To be published in *Second Language Research* (SAGE)

How Does Longitudinal Interaction Promote Second Language Speech Learning? Roles of Learner Experience and Proficiency Levels

Kazuya Saito[1]
Shungo Suzuki
Tomoko Oyama
Yuka Akiyama

Abstract

This study examined how longitudinal interaction impacts the development of second language (L2) oral proficiency in relation to learners' different experience and proficiency levels. Japanese English-as-a-Foreign-Language learners participated in weekly conversation exchanges with native speakers (NSs) in the US via videoconferencing tools over one academic semester (12 weeks). The participants' spontaneous speech, elicited from a story telling task before and after the treatment, was analyzed via a set of linguistic measures. In line with the componential view of L2 oral proficiency (De Jong et al., 2012) and development (Bundgaard-Nielsen et al., 2011), our results hinted L2 learners' experience and proficiency levels as a mediating factor for determining the link between interaction and its impact on different dimensions of L2 speech learning. While the longitudinal interaction equally improved the participants' grammatical complexity and articulation rate—a fundamental component for defining L2 oral proficiency, the development of less experienced/proficient learners was observed across a wide range of lexicogrammar and fluency features (lexical appropriateness/richness, grammatical accuracy, pause ratio). It was only more experienced/proficient learners that significantly enhanced phonological accuracies (segmentals, word stress) which are thought to gradually develop in the later stages of L2 speech learning. These findings add another piece of evidence for the differential effects of long-term interaction relative to L2 learners' developmental stages.

Key words: Interaction, Feedback, Second Language Speech, Computer Assisted Language Learning

[1] We are grateful to Second Language Research reviewers for their insightful comments on the early version of our manuscript. We also acknowledge Masaki Eguchi, Kokoro Muramoto, Takumi Uchihara, and Ethan Beaman for their help for data collection and analyses The project was funded by the Japan Foundation Research Grant.

Over the past 40 years, one of the most extensively researched topics in the field of second language acquisition (SLA) has been the role of conversational interaction in language development. The current study took a longitudinal approach to examine the extent to which two groups of Japanese English-as-a-Foreign-Language learners—those with ample experience overseas and higher second language (L2) proficiency compared to inexperienced and less proficient ones—could develop multiple dimensions of L2 speech (i.e., pronunciation, fluency, lexicogrammar), when interacting with interlocutors in the US for one academic semester through video-conferencing tools. It is crucial to note here that we refer to "experience" and "proficiency" interchangeably in this paper. As we detailed below, our intention here concurs with the assumptions underlying many L2 speech learning theories that experience and proficiency are strongly tied to each other and that more accumulative conversational experience leads to improved L2 proficiency (i.e., experience effects) (e.g., Bundgaard-Nielsen, Best, & Tyler, 2011; Mackey, 2012; Flege, 2009).[2]

Background

Interaction Effects in SLA

In the field of SLA, few researchers would disagree with the fundamental idea that L2 learners improve their proficiency (beginner → intermediate → advanced) through meaningful conversation experience with native speakers (NSs) and other non-native speakers (NNSs). Interaction provides many opportunities to impact various aspects of SLA processes, especially when interlocutors encounter communication breakdowns attributable to language and work together on solutions. NSs aim to retrieve meaning from NNSs' incomprehensible speech by using several negotiation strategies, such as repetition, confirmation checks and clarification requests (i.e., comprehensible input). In addition, they may signal comprehended yet erroneous speech by recasting NNSs' erroneous productions (i.e., interactional feedback). Finally, NNSs may actively seek assistance from NSs when it comes to linguistic features (e.g., vocabulary) that they have not understood (i.e., self-initiated negotiation for meaning) (for a summary of the interactionist paradigm in SLA, see Mackey, 2012).

From theoretical perspectives, such improved L2 oral proficiency is a multifaceted phenomenon. For instance, a componential view posits that L2 oral proficiency consists of a

[2] However, we do acknowledge certain cases, in which experience does not necessarily relate to L2 speech learning. For example, some L2 learners may choose to use their L1 (instead of L2) during their stay in an L2 speaking environment (Martinsen et al., 2010); and highly experienced L2 learners' attained speech performance becomes relatively stable and unchanged regardless of additional experience (Flege, 2009).

combination of multiple subskills in the areas of pronunciation, fluency, vocabulary and grammar (De Jong, Steinel, Florijn, Schoonen, & Hulstijin, 2012). Extant empirical evidence has pointed out that the relative contributions of these subskills to global proficiency differ depending on the learners' proficiency level. On a broad level, the appropriate and fluent use of lexicogrammar serves as a crucial linguistic element in differentiating between beginner and intermediate levels; and pronunciation accuracy is instrumental in distinguishing between intermediate and advanced levels (e.g., Iwashita, Brown, McNamara, & O'Hagan, 2008 for TOEFL; Isaacs & Trofimovich, 2012; Saito, Trofimovich, & Isaacs, 2016 for perceived comprehensibility). If we take the stance that L2 speech learning occurs on a continuum of global proficiency as a function of increased experience (beginner → intermediate → advanced), these cross-sectional findings suggest that L2 learners develop different aspects of language at different stages of L2 speech learning (enhanced lexicogrammar appropriateness/fluency → pronunciation refinements).

Turning to psycholinguistic literature, there is a consensus that L2 speech learning is initially lexically-driven, followed by an increase in phonological sophistication in the long run. One such theoretical account is Bundgaard-Nielsen, Best, and Tyler's (2011) vocabulary tuning model which states that less experienced/less proficient L2 learners mainly process word-sized units of their L2 input as minimum meaningful chunks of language. Through this, these learners can grasp the overall message of L2 speech in the most efficient and effective way, while simultaneously accelerating the expansion of their own vocabulary size (i.e., lexical explosion). When L2 learners have sufficient conversational experience (e.g., length of residence > 1 year) and/or become more proficient (e.g., vocabulary size > 6,000 word families), they start paying attention to the phonetic details of L2 input, gradually filling in gaps in their abilities with more target-like pronunciation forms. This word-to-sound re-attunement is crucial for L2 learners to realize the phonetic/articulatory features that do not exist in their first language (L1). As such, they can quickly, accurately, and reliably differentiate words that would otherwise sound identical based on their L1 phonological system (for similar theoretical accounts, see also Bradlow & Pisoni, 1999).

Empirical Evidence

To date, many empirical studies have longitudinally examined how multiple dimensions of L2 learners' oral performance change in naturalistic settings. Within periods of short immersion (e.g., study-abroad), it has been demonstrated that L2 learners' vocabulary use becomes more appropriate (Schmitt, 1998), more diverse (Muñoz, 2010), and more fluent (Segalowitz & Freed, 2004). L2 learners also tend to become more capable of conveying their intended message by using a wider variety of syntactic structures (Vercellotti, 2017) and by containing more morphologically accurate forms (Mora & Valls-Fellar, 2012). There could still be improvements in vocabulary and morphosyntax at higher proficiency. For example, even advanced L2 learners have been reported to show difficulty acquiring certain lexical features

which are less communicatively important and salient (Saito, 2019), more semantically complex and abstract (Zareva, Schwanenflugel, & Nikolova, 2012), and more infrequent and context-specific (Kyle & Crossley, 2015). When it comes to the phonological accuracy aspects of language (i.e., pronouncing sounds and words correctly without L1 substitutions), however, any changes are unlikely to happen so rapidly. This is arguably due to the possibility that L2 pronunciation learning may require an extensive amount of L2 experience (Flege, 2009; Saito & Brajot, 2013) and/or language-focused instruction (Derwing, Munro, Foote, Waugh, & Fleming, 2014).

