Accepted for publication in *Language Learning* (Wiley Blackwell)

Effects of Sound, Vocabulary and Grammar Learning Aptitude on Adult Second Language Speech Attainment in Foreign Language Classrooms

Kazuya Saito[1]

Abstract

The current study examined the relationship between different type of language learning aptitude (measured via the LLAMA test) and adult second language (L2) speech production attainment in English-as-a-foreign-language (EFL) classrooms. Picture descriptions elicited from 50 Japanese EFL learners with varied proficiency levels were analyzed by a range of pronunciation, fluency, vocabulary and grammar measures. According to the results of the statistical analyses, the participants' aptitude test scores in phonemic coding, rote and associative memory and language analytic ability were moderately predictive of the phonological/morphological accuracy, speed fluency and lexicogrammar complexity of production—linguistic features thought to be instrumental to the acquisition of advanced L2 oral ability. In contrast, such aptitude-proficiency links were not found with respect to relatively implicit and incidental learning aptitude (sound recognition) and fundamental proficiency domains (the appropriate use of frequent words).

Key words: Foreign language aptitude, Pronunciation, Fluency, Vocabulary, Grammar

[1] I am grateful to Andrea Révész, Peter Skehan, the journal associate editor, Kara Morgan-Short and *Language Learning* reviewers for their constructive feedback on earlier versions of the manuscript. I also thank Takumi Uchihara, Keiko Hanzawa, Shungo Suzuki, Masaki Eguchi, George Smith and Ze Shen 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).

Effects of Sound, Vocabulary and Grammar Learning Aptitude on Adult Second Language Speech Attainment in Foreign Language Classrooms

Over the past 50 years, second language acquisition (SLA) researchers have extensively examined the role of foreign language aptitude in determining the rate and ultimate attainment of second language (L2) morphosyntactic performance in foreign language classrooms (Li, 2016; Skehan, 2015). To date, a growing amount of attention has been directed towards conceptualizing and elaborating the assessment framework for not only L2 learners' vocabulary and grammar but also pronunciation and fluency usage during *spontaneous* speech production (e.g., Housen, Kuiken, & Vedder, 2012; Trofimovich & Isaacs, 2012). By interfacing L2 aptitude and speech research perspectives, the current study was designed to take an exploratory approach towards examining the associations between sound, vocabulary and grammar learning aptitude, as measured by the LLAMA test (Meara, 2005), and the phonological (segmentals, word stress, intonation), temporal (breakdown, speed), lexical (appropriateness, richness) and grammatical (accuracy, complexity) components of spontaneous oral speech among 50 Japanese English-as-a-Foreign-Language (EFL) learners.

Background

Different from L2 learners in naturalistic environments, who have access a considerable amount of input on a daily basis, foreign language students typically receive only a few hours of instructional input per week, and their L2 use is extremely limited outside of classrooms (Muñoz, 2014). In fact, the nature of instructional treatment (e.g., grammar-translation, rote vocabulary memorization, intensive and extensive reading) is more likely form- than meaning-oriented, especially in Japanese EFL classrooms (the main focus of the study) (Nishino & Watanabe, 2008). Not surprisingly, the ultimate performance of these students after years of classroom experience is subject to much individual variability due to factors, such as the length of instruction, the amount of study-abroad experience, the current frequency of L2 use outside of classrooms, and the age of learning and testing (e.g., Muñoz, 2014). Among the many affecting variables, foreign language learning aptitude (henceforth aptitude) has been identified as a strong predictor of general L2 proficiency achievement in foreign language classrooms.

Aptitude refers to a set of cognitive and perceptual abilities which predict how learners can quickly improve their L2 performance (rate of learning) and the extent to which they can eventually approximate native-like performance (ultimate attainment) in classroom and naturalistic settings. Such aptitude is considered to be a relatively stable *trait* (rather than *skill*), regardless of previous L2 learning experience. One of the most well-known language aptitude tests, the Modern Language Aptitude Test (MLAT), was developed by Carroll and Sapon (1959). The four cognitive tasks in the test are assumed to tap into four essentially different dimensions of foreign language aptitude: (a) phonemic coding ability (analyzing and remembering unfamiliar sounds); (b) grammatical sensitivity (identifying the functions of words in a sentence); (c) inductive language learning ability (identifying patterns based on language samples); and (d) rote and associative memory (remembering new word form-meaning pairings).

According to early validation studies, MLAT scores were found to predict students' final course grades, teachers' evaluations and SAT scores (e.g., Carroll, 1965). These studies have generally indicated that aptitude is strongly related to L2 learners' short-term success especially in the *initial* stages of L2 learning, and in foreign language classroom settings. More recently, L2 aptitude researchers have begun to conceptualize aptitude as multifaceted (rather than singular)

construct (e.g., explicit and conscious vs. implicit and incidental language learning aptitude), resulting in the redevelopment and validation of the existing test batteries themselves (Li, 2016; Skehan, 2015, 2016 for comprehensive reviews). For example, Linck, Hughes, Campbell, Silbert, Tare, Jackson, and Doughty's (2013) Hi-LAB, consisting of 11 computer-delivered cognitive tasks, was designed to test various areas of cognitive and perceptual aptitude spanning executive functioning in working memory, phonological short term memory, associative memory, long term memory retrieval, implicit learning, processing speed, and auditory perceptual acuity. The results of the three tasks—phonological short term memory (letter span), associative memory (paired association) and implicit learning (serial reaction time)—successfully distinguished very high levels of L2 reading and listening proficiency.

Another widely-cited aptitude test is the LLAMA test (Meara, 2005), which was used in the current investigation. Loosely adapted from MLAT, the LLAMA test is a computer-based aptitude test using picture visuals and verbal materials adapted from a British-Columbian and Central-American indigenous language. The test comprises of the four subtest components including rote vocabulary learning (LLAMA-B), sound recognition (LLAMA-D), sound-symbol associations (LLAMA-E) and grammar inferencing (LLAMA-F). According to Granena's (2013a) partial validation study with Spanish-Chinese bilinguals, the three components of the test (i.e., LLAMA-B, E, F) were loaded with their cognitive test scores measuring explicit language learning ability (e.g., general intelligence), and their LLAMA-D scores with their other cognitive scores measuring implicit learning ability (e.g., serial reaction time).

Although the further validation of the LLAMA test is strongly called for, scores on the LLAMA test have been found to predict a wide range of L2 phenomena in the field of SLA. For instance, the scores can be correlated with the extent to which L2 learners can improve their L2 performance in foreign language classrooms, when they receive explicit instruction (e.g., Roehr, 2008) and corrective feedback (e.g., Yilmaz, 2013). Furthermore, the LLAMA test scores have successfully predicted experienced L2 learners' proficiency levels in naturalistic settings. The Swansea LAT test (the original version of the LLAMA test) was correlated with the near-nativelike performance of Swedish-Spanish bilinguals who started learning their L2 after puberty (Abrahamsson & Hyltenstam, 2008). Using the LLAMA-D, Granena (2013b) found that their sound learning ability was a significant predictor for determining early Spanish-Chinese bilinguals' morphosyntactic comprehension ability.

