a brief explanation on the definitions of comprehensibility; for training scripts and onscreen labels, see Appendix A). Such *intuitive* (rather than *deliberate*) judgement of extemporaneous L2 speech is believed to well reflect what native speakers do in real life when communicating with non-native speakers (Derwing & Munro, 1997). Five native speakers of English (2 males, 3 females) were recruited at an English-speaking university in Montreal, Canada. Their mean age was 22.5 years. They were born and raised in English-speaking homes in Montreal. All of the raters were undergraduate students with non-linguistic backgrounds (e.g., business, psychology) and reported no previous teaching experience in L2 classrooms. They reported relatively low familiarity with Japanese-accented English (M = 1.5 from 1 = *Not at all* to 6 = *Very much*). None of the raters reported any hearing problems.

Procedure. The 120 speech samples were randomly presented to the raters via the MATLAB software. After listening to each file in its entirety, the raters used a free moving slider on a computer screen to assess comprehensibility. 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." To tap into the initial intuitions and impressions of comprehensibility, each sample was played only once for the raters' judgment.

The rating session took place for approximately 1.5 hours in a quiet room at the university. First, the raters familiarized themselves with the picture prompts and key words, and received a brief explanation of comprehensibility from a trained research assistant. Next, they practiced the procedure with five speech samples (not included in the dataset), and then proceeded to the global judgement of the tokens.

Inter-Rater Reliability. Similar to previous research (e.g., Derwing & Munro, 1997), the five inexperienced raters showed high inter-rater agreement for their comprehensibility ratings (Cronbach's alpha = .94). Thus, mean scores were calculated by pooling the scores of the five inexperienced raters. The mean scores were then applied to each token produced by the participants.

Segmental, Syllabic and Prosodic Analyses

In conjunction with Isaacs and Trofimovich's (2012) coding scheme for L2 pronunciation accuracy, three domain-specific measures were devised to analyze the segmental, syllabic and prosodic qualities of L2 speech.

Experienced Coder. The pronunciation accuracy analyses were conducted by an L1 Japanese researcher with a great deal of L2 speech analysis experience, and near-nativelike proficiency in

L2 English. Different from many previous studies, where native speakers have typically been recruited as coders (e.g., Derwing & Munro, 1997; Trofimovich & Isaacs, 2012), our unique decision here (i.e., recruiting the coder who had much experience with both L1 and L2) was made for the following reasons. Since the goal of L2 speech learning concerns comprehensible and intelligible pronunciation rather than nativelike accuracy, the validity of the native speakers' subjective, dichotomous judgements of non-native speech ("targetlike" vs. "non-targetlike") remains open to discussion (Jenkins, 2002). Instead, the pronunciation quality of L2 speech in the study was assessed based on whether the participants continued to use L1 Japanese forms or demonstrated any effort to use L2 English forms in obligatory contexts (rather than to what degree their pronunciation simply approximated the native baseline). As seen in some L2 speech studies with similar FL learners (e.g., Riney, Takada, & Ota, 2000 for college-level Japanese FL learners), we believe that a L1 Japanese speaker with near nativelike L2 English proficiency (rather than native speakers of English) would be better qualified as a coder to capture the subtle changes in L2 learners' interlanguage pronunciation forms.[5]

Procedure. The Japanese coder carefully listened to each speech sample, and analyzed for the following three categories:

Segmental Error Ratio. This category was analyzed by dividing the number of L1 substitution errors (e.g., Japanese /ɾ/ for English /ɹ/, Japanese /s/ for English /θ/) by the total number of segments articulated (for more details, see also Appendix B).

Syllable Error Ratio. This category was analyzed by dividing the number of consonant and vowel insertions (e.g., "blue" pronounced as /bəlu/) by the total number of syllables articulated.

Prosodic Error Ratio. This category was analysed by dividing the number of the absence of primary stress errors (e.g., "DRIVEway" pronounced as "driveway" [no primary stress] or "DRIVEWAY" [equal primary stress]) and the misplacement of primary stress errors (e.g., "DRIVEway" pronounced as "driveWAY" by the total number of multisyllabic words.

Inter-Coder Reliability. To check the reliability of the Japanese coder's analyses, she and another coder (L1 Japanese speaker with coding experience) first separately analyzed a total of 40 similar speech samples produced by Japanese learners of English which were not a part of the current study. According to the inter-coder sessions, relatively high correlations were found for their judgements of segmentals, syllables and word stress (Cronbach alpha > .90). Afterwards, the first coder proceeded to the analysis of the main dataset on her own (a total of 120 speech samples).

Temporal Analyses

Following De Jong, Steinel, Florijn, Schoonen, and Hulstijn's (2012) notion of fluency, three measures were developed to objectively analyze the three temporal aspects of L2 pronunciation proficiency: (a) breakdown (how effortlessly speech is articulated without many pauses and hesitations), (b) speed (how many words/syllables are produced within a certain period of time)

and (c) repair (how often corrections and repetitions are present in speech) (see also Bosker et al., 2013).

Breakdown Fluency (Pause Ratio). Breakdown fluency was measured by dividing the total number of filled pauses (e.g., eh, ah, oh) and unfilled pauses (i.e., silence) over the total number of words. The number of filled pauses was counted based on raw transcripts, and the number of unfilled silent pauses was automatically calculated via a script programmed in *Praat* with the minimum length of silence set at 250 milliseconds (for a similar decision, see De Jong et al., 2012).

Speed Fluency (Articulation Rate). One dimension of speed fluency—articulation rate—was measured by dividing the total phonation time (without all filled pauses) by the total number of syllables.