Though revealing, the results reviewed above need to be interpreted with caution. Although many studies have monolithically quantified participants' experience profiles according to the overall length of immersion in an L2 speaking environment, the extent to which, with whom, where and how learners use a target language varies widely (Martinsen, Baker, Dewey, Bown, & Johnson, 2010). Certain studies have attempted to document both the quantity and quality of the interaction that L2 learners actually experienced via interviews and self-reports. Yet, such results are purely based on participants' retrospection, and thus could be subject to error. As Flege (2009) pointed out, due to much variability among the participants themselves, the exact nature of L2 interaction is extremely difficult to track when investigations last for a prolonged period of time (cf. Ranta & Meckelborg, 2013).

Motivation for Current Study

Recently, SLA researchers have begun to explore how L2 learners can develop their linguistic performance through interacting with NS interlocutors using video-conferencing tools. Such computer-assisted conversational activities have become increasingly popular, especially in foreign language classrooms (for a comprehensive review, see Chun, Kern, & Smith, 2016). From a methodological point of view, this specific research setting—video-based interaction in foreign language classrooms—could be considered as a unique testing ground for the longitudinal analysis of L2 interaction. Unlike naturalistic settings, where L2 learners access ample opportunities to use a target language on a daily basis, under foreign language conditions L2 use in communicatively authentic contexts is limited. Thus, by introducing video-based conversation activities to foreign language students (who rarely engage in interactions outside classrooms), researchers can experimentally control and track the quantity and quality of L2 conversation experience throughout their research project.

To advance our knowledge on this topic, we conducted a preliminary study concerning the effect of longitudinal video-based interaction on the L2 oral proficiency development of inexperienced Japanese learners of English (Saito & Akiyama, 2017). The learners participated in weekly, dyadic task-based interaction activities with NS interlocutors in the US over one academic semester. According to the results of this study, the participants significantly developed in their overall fluency, vocabulary and grammar skills, but failed to show significant changes surrounding phonological accentedness.

Though revealing, the research raised several issues worthy of future investigation: To provide a full-fledged picture of the experience effects on SLA, it is important to acknowledge that the findings in the precursor study were exclusively concerned with inexperienced and low-proficient Japanese learners (with little background for using L2 English for conversational purposes nor any experience overseas). The generalizability of the findings should be further tested with different L2 populations, such as more advanced L2 learners with more experience of L2 oral communication.

As reviewed earlier, NNSs selectively work on different areas of L2 oral proficiency as a result of their increased L2 experience (lexicogrammar appropriateness and fluency → pronunciation refinements) (i.e., Bungaard et al., 2011; Isaacs & Trofimovich, 2012; Saito et al. 2016). Previous research has also pointed to learners' proficiency levels as a significant predictor for determining the level of effectiveness surrounding interactions. (e.g., Mackey & Philp, 1998).

Consequently, it is predicted that L2 interaction could "differentially" impact the development of L2 learners' speech according to their different experience and proficiency levels. In regards to less experienced/less proficient learners, interactional gains may be clearly achieved in lexicogrammar and fluency features that likely entail a great deal of acquisitional potential during the early phases of L2 speech learning—lexical appropriateness and richness, grammatical accuracy and complexity, and fluency (Issacs & Trofimovich, 2012; Saito et al., 2016). Regarding more experienced/proficient learners, some learning may occur with those linguistic features that are hypothesized to develop later in the developmental trajectory—refinements in pronunciation (Flege, 2009; Saito & Brajot, 2013).

To further pursue this crucial topic for theory building and pedagogical relevance, the current study aimed to reveal how L2 learners with different experience/proficiency levels could benefit from increased conversational input. Pairing Japanese learners of English (NNS learners) with American learners of Japanese (NS interlocutors), our overall goal was to analyze the phonological (segmentals, suprasegmentals), temporal (speed, breakdown) and lexicogrammatical (appropriateness, complexity) development of two groups of Japanese learners (Experienced vs. Inexperienced), who engaged in video-mediated interactions with NSs over one academic semester. Their development patterns were compared to a comparison group who did not participate in such long-term interaction activities. The following research questions were formulated:

1. Does the nature of interaction (the number of linguistic errors, feedback, uptake) vary when more and less experienced/proficient L2 learners engage in conversation activities with NS interlocutors?

2. Does longitudinal interaction differentially impact pronunciation, fluency and lexicogrammar aspects of more and less experienced/proficient L2 learners' oral proficiency?

Method

In the field, a range of laboratory studies have been conducted to test the effectiveness of face-to-face and video-based interaction activities in foreign language settings. According to Mackey and Goo's (2007) research synthesis, however, one crucial methodological limitation of previous studies concerns the brevity of interactional treatment (*M* = 31.9 min, *Range* = 5-60 min). Given that the effectiveness of interaction appears to be larger in delayed rather than immediate tests (e.g., Gass & Varonis, 1994), more research is needed to capture the potentially substantial impact of L2 interaction on the long-term L2 development from a longitudinal perspective. The project was set up as a semester-long language exchange program between universities in Japan and the US (12 weeks: 30 minutes × 9 sessions + one orientation + pre/post sessions). Following Ortega and Iberri-Shea's (2005) guidelines, the study could be considered as "longitudinal" in nature, as it meets the three crucial conditions of such research design:

- Multiple sessions: The participants in the current study were involved in multiple sessions over time (i.e., 9 weekly sessions over one academic semester), as opposed to previous L2 interaction studies which typically involved only a brief amount of interactional treatment (*M* = 30min) (Mackey & Goo, 2007).

- Multiple data collection points: The current study adopted multiple data collection points with pre- and post-tests for measuring acquisition (Weeks 1 and 12) and video-recordings of the first and last conversational sessions for measuring interactional patterns (Weeks 4 and 10; 2nd and 8th interaction sessions) (for session schedule, see the section of Treatment).

- Multiple types of analyses: The analyses were designed to tap into both process (i.e., how the participants engaged in interaction) and product (i.e., how much interaction was facilitative of the participants' speech learning).

Participants

As a part of a larger project, the current study recruited 30 Japanese learners of English at universities in Tokyo (as NNS learners) and 20 undergraduate students at universities in the USA (as NS conversational partners). As detailed below, the *n* = 10 inexperienced NNSs were the same as those in our precursor research (Saito & Akiyama, 2017); and the data was used as a point of comparison. In the current study, two new groups (*n* = 20 experienced and comparison NNSs) participated. In addition, all the participants provided new production data (sufficiently long for robust lexicogrammar and fluency analyses) via a story telling task.

Experienced vs. Inexperienced NNSs. A total of 20 Japanese students were carefully selected and divided into two groups (Experienced, Inexperienced). The group distinction was determined through two different standards: (a) immersion experience; and (b) general proficiency.

First, the participants were categorized into "experienced" or "inexperienced" depending on the presence/absence of immersion experience in an English-speaking environment. This standard was necessary as immersion experience allowed learners to access ample opportunities to use the target language for meaningful purposes (for a similar methodological decision, see Flege, 2009). As observed in many instructed SLA studies (e.g., Lambert, Kormos, & Minn, 2017), participants' general proficiency scores, as measured by TOEIC, were also used as another index for the group distinction (for the importance of proficiency test scores in L2 interaction research, see also Plonsky & Kim, 2016). In comparison with the CEFR benchmarks, the threshold was set at 700 (out of 990). The TOEIC scores of those in the experienced group were the equivalent of those of Independent to Proficient Users (B2-C1), while those in the inexperienced group were considered to be Basic to Independent Users (A2-B1).

