The Effect of Oral Reading Practice Using L1 Leveled Readers on the Speaking Ability of Science Majors

Sae Matsuda, Makoto Imura, Noriko Nakanishi, and Michael Herke

［要約］

本稿では、英語母語話者用の児童絵本を用いたインプットとアウトプットの反復練習が理系大学生のスピーキング能力の向上に及ぼす影響について本学で調査した結果を報告する。17 名の参加者は、全 9 回の各セッションの中で *Oxford Reading Tree* (Hunt, 2011)を用いて聴き読みを行った後にシャドーイング、リピーティング、音読を行った。本調査（2017 年）は、2016 年に実施した予備調査の再現であるが、より多くの繰り返し練習時間を確保するために、ブックトーク・再話を音読に差し替えた。アセスメントとして活動期間の開始前と終了後に 3 つの商用オンラインテスト（Progress、Versant、OPIc）を実施して比較した結果、Versant の「全体」スコアを除いて統計的に有意な変化はなかったが、Progress と Versant の両方でいくつかの点で中程度の効果量が観察された。さらに、学生の発話内容を詳しく調べるために、独自のスピーキングテストを行って音声データを収集し、音素カウンター、Praat、および Lextutor を用いて音韻、発話速度、語彙使用の各側面について分析を行った結果、スピーキング能力のいくつかの側面で改善が見られた。これらの調査結果は、反復的なインプットとアウトプットの練習が少なくとも発話の表層面での流暢さにプラスの影響を与えた可能性があることを示唆している。自発的な発話を構築するために必要な認知レベルの流暢さをいかに習得することができるか、そして教師はそのプロセスを促進するために何ができるかを見極めることが今後の課題である。

1. Introduction

Many teachers have experienced Asian students' reluctance to participate in class and attributed their reticent behavior to culture and previous education (cf. Flowerdew & Miller, 1995; Turner & Hiraga, 1996; Cheng, 2000; Greer, 2000). The Ministry of Education, Culture, Sports, and Science and Technology (2018) reported that 87.2 % of senior students in Japanese high school fell into the Common European Framework of Reference (CEFR) A1 category or below in terms of speaking. Those students enter college but display excessive self-monitoring when they have to speak in English. Research suggests that a lack of experience in speaking English, lack of confidence in their spoken English, and anxiety from high performance expectations and perception of the learner role as possible causes of students' reticence (cf. Tsui, 1996; Liu & Littlewood, 1997; Cheng, 2000; Jackson, 2002). On the other hand, Liu and Littlewood (1997) report that after having had practice in speaking, students felt more confident about their perceived proficiency, and, ultimately, frequency of speaking had a significant effect on actual proficiency. The reason for particularly focusing on speaking in this study is that this research was motivated by the existence of a gap between the social demand for English speaking ability expected of science students and their actual proficiency (Matsuda, Imura, & Nakanishi, 2018).

Measuring learners' speaking ability remains a big challenge. Hato, Takei, Healy, Kamizawa, and Ito (2015) tried to assess examinees' speaking ability by conducting an original interview test. However, they concluded that it would be impractical as they designed it to carry on the same procedures with limited human resources in a national university. In response to the results, Hato et al. (2016, 2018) have been conducting large-scale research to develop rating scales for a computer-based speaking test.

Now the government and universities are also looking into outside sources for testing speaking ability. *Versant* – an online speaking test – has been used to electronically measure learners' speaking ability before and after studying abroad (cf. Yoshida, Kagata & Ikuma, 2012; Shimizu, Kirimura & Nozawa, 2014). Yet, to the authors' knowledge, no research has been conducted using online speaking tests to measure the effects of extensive reading on spoken output, let alone a detailed analysis of the students' utterance data. Although Nakanishi and Ueda (2011) reported positive effects of reading aloud and shadowing on Secondary Level English Proficiency Test (SLEP) using extensive reading books, the participants' speaking skills were not measured.

It is quite difficult to define speaking ability comprehensively because there are so many

valuables involved: productive vocabulary; grammar; pronunciation and other prosodic features including stress, rhythm and intonation; phonetic changes including linking, deletion and assimilation; utterance fluency features including the rate of speech and the number of pauses; cognitive fluency features including the construction of a sentence and larger units of discourse. There will also be psychological factors influencing the speaking performance such as a volition to speak, anxiety, etc. As an operational definition, however, this study defines speaking ability broadly as a descriptive ability that is specified in the phased manner in the speaking "can-do" list of CEFR-J (explained hereinafter). Each online test has its own measures for speaking ability but their assessments all conform to the CEFR standard and thus considered valid as assessment tools (Pearson, 2015; Pearson Education Inc., 2019; Language Testing International, n.d.).

2. Literature Review

2.1 Extensive Reading and Leveled Readers

Research has shown that extensive reading not only enhances reading competency (cf. Elley, 1991; Mason & Krashen, 1997; Day & Bamford, 1998; Rodrigo, Krashen & Gibbons, 2004) but also improves other skill areas such as writing (cf. Hafiz & Tudor, 1989; Lai, 1993; Mason & Krashen, 1997), vocabulary (cf. Pitts, White, & Krashen, 1989; Day, Omura & Hiramitsu, 1991; Cho & Krashen 1994; Nation, 2001; Horst, 2005; MacQuillan, 2019), grammar (Yoshizawa, Takase & Otsuki, 2017) and speaking and listening (cf. Hafiz & Tudor, 1989; Cho & Krashen, 1994; Oya, Manalo, & Greenwood, 2009). Moreover, extensive reading enhances learners' motivation (cf. Elley, 1991; Cho & Krashen, 1994; Takase, 2008). Reading books that learners can easily understand and enjoy is motivating, and Nuttall's (1996) virtuous cycle shows that if they can understand better, they enjoy reading, then read faster and more, and eventually they become fluent readers. Finding reading materials that meet all these requirements can be a challenge. Leveled readers—easy picture books aimed for L1 children—can be suitable materials.

Leveled readers are originally targeted at English-speaking children who are learning to read. Yet, as Takase (2008) and Nishizawa and Yoshioka (2016) report, those easy L1 readers such as *Oxford Reading Tree* (hereafter *ORT*) and *Longman Literacy Land* are appealing to Japanese university EFL students and have been considered as appropriate materials for particularly inexperienced readers. They are written with simple vocabulary and syntax, and more importantly, contain another crucial element for fluency, i.e. repetition. Nation (2013) and Kadota (2015) emphasize that repetition is a key element for success. Imura (2020) compiled an *ORT* corpus

(73,550 words [tokens] /2,080 headwords [lemmas]) and verified that *ORT* has plenty of examples of "varied repetition where the same language features re-occur in changing contexts" (Nation, 2013: 38).

2.2 Reading Aloud and Shadowing

Nation (2013:10) emphasizes that every language class should include "fluency development" such as speed reading, easy extensive reading, and repeated listening/speaking. He also points out that those fluency activities should have four important features: 1) easy materials, 2) pressure to go faster, 3) message focused, and 4) quantity of practice.

Kormos (2006) explored speech production in second language acquisition and discussed different fluency measures. Likewise, Segalowitz (2016:14) examined L2 fluency "from a componential viewpoint, as comprising different dimensions of performance." Segalowitz (2010) claims that there are two types of L2 fluency. Utterance fluency refers to the fluidity of the observable speech including measurable features such as speech rate, articulation rate, and pauses while cognitive fluency refers to the fluid operation of cognitive processes dealing with semantic retrieval, utterance construction, and working memory.

Takeuchi (2003) claims that reading aloud is an effective way to learn a language. The science majors who are lacking output practice could benefit from reading aloud. It is argued that the language is internalized when learners read aloud after understanding the content and pronunciation. Kadota (2015) also claims that oral reading facilitates automatic phonological coding and improves reading skill. Similarly, shadowing facilitates automatic perception of input speech, the lower level decoding process which precedes comprehension of message, and improves listening skill. Both oral reading and shadowing enhance vocal rehearsal rate of speech input and promote internalization (or memorization) of words or formulaic chunks, etc. There are many steps to automatize the decoding process. In other words, to become a fluent reader or speaker, students have to automatize these decoding steps so that they can concentrate on comprehension rather than decoding.