Repair Fluency (Repair Ratio). Repair fluency was calculated by dividing the total number of repetitions (repeating words and phrases) and reformulations (self-correcting nontargetlike forms) by the total number of words (based on raw scripts).

RESULTS

Aptitude and Experience Profiles

The first objective of the statistical analyses was to report the relationship within and between participants' aptitude and experience profiles (descriptively summarized in Table 1). For the sake of comparison, the participants' aptitude scores (LLAMA-D, -B, -E) were converted into z-scores. With respect to the experience variables, the participants widely differed in terms of the number of English classes they had received (i.e., form- and content-based classes) and the number of hours they had spent conversing outside classrooms (with native and non-native speakers) in Semesters 1 and 2, respectively.

Next, a set of Pearson correlation analyses were performed on the aptitude scores (LLAMA-D, -B, -E) and experience variables (the amount of L2 use during Semesters 1 and 2). According to the results summarized in Table 2, the participants' scores on the three aptitude tests were not significantly correlated with each other ($p > .05$). As proposed by Meara (2005), the findings here indicate

Table 1. Descriptive Statistics of Learner Aptitude and Experience Profiles

	M	*SD*	95% CI	
			Lower	Upper
Language aptitude[a]				
LLAMA-D	0.0	1.0	-.21	.23
LLAMA-B	0.0	1.0	-.20	.18
LLAMA-E	0.0	1.0	-.39	.09
Experience[b]				
Inside classroom (Semester 1)	116.9	58.1	98.3	135.5
Outside classroom (Semester 1)	2.9	7.1	0.6	5.2

Inside classroom (Semester 2)	89.3	49.3	73.5	105.1
Outside classroom (Semester 2)	6.8	12.2	2.9	10.8

Note. [a]All aptitude scores were converted to z scores. [b]The experience factors refer to the total hours of L2 use inside and outside FL classrooms per semester.

that the LLAMA test in the study appeared to tap into three different constructs of cognitive abilities: sound sequence recognition (LLAMA-D), associative memory (LLAMA-B), and phonemic coding (LLAMA-E). In terms of their FL experience, significant correlations were observed between L2-use inside and outside classrooms during both Semesters 1 and 2, suggesting that participants who took more English classes likely spent more time on conversation activities with native and non-native speakers outside classrooms. Overall, the strength of the aptitude-experience link was relatively weak ($r < .2$) and non-significant ($p > .05$). The exception was their performance on the LLAMA-E, which was significantly correlated with the number of FL classes they had received during Semester 1 ($r = .32$, $p < .05$).

Pronunciation Proficiency Profiles

The second objective of the statistical analyses was to illustrate how the global (comprehensibility) and specific (segmentals, syllables, prosody, fluency) dimensions of L2 pronunciation proficiency were interrelated. To this end, the relative weights of the participants' segmental, syllabic, prosodic and temporal scores in terms of their global comprehensibility scores at three different points of time (T1, T2, T3) were examined via a linear mixed-effects regression analysis. Their comprehensibility scores were used as a dependent variable relative to six

Table 2. Interrelationships between Aptitude and Experience Factors
Note. indicates statistical significance at $p < .05$; † indicates marginal significance at $p < .10$.

	LLAMA-B		LLAMA-E		L2 use inside (Semester 1)		L2 use outside (Semester 1)		L2 use inside (Semester 2)		L2 use outside (Semester 2)	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
LLAMA-D	.112	.492	-.118	.474	.063	.698	.159	.326	.033	.840	-.007	.968
LLAMA-B			.151	.351	.223	.167	.066	.684	.238	.139	.112	.490
LLAMA-E					.320*	.044	.213	.188	.248	.123	.077	.638
L2 use inside (Semester 1)							.278†	.083	.926*	< .001	.367*	.020
L2 use outside (Semester 1)									.243	.131	.654*	< .001
L2 use inside (Semester 2)											.350*	.027

independent variables (segmentals, syllables, word stress, breakdown/speed/repair fluency).

The model identified three significant fixed effects for segmentals, B =-0.89, t =, -2.69, p = .008, syllables, B =-0.89, t =, -2.62, p = .010, and breakdown fluency, B =-0.41, t =, -4.36, p < .001. The results here suggest that the native judges relied on, in particular, three aspects of pronunciation information (segmentals, syllables, fluency) during their overall comprehensibility judgements at T1, T2 and T3.

To further examine the interrelationships between the six specific pronunciation measures (segmentals, syllables, prosody, breakdown/speed/repair fluency), another set of mixed-effects model analyses were performed on each of the six specific pronunciation scores as dependent variables and the other measures as predictor variables. According to the results, strong associations were found between segmentals and syllables, B = 0.27, t =, 4.27, p < .001; syllables and breakdown fluency, B = -0.62, t = -2.06, p = .047; and syllables and speed fluency, B = -0.01, t =, -3.80, p = .001.

In sum, a total of seven global and specific pronunciation measures adopted in the study appeared to correspond to four different aspects of the participants' pronunciation proficiency. This included their ability to produce correct pronunciation forms (segmentals, syllables) with adequate prosody (word stress), and connect them to form sentences at optimal speed (speed fluency, syllables) by making fewer long pauses (breakdown fluency, syllables), all of which were somewhat tied to the global impression of L2 speech (comprehensibility).

Pronunciation Development

The third objective of the statistical analyses was to probe how the participants' various dimensions of L2 pronunciation proficiency—comprehensibility, segmentals, word stress, fluency—changed over one academic semester. The participants' global and specific pronunciation scores at three different time points (T1, T2, T3) are summarized in Table 3. The descriptive statistics showed that the participants' improvement patterns were involved with a great deal of individual variability across all the pronunciation measures over one academic year. The results here were not surprising, as these students were substantially different in terms of aptitude (the extent to which they were adept at L2 pronunciation learning) and experience(how they practiced the target language in Semesters 1 and 2) (see Table 1).