- **Experienced NNSs.** As summarized in Table 1, the participants in the experienced group enrolled in several hours of EFL classes in their institutes each week at the same time as the project, with the exception of one student who did not take any language-related lessons. None of them reported any experience at private language schools where they could practice conversational English with NSs, indicating that their L2 use outside of the classrooms was substantially limited—common learner profiles of Japanese (and many other East Asian) EFL students. All the experienced learners had resided in English speaking countries for longer than one month (e.g., US, UK, Australia) (*Range* = 1-48 months). Their general proficiency scores were relatively high (*Range* $_{\text{TOEIC}}$ = 700-950).

- **Inexperienced NNSs.** The participants in this group were the same as those in our precursor research (Saito & Akiyama, 2017), wherein a total of 15 inexperienced learners were originally recruited. All the participants in this group had learned English just through EFL education in Japan and had no experience abroad prior to the project. However, we had to eliminate five participants (out of 15) for the subsequent analyses, because they did not fit the definition of "Inexperienced" in the current study (TOEIC scores > 700).

Table 1
Learner Profiles of 30 NNS learners

	Experienced (*n* = 10)			Inexperienced (*n* = 10)			Comparison (*n* = 10)		
	M	*SD*	*Range*	*M*	*SD*	*Range*	*M*	*SD*	*Range*
Age	20.5	3.0	18-26	18.8	0.6	18-20	18.6	0.8	18-20
Age of learning (years)	11.5	1.9	6-13	11.7	0.4	11-12	11.5	0.4	11-12
Total hours of L2 English classes per week during the project	4.3	4.3	0-10	3.9	1.1	2.5-4.5	2.7	0.6	1.5-3
Length of residence in English speaking countries (months)	12.3	17.3	1-48	0	*n.a.*	*n.a.*	0	*n.a.*	*n.a.*
General English proficiency test scores (TOEIC)[a]	815.8	77.7	700-950	596.5	72.4	400-650	497	84.5	400-550

Note. [a]TOEIC consists of reading and listening components with a total score of 990 points.

Comparison NNSs. In this current project, including a comparison group was crucial for the following reasons. As we detailed below, all the participants used the same materials at both pre- and post-tests (so as to provide sufficiently comparable L2 speech for linguistic analyses). Examining the comparison group, who took pre- and post-tests without any interaction treatment, was assumed to reveal the presence and absence of test-retest effects in such research design. Additionally, any improvement observed in the experimental groups of experienced and inexperienced learners could be ascribed not only to the interaction treatment, but also to the EFL instruction that the participants received within the timeframe of the project. Therefore, to check and separate the effects of one-semester's EFL instruction, we analyzed the L2 development of a similar population as a comparison group—i.e., a total of 10 Japanese university students who studied English only in EFL classrooms within the same timeframe. All the participants were enrolled in approximately three hours of EFL classes per week (M = 2.7 hours, *Range* = 1.5-3) but none of them engaged in video-mediated oral communication.

In light of the aforementioned standards, the comparison group's proficiency could be considered comparable to the Inexperienced Group as none of them had any experience abroad and they reported relatively low scores in TOEIC (400-550). Kruskal-Wallis tests showed that the three groups (Experienced, Inexperienced, Comparison) were comparable in terms of age (z = -1.34, p = .218), age of learning (z = -1.19, p = .393), and total hours of EFL classes during the project (z = -.231, p = .858). According to the participants' self-reports, the quality of EFL instruction was considered similar, with their classes being highly language-focused and void of conversation/speaking activities without interaction—typical of EFL education in Japan and generally referred to as focus on formS in instructed SLA literature (Loewen, 2014).

NS Interlocutors. A total of 20 NSs of American English (M_{age} = 20.65) who were studying Japanese as a foreign language at US universities at the time of the project participated . While some registered as a part of a requirement for a one credit course, others volunteered in order to increase the amount of Japanese conversations they were able to have outside of the classroom.. Their proficiency levels in Japanese varied widely (beginner to advanced).

Interaction Treatment

In Week 1, all of the participants completed the pre-test in the researcher's office. They then participated in an orientation session to learn about the procedure for the video-based conversation activities (Week 2). Afterwards, they engaged in nine weekly sessions (Weeks 3-11) with the same conversational partner via *Google Hangouts*. One week after completing the interaction sessions (Week 12), they revisited the researcher's office to take the post-test.

In each session, the Japanese and American students scheduled to meet for 60 minutes and completed the conversation activities; they used their own computers due to the large time difference between Japan and the US. For the conversation task, each participant brought two visuals related to a theme of the week (e.g., sports, pop culture) representing

either Japan or the US, and prepared two discussion questions for each photo. This kind of two-information exchange (minimally structured, on real-life topics) can be an ideal space for L2 learners to practice a range wide array of conversational skills, such as describing, narrating, and expressing opinions (Lee, 2002). With the participants' primary focus on meaning conveyance throughout the conversational sessions, the present EFL learning context could be labelled as focus on meaning (Loewen, 2014).

As part of the language-exchange program, they spent 30 minutes performing the task in English, and then switched their roles (Japanese students as NSs; American students as NNSs) to complete the second task with another visual for the remaining 30 minutes in Japanese. The participants were specifically asked to not use the multimodal features of *Google Hangout* (e.g., text chat, screen sharing); this was done to control their potentially different amount of familiarity with technology and video-mediated conversation which may influence the nature of computer-mediated L2 learning experience (Develotte, Guichon, & Vincent, 2010). To ensure the participants' regular and consistent attendance, they recorded and submitted their own sessions to the researchers (by using a function of *Google Hangouts*) every week.

Similar to previous L2 interaction research (e.g., Mackey, Gass, & McDonough, 2000; Mackey & Philp, 1998), the NS interlocutors in this study were encouraged to provide interactional feedback where natural and appropriate. They received guidance during the orientation (Week 2) regarding when and how to provide recasts in response to certain linguistic errors that may cause difficulty in message comprehension. Specifically, they were encouraged to give recasts as a part of negotiation (e.g., confirmation requests, clarification requests) after a communication breakdown actually occurred, and/or when the NSs perceived the NNSs' errors as potentially threatening to successful communication in future situations.

Coding of Interactional Features. To provide suggestive patterns on the nature of the interaction that the participants had actually experienced during the semester-long project, we coded the second (T1 = Week 4) and eighth (T2 = Week 10) sessions of 20 dyads (20 dyads × 0.5 hours × T1 & T2 = 20 hours), when the conversational themes were counter-balanced. In keeping with the coding scheme developed by Lyster and Ranta (1997), the video-recordings were analyzed for three key elements of L2 interaction—triggers, feedback and uptake (for examples, see Supporting Information-A):

1. *Triggers* referred to the linguistic errors that NNSs made in pronunciation (mispronunciation of segmentals and prosody), vocabulary (wrong word, collocation and preposition choice), and grammar (morphology and word order errors).

2. *Feedback* referred to the recasts and negotiation strategies (confirmation checks, repetition, clarification requests) that NS interlocutors provided in response to NNSs' errors.

3. *Uptake* referred to NNSs' successful repairing of their original errors (successful repair), failure in self-correcting the errors (needs-repair), or no reaction to the feedback move (no uptake)[3].