2.3 CEFR-J and Speaking Descriptors

How can learners' progress be measured? Common European Framework of Reference (CEFR) "describes in a comprehensive way what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively." The Framework also defines 6 levels of proficiency (A1-C2) which allow learners' progress to be measured at each stage of learning. Furthermore, CEFR

describes what learners can do across five language skills: spoken interaction, spoken production (presentation), listening, reading, and writing. For all five skills at each level, there are sets of detailed "can do" statements called "can-do descriptors." By dividing speaking in two, "the CEFR focuses both on the learner's production and their ability to take part in conversations and discussion" (https://www.coe.int/en/web/common-european-framework-reference -languages/uses-and-objectives).

Finding that "the population of Japanese EFL learners skews towards the lower levels," Negishi (2012:105) subdivided CEFR A1-B2 levels and created the modified Japanese version (CEFR-J). According to the CEFR-J speaking "can-do" list, the presentation descriptors include using pictures, simple phrases and sentences, and a short story as follows (underlined by the authors):

A2.1 I can give a brief talk about familiar topics (e.g. my school and my neighborhood) supported by visual aids such as photos, pictures, and maps, using a series of simple phrases and sentences.

A2.2 I can make a short speech on topics directly related to my everyday life (e.g. myself, my school, my neighborhood) with the use of visual aids such as photos, pictures, and maps, using a series of simple words and phrases and sentences.

B1.2 I can give an outline or list the main points of a short story or a short newspaper article with some fluency, adding my own feelings and ideas.

As for the interaction skills, A2.1 level indicates "I can get across basic information and exchange simple opinions, using pictures or objects to help me." In this respect, leveled readers are expected to be a suitable material for practicing speaking while fulfilling those CEFR-J criteria.

2.4 Repetitive Input and Output Using L1 Readers: The Pilot Study

In 2016, Matsuda, Imura, and Nakanishi (2018) conducted a pilot study. The purpose of the study was to find whether repetitive input and output using leveled readers could enhance participants' speaking ability to the extent that could be measured by various types of commercial online tests. Twenty science majors—15 students in the experimental group and 5 students in the control group—participated in the study. The experimental group had ten 90-minute classes where they read and listened to *Oxford Reading Tree*, followed by a series of

practice including shadowing, repeating, reading aloud, and a book talk/retelling the story.

One of the instructors administered the training sessions alone from Week 4 to the end, but there were no explicit instructions provided about pronunciation, grammar, etc. When the students finished reading all the books on the level they were working on, they came to the instructor and performed a book talk or retold the story to her using their favorite book from that level. Mid-term, one-to-one conferences were also conducted in Week 8 to see how the students were doing.

Three types of online test—*Progress, Versant*, and *OPIc* (as will hereinafter be described in detail)—were also conducted before and after the treatment, and the results of the pre- and post-tests were compared within the experimental group. The control group also took the same tests.

Despite the limited period, most students in the experimental group showed improvements in some elements measured by the three online tests. The control group, however, failed to maintain their original level.

Based on the pilot study in 2016, the following research questions are explored in this study:

1. Can repetitive input and output with L1 readers enhance L2 science majors' speaking skills? If so, can the changes be measured by online tests?
2. Will the analysis of students' recorded sound data show any positive changes in terms of pronunciation, fluency, and vocabulary?

3. Method

3.1 Participants

The participants in the 2017 project were seventeen science majors who volunteered to take part in the project as an extracurricular activity. Since we were unable to form the control group in 2017, we have alternatively used the data of the experimental group in 2016 as a reference. Both groups have a similar construction except for a slight difference in their assignments.

Similar to the pilot study in 2016, the project in 2017 was advertised through the network of the Faculty of Science and Engineering. Two information meetings were held in June, which drew twenty-three students. Some students attended voluntarily while others were encouraged by their seminar teachers. At the information meetings, the students learned the content of the activities, however, the purpose of the study was not informed. After hearing the explanation, the students submitted their available time slots, and the sixth period on Monday was selected

because it was the only slot when most students were available.Due to time conflicts or some other reasons, six students decided to withdraw from the project. Finally, at the beginning of September 2017, the project was launched with seventeen students. Table 1 shows the number of participants in 2016 and 2017, their majors, and the year of study.

Table 1 *Number of Participants*

Department/Year	2016 (*n*=15)				2017 (*n*=17)				
	1	2	3	4	1	2	3	4	M
Architecture								1	
Civil & Environmental Engineering		1			2**	2*			
Electrical & Electronic Engineering	6**	1*			1**				
Life Science	2**	2	2		1**	5*			
Mechanical Engineering		1*						4	1

All the first-year students (indicated with two asterisks) were attending two 90-minute English classes a week as part of their core curriculum while seven second-year students (indicated with one asterisk) were taking one weekly 90-minute English class. The rest of the students (without asterisks) were not taking any university English classes during the treatment period. Two out of the five students from the Life Science department participated in the pilot study in 2016. While there were no fourth-year or graduate students in 2016, five fourth-year and one graduate students joined the project in 2017.

3.2 Instruments

3.2.1 Online tests

The commercial online tests *Progress, Versant,* and *OPIc* were used as pre- and post-tests to detect any effects from the treatments.

Instruments

The following excerpts are the official descriptions of the online tests used in this study:

1) *Progress* (Pearson)

> *Progress* is an online, integrated skills English language proficiency test package reporting scores on the Global Scale of English for listening, reading, writing, speaking, grammar, and vocabulary. The tests feature a combination of both

adaptive and linear sections which include multiple versions of the linear tests and a large number of questions in the item bank for the adaptive section (Pearson, 2015).

2) *Versant* (Pearson)

The *Versant* testing system, based on the patented Ordinate® technology, uses a speech processing system that is specifically designed to analyze speech from native and non-native speakers of the language tested. In addition to recognizing words, the system also locates and evaluates relevant segments, syllables, and phrases in speech. The *Versant* testing system then uses statistical modeling techniques to assess the spoken performance. The Versant English Test score report is comprised of an Overall score and four diagnostic subscores (Sentence Mastery, Vocabulary, Fluency, and Pronunciation).

(https://www.versanttests.com/technologyresearch)

3) *OPIc* (Global 8)

The OPIc is an internet-delivered test which provides valid and reliable oral proficiency testing on a large scale. The computer-delivered assessment emulates the "live" *OPI*, but the delivery of questions is through a carefully designed computer program and via a virtual avatar, allowing the test to be taken on demand and at a time convenient to the candidate and proctor (Language Testing International,n.d.).

3.2.2 The original speaking test

An original speaking test was added in 2017 in order to analyze the speech samples of the students and see what is actually happening in their utterance. It consists of four tasks: 1) oral reading, 2) describing a picture A, 3) describing a picture B, and 4) describing a picture sequence. The oral reading is a 65-word excerpt from *The Story of Dr. Dolittle* (Lofting, 1988). Picture A is a page from *Curious George* (Ray, 1973).The picture B and the picture sequence are scenes from *ORT,* but from the books that the students haven't read. The students had 30 seconds to prepare for all the sections, and they read aloud or described a picture (for 30 seconds) or described a picture sequence (for 1 minute). They submitted the sound data to the designated folder in the CALL system.

The further analyses of the pre- and post-original speaking tests were carried out using the following software:

1) *Phoneme Counter*

Phoneme Counter is an online system that converts English texts into phonetic transcription and calculates the number of each phoneme. It also has a function to compare the phoneme counts of two English texts (Nakanishi, 2018a; 2018b). In this study, the students' sound data collected from the original speaking test 1) oral reading were analyzed in the following procedure: First, the sound data were transcribed into texts using the iOS speech recognition application (Bodell, 2011). Secondly, the transcribed texts were converted into phonetic transcription, and then the phoneme counts were compared with those of the original script.