Table 3. Descriptive Statistics of Pronunciation Proficiency at the Beginning (T1), Mid (T2) and End (T3) of the Project

	T1				T2				T3			
	M	*SD*	95% CI		*M*	*SD*	95% CI		*M*	*SD*	95% CI	
			Lower	*Upper*			*Lower*	*Upper*			*Lower*	*Upper*
A. Global proficiency												
1. Comprehensibility (1000 points)	416	155	368	464	457	138	414	500	466	138	423	509
B. Segmental/syllabic/prosodic proficiency												
1. Segmental error ratio	.050	.037	.038	.061	.043	.038	.032	.055	.044	.037	.033	.056
2. Syllables error ratio	.069	.084	.043	.095	.093	.109	.060	.127	.058	.078	.034	.082
3. Word stress error ratio	.277	.210	.212	.343	.291	.209	.226	.356	.308	.222	.239	.377
C. Temporal fluency												
1. Pause ratio (breakdown)	.314	.127	.275	.353	.259	.131	.219	.300	.203	.110	.169	.237
2. Articulation rate (speed)	175	30.9	166	185	178	31.5	168	188	188	31.1	179	198
3. Repair ratio (repair)	.057	.064	.037	.077	.059	.063	.040	.079	.041	.055	.023	.058

Aptitude, Experience and Proficiency Link

The final objective of the statistical analyses was to delve into the extent to which the participants' individual differences in their L2 pronunciation development could be related to their aptitude and experience profiles.

Correlation Analyses. To provide a general picture of the relationship between aptitude, experience and L2 speech learning, we conducted a set of partial correlation analyses. The participants' gain scores in Semester 1 (T2 minus T1 scores) and Semester 2 (T3 minus T2 scores) were used as dependent variables, and the participants' aptitude (LLAMA-D, B, E) and experience (L2 use inside and outside classrooms for Semesters 1 and 2) were used as independent variables.

Although our main interest lies in the participants' gain scores over Semester 1 (T1 → T2) and Semester 2 (T2 → T3), it is noteworthy that their initial pronunciation performance was substantially different between individuals even at the beginning of the term (T1, T2). This could be due to a combination of learner-extrinsic (how the participants had practiced the target language prior to the data collection) and learner-intrinsic (the extent to which their different levels of aptitude, motivation and affect promoted or debilitated their SLA) factors, a result whose discussion was beyond the scope of the study, and closely examined in our precursor research (Saito & Hanzawa, 2017).

As has been the case with similar longitudinal aptitude research (Yalçin & Spada, 2016), the decision was made to use the participants' initial performance (T1 for Semester 1, T2 for Semester 2) as a covariate, and statistically remove its effect from the rest of the subsequent analyses. In this way, we were able to focus on analyzing the participants' L2 pronunciation learning within specific timescales—Semester 1 (T1 → T2) and Semester 2 (T2 → T3)—without any conflation with their previous FL experience outside of the project.[6]

As shown in Table 4, three aspects of the participants' pronunciation proficiency improvement (comprehensibility, syllables, speed fluency) were correlated with the number of classes they had received during Semester 1. In contrast, their gains in prosodic and temporal abilities were weakly associated with explicit sound learning aptitude— associative memory (LLAMA-B) for breakdown fluency, and phonemic coding ability (LLAMA-E) for word stress and speed fluency. These results indicate that a great deal of L2 use may have initially facilitated the development of comprehensibility, as the participants could enhance the syllabic

Table 4. Results of Partial Correlations between Aptitude, Experience and Pronunciation Scores at Semester 1 (T1 → T2)

	LLAMA-D		LLAMA-B		LLAMA-E		L2 use inside (Semester 1)		L2 use outside (Semester 1)	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
A. Global proficiency[a]										
1. Comprehensibility (1000 points)	-.218	.183	-.005	.978	.226	.167	.429*	.006	-.003	.986
B. Segmental/syllabic/prosodic proficiency[a]										
1. Segmental error ratio	-.079	.631	-.139	.397	.022	.896	-.146	.376	-.139	.400
2. Syllables error ratio	.030	.857	-.181	.270	-.233	.154	-.444*	.005	-.102	.538
3. Word stress error ratio	-.120	.466	-.027	.871	-.302†	.062	-.188	.252	-.130	.431
C. Temporal proficiency[a]										
1. Pause ratio (breakdown)	-.089	.591	-.398*	.012	-.063	.705	-.174	.290	-.029	.861
2. Articulation rate (speed)	-.042	.799	.256	.115	.402*	.011	.360*	.025	.048	.772
3. Repair ratio (repair)	.001	.997	.210	.200	-.040	.809	.133	.419	.039	.814

Note. [a]Their pronunciation scores at T2 were used as dependent variables and their T1 scores were used as a covariate. *indicates statistical significance at $p < .05$; † indicates marginal significance at $p < .10$.

qualities of their L2 speech (i.e., reducing schwa insertion errors) in relation to an increasing amount of classroom experience. Furthermore, explicit aptitude— associative memory (LLAMA-B) and phonemic coding (LLAMA-E)—appeared to help certain L2 learners further improve their prosodic and temporal abilities, which are considered to be relatively difficult aspects of L2 speech learning.