Based on his synthesis of an extensive body of aptitude literature, Skehan (2016) proposed that different constructs of aptitude can be uniquely tied to putative stages in SLA—input processing, noticing, pattern identification, automatization and lexicalization. By aligning exiting aptitude constructs to how L2 learners actually acquire target language during naturalistic interaction (processing input for meaning → relating new patterns to existing systems → producing language in response to task demands), Skehan attempted to further enhance the validity of the construct. For example, phonemic coding may help L2 learners hold and analyze unfamiliar sounds in an efficient, timely fashion (Carroll, 1965), suggesting that this aptitude is relevant with L2 learners' ability to process, notice and analyze input at the initial stages of SLA. As language analytic ability (covering both grammatical inferencing and sensitivity) may allow L2 learners to grasp some wider grammatical and lexical structures in language, this aptitude is hypothesized to promote the identification of patterns and restructuring of the existing system in the mid stages of SLA (e.g., Yilmaz, 2013). Finally, Skehan pointed out that the later stage of SLA, characterized by the automatization of existing language knowledge, may be related to rote and associative memory and/or implicit learning aptitude (e.g., Linck et al., 2013).

Motivation for Current Study

Although much attention has been given towards the complex relationship between various kinds of aptitude (explicit vs. implicit), L2 proficiency (beginner, intermediate, advanced) and contexts (classroom vs. naturalistic settings), it is noteworthy that most of the relevant research evidence has been exclusively concerned with the role of aptitude in the learning of L2 morphosyntax. Linguistic proficiency has been typically measured via comprehension tasks (e.g., grammatical judgements: Abrahamsson & Hyltenstam, 2008; Granena, 2013b) and/or controlled production tasks (e.g., sentence readings: Roehr, 2008). To date, very few studies have examined the extent to which aptitude can contribute to determining L2 adult learners' phonological (pronunciation and fluency) *and* morphosyntactic (vocabulary and grammar) abilities while speaking spontaneously (Granena, 2013a).

In the field of L2 speech research, a growing number of scholars have attempted to conceptualize adult L2 speech learning model which can explain how L2 learners are to develop various areas of their oral abilities as a function to their increased experience and practice (e.g., Saito, 2015). Following the componential view of L2 speech, L2 oral ability in this line of research is generally conceptualized as a composite phenomenon of numerous linguistic skills spanning pronunciation (segmentals, prosody), fluency (breakdown, speed), vocabulary (appropriateness, richness) and grammar (accuracy, complexity) (De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012). According to the recent literature, there is ample longitudinal evidence showing that L2 learners selectively work on certain linguistic features with more communicative and learning value. Inexperienced L2 learners tend to focus, in particular, on the phonological intelligibly of frequent words (Munro & Derwing, 2008) without too many dysfluencies (Derwing, Munro, Thomson, & Rossiter, 2009) and on their appropriate use in context (Schmitt, 1998) for daily communication purposes, given that they cover a great deal of spoken discourse (Adolph & Schmitt, 2000). In order to attain advanced L2 oral proficiency, learners are expected to acquire more varied and infrequent words without relying on the simple repetition of frequent ones (Crossley, Salsbury, & McNamara, 2015). In addition, they need to produce them with more refined pronunciation (Bundgaard-Nielsen, Best, Kroos, & Tyler, 2012) and more accurate/complex morphology at more optimal tempo (Mora & Vall-Ferrars, 2012)—linguistic features more strongly tied to native speakers' impressionistic judgements of nativelikeness (Trofimovich & Isaacs, 2012).

Building on this line of L2 aptitude and speech research, the current study was designed to revisit the extent to which different types of aptitude relates to different dimensions of long-term L2 oral ability development in foreign classrooms. Using the LLAMA test battery, the current study drew on Meara's (2005) aptitude for L2 learning framework, which comprises rote and associative memory (LLAMA-B), sound recognition (LLAMA-D), phonemic coding (LLAMA-E) and grammatical inferencing (LLAMA-F). These components of the LLAMA differ from a methodological point of view. LLAMA-B, E and F intend to measure L2 learners' ability to learn vocabulary, pronunciation and grammar with awareness and intention, as learners are explicitly encouraged to practice tasks by using any problem-solving strategies prior to the tests. On the other hand, LLAMA-D was designed to measure L2 learners' ability to learn pronunciation without awareness and intention, as they proceed with the testing sessions without being told the focus of the test, and do not go through any study phase.

In the current study, adult L2 learners' speech was first elicited via a spontaneous speaking task (picture narrative) and then analyzed via a set of comprehensive measures tapping into various domains of L2 speech, such as pronunciation (segmentals, word stress, intonation:

Trofimovich & Isaacs, 2012), fluency (speed, breakdown: De Jong et al., 2012), vocabulary (appropriateness, richness: Reed, 2000) and grammar (accuracy, complexity: Housen et al., 2012). Through this, the current study took an exploratory approach to providing preliminary data on the potential relationship between aptitude and L2 speech achievement in foreign language classrooms. The study's research question and hypotheses were thus formulated as follows:

Research question: Whether, to what degree and how are rote and associative memory (LLAMA-B), sound recognition (LLAMA-D), phonemic coding (LLAMA-E) and grammatical inferencing (LLAMA-F) associated with different dimensions of L2 speech achievement in foreign language classrooms?

Hypotheses: In light of Skehan's (2016) process-oriented model of aptitude reviewed above, certain types of aptitude may directly help L2 learners encode and access phonological, lexical and grammatical information during speech production in an effective and efficient fashion. Given that the aforementioned L2 speech literature has indicated that various dimensions of L2 oral ability development entail different levels of learning difficulty (e.g., Saito, 2015), it was thus predicted that aptitude effects could be clearly observed especially in the acquisition of more difficult, nativelikeness-related features (the accurate, fluent and complex use of phonological and morphological forms) rather than more learnable, comprehensibility-related features (the appropriate use of frequent words without too many pauses). More specifically, the following predictions were formulated: (a) participants with higher phonemic coding ability (LLAMA-E) would make the most of limited, language-focused input in foreign language classrooms, and thus attain accurate use of pronunciation and morphology in L2 speech; (b) those with higher language analytic ability (LLAMA-F) would demonstrate more diverse, sophisticated and complex lexicogrammar usage beyond an overreliance on frequent words and simple grammatical structures; and (c) those with high-level rote and associative memory (LLAMA-B) and incidental sound recognition (LLAMA-D) would attain enhanced fluency via increased control over already acquired knowledge.

Method

Participants

The project was widely advertised at a university in Tokyo, Japan with the primary purpose of surveying Japanese college-students' English oral proficiency across the school. A total of 50 young adult L2 learners (23 males, 27 females) were recruited from various disciplines at the university (e.g., business, economics, international communication, liberal arts). All of them were native speakers of Japanese (they were raised by Japanese speaking parents from birth onward), and second year university students at the time of the project who had studied English for seven years since Grade 7 exclusively in foreign language classroom settings without any experience abroad. The foreign language learning profiles of the participants were similar in terms of the length of instruction (seven years), the age of learning (since Grade 7), the chronological age at the time of testing (18-19 years old), the amount of study-abroad (zero), and the current frequency of L2 use (highly limited outside of classrooms) (Muñoz, 2014).

According to their general English proficiency test scores (i.e., TOEIC), however, these participants showed much variability in listening and reading (M = 653.4 out of 990, SD = 97.6, range = 500 to 890). The distribution of TOEIC scores indicated that their CEFR bands could be considered from B1/B2 (Independent users) to C1 (Proficient users).