2) *Praat*

Praat is a widely used speech analyzer developed by Paul Boersma and David Weenink of the University of Amsterdam, which is free software downloadable at: http://www.fon.hum.uva.nl/praat/. With a display of spectrograms, it allows users to analyze various features of speech including pitch, formants, intensity, and pauses. Also by incorporating a specifically designed script (https://sites.google.com/site/speechrate/), Praat can show us the following indicators to measure speech fluency.

(A) Speech rate: the speed of speech (including pauses) indicated by the number of syllables per minute.
(B) Articulation rate: the speed of speech (excluding pauses) indicated by the number of syllables per minute. In other words, it indicates the speed of utterance itself.
(C) Phonation-time ratio: the ratio of utterance time to the total duration. It may be understood as utterance density (how wordy the speech is), or the percentage of the time occupied by the sound of speech.
(D) Mean length of runs: the average number of syllables between pauses. In other words, it means how long the speaker keeps talking without pausing.
(E) The number of silent pauses per minute (In default, a silence over 300 milli-seconds is counted as a silent pause.)
(F) The mean length of pauses.

3) *Lextutor*

Meara (1996) claims that vocabulary is "at the heart of communicative competence." He and many other researchers have described the relationship between vocabulary and

language skills. Our rationale for choosing vocabulary as one of the dependent variables was in part based on Read's analysis (2005) of the oral production of IELTS candidates in which he found a correlation between their vocabulary scores and their test band levels. We hypothesized that measuring participants' vocabulary pre- and post- treatment may reveal changes that corresponded with other measurements. In addition, De Jong, Steinel, Florijin, Schoonen, and Hulstijn (2012) based their proposed model of speaking proficiency on the functional accuracy of participants' responses, a construct that was partially comprised of vocabulary knowledge. The CEFR "can do" framework is based on a similar functional notion of language competence, as is the goal of science majors' language use of English in a professional capacity.

The Compleat Lexical Tutor (hereafter Lextutor) is a web-based set of tools for analyzing and researching vocabulary. We chose it based on its ease of use and relative popularity among vocabulary researchers. This project used the Web Vocabulary Profile (WebVP) application. WebVP provides comparisons of a user's text file to standard research lists including the Academic Word List and General Service List as well as computing the number of word tokens, types, families and type-token ratio (TTR).

3.3 Procedures

Table 2 shows the weekly schedule. Since the pilot study in 2016 began a week earlier and there was no original speaking test then, the students had 10 weeks of training while in 2017, the students had a little less than 9 weeks of training.

Table 2 *Weekly Schedule*

	2016	2017
Week 0	*Progress* 1	×
Week 1	*OPIc* 1	*Versant* 1 & Orientation
Week 2	*Versant* 1 & Orientation	*OPIc* 1
Week 3	Reading, listening, speaking activities	*Progress* 1
Week 4	Reading, listening, speaking activities	Original Speaking Test 1 & Reading, listening, speaking activities
Weeks 5-11	Reading, listening, speaking activities	Reading, listening, speaking activities
Week 12	Reading, listening, speaking activities	Reading, listening, speaking activities & *Versant* 2
Week 13	*Progress* 2	*OPIc* 2
Week 14	*OPIc* 2 & *Versant* 2	Original Speaking Test 2 & *Progress* 2
Week 15	Feedback	Feedback

In Week 1 after the *Versant* test, the orientation was provided in which the authors explained the procedures and demonstrated how to do shadowing and repeating activities. After students took three online pre-tests (*Progress, OPIc,* and *Versant*) and the original speaking test in Weeks 1-4, the students engaged in reading and speaking activities in each session thereafter, but there were no explicit instructions provided about pronunciation, grammar, etc. In Weeks 12-14, they took online post-tests, and in Week 15 they received their results and feedback.

In Weeks 4-12, the students read the books while listening to the audio, then they practiced shadowing and repeating, and after that, they read the book aloud to themselves. In the pilot study, after they read three books, they recorded a book talk, or later, retold the story about one of the three books. However, it was observed that students took too much time thinking about what to say before recording, and it slowed down their reading activities. Thus, in 2017, slight changes were made in the procedures. The book talk/retelling steps were eliminated and oral reading was added instead. That means they repeated each book 5 times. The comparison of the procedures is as follows:

Table 3 *Procedures*

Step	2016	2017
1	Silent reading with audio	Silent reading with audio
2	Shadowing	Shadowing
3	Repeating	Repeating
4	Oral reading	Oral reading (rehearsing)
5	Recording a book talk/retelling a story	Oral reading (recording)
6	×	Listening to the three recordings and submitting one

The main materials were *Oxford Reading Tree special packs ORT tadoku pack* (all packs from stage 1+ to stage 9), which consists of 30 packs (180 books), and three sets were purchased for this study. With OUP permission, the audio files had been installed into the CALL system prior to the training.

4. Results

4.1 Completion Rate and the Amount of Reading

The completion rate dropped to 70.5% (12 out of 17 students) in 2017 compared to 93.3% (14 out of 15 students) in 2016. Table 4 shows the average number of books the students

read. In nine weeks, the students read 61 books (4,256 words) on average. When the repetition was counted, they read/spoke 304 books (21,236 words) on average. The slowest students barely finished Level 2 while the fastest student completed Level 4.

Table 4 *Amount of Input and Output*

	Type (Repetition Uncounted)				Token (Repetition Counted)			
	Books		Words		Books		Words	
	2016	2017	2016	2017	2016	2017	2016	2017
max	96	78	6,002	5,429	384	390	24,008	27,145
min	37	32	1,825	2,536	148	160	7,300	12,680
M	67.9	60.8	4,243.6	4,257.0	271.4	303.5	16,974.6	21,236.3
sd	14.73	16.62	1,203.35	1,105.06	58.93	82.83	4,813.42	5,481.05

4.2 Online Tests

1) *Progress*

One student did not take the *Progress* post-test in 2017. Therefore, the results of 14 students in 2016 and 11 students in 2017 were compared. The bottom score "<10" was calculated as "9" in this study to distinguish it from "10." The box charts below give a glance at actual distribution (Figures 1, 2).

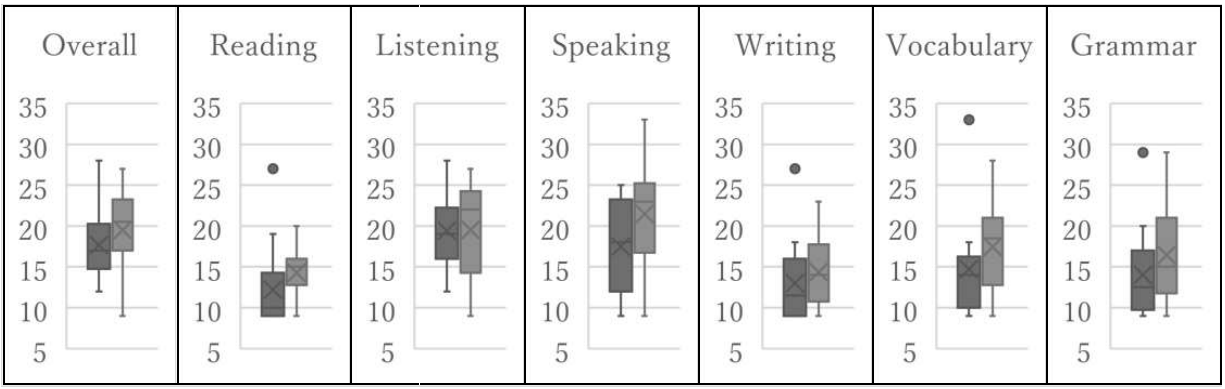

Figure 1: Results of *Progress* pre- and post-tests 2016.