For Semester 2 (summarized in Table 5), the extent to which the participants continued to develop their L2 pronunciation skills appeared to be unrelated to any of the experience factors. Rather, the amount of global (comprehensibility) and specific (segmentals, syllables, breakdown fluency) improvement in pronunciation proficiency was moderately correlated with their incidental learning aptitude (sound sequence recognition: LLAMA-D) and weakly with their intentional learning aptitude (associative memory: LLAMA-B). Taken together, the results here suggest that when L2 learners' pronunciation development is unrelated to any experience variables, these learners may ultimately need some form of innate language learning talent—e.g., sound sequence recognition—to further improve and attain advanced-level L2 comprehensibility (relative to other participants) by reducing not only schwa insertion errors, (syllables) but also L1 substitution errors (segmentals).

Multiple Regression Analyses. To confirm the suggested relationship between aptitude, experience and L2 pronunciation development, and to further examine the potentially different contributions of the aptitude and experience effects witnessed during Semesters 1 and 2, we next performed a set of multiple regression analyses. The participants' gain scores in comprehensibility, segmentals, syllables, prosody and fluency were used as dependent variables, while their aptitude and experience scores were used as independent variables. Similar to the aforementioned partial correlation analyses, the participants' pronunciation scores at T1 and T2 were also selected as an independent variable as a way to control for the influence of the participants' previous FL experience prior to Semesters 1 and 2.

To avoid multicollinearity problems, the decision was made to reduce the number of predictors by using only significant and marginally significant predictors identified in the partial correlation analysis (i.e., LLAMA-D, B, E; L2 use inside classrooms). The power of the dataset ($N = 40$ with five predictors) to find a medium effect size was .68, which could be considered as beyond the minimum requirement in the field of SLA research (> .50) (Larson-Hall, 2010).

As summarized in Table 6, the regression models showed that the participants' pronunciation learning during Semester 1 were substantially impacted by their initial performance—an indication of their six years of EFL experience. As for

Table 5. Results of Partial Correlations between Aptitude, Experience and Pronunciation Scores at Semester 2 (T2 → T3)

	LLAMA-D		LLAMA-B		LLAMA-E		L2 use inside (Semester 2)		L2 use outside (Semester 2)	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
A. Global proficiency[a]										
1. Comprehensibility (1000 points)	.438*	.005	.288†	.075	-.060	.717	.129	.433	.081	.622
B. Segmental/syllabic/prosodic proficiency[a]										
1. Segmental error ratio	-.328*	.042	.017	.919	-.197	.228	-.086	.601	-.043	.793
2. Syllables error ratio	-.369*	.021	-.256	.116	-.032	.846	-.079	.633	.116	.480
3. Word stress error ratio	-.002	.991	-.047	.777	.149	.365	.082	.619	-.040	.890
C. Temporal proficiency[a]										
1. Pause ratio (breakdown)	-.111	.501	-.310†	.055	.026	.877	-.191	.245	-.069	.677
2. Articulation rate (speed)	.151	.360	.063	.702	-.038	.819	-.125	.450	.008	.963
3. Repair ratio (repair)	.058	.724	-.196	.233	.062	.708	.201	.220	-.109	.509

Note. [a]Their pronunciation scores at T3 were used as dependent variables and their T2 scores were used as a covariate. *indicates statistical significance at $p < .05$; † indicates marginal significance at $p < .10$.

Table 6. Significant Results of Multiple Regression Analyses Using Aptitude, Experience and Initial Performance as Predictors of L2 Pronunciation Gains during Semester 1

Predicted variables	Predictor variables	*Adjusted* R^2	R^2 *change*	*F*	*p*
Comprehensibility	T1 scores	.461	.461	32.51	< .001
	L2 use inside class	.560	.099	23.55	< .001
Segmentals	T1 scores	.395	.395	24.83	< .001
Syllables	L2 use inside class	.389	.389	24.22	< .001
Word stress	T1 scores	.258	.258	13.21	.001
Breakdown fluency	T1 scores	.409	.409	26.24	< .001
	LLAMA-B	.504	.095	18.78	< .001
Speed fluency	T1 scores	.477	.477	34.61	< .001
	LLAMA-E	.561	.085	23.67	< .001
Repair fluency	T1 scores	.106	.106	4.52	.040

Note. The variables entered into the regression equations included LLAMA-D, LLAMA-B, LLAMA-E, L2 use inside class, and T1 pronunciation scores.

experience effects, some variance in L2 comprehensibility (9.9%) and syllable (38.9%) development was significantly explained by the amount of L2 use inside classrooms throughout the first term. In terms of aptitude effects, two aspects of improved fluency—speed and breakdown (8.5, 9.5%)—were moderately related to LLAMA-E and LLAMA-B.

The results of the significant predictors for the participants' L2 speech learning during Semester 2 are summarized in Table 7. Different from Semester 1, neither explicit aptitude nor the experience factors significantly accounted for the variance in the participants' gains across the second term. It was rather the participants' incidental learning aptitude—LLAMA-D—that played a pivotal role in their continuous development of comprehensibility (12.7%), segmentals (5.9%) and syllables (7.7%) at the later phase of the project.

Table 7. Significant Results of Multiple Regression Analyses Using Aptitude, Experience and Initial Performance as Predictors of L2 Pronunciation Gains during Semester 2

Predicted variables	Predictor variables	*Adjusted* R^2	R^2 *change*	*F*	*p*
Comprehensibility	T2 scores	.338	.338	19.44	< .001
	LLAMA-D	.465	.127	16.10	< .001
Segmentals	T2 scores	.447	.447	30.69	< .001
	LLAMA-D	.506	.059	18.96	
Syllables	T2 scores	.432	.432	28.87	< .001
	LLAMA-D	.509	.077	19.19	
Word stress	T2 scores	.610	.610	59.41	.001
Breakdown fluency	T2 scores	.446	.446	30.57	< .001
Speed fluency	T2 scores	.548	.548	46.09	< .001
Repair fluency		*n.s.*			

Note. The variables entered into the regression equations included LLAMA-D, LLAMA-B, LLAMA-E, L2 use inside class, and T2 pronunciation scores.