All of the data collection sessions took place individually in a quiet room at the university. They first took the aptitude test (LLAMA), then engaged in a speaking task (picture narrative), and finally completed a language background questionnaire. The entire session took approximately 50 minutes per participant.

Aptitude Test (LLAMA)

The LLAMA test consists of four subtests measuring four domains of L2 aptitude. The entire session took approximately 30 minutes in the following order: LLAMA-D → B → E → F. The tests were automatically scored out of 75 for the LLAMA-D and 100 for LLAMA-B, E and F.

LLAMA-D. The LLAMA-D measures the ability to recognize items after listening to sound strings only once without any practice. To avoid any intentional learning during the listening sessions, the participants were asked to listen to 10 sound strings just to check if they could normally hear them played from computer. Unlike the other subtests (LLAMA-B, E, F), the participants were not notified that they would be tested for recall.[1] Next, they moved onto a recognition test, where they listened to 30 items and then detected whether they had heard them during the sound check session. A casual interview was conducted with each participant, and none reported their intention and action to learn new sound strings while they were going through the sound check session, as they had not been given any explicit instruction to do so.

LLAMA-B. The LLAMA-B measures the ability to learn written forms of new vocabulary items by associating word strings with pictures (similar to the paired-associates test in MLAT). Unlike the LLAMA-D, the participants were explicitly told about the purpose of the test (i.e., vocabulary learning followed by recollection). The participants first learned as many words as possible by drawing on their rote and associative memory under time pressure (two minutes), clicking on images to display the names of various objects. In the following testing phase, they were asked to correctly associate the names of randomly chosen objects with the correct picture.

LLAMA-E. The LLAMA-E measures the ability to learn new sound-symbol correspondences (phonemic coding ability) by associating sound strings with unfamiliar alphabetical symbols (similar to the phonetic script test in MLAT). The participants first engaged in a two-minute timed study phase where they clicked on 24 different symbols and were asked to remember their corresponding sound strings (one syllable per symbol). Subsequently, they took a test to see whether they could correctly identify orthographic representations after listening to a combination of two syllables.

LLAMA-F. The LLAMA-F measures the ability to induce the grammatical rules of an unfamiliar language (similar to the grammatical sensitivity task in the MLAT). During the study phase, the participants were given five minutes to infer grammatical rules based on 20 pictures with sentences; then they were to choose a grammatically correct sentence (out of two choices) for each of 20 pictures randomly displayed by the program.

Internal consistency. The test scores were automatically scored for the LLAMA-D (out of 75) and LLAMA-B, E and F (out of 100). According to Cronbach alpha analyses, the internal consistency of the four subtests were .72 for the LLAMA-D (k = 30), .69 for the LLAMA-B (k = 20), .73 for the LLAMA-E (k = 20), and .77 for the LLAMA-F (k = 20). In line with L2 research standards, the Cronbach's alpha values here could be considered as an “acceptable” level (Larson-Hall, 2010).

Oral Task

To elicit spontaneous speech, SLA researchers have claimed that tasks should be designed to induce L2 learners to pay a primary attention to message conveyance rather than the accurate production of linguistic forms (Spada & Tomita, 2010). Examples of such free constructed tasks include story retellings (Yilmaz & Granena, 2015), oral interviews (Abrahamsson & Hyltenstam, 2008), and monologues (Derwing, Rossiter, Thomson, & Munro, 2004). In the current project, the participants' spontaneous speech was elicited via one of the most commonly-used speaking tasks—picture narrative. This task was selected for the purpose of comparisons, as it has been used extensively in previous L2 speech studies as a way to elicit sufficiently long spontaneous speech samples for pronunciation (Trofimovich & Isaacs, 2012), fluency (Derwing et al., 2004) and lexicogrammar (e.g., Saito, Webb, Trofimovich, & Isaacs, 2016) analyses.

Procedure. After completing the LLAMA tests, the participants proceeded to the oral narrative task. First, the participants were allowed to spend one minute familiarizing themselves with an eight-image picture sequence about two strangers bumping into each other on a city street corner and switching their suitcases by accident. Then, they described the picture cartoon without any time constraint. All 50 picture descriptions were recorded using a digital Marantz PMD 660 audio recorder (44.1-kHz sampling rate with 16-bit quantization), and normalized for peak amplitude.

Material preparation. With respect to L2 pronunciation analyses, the first 30 seconds of each speech sample were extracted and stored in a single WAV file. These audio files were assessed by human judges for segmental and prosodic (word stress, intonation) accuracy. This methodological decision (30 sec per file) was chosen to avoid any conflating effects of listener fatigue (which could influence the quality of L2 subjective ratings: Flege & Fletcher, 1993) and is consistent with standards in L2 speech research (e.g., Trofimovich & Isaacs, 2012).

To provide enough linguistic information for the objective analyses of L2 fluency, vocabulary and grammar usage, the full-length recordings (M = 2 min 42 s, range = 2 min 5 sec-4 min 24 sec) were orthographically transcribed. The raw transcripts were then cleaned by removing obvious mispronunciations based on contextual information available in the pictures (e.g., "on the *load"* was transcribed as "on the *road*"), and orthographic markings of pausing (e.g., uh, um, oh, ehh). The word length of the 50 picture descriptions substantially varied (M = 117.8 words, range = 57-208 words).

Pronunciation Analyses

In the current study, I chose not to use objective measures (e.g., acoustic analyses) to analyze the speech produced in the L2 picture narrative task, as they tend to be highly sensitive to variability in phonetic context (e.g., following and preceding vowels) and talker characteristics (e.g., anatomical difference in vocal tract). An alternative option to evaluating the phonological qualities of *spontaneous* L2 speech involves using linguistically trained raters' subjective scalar judgements. Such a procedure has been commonly used for analyzing segmentals (e.g., Piske, Flege, MacKay, & Meador, 2011) and prosodic (e.g., Derwing et al., 2004) aspects of spontaneous L2 speech. In Saito, Trofimovich and Isaacs' (2015) validation study, a training procedure was elaborated in order to help experienced native speaking raters (with much linguistic and pedagogical experience) evaluate three rater-based categories: (a) segmentals (substitution, omission, or insertion of individual consonant and vowel sounds); (b) word stress (misplaced or missing primary stress); and (c) intonation (appropriate, varied use of pitch moves).

Expert raters. According to the definition of expert raters in Isaacs and Thomson (2013), five native speaking raters (3 females, 2 males) were recruited at an English speaking university in Montreal. All of them were graduate students in Applied Linguistics (M_{age} = 29.5 years, *range* = 26-37 years) and reported a great deal of experience with L2 pronunciation analyses (either via enrollment in a semester-long course on applied phonetics and pronunciation teaching or participation in L2 speech projects as research assistants). They had taught English for several years in various second and foreign language classrooms ($M_{teaching\ experience}$ = 6 years, range = 3-14.5 years). None of them reported any hearing problems. Their familiarity with Japanese accented speech was relatively high ($M_{familiarity}$ = 5.5, range = 5-6) on a 6-point scale (*1 = not at all, 6 = very much*).