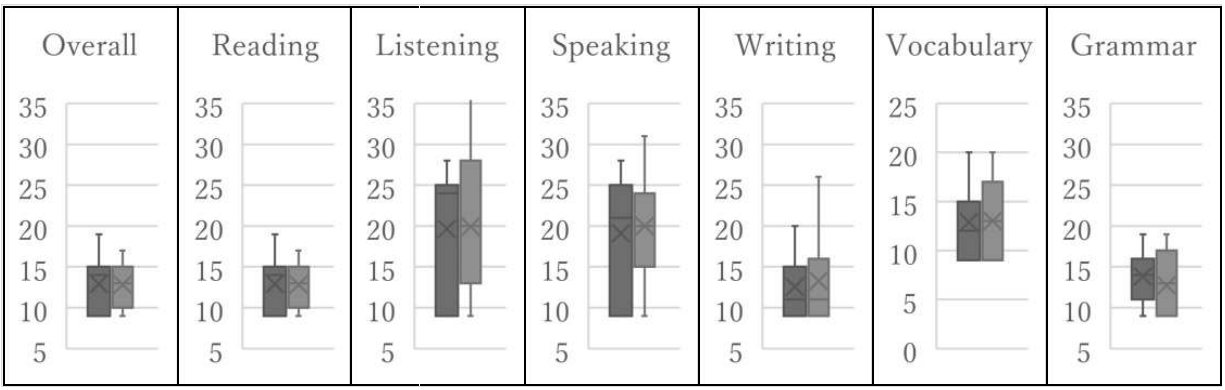

Figure 2: Results of *Progress* pre- and post-tests 2017.

Since the sample size was very small and we could not expect a normal distribution, we have decided to use Wilcoxon's signed rank sum test (nonparametric), which is good for a "paired" group. The test results for both 2016 and 2017 are shown in Table 5.

Table 5 *Comparison of Progress pre- and post-tests in 2016 and 2017* (Wilcoxon signed- rank test)

Pre-post test skills	2016 (n=14)		2017 (n=11)	
	Z-value	Effect size r	Z-value	Effect size r
Overall	1.65 *n.s.*	.312	0.48 *n.s.*	.101
Reading	1.56 *n.s.*	.295	0.12 *n.s.*	.025
Listening	0.22 *n.s.*	.042	0.37 *n.s.*	.078
Speaking	2.00 †	.378	0.20 *n.s.*	.044
Writing	0.95 *n.s.*	.179	0.12 *n.s.*	.025
Vocabulary	1.68 *n.s.*	.318	0.26 *n.s.*	.055
Grammar	1.38 *n.s.*	.260	0.67 *n.s.*	.142

†p<.05

In 2016, there was a statistically significant difference between the pre-test (Mdn=18.0) and the post-test (Mdn=23.0, Z=2.00, p<.05, r=.378) in "speaking." We found no significant differences in the other areas but medium-sized effects were observed in "overall" (r=.312) and "vocabulary" (r=.318), and small-sized effects in "reading" (r=.295) and "writing" (r=.179) and "grammar" (r=.260). In 2017, no statistically significant differences were found in any areas but small-sized effects were observed in "overall" (r=.101) and "grammar" (r=.142).

2) *Versant*

Three students in 2016 and 2017 respectively received "not scored" results: i.e. "The candidate's responses may have been in a language other than English, produced with poor/unintelligible pronunciation, or irrelevant," according to Pearson (2015). Therefore, the results of 11 students in 2016 and 9 students in 2017 were compared. The box charts below give a glance at the actual distribution (Figures 3 and 4).

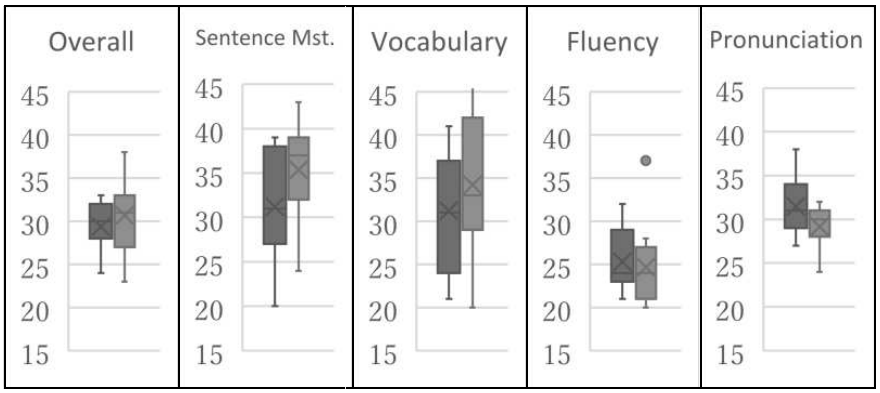

Figure 3: Results of *Versant* pre- and post-tests 2016.

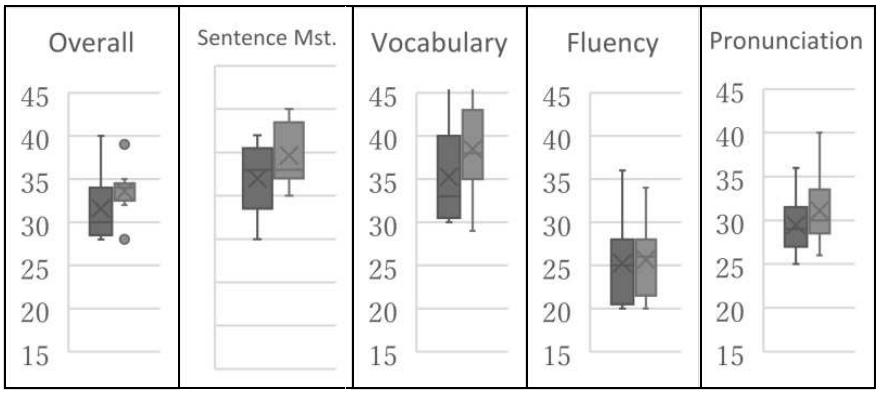

Figure 4: Results of *Versant* pre- and post-tests 2017.

Again, Wilcoxon's signed-rank tests were conducted for both 2016 and 2017 (Table 6).

Table 6 *Comparison of Versant pre- and post-tests in 2016 and 2017* (Wilcoxson signed-rank test)

Pre-post test skills	2016 (n=11)		2017 (n=9)	
	Z-value	Effect size r	Z-value	Effect size r
Overall	1.18 *n.s.*	.252	1.98†	.466
Sentence Mastery	2.49†	.531	1.84 *n.s.*	.435
Vocabulary	1.28 *n.s.*	.272	1.25 *n.s.*	.294
Fluency	0.63 *n.s.*	.134	0.95 *n.s.*	.225
Pronunciation	1.84 *n.s.*	.393	1.87 *n.s.*	.441

†p<.05

In 2016, there was a statistically significant difference between the pre-test (Mdn=31.0) and the post-test (Mdn=37.0, Z=2.49, p<.05, r=.531) in "sentence mastery." There were no significant differences in the other areas; however, a medium-sized effect was observed in "pronunciation" (r=.393), and small-sized effects in "overall" (r=.252) and "vocabulary" (r=.272). In 2017, there was a significant difference between the pre-test (Mdn=30.0) and the post-test (Mdn=34.0, Z=1.98, p<.05, r=.466) in "overall." Although no significant differences were found in the other areas, medium-sized effects were observed in "sentence mastery" (r=.44) and "pronunciation" (r=.441), and small-sized effects in "vocabulary" (r=.29) and "fluency" (r=.231).

3) *OPIc*

One student in 2016 missed the post-test, thus her data were excluded from the analysis. Figure 5 shows the transition of levels demonstrated by the students in 2016 (n=13) and in 2017 (n=12).

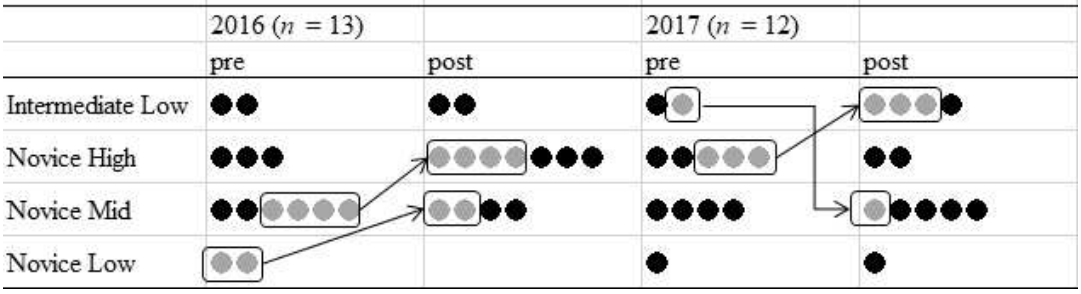

Figure 5: The transition of *OPIc* levels from pre-test to post-test. The upper part of the figure indicates higher levels.