DISCUSSION AND CONCLUSION

In the context of 40 Japanese first-year university students in FL settings, the current study examined how different constructs of aptitude (sound sequence recognition, associative memory, phonemic coding) interact to affect the longitudinal development of L2 global, segmental, syllabic, prosodic and temporal proficiency while they differently practiced the target language in quantity and quality over one academic year. On the whole, the results showed that the participants' aptitude and experience scores uniquely predicted L2 pronunciation development at different time points (T1 → T2 → T3). With respect to Semester 1 (T1 → T2), aptitude demonstrated relatively small effects on L2 pronunciation learning. Explicit learning aptitude—associative memory (LLAMA-B) and phonemic coding (LLAMA-E)—weakly predicted the development of prosody and fluency. The development of global comprehensibility was strongly associated with the amount of L2 experience especially inside the FL classroom, as the participants significantly enhanced the syllabic aspects of their L2 pronunciation proficiency. With respect to Semester 2 (T2 → T3), the amount of the participants' L2 pronunciation performance was not significantly related to any experience factors. It was rather the incidental learning ability (sound sequence recognition measured by LLAMA-D) that significantly predicted the extent to which certain learners could continue to improve and ultimately attain advanced-level L2 comprehensibility (relative to the other learners), especially thanks to the refinement of segmental accuracy.

The results here in turn indicate several possible interpretations of the complex relationship between aptitude, experience and L2 pronunciation development. When L2 learners start learning the target language under new classroom conditions, their pronunciation proficiency quickly develops as a function of increased L2 use. This could be true for even experienced L2 learners, notably for learners such as the first-year university students in the current study, who had six years of FL experience prior to the project. In this case, experience effects are readily observable in the acquisition of relatively easy phonological features, such as reduced vowel insertions in complex syllables—a primary interlanguage strategy that many FL students have been reported to use to attain minimally successful comprehensibility in L2 speech (Lin, 2001). In the long run, however, the aptitude factor may steadily impact the acquisition of relatively difficult pronunciation features, which otherwise require a tremendous amount of input, interaction and practice, such as segmentals (Schmid et al., 2014), word stress (Trofimovich & Baker, 2006) and fluency (Lahmann et al., 2016).

In essence, the findings presented here supported the predictions which we presented earlier based on Skehan's (2016) aptitude-acquisition model: phonemic coding for accuracy; associative memory for fluency; and sound sequence recognition for continuous, long-term development. According to the results, explicit learning aptitude—associative memory and phonemic coding—allowed L2 learners to show not only quick, but also robust improvement in their L2 pronunciation proficiency even over a short period of time (i.e., Semester 1). The results suggest that L2 learners with greater associative memory may have the capacity to hold a substantial amount of information in their phonological store which will, as a result, free up their cognitive resources for use in delivering L2 utterances fluently (see O'Brien et al., 2008 for the effect of phonological short term memory—a similar construct to associative memory). L2 learners with greater phonemic coding ability may also have the capacity to identify and analyze the prosodic information of words in the incoming input to produce individual words with correct stress patterns (see Cerbian, 2006 for the effect of phonological awareness).

Whereas associative memory may continue to make marginal contributions to the development of fluency beyond Semester 1—a crucial indication of automatization, L2 learners may need not only explicit but also incidental learning aptitude (sound sequence recognition) in order to attain more refined segmental accuracy—a crucial characteristic of advanced-level L2 comprehensibility (Saito, Trofimovich, & Isaacs, 2016). Echoing Skehan's (2016) proposal, the participants in the current study could have done this (i.e., attaining high-level proficiency) by exploiting the full acquisitional value of the L2 input in FL classrooms, especially during Semester 2. According to the L1 acquisition literature (e.g., Saffran, 2014), sound sequence recognition ability plays a key role in successful implicit language learning in several respects. For one, it is believed to induce humans to statistically analyze characteristics of native phoneme categories, phonotactic regularities, and word stress patterns without awareness. Such knowledge is stored in long-term memory, enables learners to segment speech streams into word units, and helps learners recognize and produce these units more automatically (Mattys, Jusczyk, Luce, & Morgan, 1999). Though few in number, some SLA studies have provided preliminary findings that adult L2 learners with higher sound sequence recognition ability could establish more robust lexical (Speciale et al., 2004) and morphosyntactic (Granena, 2013) representations in the target language.

When it comes to L2 pronunciation learning, the current study added that the sound sequence recognition ability may relate to, in particular, the refinement of L2 segmental representations. According to recent L2 speech learning models (e.g., Bundgaard-Nielsen et al., 2012 for L2 Vocab Model), L2 learners tend to start paying special attention to improving their segmental accuracy as they move onto more advanced L2 pronunciation proficiency, after meeting the minimum requirements for successful social interaction (e.g., vocabulary size > 6,000-7,000 word families). These theoretical accounts agree that achieving nativelike segmental representations is crucial in the later stages of L2 speech learning, given that this phonological re-attunement is believed to push L2 learners to develop, expand and access a large number of L2 lexicons more accurately and quickly without becoming confused by phonologically similar words—i.e., vocabulary spurt (Bundgaard-Nielsen et al., 2012).