Pre-/Post-Test Materials

To elicit sufficiently long speech data, we adopted a story telling task—the format widely used in L2 speech research (Derwing & Munro, 1997) and L2 vocabulary research (Uchihara & Clenton, 2018). The participants first familiarized themselves with an eight-frame cartoon picture (1 min), and then explained the sequence of the events that were depicted. The task was considered suitable, as it adequately reflects what the NNSs did in the videoconferencing tasks (i.e., accurately describing a visual image of their choice). However, there was no single conversational session, whereby the NNSs practiced any similar vocabulary and theme that were used in the pre- and post-test materials. In Weeks 1 and 12, all of the recordings were conducted individually in a quiet room at the university using a Roland-05 audio recorder (set at a 44.1 kHz sampling rate and 16-bit quantization) and a unidirectional condenser microphone.

To provide comparable speech samples for pronunciation, lexicogrammar and fluency analyses, a decision was made to use the same materials for the pre- and post-tests (for the same decision in L2 longitudinal research, see Derwing et al., 2014). We considered the test-retest effects to be minimum. There is evidence that using different prompts may result in different speech behaviours especially in fluency and vocabulary use even within the same task design (e.g., picture narratives) (De Jong & Vercellotti, 2016). Comparatively, Derwing, Thomson and Munro (2006) showed completing the same story telling task (used in the current study) twice with an interval of two months did not change L2 learners' oral proficiency (comprehensibility, fluency, and accentedness). To support this, the comparison group in the current study indeed changed only grammatical complexity aspects of L2 speech, when they took the pre- and post-tests without any interaction treatment. This in turn suggested that the other areas of L2 speech (pronunciation, lexicogrammar accuracy and richness, fluency) elicited from this particular task format (story telling) were resistant to change thanks to task repetition at least within the timeframe and context of the current study (10 weeks of EFL instruction) (see the Results section below).[4]

[3] In this project, we did not conduct any follow-up analyses of whether participants had understood the errors or not during the interaction treatment. Notably, it has remained considerably difficult and controversial (a) whether, to what degree and how we can measure learners' understanding of errors, and (b) whether, to what degree and how the awareness, noticing, and understanding of errors can be directly related to acquisition (see Lyster, Saito, & Sato, 2013). Rather, our main focus lay in examining the extent to which they could actually modify their own errors (i.e., process data), and the extent to which such self-modification can lead to acquisition (i.e., product data).

[4] Notably, it has been shown that L2 learners' speech (fluency in particular) is susceptible to change, when they repeat the same speaking task *immediately* (e.g., Lambert, Kormos, & Minn, 2018). Our pilot data showed that using different story telling tasks resulted in different speaking behaviours within the same speaker (especially in terms of pronunciation) (see also De Jong & Vercellotti, 2016). We call for future studies which will probe the complex relationship between different types of

The length of the speech samples in the current study was substantially longer (*M* = 142 sec ranging from 95 to 301 sec) than our precursor research (Saito & Akiyama, 2017), which used relatively short speech samples elicited via picture descriptions (*M* = 30 seconds, 40 words), and thus the current study included a sufficient number of words for robust lexical analyses (*M* = 105.3 words ranging from 55 to 206 words). The task demand of the story telling task could be considered relatively high enough to elicit supposedly different levels of speech performance from Experienced and Inexperienced NNSs (Isaacs & Trofimovich, 2012).

Lexicogrammar and Fluency Analyses

The lexicogrammar and fluency of the NNS speech samples were analyzed in line with Read's (2000) model of L2 vocabulary use (i.e., appropriateness, richness) and Housen, Kuiken and Vedder's (2012) framework for L2 grammar knowledge (i.e., complexity, accuracy, fluency). For the analysis of fluency, raw speech samples were used. For the lexicogrammar analysis, these were transcribed and cleaned up by removing orthographic markings of filled pauses (e.g., *uh, um, oh, ehh*). While lexical richness was analyzed via the *Coh-Metrix* software (McNamara, Graesser, McCarthy, & Cai, 2014), the other dimensions (see below) were manually analyzed by two linguistically trained coders (one of them was the researcher). Before the analysis took place, the coders agreed on a clear understanding of what constituted "fluency" (speed and breakdown), "lexical appropriateness" (adequate word choice), and "grammatical accuracy" (the accurate use of morphology) and complexity (subordination and sub-clausal complexification). Subsequently, they analyzed a training set of 10 non-native transcripts from our previous research (using the story telling task). Their reliability was relatively high for each measure ($r > .85$). While the first coder proceeded to the analysis of the experienced and inexperienced groups (40 samples), the other coder analyzed the comparison group (20 samples). Below, we define and describe how each linguistic measure was operationalized in the study.

Lexical Appropriateness. This category refers to how L2 learners are able to choose appropriate vocabulary in context, and was calculated based on the ratio of vocabulary errors which included (a) false cognates (e.g., "Rimokon" instead of "remote control") and (b) imprecise word choice (e.g., "drop on the ground" instead of "fell on the ground") to the total number of words (Isaacs & Trofimovich, 2012) and Analysis of Speech (AS) units (Foster, Tonkyn, & Wigglesworth., 2000)[5].

Lexical Richness. This category refers to how L2 learners can access a wide range of sophisticated, infrequent words, and was automatically analyzed via *Coh-Metrix* (McNamara et al., 2014) in terms of the Measure of Textual Lexical Diversity (MTLD) (for diversity) and

repetition (e.g., task vs. procedural; immediate vs. delayed) and their impact on multiple dimensions of L2 speech development (cf. Mora & Levkina, 2017).

[5] AS-units provide segmentation criteria for the linguistic analyses of spoken discourse, defined as "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause (s) associated with either" (Foster et al., 2000, p. 365).

the average log frequency of all words based on the CELEX corpus of English (for frequency).

Grammatical Accuracy. This category refers to L2 learners' competence using conceptually and contextually accurate morphological markers in verbs (tense, aspect, modality, and subject-verb agreement), nouns (plurals) and articles (definite, indefinite, and non-articles). L2 grammatical accuracy was analyzed through the ratio of morphological errors to the total number of words (Isaacs & Trofimovich, 2012) and AS-units (Mora & Valls-Ferrars, 2012).

Grammatical Complexity. This category refers to L2 learners' willingness and capacity to convey large amounts of information within one sentential unit by using advanced syntactic structures (Skehan, 1998). Following Norris and Ortega's (2009) recommendation, this dimension was analyzed via the subordination measure (i.e., the clause to AS-unit ratio) and the subclausal length measure (i.e., the total number of words per clause).

Fluency. This category refers to how many words are produced effortlessly, and was analyzed in conjunction with Tavakoli and Skehan's (2005) notion of breakdown and speed fluency. The former dimension was analyzed by dividing the number of filled and unfilled (silence > 250ms) pauses over the number of syllables (i.e., pause frequency); the latter dimension was calculated by dividing the speaking time (without filled and unfilled pauses) by the total number of syllables (i.e., articulation rate).

Pronunciation Analyses

In accordance with other L2 pronunciation research, we used linguistically trained raters' subjective scalar judgements in order to evaluate the segmental and prosodic qualities of the participants' *spontaneous* L2 speech (e.g., Derwing & Munro, 1997). Corresponding to the standards seen in previous research (e.g., Isaacs & Trofimovich, 2012), the first 30 seconds of each speech sample were cut and used for the pronunciation analysis, which is detailed below.