In the year 2016 group, the post-test results revealed that six students out of 13 reached one level higher than their original level. However, all the five students who started with NH (Novice High) or IL (Intermediate Low) stayed at the same level. Meanwhile, in 2017, three students in NH reached IL in the post-test. One student who fell two levels from IL to NM had a unique situation: She had just come back from a two-week summer program in Canada when she took the pre-test and she was probably better at answering OPIc's open-ended questions, but unfortunately, the repetition practice did not help maintain her spontaneous speech skill.

4.3 Original Speaking Tests

1) *Phoneme Counter*

Pre- and post- test recordings collected in 1) oral reading task in 2017 original test were compared. One of the students' data was discarded due to its poor sound quality both in pre- and post-tests. Thus, 20 recordings (pre- and post- files by ten students) were transcribed into texts by Automatic Speech Recognition (ASR) using Speech API (Nakanishi, 2019), and then the phoneme counts were compared with the original script. Table 7 is a summary of the word and phoneme counts in the original script and the transcription of the pre- and post- recordings.

Table 7 *Number of words and phonemes in three data sets.*

	Words		Vowels		Consonants	
ASR pre *M (sd)*	65.7	(3.9)	77.9	(5.7)	119.5	(10.6)
ASR post *M (sd)*	60.5	(7.1)	71.5	(9.7)	113.7	(16.0)
Original	61		76		125	

The ASR results in the pre-test had more words (*M*=65.7) than the original script, with more vowels (*M*=77.9) and fewer consonants (*M*=119.5), which may indicate the tendency to read slowly and add unnecessary vowels in between consonants.

A typical example is one student's recording for "and when he walked down the street in his tall hat (11 words, 11 vowels, and 22 consonants)," which was transcribed as "on the man here it won't take down the story of him if tow path (15 words, 16 vowels, and 23 consonants)." As can be seen in Figure 6, "and when" was recognized as "on the man," suggesting that this student inserted a vowel after "and". The possibility of vowel insertion was also indicated in the ASR results of other phrases such as "here it" for "he"; "won't take" for "walked"; "story" for "street"; and "of him if" for "in his".

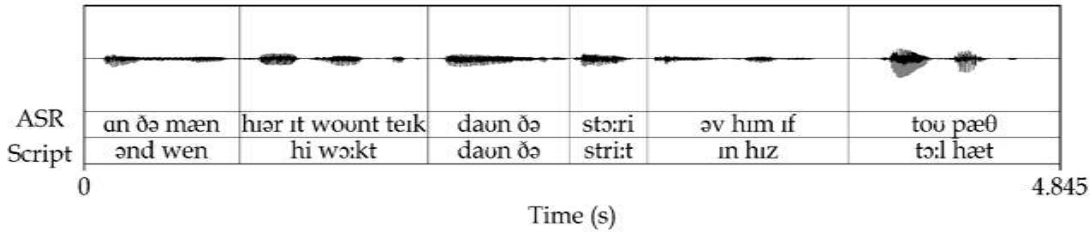

Figure 6: Phonemes included in ASR result and original script (pre-test).

On the other hand, in the post-test, the average number of the words (*M*=60.5) was a little less than the original script, with fewer vowels (*M*=71.5) and consonants (*M*=113.7). In some of the students' recordings, the total number of words, vowels, and consonants were smaller than those in the original script because the ASR could not catch up with their speed of oral reading. The transcription of the post-test recorded by the student introduced above was "on the way he walked down just wait in case toll hacked (12 words, 12 vowels, and 22 consonants)." The word "and" was still recognized as "on the", but the phoneme counts of the rest of the phrases are closer to the original script than in the pre-test.

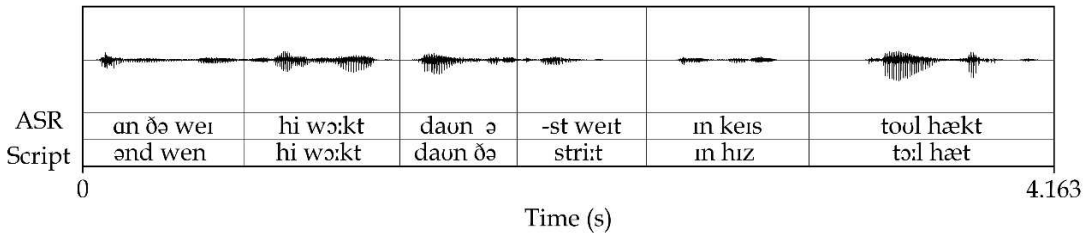

Figure 7: Phonemes included in ASR result and original script (post-test).

2) *Praat*

Using *Praat*, the students' pre- and post-tests utterances of all tasks were compared in terms of the six fluency indicators: speech rate (SR: the number of syllables per minute including pauses), articulation rate (AR: the number of syllables per minute excluding pauses), phonation time ratio (PTR: the ratio of utterance time against total duration), mean length of run (MLR: the number of syllables between pauses), silent pauses per minute (SPM: the number of pauses per minute), and mean length of pauses (MLP: the average length [seconds] of pauses). Prior to the processing, the noise was removed and the data were trimmed to

include sounds from the beginning of the utterance to the end of the utterance.

Results are presented in the box charts below (Figures 8-11).

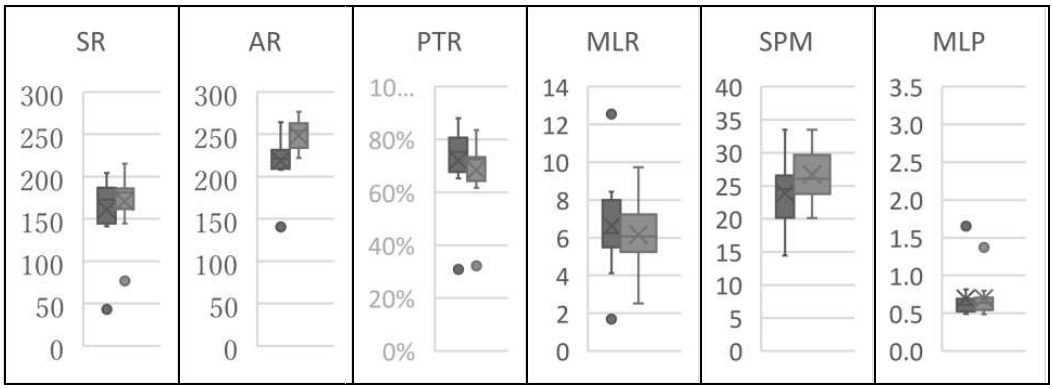

Figure 8: Fluency measures in Task 1 pre- and post-tests.

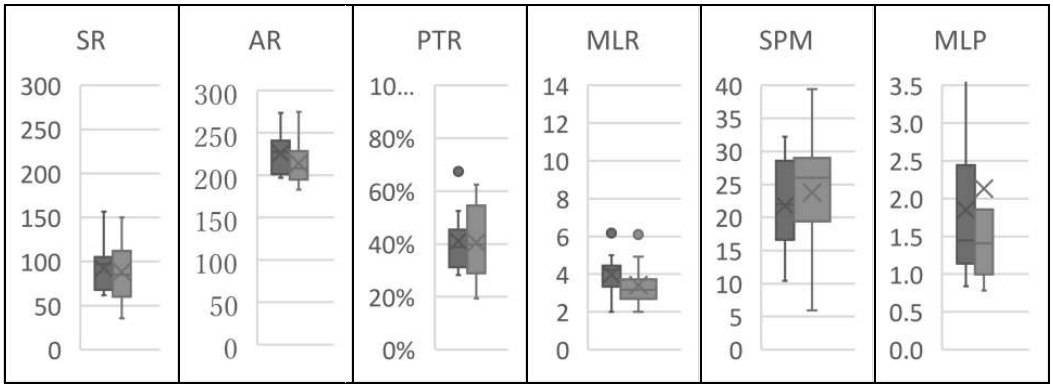

Figure 9: Fluency measures in Task 2 pre- and post-tests.