Finally, it is intriguing to remember that the results of the current study (which were longitudinally obtained) did not concur with those of the two cross-sectional studies, where sound sequence recognition ability (LLAMA-D) was unrelated to any dimensions of L2 pronunciation proficiency in naturalistic (Granena & Long, 2013) and FL classroom (Saito, 2017) settings. One reason for the discrepancy in the results could be due to the different nature of the dataset in the individual studies (longitudinal vs. cross-sectional methods). In the cross-sectional studies (Saito, 2017; Granena & Long, 2013), the aptitude factor was linked to L2 learners' final attainment after years of L2 learning experience, which could have resulted from a combination of explicit and implicit learning at different time points in various learning contexts. However, the longitudinal design of the current study allowed us to *pinpoint* the effect aptitude has during different stages of L2 speech learning. That is, our study revealed an aptitude-proficiency link at two different time points—(a) explicit learning aptitude was a significant predictor during Semester 1, during the rapid stage of L2 pronunciation development tied to an increasing amount of experience (i.e., experience phase); and (b) incidental learning aptitude was a significant predictor during Semester 2, when L2 pronunciation learning was unrelated to additional experience.

Taken together, these results provide two tentative hypotheses as to the predictive power of explicit and incidental learning aptitude in accordance with the timing of L2 speech leaning. First, cross-sectional studies have hinted that adult L2 speech learning mainly involves explicit and intentional learning over an extensive period of time (Saito, 2017; Granena & Long, 2013). The

present longitudinal study suggests that a certain amount of incidental learning occurs in the relatively later stages of L2 speech learning, where the relationship between experience and proficiency is relatively weak.

To close, two limitations to this exploratory study should be acknowledged for future research. First, the analyses of L2 pronunciation proficiency in this study were based on a particular group of L2 learners (40 Japanese students learning L2 English in FL settings) who engaged in a single task—a timed picture description. Since the linguistic characteristics of L2 learners' speech have been found to vary according to different types of task conditions (e.g., Ahmadian & Tavakoli, 2011 for task repetition; Yuan & Ellis, 2003 for pre-task and online planning time) and learner factors (e.g., De Jong et al., 2012 for proficiency levels; Lahmann et al., 2017 for age of testing), the findings of the current study need to be replicated with larger samples of participants with various ages (young/adult) and linguistic (L1/L2) backgrounds while adopting multiple task modalities and demands.

Second, the incidental learning aptitude in this study was analyzed via only LLAMA-D. Thus, another promising future direction is concerned with the development, elaboration and sophistication of new aptitude test batteries for measuring, in particular, incidental and implicit pronunciation learning. Although the LLAMA-D was used to measure sound sequence recognition, the validity of the test has been questioned (cf. Granena, 2013). Whereas serial reaction time has also been used to measure L2 learners' sequence learning ability without any awareness (implicit learning aptitude), the relationship between tests using non-linguistic materials (digits, symbols) and L2 pronunciation learning remains unclear. According to Skehan's (2016) review, the LLAMA-D is considered to be one of the few aptitude tests to measure relatively incidental learning aptitude based on natural language materials (North American indigenous languages). Thus, it would be intriguing to expand the current design of the LLAMA-D to more closely align with the way L2 learners actually acquire complex sound systems in new languages in an incidental (and implicit) manner. To this end, for example, speech samples that L2 learners are exposed to during the test should include not only simple monotonous sound sequences (CV), with but also complex ones with varied syllable structures (not only CV, but also CVC, CCVCC) and varying prosodic patterns (signalling of word stress via various pitch contours). For the recall section of the test, participants should be tested not only on their memory of old items, but also on their intuitive judgements of the "wordlikeness" of novel items in the language, which has been previously identified as a firm psycholinguistic phenomenon directly related to incidental and implicit language learning in L1 acquisition (e.g., Gathercole, Frankish, Pickering, & Peaker, 1999) and L2 acquisition (e.g., Gullberg, Roberts, Dimroth, Veroude, & Indefrey, 2010).

NOTES

1. Given that associative memory in the current study was measured via written materials in LLAMA B without any audio information (see the Method section), one may question the relationship between such aptitude and sound learning. As we revisit and detail this issue in the Discussion section, it indeed remains controversial the extent to which foreign language aptitude is domain-specific and -general (Skehan, 2016). In this current study, we did not have any specific stance on this issue. Our hypothesis was that associative memory would relate to the fluency aspect of L2 speech, whether its test format constitutes written or audio materials (cf. Silbert et al., 2015).

Notably, there is some evidence that L2 learners' cognitive abilities measured via even non-verbal materials are significantly related to their L2 speech performance (e.g., Darcy, Park, & Yang, 2015 for the link between speed naming and L2 vowel perception).

2. We conducted preliminary acoustic analysis of the stimuli used in LLAMA-D. To segmentize and recognize words successfully, the participants were expected to find two different levels of phonetic regularities. The stimuli constituted not only (a) a set of certain consonantal and vocalic sounds but also (b) a combination of five syllable patterns (i.e., V, CV CCV, CVC and CCVC).

3. According to our casual interview after the LLAMA-D test, the participants equally pointed out that they had not attempted to remember the sound strings at all during the sound check session, simply because they were instructed to engage in the initial listening session for the purpose of a sound check without any notification of the testing phase later on.

4. Although the same material was used at T1, T2 and T3, the test-retest effect in this study could be considered minimum in the current study, given the relative long interval of the tests (1 academic semester).

5. In the current study, segmental errors were operationalized as the lack of any learner effort to pronounce L2 sounds (substituting the L1 counterparts). To this end, we intentionally chose an L1 Japanese speaker with high-level proficiency in L2 English who could judge and differentiate the following three categories: (a) accurate and intelligible L2 English pronunciation, (b) unintelligible interlanguage forms, and (c) L1 Japanese pronunciation. It was only the last category that we coded as "errors." In other studies (e.g., Trofimovich & Isaacs, 2012), however, segmental errors have been defined as any deviations from native norms. In that case, L1 speakers could be a preferable option for coders, as they could intuitively but reliably make such dichotomous assessments (nativelike or non-nativelike).