Procedure. Each session took place individually in a quiet room at the university. First, each rater received thorough explanation from a trained research assistant on the three categories—segmentals, word stress and intonation. The raters listened to speech samples played in a randomized order via a custom software (developed using MATLAB 8.1, The MathWorks Inc., Natick, MA, 2013), and then used the moving slider to rate them on a 1000-point scale for segmental errors (*0 = frequent, 1000 = infrequent or absent*); word stress errors (*0 = frequent, 1000 = infrequent or absent*), and intonation (*0 = unnatural, 1000 = natural*)—see a similar approach often used in L2 speech research (e.g., Flege, Munro, & MacKay, 1995). Due to the demands of the task, they were allowed to listen to each speech sample as many times as they wanted to. For training scripts, see Supporting Information.

After familiarizing themselves with the picture sequence, the raters practiced the rating procedure with three samples (not used in the main analyses). For each sample, they were asked to explain their decisions, and received feedback from a trained research assistant. Finally, they moved onto the pronunciation judgements of the main dataset (i.e., 50 speech samples). The length of each rating session lasted for two hours with a 5-minute break halfway through.

Post-task questionnaire. After completing all rating sessions, the raters assessed the extent to which they understood the rated categories on a 9-point scale (*1 = "I did not understand at all", 9 = "I understand this concept well"*). Their understanding of the rated categories was relatively high (M = 9 for segmentals, 8.2 for word stress and 8.5 for intonation).

Rater consistency. According to Cronbach's alpha analyses, the five expert raters yielded relatively high alpha values for segmentals (α = .91), word stress (α = .90), and intonation (α = .89). Their scores were therefore considered sufficiently consistent, and were averaged to derive a single score per rated category for each speaker.

Fluency Analyses

The 50 full-length audio files and written transcripts were scrutinized for the temporal qualities of L2 speech according to the oft-used notion of breakdown and speed in L2 fluency research (e.g., De Jong et al., 2012).

Breakdown fluency. This subcategory refers to how effortlessly speech is articulated (i.e., without many pauses and hesitations). Two types of pause ratio were measured by dividing the number of filled (lexical fillers such as eh, oh) and unfilled (silence) pauses by the total number of words. Whereas the number of filled pauses was counted based on raw transcripts, the number and length of unfilled silent pauses was automatically calculated via a script programmed in *Praat* with the minimum silence duration set to 350 milliseconds.

Speed fluency. This subcategory refers to how many words/syllables are produced within a certain period of time. Speech rate was measured by dividing the total length of each audio file by the total number of words. Articulation rate was measured by dividing actual speaking time

by the total number of syllables. The speaking time was measured by subtracting all these silent pauses from the total length of each audio file.

Vocabulary and Grammar Analyses

The 50 cleaned transcripts were scrutinized in line with Read's (2000) model of L2 vocabulary use (i.e., appropriateness, richness), and Housen et al.'s (2012) framework for L2 grammar knowledge (i.e., complexity, accuracy). Whereas the lexical appropriateness, morphological accuracy and grammatical complexity analyses were conducted by a linguistically trained coder, the lexical richness analyses were performed via *Coh-Metrix* (McNamara, Graesser, McCarthy, & Cai, 2014).

For the former analyses (lexical appropriateness, morphological accuracy and grammatical complexity), two trained coders first received thorough instruction on each lexicogrammar category. Then, both of them separately practiced the coding procedure by using 10 oral narratives which were not included in the current study. For the practice set, the coders' performance was submitted to interrater agreement for lexical appropriateness (r = .95), morphological accuracy (r = .93), and grammatical complexity (r = .88). One of them thus proceeded to analyze the main dataset (50 oral narratives).

Lexical appropriateness. This subcategory refers to how L2 learners use vocabulary in a contextually and conceptually appropriate manner. Lexical appropriateness was measured based on the ratio of lexical errors (e.g., "attach" instead of "bump into," "uniform" instead of "clothes") to the total number of words.

Lexical richness. This subcategory refers to L2 learners' capacity to use a wide range of sophisticated and infrequent words at the surface level. The richness domain of L2 lexical competence was automatically measured via textual lexical diversity (MTLD) and the average frequency of all words based on the CELEX corpus of English in *Coh-Metrix* (McNamara et al., 2014).[2]

Grammatical accuracy. This subcategory refers to how accurately L2 learners use morphological markers in context. The accurate use of morphology was measured based on the ratio of morphological errors in verb (tense, aspect, modality, and subject-verb agreement), noun (plurals) and article (definite, indefinite, and non-articles) to the total number of words.

Grammatical complexity. This subcategory refers to L2 learners' ability to deliver information by way of advanced and complex syntactic structures. In keeping with Norris and Ortega's (2009) suggested framework, the grammatical complexity domain was calculated based on the clauses per Analysis of Speech (AS) unit (i.e., subordination complexity) as well as the total number of words per clause (i.e., subclausal complexity).

Results

Statistical Profiles of Aptitude Test

Table 1 summarizes the 50 Japanese EFL learners' aptitude scores according to the four subtests (LLAMA-D, B, E and F). The participants attained relatively high scores (M > 83.4) on the LLAMA-E (sound-symbol correspondence) compared to other domains (M = 50-60). Next, a set of correlation analyses were conducted to examine the strength of the interrelationships between the participants' aptitude scores in the four subtests (LLAMA-D, B, E and F). Due to the relatively small sample size of participants (N = 50), any statistical analyses of the dataset (which also noted much individual variability) could be affected by any outliers. Thus, Spearman nonparametric correlations were chosen as a more conservative statistical method. According to Table 2, none of the contrasts reached statistical significance at a p < .05 level. This indicates that there were not any associations among these subtests, at least, in the content of the 50

Japanese learners in this study. As suggested by Meara (2005), therefore, their subtest scores could be considered to resemble sound recognition ability (LLAMA-D), rote and associative ability (LLAMA-B), phonemic coding ability (LLAMA-E) and language analytic ability (LLAMA-F).

Table 1
Descriptive Statistics of Aptitude Test Scores

	M	*SD*	Range		95% CI	
			Min	Max	Lower	Upper
Language aptitude						
LLAMA-D (75 points)	41.8	13.2	10	70	38.1	45.4
LLAMA-B (100 points)	60.0	17.3	25	100	55.2	64.7
LLAMA-E (100 points)	83.4	19.4	10	100	77.8	88.7
LLAMA-F (100 points)	57.4	24.0	10	100	50.7	64.0

Table 2
Interrelationships between the Four Subtests (D, B, E, F) in LLAMA

	LLAMA-B		LLAMA-E		LLAMA-F	
	r	*p*	*r*	*p*	*r*	*p*
LLAMA-D	.05	.70	-.06	.67	.11	.43
LLAMA-B			.20	.14	.10	.46
LLAMA-E					.15	.28

Aptitude-Proficiency Link

Raw proficiency scores × aptitude. As summarized in Table 3, the participants demonstrated a wide range of phonological, temporal, lexical and grammatical abilities. A set of Spearman correlation analyses were performed to examine the extent to which L2 learners' aptitude scores (LLAMA-D, B, E and F) were linked with pronunciation, fluency, vocabulary and grammar scores. To adjust for multiple comparisons, the alpha level was set at $p = .0125$ via the Bonferroni correction. Correlations with *p* values of more than .0125 but less than .05 were considered as marginally significant. The magnitude of the significant and marginally significant correlations was also considered in consultation with Plonsky and Oswald's (2014) field-specific benchmarks ($r = .25$ for small, .40 for medium, .60 for large). According to Table 4, L2 learners' LLAMA-E (sound-symbol correspondence) significantly related to one of the pronunciation measure (segmentals); LLAMA-B (rote and associative ability) to the complexity measure (clause to AS-unit ratio); and LLAMA-F (language analytic ability) to the two lexical richness measures (diversity and frequency) ($p < .0125$). Additionally, LLAMA-E was marginally significantly correlated with the other pronunciation (word stress, intonation) and morphological accuracy measures; and LLAMA-B to the fluency measure (articulation rate) ($p < .05$). These correlation coefficients could be considered within the small-to-medium range ($r = .25$-$.40$) (Plonsky & Oswald). In contrast, such significant aptitude-proficiency links were not found for LLAMA-D (sound recognition) and the lexical appropriateness (lemma error ratio) measures.