Raters. In light of the demanding nature of the rating task (analyzing four different dimensions of pronunciation proficiency) and the potential amount of listener fatigue, the dataset was divided into two listener groups (for a similar approach, see Trofimovich, Lightbown, Halter, & Song, 2009): $n = 4$ raters in Canada (Group A) and $n = 4$ raters in Japan (Group B). While Group A assessed the pronunciation qualities of the experienced and inexperienced NNSs, Group B raters analyzed those of the inexperienced and comparison NNSs. To confirm the comparability of the raters in Groups A and B, their inter-rater reliability was checked for the inexperienced group (which both of the raters evaluated).

All the Group A raters were recruited at an English-speaking university in Montreal, Canada. They were NSs from English-speaking families and had at least one parent who was a native speaker of English. They were all graduate students in Applied Linguistics and reported experience of ESL/EFL teaching ($M = 3.2$ years). Their familiarity with Japanese-

accented English was also relatively high (*M* = 5.8 on a 6-point scale, where *1 = not at all familiar* and *6 = very familiar*). For Group B, a total of two Japanese raters with high-level L2 English proficiency and two L1 English speakers were recruited in Japan. All of them had a degree in applied linguistics at a BA/MA level with an extensive amount of ESL/EFL experience (*M* = 4.0 years). Thanks to their residence in Japan, their familiarity with Japanese accented English was invariably "6" (*1 = not at all familiar* and *6 = very familiar*).

Procedure. This study employed the training paradigm elaborated in Saito and Akiyama's (2017) validation research. The raters first received thorough instruction from a trained research assistant on two categories: (a) segmentals (substitution, omission, or insertion of individual consonant and vowel sounds) and (b) word stress (misplaced or missing primary stress). The raters listened to speech samples presented in a randomized order via a MATLAB custom software, and then used a moving slider to rate them on a 1000-point scale for segmental errors and word stress errors (*frequent – infrequent/absent*). The raters were allowed to listen to each speech sample as many times as they wanted, until they were satisfied with their judgements. For the details of the validation study, training scripts and onscreen labels, see Supporting Information-B.

All rating sessions took place individually in the researcher's office (Group A in Canada; Group B in Japan). Each session lasted for one hour. The raters first practiced L2 pronunciation analyses for each task with three samples (not included in the main dataset). Upon hearing each practice sample, the raters evaluated them and were asked to explain their decisions, receiving feedback from the research assistant. After the assistant confirmed the raters' adequate understanding of the rating procedures, they proceeded to the judgement of the main dataset. Each session lasted for approximately 60 minutes with a 10-minute intermission halfway through.

Inter-rater Agreement. According to Cronbach's alpha analyses, the raters in Group A demonstrated consistent agreement for segmentals (α = .95) and word stress (α = .91); and those in Group B demonstrated similarly high-level reliability for segmentals (α = .90) and word stress (α = .89). Next, Cronbach's alpha was calculated for all the raters focusing on the recording of the inexperienced group which both of the raters in Groups A and B evaluated. Once again, the inter-rater agreement was strong for segmentals (α = .88) and word stress (α = .87). Thus, a decision was made to average the raters' scores to derive a single score for the perceived segmental and word stress accuracy of each sample at each testing point (pre/post).

Results

Details of L2 Interaction

The first objective of the statistical analyses was to examine whether, and to what degree, the nature of the interaction treatment differed between the experienced and inexperienced groups, as such variability could have affected the participants' development patterns. Thus, we explored (a) how often NNSs made pronunciation, vocabulary and grammar errors, (b) how often NSs provided interactional feedback (recasts after communicatively salient errors and negotiation after communication breakdown), and (c) how often NNSs produced self-modified output (successful repair, needs repair, and no uptake). Since our video coding was applied only to the second session (T1) and the eighth session (T2) of the project, the following results are suggestive of how the NNSs interacted with their NS partners.

Triggers. The average amount of pronunciation, vocabulary and grammar errors that the experienced and inexperienced NNSs made are summarized in Table 2. A three-way repeated ANOVA was conducted with Group (Experienced, Inexperienced) as a between-subjects factor, and Linguistic Category (pronunciation, vocabulary and grammar errors) and Time (T1, T2) as within-subjects factors. Although the interaction effect of Group × Linguistic Category × Time did not reach statistical significance, $F(2, 36) = 0.246$, $p = .783$, $\eta_p^2 = .086$, a Group × Linguistic Category interaction effect was found to be significant, $F(2, 36) = 4.470$, $p = .018$, $\eta_p^2 = .730$. According to Bonferroni multiple comparisons, the inexperienced group made significantly more pronunciation errors ($M = 29.6$ errors per participant within one session) than the experienced group did ($M = 16.2$ errors) ($p = 0.35$); both of the groups produced significantly less vocabulary errors than pronunciation and grammar errors ($p = 0.18$, 0.47, respectively).

Feedback. According to the descriptive statistics (Table 2), the NS interlocutors corrected only a small portion of pronunciation (2.4-14.7%), vocabulary (15.7-39.1%) and grammar (7.7-17.7%) errors. A four-way ANOVA was conducted with Group (Experienced vs. Inexperienced) as a between-subjects factor, and Feedback (recasts, negotiation), Linguistic Focus (pronunciation, vocabulary and grammar), and Time (T1, T2) as within-subjects factors. The results found significant main effects for Group, $F(1, 18) = 6.750$, $p = .018$, $\eta_p^2 = .691$., for Feedback, $F(1, 18) = 13.577$, $p = .002$, $\eta_p^2 = .938$, and for Linguistic Focus, $F(1, 18) = 6.646$, $p = .003$, $\eta_p^2 = .889$; but not for Time, $F(1, 18) = 3.044$, $p = .098$, $\eta_p^2 = .379$. Furthermore, the results yielded a significant interaction effect of Linguistic Focus and Group, $F(1, 18) = 5.734$, $p = .007$, $\eta_p^2 = .836$. According to Bonferroni multiple comparison analyses, the inexperienced group received more feedback (recasts) on pronunciation errors ($M = 1.8$ times) than the experienced group ($M = 0.4$ times) ($p = .005$), and more feedback on pronunciation errors than vocabulary errors ($M = 0.5$ times, $p = .002$) and grammar errors ($M = 1.4$ times, $p = .009$).

Table 2 *Mean Scores and Percentage of Trigger, Feedback, and Uptake per Participant at T1 (Week 4) and T2 (Week 10)*

A. Experienced Group (n = 10)

	T1						T2					
	Pronunciation		Vocabulary		Grammar		Pronunciation		Vocabulary		Grammar	
Interaction patterns	*M*	*%*	*M*	*%*	*M*	*%*	*M*	*%*	*M*	*%*	*M*	*%*
Error triggers	17.2	47.3	3.3	9.0	15.8	43.5	15.2	46.4	3.2	9.7	14.3	4.3
Feedback												
No feedback	16.8	97.6	2.4	72.7	12.3	77.8	14.0	92.7	2.7	84.3	13.2	92.3
Recasts	0.2	1.2	0.6	18.2	1.5	9.5	1.1	7.3	0.2	6.3	0.7	4.9
Negotiation	0.2	1.2	0.5	15.1	0.4	2.5	0.1	0.0	0.3	9.4	0.4	2.8
Uptake after recasts												
Repair	0.0	0.0	0.2	33.3	0.2	13.3	0.1	9.0	0.1	50.0	0.1	14.3
Needs repair	0.1	50.0	0	0	0.1	6.7	0.4	36.4	0	0	0.1	14.3
No uptake	0.1	50.0	0.4	66.7	1.2	80.0	0.6	54.5	0.1	50.0	0.5	71.4
Uptake after negotiation												
Repair	0.0	0.0	0.1	20.0	0.2	50.0	0.0	0.0	0.1	33.3	0.1	25.0
Needs repair	0.0	0.0	0.2	40.0	0.1	25.0	0.1	100.0	0	0	0.3	75.0
No uptake	0.2	100.0	0.2	40.0	0.1	25.0	0.0	0.0	0.2	66.7	0	0