6. When it comes to longitudinal studies in the field of SLA, many researchers' main interests lie in examining the extent to which any change in dependent variables between different data collection points could be ascribed to a range of independent variables. However, L2 learners' linguistic performance and learning backgrounds likely vary to a great degree at the beginning of projects. To statistically control for L2 learners' *initial* individual differences, and restrict all relevant analyses to their L2 development within specific time frameworks during particular research projects (e.g., pre- to post-tests), the learners' initial scores are typically used as a covariate (for a comprehensive overview, see Plonsky & Oswald, 2017). In the current study, we focused exclusively on the effects of aptitude and experience on the participants' pronunciation learning during Semesters 1 and 2 from such a longitudinal perspective. For a cross-sectional examination of the link between aptitude and L2 learners' attained L2 pronunciation performance, see Saito, 2017.

REFERENCES

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

Ahmadian, M. J., & Tavakoli, M. (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research*, *15*, 35-59.

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.

Bundgaard-Nielsen, R., Best, C., Kroos, C., & Tyler, M. (2012). Second language learners' vocabulary expansion is associated with improved second language vowel intelligibility. *Applied Psycholinguistics, 33,* 643-664.

Cebrian, J. (2006). Experience and the use of non-native duration in L2 vowel categorization. *Journal of Phonetics, 34*, 372-387.

Carroll, J. B. (1965). The contributions of psychological theory and educational research to the teaching of foreign languages. *The Modern Language Journal*, *49*, 273-281.

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

Darcy, I., Park, H., & Yang, C. L. (2015). Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing. *Learning and Individual Differences*, *40*, 63-72.

De Jong, N.H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2012). The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and nonnative speakers. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA* (pp. 121-142). Amsterdam: John Benjamins.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *12*, 1-16.

Derwing, T. M., Thomson, R. I., & Munro, M. J. (2006). English pronunciation and fluency development in Mandarin and Slavic speakers. *System*, *34*, 183–193.

Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics*, *29*, 359-380.

Gathercole, S. E., Frankish, C. R., Pickering, S. J., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 84-95.

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

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

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.

Gullberg, M., Roberts, L., Dimroth, C., Veroude, K., & Indefrey, P. (2010). Adult language learning after minimal exposure to an unknown natural language. *Language Learning*, *60*, 5-24.

Hulstijn, J. H. (2003). Incidental and intentional learning. In C. Doughty & M. H. Long (Eds.), *The handbook of second language research* (pp. 349–81). London, England: Blackwell.

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

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics*, *23*, 83-103.

Lahmann, C., Steinkrauss, R., & Schmid, M. S. (2017). Speed, breakdown, and repair: An investigation of fluency in long-term second-language speakers of English. *International Journal of Bilingualism*, *21*, 228-242.

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

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

Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition, 38*, 801-842.

Lin, Y. (2001). Syllable simplification strategies: A stylistic perspective. *Language Learning, 51,* 681-718.

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., & Doughty, C. J. (2013). Hi-LAB: A New Measure of Aptitude for High-Level Language Proficiency. *Language Learning*, *63*, 530-566.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive psychology*, *38*, 465-494.

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

Moyer, A. (2014). Exceptional outcomes in L2 phonology: The critical factors of learner engagement and self-regulation. *Applied Linguistics*, *35*, 418-440.

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

Muñoz, C. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Applied Linguistics, 35*, 463-482.

O'Brien, I., Segalowitz, N., Collentine, J., & Freed, B. (2006). Phonological memory and lexical, narrative, and grammatical skills in second language oral production by adult learners. *Applied Psycholinguistics*, *27*, 377-402.

Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition*, *39*, 579-592.

Riney, T., Takada, M., & Ota, M. (2000). Segmentals and global foreign accent: The Japanese flap in EFL. *TESOL Quarterly, 34*, 711-737.

Saffran, J. (2014). Sounds and meanings working together: Word learning as a collaborative effort. *Language learning*, *64*, 106-120.

Saito, K. (2014). Experienced teachers' perspectives on priorities for improved intelligible pronunciation: The case of Japanese learners of English. *International Journal of Applied Linguistics, 24*, 250-27

Saito, K. (2017). Effects of sound, vocabulary and grammar learning aptitude on adult second language speech attainment in foreign language classrooms. *Language Learning, 67*, 665-693.

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

Saito, K., & Hanzawa, K. (2017). The role of input in second language oral ability development in foreign language classrooms: A longitudinal study. *Language Teaching Research*. DOI: 10.1177/1362168816679030

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics, 37*, 217-240.

Schmid, M. S., Gilbers, S., & Nota, A. (2014). Ultimate attainment in late second language acquisition: Phonetic and grammatical challenges in advanced Dutch–English bilingualism. *Second Language Research*, *30*, 129-157.

Schneiderman, E. I. & Desmarais, C. (1988). The talented language learner: Some preliminary findings. *Second Language Research, 4*, 91-109.

Silbert, N. H., Smith, B. K., Jackson, S. R., Campbell, S. G., Hughes, M. M., & Tare, M. (2015). Non-native phonemic discrimination, phonological short term memory, and word learning. *Journal of Phonetics*, *50*, 99-119.

Skehan, P. (2016). Foreign language aptitude, acquisitional sequences, and psycholinguistic processes. In G. Granena, D. Jackson & Y. Yilmaz (Eds.), *Cognitive individual differences in L2 processing and acquisition* (pp. 15-38). Amsterdam: John Benjamins.

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

Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics, 25*, 293-321.

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

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

Yalçın, Ş., & Spada, N. (2016). Language aptitude and grammatical difficulty. *Studies in Second Language Acquisition*, *38*, 239-263.