Table 3

Descriptive Statistics of Oral Ability Measures (1000 points)

	M	*SD*	Range		95% CI	
			Min	Max	Lower	Upper
Pronunciation						
Segmentals	423	155	203	844	381	466
Word stress	439	137	218	777	402	478
Intonation	438	149	240	778	397	479
Fluency						
Filled pause ratio	0.08	0.07	0	0.30	0.06	0.10
Unfilled pause ratio	0.60	0.20	0.28	1.16	0.55	0.66
Speech rate	62.7	13.8	36.9	93.8	58.8	66.7
Articulation rate	107.0	25.8	50.9	176.2	99.6	114.4
Lexical appropriateness						
Vocabulary error ratio	.091	.050	.011	.205	.077	.105
Lexical richness						
Diversity	36.4	9.5	19.2	63.0	33.8	39.1
Average frequency levels	2.52	.134	2.27	3.05	2.49	2.56
Grammatical accuracy						
Morphology error ratio	.058	.026	.017	.152	.051	.066
Grammatical complexity						
Subordination	10.9	2.87	6.2	18.8	10.15	11.72
Clause length	5.71	.77	3.79	7.22	5.49	5.93

Factor proficiency scores × aptitude. Although the results above provided an overall picture of the relationship between aptitude and proficiency, they need to be interpreted with caution, because the 13 oral ability measures may have overlapped with each other, resulting in multicollinearity problems and increasing the risk of type one error. In this subsection, an exploratory factor analysis was first conducted with Oblimin rotation to identify any patterns underlying the multiple proficiency measures. Following Loewen and Gonulal's (2015) field-specific recommendations, two steps were followed to determine the number of groupings including the utmost variance in the participants' oral abilities. Given that the cumulative percentage of explained variance reported in L2 research is typically 60-65% and considered relatively low, the threshold for the current analyses was set to 80%. Accordingly, a cut-off point for eigenvalues was set to .07—the Jolliffe criterion (explaining 82.0%) rather than 1.0—the Kaiser criterion (explaining 66.7%). The factorability of the entire dataset was examined and validated via two tests: the Bartlett's test of sphericity ($\chi^2 = 389.58$, $p < .001$) and the Kaiser-Meyer-Olkin measure of sampling adequacy (.752). A decision was made to identify a "five-factor" solution with eigenvalues beyond .07 which accounted for 82.0% of the total variance in the participants' L2 speech performance.

Table 4

Spearman Correlations between Aptitude and Raw Proficiency Scores

	LLAMA-D		LLAMA-B		LLAMA-E		LLAMA-F	
Pronunciation								
Segmentals	r = .18	p = .20	r = .18	p = .20	r = .40	p < .01**	r = .16	p = .24
Word stress	r = .17	p = .22	r = .18	p = .20	r = .28	p = .04*	r = .17	p = .22
Intonation	r = .23	p = .09	r = .24	p = .87	r = .30	p = .03*	r = .23	p = .09
Fluency								
Filled pause ratio	r = .03	p = .78	r = -.13	p = .35	r = -.15	p = .27	r = .04	p = .74
Unfilled pause ratio	r = -.09	p = .49	r = -.20	p = .15	r = -.09	p = .52	r = -.01	p = .96
Speech rate	r = .08	p = .56	r = .22	p = .11	r = .11	p = .41	r = .06	p = .66
Articulation rate	r = .20	p = .16	r = .31	p = .02*	r = .19	p = .18	r = .24	p = .08
Lexical appropriateness								
Vocabulary error ratio	r = -.13	p = .33	r = -.03	p = .78	r = -.08	p = .54	r = -.12	p = .38
Lexical richness								
Diversity	r = .08	p = .54	r = .15	p = .28	r = .07	p = .60	r = .36	p < .01**
Average frequency levels	r = -.07	p = .61	r = -.16	p = .24	r = -.24	p = .09	r = -.37	p < .01**
Grammatical accuracy								
Morphology error ratio	r = .22	p = .11	r = .01	p = .98	r = -.28	p = .04*	r = -.02	p = .85
Grammatical complexity								
Subordination	r = .20	p = .16	r = .38	p < .01**	r = .27	p = .05	r = .07	p = .61
Clause length	r = .04	p = .75	r = .08	p = .56	r = .12	p = .41	r = -.10	p = .45

Note. * denotes marginal significance (p < .05) and ** denotes statistical significance (p < .0125)

Each factor was carefully interpreted based on the strongest loadings which are summarized in Table 5. In line with Hair, Anderson, Tatham, and Black's (1998) guidelines, the cut-off value for the "practically" significant factor loadings was set to 0.4. Factor 1 was labeled as "pronunciation," as the first three strongest loadings comprised the pronunciation measures. Factor 2 was labeled as "fluency" (covering all the fluency measures). Factor 3 was labeled as Lexicogrammar accuracy and complexity, because the relevant measures (grammatical accuracy/complexity, lexical appropriateness) were loaded to this category. Factors 4 and 5 were labeled as "lexical frequency" and "lexical diversity," respectively, reflecting the loadings of each factor (lexical frequency/diversity). The five groups suggested here roughly correspond to the broad conceptualization of L2 oral ability in the study (pronunciation, breakdown and speed fluency, lexical appropriateness and richness, and grammatical accuracy and complexity).

Table 5. *Summary of a Five-Factor Solution Based on a Factor Analysis of the 13 Oral Ability Variables*

Factor 1 (Pronunciation)	Segmental errors (.927), word stress (.896), intonation (.881), complexity subordination (-.770)
Factor 2 (Fluency)	Unfilled pauses (.854), articulation rate (-.827), speech rate (-.808), filled pauses (.524)
Factor 3 (Lexicogrammar accuracy and complexity)	Grammatical accuracy (.882), grammatical complexity length (-.771), Lexical appropriateness (.429)
Factor 4 (Lexical frequency)	Lexical frequency (.846)
Factor 5 (Lexical diversity)	Lexical diversity (.998)

Note. All eigenvalues > .07.

Finally, spearman correlation analyses were conducted to explore the relationship between the participants' aptitude scores (LLAMA-D, B, E and F) and five factor oral ability scores (Factors 1-5). As shown in Table 6, significant associations were found between LLAMA-E and Factor 1 (pronunciation) and LLAMA-F and Factor 5 (lexical diversity) ($p < .0125$, Bonferroni corrected). The strength of the relationship between LLAMA-B and Factor 2 (fluency) was marginally significant ($p = .03$). These aptitude-proficiency links could be considered as small-to-medium (Plonsky & Oswald, 2014). On the contrary, LLAMA-D was not significantly related to any oral ability factors ($p > .0125$).