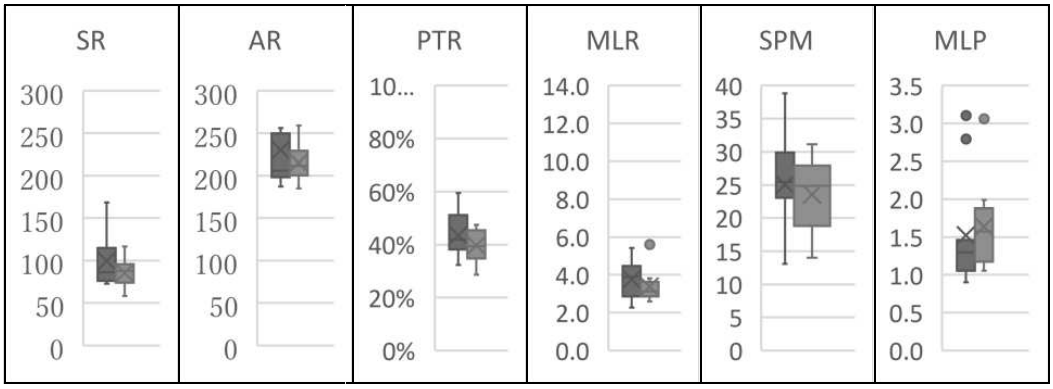

Figure 10: Fluency measures in Task 3 pre- and post-tests.

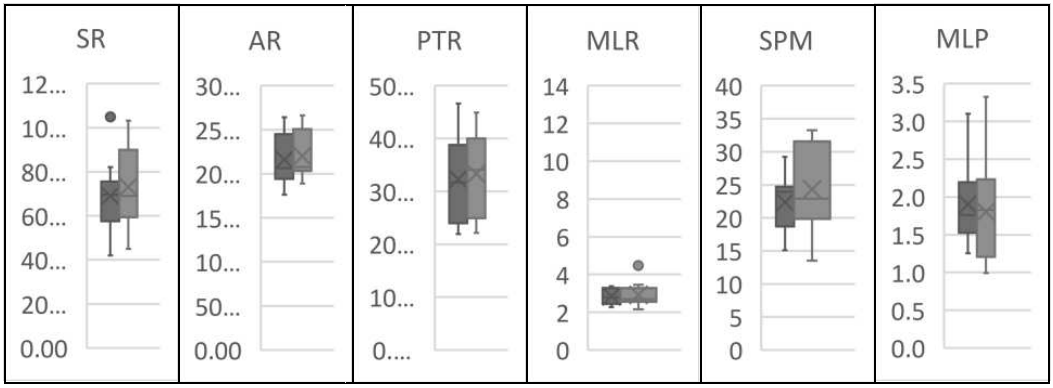

Figure 11: Fluency measures in Task 4 pre- and post-tests.

Wilcoxon's sign-rank test was conducted (Table 8).

Table 8 *Comparison of fluency measures between pre and post tests* (Wilcoxson signed-rank test, n=11).

		SR	AR	PTR	MLR	SPM	MLP
Task 1 pre-post	*Z*	1.27 *n.s.*	2.58 †	1.26 *n.s.*	0.62 *n.s.*	1.45 *n.s.*	0.21 *n.s.*
	r	.272	.551	.270	.133	.309	.045
Task 2 pre-pst	*Z*	0.82 *n.s.*	1.47*n.s.*	0.30 *n.s.*	1.64 *n.s.*	0.91 *n.s.*	0.00 *n.s.*
	r	.174	.313	.064	.349	.193	.000
Task 3 pre-post	*Z*	0.91 *n.s.*	0.15 *n.s.*	0.82 *n.s.*	1.20 *n.s.*	0.30 *n.s.*	0.21 *n.s.*
	r	.193	.033	.174	.256	.064	0.04
Task 4 pre-post	*Z*	0.58 *n.s.*	0.21 *n.s.*	0.39 *n.s.*	0.62 *n.s.*	1.17 *n.s.*	0.62 *n.s.*
	r	.123	.045	.082	.133	.025	.133

†p<.05

Although we could not observe any significant differences in most of the cases, the articulation rate in Task 1 (oral reading task) indicated a statistically significant gain (Z=2.58, p<.05). Since the articulation rate refers to the amount of speech per unit time (excluding pauses), the result means that the students' utterance speed in the oral reading task increased.

3) Lextutor

Pre- and post-test recordings of Tasks 2 to 4 were transcribed and analyzed using the Web Vocabulary Profile (WebVP) application available at www.lextutor.ca (Cobb, n.d.). Participants' transcriptions were analyzed for any changes in vocabulary output, with focus on fluency (measured in words per minute), tokens, types and families and type-token ratio (TTR). Tokens refers to all the running words in a text including repetitions. Types refers to all the

different words in a text. Families means all the variants of a word plus relevant derivations. TTR indicates lexical variation in a text. Before being uploaded for analysis, transcriptions were checked by two researchers for errors. The results are displayed in Tables 9, 10, and 11 below.

Table 9 *Vocabulary profile* (*n*=11)

	Pre	Post	Change	Percent
Tokens	679	792	113	16.6%
Types	161	180	19	11.8%
Families	115	124	9	7.8%

Overall, there were increases in tokens, types, and word families. Table 10 below shows the vocabulary profile as calculated by WebVP. Arithmetically, participants produced on average 0.82 more word families, 1.73 more types and 10.3 more tokens per 90 minutes of treatment.

Type-token ratio (TTR) is a common measure of lexical density. The closer the number is to 1 (100%), the richer the lexis. TTR is displayed in Table 10 below.

Table 10 *Lexical density* (*n*=11)

	Pre	Post	Change
TTR	161/679	180/792	
	23.70%	22.70%	-4.20%

The slight drop may be due to the lower number of new words being added to the already-existing narrative, which in this case were the three original speaking tests.

Fluency was also measured by calculating the number of words spoken per minute (WPM) as shown in Table 11 below. Words in this case refers to tokens or the total number of running words.

Table 11 *Average words-per-minute* (*n*=11)

	Pre	Post	Change	Percent
WPM	37.8	44.1	6.3	16.7%

One participant's data was excluded from the WPM calculation because he only uttered one word in the pre-test and five words in the post-test. Such a low score would have skewed

the overall result.

The participants uttered 37.8 WPM in the pre-test and 44.1 WPM in the post-test on average for an increase of 6.3 WPM or 16.7%. This increase in spoken fluency as measured by WPM positively corresponds with participants' increase in utterance speed in the oral reading task as measured by *Praat*. Both measures seem to indicate that the treatment had a positive effect on spoken fluency.

5. Discussion

5.1 Implications of the study

Since the project started a week later and the pre- and post-original speaking tests were added in 2017, the treatment period was shorter than in 2016. Thus, although the re-telling stage was eliminated to reduce students' cognitive load, the average numbers of books and words read were about the same as in the previous year. However, when the repetition was counted, the 2017 group had more input and output.

The results of the pre- and post-online tests displayed inconclusive results. The 2017 participants scored lower than the 2016 counterparts in almost all the skill areas in *Progress*. In contrast, *Versant* showed that the students scored higher in sentence mastery, vocabulary, and pronunciation in 2017. It should also be noted that the pronunciation score demonstrated a positive change in 2017 while the opposite was the case in 2016. Although the activities looked comparatively monotonous, the 2017 treatment with more repetition may have had some positive effects on pronunciation.