Yilmaz, Y., & Granena, G. (2016). The role of cognitive aptitudes for explicit language learning in the relative effects of explicit and implicit feedback. *Bilingualism: Language and Cognition*, *19*, 147-161.

Yilmaz, Y. & Koylu, Y. (2016). The interaction between phonetic coding ability and feedback exposure condition. In G. Granena, D. Jackson, & Y. Yilmaz (Eds.), *Cognitive individual differences in L2 processing and acquisition* (pp. 303-326). Amsterdam: John Benjamins.

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.

Zhang, X., & Lu, X. (2013). A longitudinal study of receptive vocabulary breadth knowledge growth and vocabulary fluency development. *Applied Linguistics*, *35*, 283-304.

APPENDIX A

Training materials and onscreen labels for comprehensibility judgement

Comprehensibility	This term refers to how much effort it takes to understand what someone is saying. If you can understand with ease, then a speaker is highly comprehensible. However, if you struggle and must listen very carefully, or in fact cannot understand what is being said at all, then a speaker has low comprehensibility.

APPENDIX B

Segmental Analysis Procedure

As a part of a large project, we surveyed a list of challenging vowels and consonants for Japanese learners of North American English in conjunction with previous literature and expert teachers' opinions. For the current study, after the coder received instruction on these challenging English sounds for Japanese learners, she carefully listened to each speech sample to check whether a talker used a Japanese counterpart (the Japanese tap) or made any effort to pronounce target feature (e.g., English /ɹ/). In the table below, we summarized the number of the participants who made such substitution errors at three different testing points (T1, T2 T3).

In the existing assessment literature, L2 speech judgements are well-known to be influenced by listeners' previous relevant experience to a great degree, such as their familiarity with particular accents (e.g., Kennedy & Trofimovich, 2008) and the quantity and quality of prior L2 learning backgrounds (e.g., Winke, Gass, & Myford, 2013). In particular, non-native listeners (especially with higher L2 proficiency and experience levels) are found to demonstrate more sensitivity and leniency towards same L1 accented speech than native listeners (e.g., Imai, Walley, & Flege, 2005; Major, Fitzmaurice, Bunta, & Balasubramanian, 2002).

Following this line of thought, it is reasonable to assume that our listeners—native speakers of Japanese with high-level L2 English proficiency/experience (relative to native speakers of English and Japanese speakers with low-level L2 English proficiency/experience) would be considered suitable when it comes to assessing the Japanese college students' abilities to speak L2 English spontaneously. Based on their own L2 English learning experience, our L1 Japanese listeners could make reliable judgements on whether the Japanese students continued to use Japanese counterparts (e.g., substituting the Japanese tap sound for English /ɹ/ and /l/) or tried to

use any interlanguage forms (e.g., producing more tongue retraction and longer phonemic length for distinguishing English /ɹ/ from English /l/).

Note that the target of this global segmental analysis is comprehensive in nature (featuring various kinds of substitution errors). This is essentially different from focused analyses (see Saito, 2013 for more information on how to elicit and analyze specific segmentals at spontaneous speech levels). As has been observed in L2 grammar studies, denominators of such "general/comprehensive" (rather than "specific/focused") accuracy measures have been operationalized based on global units, such as the proportion of errors per 100 words (e.g., Révész, Ekiert, & Torgersen, 2016) and per clause (e.g., Yuan & Ellis, 2003).

Importantly, this global segmental analysis (i.e., the number of L1 substitution errors divided by the total number of segments articulated) has been widely used in L2 speech literature (e.g., Derwing & Munro, 1997; Isaacs & Trofimovich, 2012; Kang & Moran, 2014). For example, Trofimovich and Isaacs (2012) explained their global segmental analysis by using an example very specific to French learners of English (/θ/ in "think" mispronounced as /t/ in "tink"). In Trofimovich and Isaacs (2012) and our study alike, we targeted numerous target features and considered all segments to be potential contexts for substitution. Thus, we counted phonemic substitutions and divided by the total number of segments.

As one reviewer suggested, however, the denominators for the segmental error analyses could be the number of contexts/segments that could potentially be substituted (i.e., obligatory contexts); and we do agree that more methodological studies are called for with more sophisticated research designs and completely new dataset with a view of identifying the adequate denominators for the segmental analyses.

Table 1. Summary of Substitution Errors at the Beginning, Mid and Final Points of the Project

Problematic segmentals	T1		T2		T3	
	No. of participants	No. of errors	No. of participants	No. of errors	No. of participants	No. of errors
Vowels /ɪ, ʊ, æ, ʌ/	9	10	9	9	10	10
Diphthongs /aʊ, aɪ, oʊ, ɔɪ, eɪ/	3	3	1	1	4	4
Approximants /w, ɹ, l/	29	63	28	53	29	66
Nasals /n, ŋ/	0	0	0	0	0	0
Stops /p, t, k, b, d, g/	1	1	0	0	2	2
Fricatives /θ, f, ð, v, s, z/	29	84	29	88	30	87
Affricates /ʃi, ti/	0	0	0	0	0	0

References

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *12*, 1-16.

Imai, S., Walley, A., & Flege, J. (2005). Lexical frequency and neighborhood density effects on the recognition of native and Spanish–accented words by native English and Spanish listeners. *Acoustical Society of America*, *117*, 896-907.

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

Kang, O., & Moran, M. (2014). Functional loads of pronunciation features in nonnative speakers' oral assessment. *TESOL Quarterly*, *48*, 176-187.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *The Canadian Modern Language Review, 64,* 459-489.

Major, R., Fitzmaurice, S., Bunta, F., & Balasubramanian, C. (2002). The Effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*, *36,* 173-190.

Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, *37*, 828-848.

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

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.