Table 6

Spearman Correlations between Aptitude and Factor Proficiency Scores

	LLAMA-D		LLAMA-B		LLAMA-E		LLAMA-F	
Pronunciation	r = .15	p = .29	r = .24	p = .08	r = .33	p = .01**	r = .27	p = .06
Fluency	r = .13	p = .35	r = .29	p = .03*	r = .25	p = .07	r = .10	p = .48
Lexicogrammar accuracy/complexity	r = .14	p = .32	r = .02	p = .84	r = -.24	p = .09	r = .26	p = .25
Lexical frequency	r = .06	p = .64	r = .08	p = .56	r = .22	p = .12	r = .13	p = .36
Lexical diversity	r = .02	p = .88	r = .09	p = .51	r = .08	p = .56	r = .41	p < .01**

Note. * denotes marginal significance ($p < .05$) and ** denotes statistical significance ($p < .0125$)

Discussion

Whereas the role of aptitude in L2 morphosyntax learning has been extensively discussed in the field of SLA, the ultimate impact of aptitude on L2 learners' spontaneous speaking ability has remained open to investigation. Building on the L2 speech assessment framework in recent speech literature (e.g., Housen et al., 2012; Trofimovich & Isaacs, 2012), the current study was designed to examine the extent to which various components of language aptitude—incidental/implicit sound recognition (LLAMA-D), rote vocabulary memorization (LLAMA-B), sound-symbol correspondence (LLAMA-E) and language analytic ability (LLAMA-F)—could predict various dimensions of L2 oral ability. These dimensions include correct pronunciation of words and sentences (segmentals, prosody intonation), fluency (speed, breakdown), accurate and complex lexicogrammar usage (vocabulary/morphology appropriateness, the amount of subordination), and vocabulary richness (diversity, frequency). Specifically, the current study examined this topic in the context of 50 college-level Japanese learners of English in foreign language classrooms.

In keeping with Plonsky and Oswald's (2014) benchmarks, the results of the study identified the following medium-to-small correlations between different types of aptitude and L2 speech. The participants' phonemic coding ability (LLAMA-E) was correlated with their accurate use of pronunciation (r = .28-.40) as well as morphology (r = -.28). Their rote and associative ability (LLAMA-B) was correlated with the speed fluency (articulation rate) and grammatical complexity (clause to AS-unit ratio) aspects of L2 speech (r = .31, 38). Their language analytic ability (LLAMA-F) was correlated with vocabulary richness—diversity (r = .36) and average frequency levels (r = -.37). Taking the interdependence of the multiple linguistic measures into account, a factor analysis successfully reduced 13 into 5 oral ability measures—pronunciation, fluency, lexicogrammar accuracy and complexity, diversity, and frequency. Based on such factor proficiency scores, the results of the correlation analyses found the medium-to-small associations, in particular between phonemic coding (LLAMA-E) and pronunciation; and between language analytic ability (LLAMA-F) and lexical richness. The relationship between rote and associative memory (LLAMA-B) and speed fluency was marginally significant. Finally, none of the aptitude tests (including LLAMA-D) were significantly related to vocabulary appropriateness.

As predicted earlier, one possibility for the differential effects of aptitude on L2 speech learning could be well explained by Skehan's (2016) proposal on the multifaceted roles of aptitude in the psycholinguistic sequence of SLA. According to Skehan (2016), three different components of aptitude—phonemic coding (LLAMA-E), language analytic ability (LLAMA-F) and rote and associative memory (LLAMA-B)—may separately impact the development of accuracy, complexity and fluency, as each of these language abilities is differentially related to three different stages of interlanguage development (noticing → patterning/restructuring → automatization). Overall, the findings of the current study supported these predictions.

First, participants with higher LLAMA-E scores tended to have better phonological and grammatical accuracy in their speech. This could be arguably because phonemic coding ability could help L2 learners notice and intergrade new linguistic knowledge more efficiently via optimizing the processing of incoming linguistic input. As conceptualized in the aptitude literature (Carroll, 1969; Meara, 2005), phonemic coding ability allows L2 learners to deconstruct words into phonetic units, and analyze the form (pronunciation, morphology) and meaning aspects of words separately. This explicit analysis of words is believed to result in the

ability to create, refine and access robust lexical representations during the production of spontaneous speech (Nation & Webb, 2011).

The suggested interpretations here—the interaction between phonemic coding, noticing and accuracy—are pertinent to extensive research on the efficacy of various components of phonological aptitude (partially overlapping with the construct of phonemic coding) on numerous domains of L1 and L2 speech learning. In the L1 literature, it has been claimed that children substantially rely on their ability to remember phonological sequences in their short-term memory (i.e., phonological memory) in order to correctly process incoming L1 input (Baddeley, Gathercole, & Papagno, 1998). In the context of adult SLA, learners' differential capacity to mimic, analyze and remember novel sound patterns serves as a good predictor for their pronunciation and grammar accuracy in production in both foreign language (e.g., O'Brien, Segalowitz, Collentine, & Freed, 2006) and naturalistic (e.g., Granena & Long, 2013) settings.

Interestingly, the participants' phonemic coding ability was not clearly associated with their appropriate use of vocabulary relative to phonological and morphological accuracy. This in turn indicates that the absence/presence of aptitude effects could be related to the amount of learning difficulty inherent to individual linguistic structures (vocabulary vs. pronunciation/grammar). According to previous literature on the rate and ultimate attainment in SLA, even a short period of immersion (e.g., < 1 year of study-abroad) likely results in the successful learning of certain aspects of L2 speech, notably the appropriate use of frequent words (Munro & Derwing, 2008; Schmitt, 1998). Arguably, this is because these features most directly impact successful meaning conveyance in communicative situations (Nation & Webb, 2011). In contrast, other aspects of adult L2 learners' speech performance are resistant to rapid change, such as pronunciation and morphology. As suggested by many researchers (e.g., Muñoz, 2014), adult L2 learners may need special language capacities (including phonemic coding) to achieve successful L2 learning in foreign language classrooms, as they are instrumental to the acquisition of not only basic dimensions of L2 oral ability (vocabulary), but also relatively difficult features (pronunciation and morphology) in acquisition-limited conditions (a few hours of L2 input per week).

Next, the results (i.e., the association between LLAMA-F and lexical richness) suggest that language analytic ability may promote the restructuring of existing knowledge. Language analytic ability (grammatical inferencing/sensitivity) helps L2 learners quickly grasp the grammatical information of words (i.e., how words are formed in a phrase and sentence level), find broad patterns in language input, and modify their interlanguage forms accordingly (Carroll, 1965). While speaking spontaneously, high-aptitude learners are assumed to allocate their limited cognitive resources towards expanding their linguistic repertories by using more advanced and sophisticated words with complex grammatical structures (Skehan, 2016). As shown in the current study, the participants' LLAMA-F scores were correlated with the degree of lexical richness in their speech production. This relationship between language analytic ability and lexical richness has also been documented in L2 writing research, where it has been shown that grammatically sensitive L2 learners are more likely to display diverse and sophisticated word choice rather than the frequent repetition of fundamental words (e.g., Booth, 2014).