In 2016, *OPIc* results seemed to imply that the treatment worked best for beginner-level learners. In contrast, in 2017 most of the novice low/mid-level students stayed at the same level while the novice high-level students demonstrated progress. It might be partially attributable to the students' motivational differences. In 2016, the low-proficiency students enjoyed the activities using picture books so much that they even imitated the dog's voices happily. After the treatment, they became more confident in speaking and started to speak louder. On the other hand, beginner-level students in 2017 were fourth-year students who were encouraged to join by their seminar professor. There seemed to be a greater discrepancy in the 2017 group. While older students who unwillingly participated in the project remained reluctant to speak especially when open-ended questions were asked by *OPIc*, younger students who voluntarily joined the project enthusiastically engaged themselves in the activities and tried their best to answer *OPIc* questions.

Yet, when the recorded sound data were further examined using the *Phoneme Counter*, *Praat*, and *Lextutor*, there were some interesting findings. First, the recordings of Task 1 (reading a short passage aloud) were analyzed using *Phoneme Counter*. The results may indicate that the repetitive input and output helped the students to avoid inserting unnecessary vowels between the consonants. It was also suggested that their speed of oral reading in the post-test was too fast for the ASR to catch their utterance. It was partly supported by the analysis of speaking test data (Tasks 1-4) using *Praat,* which showed increased articulation rate in Task 1 although no other significant gains were observed. Repetitive input and output may have facilitated utterance fluency—the fluidity of the observable speech (including measurable features) such as speech rate, articulation rate, and pauses (Segalowitz, 2016), which might explain the analysis of *Praat*.

Finally, a vocabulary software analysis of the transcripts of the pre- and post-treatment picture description tests (Tasks 2-4) showed slight increases in the number of tokens, types, and word families used, supporting the claim that overall spoken fluency improved measurably over the course of the treatment. Lexical variation, as measured by TTR, did decrease, but this is typical when the same prompt is used as fewer new words are added to the existing narrative.

We cannot explain the changes in vocabulary usage with any certainty with the possible exception of the 16.7% increase in WPM. As far as the other results are concerned, did the participants acquire the new word families through the treatment, or were these previously known words reactivated? Are there more efficient ways to increase vocabulary usage? The original speaking tests did not test for vocabulary comprehension or usage, so there is no way to confirm how well students understood the vocabulary they used. Future studies would need to better control for existing vocabulary knowledge in order to attempt to answer these questions.

However, the fact that the utterance speed of the students has increased only in oral reading (Task 1) and not in the other tasks involving spontaneity (Tasks 2-4), may indicate that the amount of input was not enough to influence the students' cognitive fluency. As Segalowitz (2016) claims, we should not view fluency as a solid construct. Segalowitz describes two types of L2 fluency: utterance fluency and cognitive fluency. As mentioned earlier, the findings in this study may suggest that utterance fluency may be enhanced by simple oral speaking practice such as reading aloud, shadowing, and repeating. However, cognitive fluency involves more complex processing: the fluid operation of cognitive processes dealing with semantic retrieval, utterance construction, and working memory. By eliminating the re-telling process in 2017, we might have hindered students' cognitive fluency.

5.2 Limitation of the study

We should also point out the limitations of this study. The major limitations of this study are the small number of participants, the lack of the random assignment of participants into experimental and control groups and the lack of an alternative treatment group. Thus, the results need to be interpreted with caution. Variance attributable to the participants and different curricula could have somehow affected their performance as well. The artificiality of the experimental setting also may have had an effect on the outcomes. Meeting in the sixth period (18:20-19:50) in 2017 instead of the fifth period (16:40-18:10) in 2016 on Mondays after having regular classes or science experiments was exhausting, and it might have affected their concentration or motivation. Furthermore, we have not examined the validity of the speaking tests. In turn, in addition to the six indicators of *Praat*, there are several other fluency measures that have not been included in this study such as the number of filled pauses per minute, the number of disfluencies per minute, pace, and space (Kormos, 2006). To investigate these fluency measures, another computer program and qualitative research will be necessary.

6. Conclusion

This study attempted to explore whether the repetitive input and output bring measurable improvements in students' speaking skills and whether the results could be measured by three online tests and an original speaking test. While the online tests revealed no statistically significant changes specific to speaking, the recorded data of the pre- and post- original speaking tests displayed some phonemic, phonological, and lexical changes. Nevertheless, the nine- week treatment may not have been sufficient, and the next plausible step is to retrieve the re-telling stage and design a more effective training program. Further longitudinal studies may be needed to gauge the effect of repetitive input and output over an extended period. We might reconsider the teacher's role as well. For further research, we are currently analyzing the data in 2018 with an increased number of samples. Also, we are planning to conduct a longitudinal study from 2020 to 2024 supported by JSPS KAKENHI Grant Number 20K00906.

Acknowledgments

The authors would like to thank the two anonymous reviewers for providing valuable comments on this manuscript. This work was supported by JSPS KAKENHI Grant Number 16K02867.

References

Bodell, M. et al. (2011). HTML speech incubator group final report. Retrieved September 22, 2018 from, https://www.w3.org/2005/Incubator/htmlspeech/XGR-htmlspeech-20111206/

Cheng, X. (2000). Asian students' reticence revisited. *System*, *28*(3), 435-446. doi: https://doi.org/10.1016 /S0346-251X (00)00015-4

Cho, K. S. & Krashen, S. (1994). Acquisition of vocabulary from the Sweet Valley kids series. *Journal of Reading,37*(8), 662-667.

Cobb, T. *Vocabulary profiles*. [computer program]. Accessed September 22, 2018 at https://www. lextutor. ca/vp/

Common European Framework of References for Languages. Retrieved August 21, 2020 from, https://www.coe.int/en/web/common-european-framework-reference-languages/uses-and-objectives

Day, R., Omura, C. &Hiramitsu, M. (1991). Incidental EFL vocabulary learning and reading. *Reading in a Foreign Language*, *7*(2), 541-551.

Day, R. R. & Bamford, J. (1998). *Extensive reading in the second language classroom.* New York: Cambridge University Press.

De Jong, N., Steinel, M., Florijn, A., Schoonen, R., & Hulstijn, J. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*(1), 5-34. doi:10.1017/S0272263111000489

Elley, W. B. (1991). Acquiring literacy in a second language: The effect of book-based programs. *Language Learning, 41*(3), 375-411.doi:10.1111/j.1467-1770.1991.tb00611.x

Flowerdew, J. and Miller, L. (1995). On the notion of culture in L2 lectures. *TESOL Quarterly, 29*(2), 345-373. doi:https://doi.org/10.2307/3587628

Greer, D. L. (2000). "The eyes of *Hito*": A Japanese cultural monitor of behavior in the communicative language classroom. *JALT Journal*, *22*(1), 183-195.

Hafiz, F. M., & Tudor, I. (1989). Extensive reading and the development of language skills. *ELT Journal,43*(1), 4-13. doi: https://doi.org/10.1093/elt/43.1.4

Hato, Y., Kanzawa, K. Mitsunaga, H., & Healy, S. (2018). Developing a computer-based speaking test of English as a lingua franca: Preliminary results and remaining challenges. *Waseda Working Papers in ELF* (English as a Lingua Franca). 7, 87-99.

Hato, Y., Kanzawa. K., Tsubota, Y., Mitsunaga, H., and Underhill, N. (2016). Developing rating scales for a computer-based (CBT) speaking test of English as a Lingua Franca. *ETAS Journal*, 34, 33-34.

Hato, Y., Takei, T., Healy, S., Kamizawa, K., & Ito, K. (2015). Introducing a speaking test of English into the university entrance examination: Evaluation of a computer−based pilot test. Paper presented at the 54th JACET National Convention.

Horst, M. (2005). Learning L2 vocabulary through extensive reading: A measurement study. *The Canadian Modern Language Review*, *61*(3), 355-382. doi:https://doi.org/10.3138/cmlr.61.3.355

Hunt, R. (2011). Oxford reading tree special packs: ORT tadoku pack (all packs from stage 1+ to stage 9) 30 packs. Oxford: OUP.

Imura, M. (2020). Analyzing the ORT corpus: Leveled reader's potential viewed from content analysis. *JERA Bulletin, 13,* 3-23.