B. Inexperienced Group (n = 10)

	T1						T2					
	Pronunciation		Vocabulary		Grammar		Pronunciation		Vocabulary		Grammar	
Interaction patterns	*M*	*%*	*M*	*%*	*M*	*%*	*M*	*%*	*M*	*%*	*M*	*%*
A. Error triggers	28.5	56.3	2.3	4.5	19.8	39.1	30.8	56.3	2.6	4.7	21.3	38.9
B. Feedback												
No feedback	24.3	85.3	1.4	60.9	14.3	72.2	27.9	90.6	1.8	69.2	18.9	88.7
Recasts	3.0	10.5	0.8	34.8	3	15.2	2.2	7.1	0.7	26.9	1.9	8.9
Negotiation	1.2	4.2	0.1	4.3	0.5	2.5	0.7	2.3	0.1	3.8	0.5	2.3
C. Uptake after recasts												
Repair	0.5	16.6	0.2	25.0	0.4	13.3	0.5	22.7	0.3	42.8	0.5	26.3
Needs repair	0.7	23.3	0	0	0.8	26.7	1.2	54.5	0.1	21.4	0.4	21.1
No uptake	1.8	60.0	0.6	75.0	1.8	60.0	0.5	22.7	0.2	35.7	1	52.6
D. Uptake after negotiation												
Repair	0.1	8.3	0	0	0	0	0.1	14.3	0.1	100	0	0
Needs repair	0.6	50.0	0.1	100	0.3	60.0	0.5	71.4	0	0	0.3	60.0
No uptake	0.5	41.7	0	0	0.2	40.0	0.1	14.3	0	0	0.2	40.0

Uptake. The descriptive statistics in Table 2 show that whereas the amount of uptake (successful repair + needs repair) widely varied across the linguistic focus of feedback, both the experienced and inexperienced learners exhibited slightly more uptake at T2 (28.6-100%) compared to T1 (0-100%). Given that several cells in the dataset included zero uptake (violating homogeneity of variance), no inferential statistics were calculated.

Effects of Interaction on the Development of L2 Oral Proficiency

The second objective of the statistical analyses was to examine the extent to which two groups of NNSs—Experienced, Inexperienced—improved their pronunciation, fluency, vocabulary and grammar through one semester of interaction with NSs relative to the comparison group, who did not engage in consistent conversational activities. For each linguistic measure, a two-way Group (Experienced, Inexperienced, Comparison) ×Time (Pre, Post) ANOVA was performed. The source of significant interaction effects was further analyzed through post-hoc pairwise comparisons (Bonferroni-corrected).

All the results are summarized in Table 3 (i.e., the presence/absence of significant interaction effects and improvement over time). Here, three observations were made. First, the inexperienced NNSs significantly enhanced their lexical diversity (MTLD) and appropriateness (lexical error rate), grammatical accuracy (morphological error rate), and speed and breakdown fluency (articulation rate, pause ratio). Second, the experienced group showed significant or at least marginal improvement not only in grammatical accuracy and speed fluency (articulation rate), but also in their pronunciation accuracy (segmentals, word stress). Finally, not only the experimental groups, but also the comparison group demonstrated significant gains in grammatical complexity (clause to AS-unit ratio). The results indicated that improvements in the grammatical complexity of L2 speech could be subject to change, when learners engage in one semester of foreign language instruction (regardless of video-based interaction activities) or/and take the same test twice (i.e., test-retest effects).

Table 3

Results of Pre/Post-Tests: Experienced, Inexperienced & Comparison Groups

	F (2, 27)	p	η_p^2	Significant contrasts: Pre to post-tests (Bonferroni corrected p value)
Lexical appropriateness				
Error free clause ratio	5.951	.022	.181	• Inexperienced (p = .011*, d = 1.15)
Errors per AS-unit	0.111	.741	.004	*n.s.*
Lexical richness				
Frequency	0.524	.598	.058	*n.s.*
Diversity	8.620	.007	.303	• Inexperienced (p = .025*, d = 0.54)
Grammatical accuracy				
Error free clause ratio	4.723	.039	.149	• Inexperienced (p = .039*, d = 0.80) • Experienced (p = .089†, d = 0.33)
Errors per AS-unit	2.458	.105	.154	*n.s.*
Grammatical complexity				
Clause to AS-unit ratio	31.938	<.001	.558	• Experienced (p = .007*, d = 1.22) • Inexperienced (p = .002*, d = 1.37) • Comparison (p = .001*, d = 1.45)
Words per clause	0.593	.560	.042	*n.s.*
Fluency				
Articulation rate	14.975	.001	454	• Experienced (p = .008*, d = 1.25) • Inexperienced (p = .024*, d = 0.50)
Pausing ratio	4.941	.035	.216	• Inexperienced (p = .011*, d = 1.14)
Pronunciation				
Segmentals	2.541	.112	.095	• Experienced (p = .097†, d = 0.36)
Word stress	5.790	.029	.135	• Experienced (p = .006*, d = 1.20)

Note. * indicates $p < .05$ † $p < .10$

Discussion

RQ1: Nature of Interaction

The current study examined the nature and impact of L2 interaction on the development of longitudinal oral proficiency (one academic semester) of two different groups of NNSs—more and less experienced/more and less proficient Japanese EFL students. According to the descriptive analysis of video recordings at the onset (T1) and endpoint (T2) of the project, NS interlocutors provided feedback on approximately 5-15% of the NNSs' pronunciation and grammar errors, and approximately 15-40% of their vocabulary errors (via recasts after communicatively salient errors and via negotiation after communication breakdowns). The frequency of feedback presented here could be considered similar to other meaning-oriented interaction contexts (cf. Mackey et al., 2000). These findings in turn suggest that the nature of the interaction in the current study was meaning-oriented rather than form-oriented. Throughout the project, the participants communicatively and collaboratively focused on improving L2 comprehensibility while using language primarily for message conveyance, as they were explicitly trained to do.

As for the different interactional patterns between the experienced and inexperienced NNSs, the results showed that the latter group generated more pronunciation errors and thus received more pronunciation-focused feedback; but, such significant group differences were not identified with respect to lexicogrammar errors and feedback episodes. The findings here partially (at least for pronunciation) concur with other L2 interaction researchers' assumptions that L2 interaction may be beneficial for inexperienced L2 learners in particular, who are likely to encounter more communication breakdowns, receive more comprehensible input, and produce more comprehensible output (Pica, Young, & Doughty, 1987). To further examine whether and how the semester-long interaction differentially impacted the experienced and inexperienced NNSs' oral proficiency over time, we now turn our discussion to the pre/post-test data vis-à-vis different linguistic domains (pronunciation, fluency, vocabulary vs. grammar).