Different from earlier predictions, however, the participants' LLAMA-F scores were not linked to their grammatical complexity (nor accuracy) performance. The lack of any significant correlations between grammar aptitude and achievement concurs with Li's (2016) meta-analysis, which similarly showed that language analytic ability (grammar inferencing/sensitivity) has relatively low predictive validity for general L2 proficiency. This result also echoes Kormos and

Trebits's (2012) empirical findings that language analytic ability is strongly associated with L2 grammatical complexity, especially when it comes to written production, where L2 learners have ample time to use relatively complex grammar structures. Whereas the results of the current study indicated that L2 learners' efficient detection of grammatical information in words leads to lexical (rather than grammatical) restructuring, such statements are considered as tentative at best, and warrant future research using various aptitude (e.g., MLAT, PLAB) and linguistic (e.g., overall vs. subordination vs. length) measures under different task conditions (narratives vs. interviews) and modes (oral and written).

Finally, participants who displayed relatively high rote and associative memory (LLAMA-B) demonstrated their ability to retain large amounts of lexical information by using multiple phrases within a single AS-unit, and express it at a faster speed (i.e., had greater fluency). In the field of SLA, increased fluency is considered as "an automatic procedural skill" (Schmidt, 1992, p. 358), as it comprises a crucial part of automatization, where L2 behaviours are executed accurately (i.e., qualitative change) and rapidly (i.e., quantitative change) (Segalowitz, 2010). Extending the thoughts of Skehan (2016), therefore, I tentatively argue that rote and associative memory may relate to the reinforcement and automatization of already-acquired knowledge. The suggested links between rote and associative memory and automatized performance are in consistent with Linck et al.'s (2013) findings this quality of aptitude (together with phonological short-term memory and implicit language learning aptitude) could successfully predict L2 learners' high-level attainment in the domain of listening comprehension. Furthermore, the fact that aptitude was related to speed fluency (articulation rate) but not to breakdown fluency (the number of filled/unfilled pauses) also lent support to many researchers' claims that breakdown and speed fluency are essentially two different phenomenon (Housen et al., 2012) with the former dimension more strongly related to L1 traits (e.g., personal speaking styles) and the latter more clearly linked to other L2 proficiency measures (e.g., vocabulary knowledge) (De Jong, Groenhout, Schoonen, & Hulstijn, 2015).

It is important to remember here that all of the significant aptitude effects were found only in LLAMA-B, E and F, all of which engaged the participants in a study phase where they had a clear understanding of the focus of each test. Given that these tests are assumed to measure the participants' explicit and intentional language learning capacities (e.g., Yilmaz & Granena, 2013), the findings suggest a match between the nature of the participants' aptitude and their prior learning experience. Specifically, all of the participants had studied English for seven years in a foreign language setting without any opportunity to study abroad or use the language in naturalistic settings. Since decontextualized instructional techniques, such as grammar-translation and rote vocabulary memorization, are dominant in Japanese EFL classrooms (Nishino & Watanabe, 2008), the participants' L2 learning experience was inevitably explicit, language-focused, and analytical in nature. To this end, it is not surprising that only the explicit and intentional learning aptitude components (LLAMA-B, E, and F) were correlated with explicit L2 speech learning in EFL classrooms.

The question then becomes: Why was LLAMA-D unrelated to the linguistic qualities of the participants' L2 speech production? In LLAMA-D, novel sound recognition ability is tested without any explicit study phase; such a methodological aspect of the test has induced some researchers to hypothesize that the test taps into "a cognitive ability involving more implicit cognitive processes" compared to LLAMA-B, E and F (e.g., Granena, 2013a, p. 124). Therefore, the fact that LLAMA-D failed to predict successful L2 learning in EFL classrooms suggests that

implicit and incidental aptitude may have little impact on the process and product of explicit and intentional learning.

At the same time, however, any existing implicit aptitude measures (including LLAMA-D) still remain open to validation, elaboration and refinement. In the field of SLA, implicit learning is thought to occur when learners do not demonstrate any learning intention or awareness of acquired knowledge (Rebuschat, 2013). To measure L2 learners' implicit learning ability, some studies have begun to introduce aptitude tests firmly rooted in cognitive psychology, such as serial reaction time (e.g., Granena, 2013b), semantic priming (e.g., Linck et al., 2013) and phonological sequences (e.g., Speciale, Ellis, & Bywater, 2004). Though few in number, these studies have provided some evidence that such implicit aptitude scores may be predictive of, in particular, highly experienced L2 learners' ultimate attainment after years of immersion in an L2 speaking environment (Granena, 2013b).

Interestingly, Skehan's (2016) review pointed out that the above-mentioned implicit aptitude tests (e.g., serial reaction time) exclusively draw on the domain-general perspective in psychology (that language learning is a generally-learned behaviour), as these tests consist of non-linguistic materials (e.g., visual cues [asterisks, numerals]). However, Skehan argued that language learning aptitude also needs to be measured via *linguistic* materials, arguably because certain aspects of SLA (e.g., input processing, noticing, pattern identification, automatization and lexicalization) could be domain-specific. According to Skehan's classification, one candidate for such a language-based aptitude test is actually LLAMA-D, where participants hear, categorize and remember audio linguistic samples unconsciously and incidentally. This could, in theory, simulate implicit sound recognition processes in the context of first and second language acquisition. As a promising future direction for L2 aptitude and speech research, it would be intriguing to explore the extent to which a range of implicit aptitude tests, both non-linguistic (e.g., serial reaction time) and language-based (e.g., LLAMA-D), are associated with adult L2 speech learning not only under limited (classroom), but also acquisition-rich (immersion) contexts.

Limitations

Due to the exploratory nature of the study, several methodological limitations need to be acknowledged for the sake of future replication. In this study, the participants' aptitude and oral ability was measured based on a single aptitude test (LLAMA) and speaking task (picture description). Accordingly, the results of the study need to be replicated with more comprehensive aptitude test batteries (e.g., Hi-LAB: Linck et al., 2013) and various kinds of task modalities (e.g., simple/complex topics in descriptive/argumentative discourse contexts: De Jong et al., 2012). In addition, future studies of this kind need a larger sample size ($N > 100$) for more robust and sophisticated statistical analyses, such as Structural Equation Modeling. Such statistical analyses will allow us to detect any latent constructs (e.g., "aptitude," "proficiency") among the observed variables (4 aptitude scores, 13 oral ability measures), and the overall relationship between these constructs (e.g., aptitude → oral abilities).

Another limitation of the study is that the findings on the aptitude-proficiency links in the current study were exclusively based on second-year Japanese college students without any experience abroad. Any above-mentioned discussion needs to be replicated with a larger sample with various L1/L2 learning, age, immersion, instrumental and integrative motivation profiles in instructed and naturalistic settings. It is noteworthy that the study assumed the participants had homogeneous L2 backgrounds at the time of the project (7 years of EFL experience without any

experience abroad). However, we did not precisely investigate how many hours and what kind(s) of L2 instruction they had received in junior high school, high school and university, and to what degree these students were motivated to study L2 English at different points in time. Given that little is known about the long-term development of L2 oral development especially in EFL classrooms (several hours of instruction per week) (cf. Muñoz, 2014), we still need to wait for future studies which corroborate and untangle the complex relationship between the quantity, quality and intensity of instruction, learner motivation, language learning aptitude, and various linguistic domains of classroom L2 speech learning.