Jackson, J. (2002). Reticence in second language case discussions: Anxiety and aspirations. *System*, *30*(1), 65-84. doi:10.1016/S0346-251X(01)00051-3

Kadota, S. (2015). *Shadowing ondoku to eigo communication no kagaku* (Science of communication in English).Tokyo: Cosmo Pier.

Kormos, J. (2006). *Speech production and second language acquisition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Krashen, S. (2020). Lecture presented on October 24, 2020 at JACET Kansai Literature in Language Education Study Group.

Lai, F-K. (1993). The effects of summer reading course on reading and writing skills. *System*, *21*(1), 87-100.doi: https://doi.org/10.1016/0346-251X(93)90009-6

Language Testing International. (n.d.). ACTFL Speaking assessment: the Oral Proficiency Interview－computer. Retrieved Aug 16, 2020 from https://www.languagetesting.com/oral-proficiency-interview-by-computer-opic

Liu, N. & Littlewood, W. (1997). Why do many students appear reluctant to participate in classroom learning discourse? *System*, *25*(3), 371-384. doi: https://doi.org/10.1016/S0346-251X(97)00029-8

Lofting, H. (1988). *The Story of Doctor Dolittle.* NY: Dell Publishing.

MacQuillan, J. (2019). Where do we get our academic vocabulary? Comparing the efficiency of direct instruction and free voluntary reading. *The Reading Matrix, 19* (1), 129-138.

Mason, B., & Krashen, S. (1997). Extensive reading in English as a foreign language. *System*, *25*(1), 91-102. doi: https://doi.org/10.1016/S0346-251X(96)00063-2

Matsuda, S., Imura, M., & Nakanishi, N. (2018). Encouraging science students to talk using L1 readers. *The Fourth Extensive Reading World Congress Proceedings,* 145-154.

Meara, P. M. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjagr & J. Williams (Eds.), *Performance and competence in second language acquisition*, 33–51. Cambridge: CUP.

Ministry of Education, Culture, Sports, and Science and Technology. (2018) Heisei 29 nendo eigoryokuchosakekka (koukou 3nensei) no gaiyo. (The report on English skills of third-year senior high school students in 2017). Retrieved September 10, 2019, from http://www.mext.go.jp/a_menu/kokusai/gaikokugo/__icsFiles/afieldfile/2018/04/06/1403470_03_1.pdf

Mizumoto, A. (2015). Langtest (Version 1.0) [Web application]. Retrieved September 18, 2019, from http://langtest.jp

Nakanishi, N. (2018a). Development of Phoneme Counter: Open software for calculating phoneme counts in English texts. *Language Education & Technology*, 55, 199-216.

Nakanishi, N. (2018b). Phoneme Counter (Ver4.0) [computer Software]. Kobe: Kobe Gakuin University.

Nakanishi, N. (2019). Speech Saver (Ver 1.0) [Computer Software]. Kobe, Japan: Kobe Gakuin University.

Nakanishi, T., & Ueda, A. (2011). Extensive reading and the effect of shadowing. *Reading in a Foreign Language*, *23*(1), 1-16.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: CUP.

Nation, I. S. P. (2013). *What should every EFL teacher know?* Portsmouth: Compass Publishing.

Negishi, M. (2012). The development of the CEFR-J: Where we are, where we are going. Retrieved August 1,

2017, from http://www.tufs.ac.jp/common/fs/ilr/EU_kaken/_userdata/negishi2.pdf

Nishizawa, H., & Yoshioka, T. (2016). Longitudinal case study of a 7-year long ER program. *The Third Extensive Reading World Congress Proceedings*, 28-40. https://erfoundation.org/ERWC3-Proceedings.pdf

Nuttall, C. (1996). *Teaching reading skills in a foreign language* (2nd ed.). Oxford: Heinemann.

Oya,T., Manalo, E., & Greenwood, J. (2009). The influence of language contact and vocabulary knowledge on the speaking performance of Japanese students of English. *The Open Applied Linguistics Journal, 2*, 11-21.

Pearson. (2015). Test information booklet. Retrieved August, 16, 2020 from https://www.pearson.com/content/dam/one-dot-com/one-dot-com/english/SupportingDocs/Progress-test-information.pdf

Pearson. (n.d.) Versant technology. Retrieved Aug 16, 2020 from https://www.versanttests.com/technologyresearch

Pearson Education Inc. (2019). Versant English Test－Test description and validation summary. Retrieved August 17, 2020 from, https://www.pearson.com/content/dam/one-dot-com/one-dot-com/english/SupportingDocs/Versant/ValidationSummary/Versant-English-Test-Description-Validation-Report.pdf

Pitts, M, White, H. & Krashen, S. (1989). Acquiring second language vocabulary through reading: A replication of the clockwork orange study using second language acquirers. *Reading in a Foreign Language, 5*, 271-275.

Ray, H.A. (1973). *Curious George.* NY: HMH Books for Young Readers.

Read, J. (2005). Applying lexical statistics to the IELTS speaking test. *Research Notes, 20*, 12-16.

Rodrigo, V., Krashen, S., & Gibbons, B. (2004). The effectiveness of two comprehensible-input approaches to foreign language instruction at the intermediate level. *System, 32*(1), 53-60. doi: https://doi.org/10.1016/j.system. 2003.08.003

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

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics*, *54*(2), 79-95. doi: https://doi.org/10.1515/iral-2016-9991

Shimizu, Y., Kirimura, R., & Nozawa, T. (2014). Effects of short-term study abroad programs for economics students: Based on analysis of speaking tests. R*itsumeikan Koto Kyoiku Kenkyu, 14*, 91-102.

Takase, A. (2008). The two most critical tips for a successful extensive reading program. *Kinki University English Journal, 1*, 119-136.

Takeuchi, O. (2003).*Yoriyoi gaikokugogakushuho wo motomete.* (Searching for better language learning strategies: Studies on good language learners in the Japanese FL context.) Tokyo: Shohakusha.

Tsui, A. B. M. (1996). Reticence and anxiety in second language learning. In K. M. Bailey & D. Nunan (Eds.), *Voices from the language classroom* (pp. 145-167). Cambridge: CUP.

Turner, J. M., & Hiraga, M. K. (1996). Elaborating elaboration in academic tutorials: Changing cultural assumptions. In H. Coleman & L. Cameron (Eds.), *Change and language* (pp. 131-140). Clevedon: BAAL and Multilingual Matters.

Yoshida, H., Kagata, T., & Ikuma, Y. (2012). Implementing VERSANT: Spoken English test on program for promoting university education and student support based on different levels of English proficiency. (in Japanese). *The Bulletin of the English Society, 57*, 35-62.

Yoshizawa, K., Takase, A., & Otsuki, K. (2017). Does extensive reading help Japanese EFL learners to develop grammatical abilities? *JERA Bulletin, 10,* 7-28.

Abstract

This study, as a continuation of our former project, attempts to examine whether repetitive input and output using children's picture books can help university science majors improve their speaking skills. Seventeen science majors had nine sessions where they read and listened to *Oxford Reading Tree* (Hunt, 2011), followed by shadowing, repeating, and reading the story aloud. Three commercial online tests－*Progress, Versant*, and *OPIc* were also administered before and after the treatment. The project 2017 was the reproduction of the pilot study conducted in 2016 except that "retelling the story" was replaced with "reading the story aloud" to ensure more repetition. Furthermore, an original speaking test was added in order to further examine the students' utterances. The recorded sound data collected through the original speaking test were later transcribed and analyzed. Although the online test scores demonstrated no statistically significant changes except for the "overall" scores in *Versant*, some effect sizes were observed both in *Progress* and *Versant*. The phonological and lexical analyses of the recorded data using *Phoneme Counter*, *Praat*, and *Lextuter* displayed improvements in some aspects of oral proficiency. These findings suggest that repetitive input and output practice may have had a positive effect on utterance fluency, yet it may not have been sufficient to affect cognitive fluency. Further research and discussion are needed to determine how learners can acquire cognitive fluency as well as utterance fluency and what teachers can do to facilitate the process.