RQ2: Impact of Interaction

First and foremost, it is noteworthy that all the NNS learners (Experienced, Inexperienced, Comparison) enhanced their grammatical complexity (subordination). The results here indicated that L2 learners could greatly enhance the complexity of their speech, as long as they received one semester of foreign language instruction (a few hours per week) or/and they took the same speaking test twice (test-retest effects). Our discussion here is compatible with Housen et al.'s (2012) suggestion on the developmental order in L2 speech learning: Changes in the underlying L2 system initially emerge in the dimension of complexity, when new, more elaborate and sophisticated structures are internalized. Another possible scenario is the test-retest effects. At post-tests, our NNS learners were able to reduce their cognitive efforts on content planning (i.e., conceptualization process) thanks to their

increased familiarity with the speech prompt that they had already seen at pre-tests. Consequently, they were possibly more aware of the relationship between events/propositions, and thus succeeded in expressing such information using more linguistically elaborated forms (i.e., subordination; see Yuan & Ellis, 2003).

Importantly, the group effects (Experienced vs. Inexperienced) were more clearly observed in the other domains of L2 speech (pronunciation, fluency, and lexicogrammar accuracy). According to the results, (a) inexperienced NNSs improved on those linguistic features that are susceptible to quick, immediate and tangible development in the early phases of L2 speech learning: lexical appropriateness (Schmitt, 1998), lexical richness (Muñoz, 2010), grammatical accuracy (Mora & Valls-Fellar, 2012), and breakdown fluency (Segalowitz & Freed, 2004); and (b) the experienced NNSs not only enhanced morphosyntactic accuracy and fluency, but also brought about some perceptible change in pronunciation, which is thought to develop gradually and slowly over a prolonged period of L2 speech learning (Flege, 2009; Saito & Brajot, 2013).

In line with the componential view of L2 oral proficiency (De Jong et al., 2012) and development (Bundgaard-Nielsen et al., 2011), the findings led to two tentative interpretations on the interaction-acquisition link as per different levels of learner experience/proficiency and linguistic domains. As for the less experienced/proficient NNSs in the current study, the bio information (summarized in Table 1) pointed out that their initial speaking skills were rather limited, arguably due to their lack of experience using the language for communicative purposes prior to the project (no experience overseas). By focusing on L2 lexicogrammatical feedback during real-time interaction activities, the inexperienced NNSs may have selectively practiced more appropriate, rich and fluent use of L2 lexicogrammar, which is relatively important for the acquisition of adequate L2 oral proficiency in the early stages of L2 speech learning (beginner → intermediate) (Isaacs & Trofimovich, 2012; Saito et al., 2016). Even though they received pronunciation-focused feedback (more than the experienced NNSs), the participants' pre-/post-test performance did not significantly change in phonological dimensions, suggesting that they may not be developmentally ready to make the most of such feedback in order to self-repair their mispronunciations, and modify their long-term phonetic representations. Indeed, it has been shown that L2 learners need to have sufficient L2 conversational experience and/or explicit phonetic knowledge in order to actually benefit from pronunciation-focused corrective feedback (Saito, 2015).

Conversely, the experienced NNSs in this project noted an adequate amount of prior L2 experience as well as high-level proficiency test scores (see Table 1). Given that phonological accuracy is crucial for advanced L2 oral proficiency development in the later stages of L2 speech learning (intermediate → advanced) (Isaacs & Trofimovich, 2012), and that phonological feedback is particularly effective for more experienced and proficient L2 learners, the experienced NNSs may have selectively worked on developing pronunciation skills by making the most of each piece of feedback they received from the NS interlocutors (cf. Saito & Lyster, 2012).

Limitations and Future Directions

To close, some limitations to this study should be acknowledged for future scholars who wish to expand this line of L2 interaction research. Notably, our discussion was exclusively concerned with Japanese EFL students' speech learning over one academic semester in classroom settings. Since we prioritized the pedagogical nature and value of such interaction activities, we embedded this study within an ongoing language curriculum and syllabus. To this end, the number of conversation exchange sessions were limited to 4.5 hours in total (30 minutes × 9 sessions). To track the nature and amount of the participants' L2 use outside the project, we interviewed the participants in a retrospective way, leaving much room for subjectivity and inaccuracies. To make any robust effect of the video-based interaction on L2 learners' subsequent improvement, laboratory studies may be needed in order to control and isolate the typical classroom environment. In such studies, the findings in this study should be replicated with a larger sample size in different L1/L2 contexts over a longer period of time (> 1 semester).

Next, we would like to point out that one uniqueness of the study was that a rather short interaction session was delivered with an equal interval (1 week) over one semester (9 weeks). This practice schedule is different from massed learning (e.g., 4.5 hours within one session) and spaced learning (9 30-minute sessions with an increasing interval). Given that the role of timing and intensity of practice has increasingly attracted scholarly attention in the field of instructed L2 acquisition (see Suzuki, Nakata, & DeKeyser, 2019). As for learner-internal variables affecting interaction effectiveness, the current study focused on two crucial affecting variables (i.e., learner proficiency, linguistic dimensions). Thus, it would be intriguing to investigate other individual difference variables, such as cognition (e.g., Segalowitz & Freed 2004 for working memory), conation (e.g., Ranta & Meckelborg, 2013 for willingness to communicate), and affect (e.g., Dewaele & MacIntyre, 2014 for anxiety/enjoyment).

Third, we adopted a total of 12 measures to tap into pronunciation (segmentals, word stress), fluency (speed, breakdown) and lexicogrammar (accuracy, complexity) dimensions of L2 oral proficiency. However, it needs to be acknowledged that each subskill can be further analyzed at more fine-grained levels, and are subject to different developmental patterns. For a comprehensive summary of subskills and examples of relatively difficult phonological, temporal, lexical and morphosyntactic features, see Table 4. In the current investigation, for example, vocabulary use was analyzed on a broader level—accuracy (error ratio) and richness (diversity). Previous research has convincingly shown that different dimensions of L2 learners' vocabulary development take place at different rates. L2 learners quickly acquire lexical features with greater saliency and communicative value, higher frequency or/and more concrete meanings (e.g., Crossley, Salsbury, & McNamara, 2009; Foster & Wigglesworth, 2016; Saito, 2019); few L2 learners can attain nativelike L2 lexical competence, such as abilities to detect conventionalized word combinations (Foster, Bolibaugh, & Kotula, 2014) and access proverbs and idioms (Abrahamsson & Hyltenstam, 2009). To this end, future studies should closely look at the impact of interaction on the multifaceted nature of L2 lexical and morphosyntactic acquisition by using a range of comprehension and production instruments.

Table 4

Examples of Relatively Difficult Phonological, Temporal, Lexical and Morphosyntactic Features at the Later Stage of L2 Speech Learning

Broad categories	Specific categories	Difficult features for experienced advanced L2 learners	Examples
Pronunciation	Segmentals, prosody	New perceptual cues and relevant articulatory configurations	• English /r/ and /l/ acquisition by Japanese speakers in perception (e.g., Iverson et al., 2003 for third formant) and production (Flege et al., 1995 for and labial, alveolar and pharyngeal constrictions) • Mandarin lexical tone acquisition by English speakers (Wang et al, 2003)
Fluency	Speed, breakdown, repair	More automatized and less monitored speech production	• Increasing articulation rate (Saito et al., 2018) • Reducing the number of repetition and self-corrections (Lambert et al., 2017)
Vocabulary	Accuracy, breadth, depth	More infrequent, multiple, complex, abstract and polysemous words	• Contextually appropriate vocabulary use (Saito, 2019) • Nativelike collocation use (Foster et al., 2014) • Proverbs and idioms (Abrahamsson & Hyltenstam, 2009) • Abstract words with multiple senses (Crossley et al., 2009) • Infrequent and context-specific words (Kyle & Crossley, 2015)
Morphology	Plurality, tense-aspect, article, word order	Morphemes at a later stage of L1 developmental hierarchy	• Third person plurality, tense-aspect, and article but not noun plurality (Bardovi-Harlig & Comajoan, 2008)