Thirdly, L2 oral ability in the current study was broadly conceptualized and discussed in terms of fundamental/learnable linguistic features (related to comprehensibility), and sophisticated/difficult linguistic features (related to linguistic nativelikeness). However, other L2 speech studies have found that the amount of learning difficulty also varies within the domains of segmentals (dissimilation or assimilation: Bundgaard et al., 2012), suprasegmentals (melody vs. tempo: Trofimovich & Baker, 2006), vocabulary (abstract, hypernymic and polysemous relations of words: Crosseley et al., 2015) and morphological accuracy (noun and third person plurality vs. tense vs. article: Bardovi-Harlig & Comajoan, 2008). Future research at the micro level is needed to provide a full-fledged picture of the effects of aptitude on the acquisition of easy-difficult linguistic features.

Finally, it is crucial to acknowledge that the current study was correlational in nature. Thus, any suggestions regarding the influence of aptitude on L2 speech learning patterns need to be replicated using longitudinal designs (cf. O'Brien et al., 2006). For example, follow-up studies need to corroborate the extent to which high- and low-aptitude learners can actually improve their speaking proficiency at various stages of L2 speech learning over short (1-3 years) and long (> 5 years) periods of time. This longitudinal evidence would be particularly crucial for the thorough validation of the LLAMA test, since it claims to predict various aspects of SLA.

Conclusion

In the context of adult Japanese EFL learners with relatively homogeneous L2 learning experiences, the current study examined the relationship between different types of language learning aptitude (measured via the LLAMA test) and L2 spontaneous speech production. The results demonstrated that explicit sound, vocabulary and grammar learning aptitude plays a significant role in determining the extent to which these learners can attain advanced-level oral ability by improving, in particular, their command of difficult linguistic features potentially related to three different developmental stages of SLA. High-aptitude L2 learners can rely on phonemic coding ability to produce words with refined phonological and morphological accuracy (noticing of new L2 features); on language analytic ability to choose various sophisticated and infrequent words (restructuring of existing lexical routines); and on rote/associative memory for delivering much information with complex structures at a rapid speed (automatization). On the contrary, relatively implicit and incidental aptitude, such as sound recognition, may be less relevant to adult L2 speech learning in foreign language classrooms, where instructional input is limited in both quality (devoid of much meaningful use of language) and quantity (a few hours per week).

Footnote

1. This decision was made in order to correspond to a concern that Granena (2013b) raised: the LLAMA-D would encourage test takers to use conscious and intentional learning strategy if they were informed that their recollection was to be tested upon the listening session.

2. Among many other diversity measures (including D), MTLD was chosen, because it has been found to be least sensitive to sample length (Koizumi & In'nami, 2012).

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. doi: 10.1017/S027226310808073X

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics, 24*, 425-438. doi: 10.1093/applin/24.4.425

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105,* 158–173. doi: 10.1037/0033-295X.105.1.158

Bardovi-Harlig, K., & Comajoan, L. (2008). Order of Acquisition and Developmental Readiness. In B. Spolsky and F. M. Hult (Eds.). The Handbook of Educational Linguistics (pp. 383-397. Malden, MA: Blackwell.

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. doi : 10.1017/S0142716411000518

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. doi: 10.2307/322133

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

Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics, 36*, 570-590. doi: 10.1093/applin/amt056

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics, 36*, 223-243. doi: 10.1017/S0142716413000210

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. Investigating complexity, accuracy and fluency in SLA* (pp. 121-142). Amsterdam: John Benjamins.

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

Derwing, T. M., Munro, M. M., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition, 31,* 533–557. doi: 10.1017/S0272263109990015

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. doi: 10.1121/1.402780

Flege, J., Munro, M, & MacKay, I. R. A. (1995). Factors affecting degree of perceived foreign accent in a second language. *Journal of the Acoustical Society of America, 97*, 3125-3134. doi: 10.1121/1.413041

Foster, P, Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language. *Applied Linguistics, 21*, 354–375. doi: 10.1093/applin/21.3.354

Granena, G. (2013a). Cognitive attitudes for second language learning and the LLAMA Language Aptitude Test. In G. Granena & M. Long (Eds.), *Sensitive periods, language aptitude and ultimate attainment* (pp. 3-41). Amsterdam: John Benjamins**.**

Granena, G. (2013b). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning, 63*, 665-703. doi: 10.1111/lang.12018

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. doi: 10.1177/0267658312461497

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.) Upper Seddle River, New Jersey, USA: Prentice-Hall International, Inc.

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. John Benjamins Publishing.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10*, 135-159. doi: 10.1080/15434303.2013.769545

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis. I: Artificial data. *Applied statistics,* 160-173.

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. doi: 10.1016/j.system.2012.10.012

Kormos, J., & Trebits, A. (2012). The role of task complexity, modality, and aptitude in narrative task performance. *Language Learning, 62*, 439-472. doi: 10.1111/j.1467-9922.2012.00695.x

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.* doi: 10.1017/S027226311500042X

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. doi: 10.1111/lang.12011/abstract

Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principal components analysis. In Plonsky, L. (Ed), *Advancing quantitative methods in second language research*. New York: Routledge.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

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

Muñoz, C. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Applied Linguistics, 35*, 463-482. doi: 10.1093/applin/amu024

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly, 46*, 610-641. doi: 10.1002/tesq.34

Munro, M. & Derwing, T. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. *Language Learning, 58,* 479-502. doi: 10.1111/j.1467-9922.2008.00448.x

Nation, I.S.P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle.

Nishino, T., & Watanabe, M. (2008). Communication-oriented policies versus classroom realities in Japan. *TESOL Quarterly, 42*, 133-138.

Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*, 555–578. doi: 10.1093/applin/amp044

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. doi: 10.1017/S0142716406060322

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

Plonsky, L., & Oswald, F. L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning, 64,* 878-912. doi: 10.1111/lang.12079

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning, 63*, 595-626. doi: 10.1111/lang.12010

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Roehr, K. (2008). Metalinguistic knowledge and language ability in university-level L2 learners. *Applied Linguistics, 29*, 173-199. doi: 10.1093/applin/amm037

Saito, K. (2015). Experience effects on the development of late second language learners' oral proficiency. *Language Learning, 65*, 563-595.

Saito, K., Trofimovich, P., & Isaacs, T. (2015). Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*. doi: 10.1093/applin/amv047

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness and sense relations. *Studies in Second Language Acquisition, 37*, 677-701. doi: 10.1017/S0272263115000297

Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in second language acquisition, 14*, 357-357.

Schmitt, N. (1998). Tracking the incremental acquisition of a second language vocabulary: A longitudinal study. *Language Learning, 48*, 281–317. doi: 10.1111/1467-9922.00042

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Skehan, P. (2015). Foreign language aptitude and its relationship with grammar: A critical review. *Applied Linguistics, 36*, 367-384. doi: 10.1093/applin/amu072

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. doi: 10.1111/j.1467-9922.2010.00562.x

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. doi: 10.1017/S0142716404001146

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. doi: 10.1017/S0272263106060013

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition, 15*, 905-916. doi: 10.1017/S1366728912000168

Yilmaz, Y. (2013). Relative effects of explicit and implicit feedback: The role of working memory capacity and language analytic ability. *Applied Linguistics, 34,* 344-368. Doi: 10.1093/applin/ams044