Relatedly, L2 segmental pronunciation skills in this study were operationalized as trained raters' impressionistic judgements, where they were supposed to evaluate the overall quality of numerous consonantal and vocalic sounds in spontaneous speech (see Supporting Information-A). Notably, extant literature has demonstrated that L2 learners quickly enhance the intelligibility of their L2 segmental production by prioritizing the acquisition of certain phonological contrasts with higher functional load than those with lower functional load (e.g., Munro & Derwing, 2008 for English /i/-/ɪ/ vs. /u/-/ʊ/). In the context of English /r/ acquisition, Japanese learners tend to acquire the durational aspect of the sound (> 50ms) within a very short period of immersion (< 1 year). Yet, these learners may need an extensive amount of experience (> 10 years) to enhance sensitivity to the primary acoustic correlate of the sound (i.e., third formants) and acquire relevant articulatory configurations (i.e., labial, alveolar and pharyngeal constrictions) (Flege, Takagi, & Mann, 1995; Saito & Brajot, 2013). It would be interesting for future studies to examine in depth how various acoustic dimensions of L2 learners' specific segmental pronunciation change in accordance with the different amount of feedback, uptake and self repair that they process during their interaction with NS interlocutors (cf. Lee & Lyster, 2017; Saito, 2015).

Finally, we need to remember that we did not find any significant improvement in certain domains of vocabulary (frequency) and grammar (subclausal complexification) regardless of group conditions. In order to further examine whether interaction can impact on the acquisition of these domains, we must first wait for future studies to answer two vital questions: (a) how do the domains actually relate to overall L2 oral proficiency, such as comprehensibility (Isaacs & Trofimovich, 2012; Saito et al., 2016) and communicative adequacy (De Jong et al., 2012; Révész, Ekiert, & Torgersen, 2016); and (b) how do the domains show "greater" change, and thus reach near-nativelikeness as a result of L2 learners' extensive residence in an L2 speaking environment (1 to 10 years) (cf. Flege, 2009).

References

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language learning*, *59*, 249-306.

Bradlow, A., & Pisoni, D. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener- and item-related factors. *Journal of the Acoustical Society of America, 106*(4 Pt 1), 2074–2085.

Bundgaard-Nielsen, R., Best, C., & Tyler, M. (2011). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in Second Language Acquisition, 33*(3)*,* 433–461.

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.

Chun, D., Kern, R., & Smith, B. (2016). Technology in language use, language teaching, and language learning. *Modern Language Journal, 100*(S1), 64–80.

Crossley, S., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, *59*, 307-334.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(1), 5–34.

De Jong, N., & Vercellotti, M. L. (2016). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research*, *20*, 387-404.

Derwing, T., & Munro, M. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*(1), 1–16.

Derwing, T. M., Munro, M. J., Foote, J. A., Waugh, E., & Fleming, J. (2014). Opening the window on comprehensible pronunciation after 19 years: A workplace training study. *Language Learning*, *64*(3), 526–548.

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

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

Flege, J., Takagi, N., & Mann, V. (1995). Japanese adults learn to produce English /ɹ/ and /l/ accurately. *Language and Speech, 38*, 25–55.

Foster, P., Bolibaugh, C., & Kotula, A. (2014). Knowledge of nativelike selections in a L2: The influence of exposure, memory, age of onset, and motivation in foreign language and immersion settings. *Studies in Second Language Acquisition*, *36*, 101-132.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*(3), 354–375.

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, *36*, 98-116.

Goo, J., & Mackey, A. (2013). The case against the case against recasts. *Studies in Second Language Acquisition*, *35*, 127–165.

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., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34*(3)*,* 475–505.

Iwashita, N., Brown, A., McNamara, T. & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*, 29–49.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*, 757-786.

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition*, *39*(1), 167–196.

Lee, A. H., & Lyster, R. (2017). Can corrective feedback on second language speech perception errors affect production accuracy? *Applied Psycholinguistics*, *38*(2), 371-393.

Loewen, S. (2014). *Introduction to instructed second language acquisition*. New York, NY: Routledge.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition, 19*(1)*,* 37–66.

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching, 46*, 1-40.

Mackey, A. (2012). *Input, interaction, and corrective feedback in L2 learning*. Oxford, UK: Oxford University Press.

Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive interactional feedback?. *Studies in second language acquisition*, *22*(4), 471-497.

Mackey, A., & Philp, J. (1998). Conversational interaction and second language development: Recasts, responses and red herrings? *Modern Language Journal, 82*(3)*,* 338–356.

Martinsen, R. A., Baker, W., Dewey, D. P., Bown, J., & Johnson, C. (2010). Exploring diverse settings for language acquisition and use: Comparing study abroad, service learning abroad, and foreign language housing. *Applied Language Learning, 20*, 45–66.

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

Mora, J. C., & Levkina, M. (2017). Task-based pronunciation teaching and research: Key issues and future directions. *Studies in Second Language Acquisition*, *39*, 381-399.

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly*, *46*(4), 610–641.

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

Muñoz, C. (2010). Staying abroad with the family: A case study of two siblings' second language development during a year's immersion. *ITL-International Journal of Applied Linguistics*, *160*, 24–48.

Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555–578.

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics*, *25*, 26–45.

Plonsky, L., & Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics*, *36*, 73–97.

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

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

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*(6), 828–848.

Saito, K. (2015). Communicative focus on L2 phonetic form: Teaching Japanese learners to perceive and produce English /r/ without explicit instruction. *Applied Psycholinguistics, 36*, 377-409.

Saito, K. (2019). To what extent does long-term foreign language education help improve spoken second language lexical proficiency? *TESOL Quarterly*, *53*, 82-107.

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.

Saito, K., Ilkan, M., Magne, V., Tran, M., & Suzuki, S. (2018).Acoustic characteristics and learner profiles of low, mid and high-level second language fluency. *Applied Psycholinguistics, 39*, 593-617.

Saito, K., & Akiyama, Y. (2017). Video-based interaction, negotiation for comprehensibility, and second language speech learning: A longitudinal study. *Language Learning, 67*, 43-74.

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

Saito, K., & Lyster, R. (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /ɹ/ by Japanese learners of English. *Language Learning, 62*, 595-633.

Schmitt, N. (1998). Tracking the incremental acquisition of a second language vocabulary: A longitudinal study. *Language Learning, 48*(2), 281–317.

Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, *26*(2), 173–199.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Suzuki, Y., Nakata, T., & Dekeyser, R. (2019). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *The Modern Language Journal, 103*, 713-720.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). Amsterdam: John Benjamins.

Trofimovich, P., Lightbown, P. M., Halter, R., & Song, H. (2009). Comprehension-based practice: The development of L2 pronunciation in a listening and reading program. *Studies in Second Language Acquisition, 31*, 609–639.

Uchihara, T., & Clenton, J. (2018). Investigating the role of vocabulary size in second language speaking ability. *Language Teaching Research*.

Vercellotti, M. L. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, *38*, 90–111.

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.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the acoustical society of America*, *106*, 3649-3658.

Ziegler, N. (2016). Synchronous computer-mediated communication and interaction: A meta-analysis. *Studies in Second Language Acquisition*, *38*, 553-586.