The Effects of Spoken Input on Learning the Spoken Forms of Second Language Words: Studies of Frequency of Exposure, Acoustic Variability, and Mode of Input

Takumi Uchihara, *The University of Western Ontario*

Supervisor: Webb, Stuart, *The University of Western Ontario*
Co-Supervisor: Saito, Kazuya, *University College London*

A thesis submitted in partial fulfillments of the requirements for the Doctor of Philosophy degree in Education

© Takumi Uchihara 2020

Follow this and additional works at: https://ir.lib.uwo.ca/etd

**Recommended Citation**
https://ir.lib.uwo.ca/etd/7402

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlsadmin@uwo.ca.
Abstract
This dissertation investigates the effects of three input factors—frequency of exposure, acoustic variability, mode of input—on learning productive knowledge of spoken forms of second language (L2) words. This thesis takes an integrated article format and is organized into (a) introduction (Chapter 1), (b) three main studies (Chapters 2, 3, & 4), and (c) conclusion (Chapter 5). The three studies involved Japanese university students learning 40 unfamiliar English words through encountering their spoken forms (and written forms in Study 3) while viewing pictures that conveyed their meanings. A picture-naming test was administered before, immediately after, and approximately one week after the treatment and the elicited speech samples were assessed for pronunciation and form-meaning connection. Study 1 (Chapter 2) investigated the effects of repetition. Seventy-five participants were randomly assigned to one of three conditions (1, 3, and 6 encounters), and their performance was assessed for pronunciation (accentedness, comprehensibility, processing time) and form-meaning connection (spoken form recall). Results showed that the number of exposures positively affected measures of form-meaning connection and pronunciation. Measurable learning gains occurred for comprehensibility after three encounters, while six encounters were necessary for foreign accent to be significantly reduced. Study 2 (Chapter 3) investigated acoustic variability and frequency of exposure. Eighty participants were randomly assigned to one of four conditions (3 encounters, 6 encounters 3 encounters with talker variability, and 6 encounters with talker variability). Spoken form recall and word stress accuracy were assessed. Results suggested that frequency of exposure promoted form-meaning mapping to a greater extent than talker variability, whereas talker variability had a stronger influence on word stress accuracy than frequency effects. Study 3 (Chapter 4)
investigated input modality. Seventy-five participants were randomly assigned to one of three conditions (reading-while-listening, reading-only, listening-only). The elicited speech was assessed for spoken form recall, accentedness, and comprehensibility. Results showed that the reading-while-listening group outperformed the listening-only group in form recall. The reading-while-listening and listening-only groups sounded more nativelike and comprehensible compared to the reading-only group. This dissertation concludes with implications for researching and teaching L2 vocabulary as well as suggestions for future studies.

**Keywords**: Frequency, Acoustic variability, Mode of input, Second language vocabulary learning, Second language pronunciation learning, Accent, Comprehensibility, Word stress

**Summary for Lay Audience**

Learners are likely to pick up second language (L2) words through seeing or hearing the forms of new words while reading books, watching television, and listening to songs. Encountering words in speech and writing therefore is an important source of input for learners to build L2 word knowledge. This dissertation explores the best ways to optimize input for enhancing productive knowledge of spoken forms (i.e., pronunciation) of unknown L2 words. It consists of three empirical studies focusing on Japanese university students studying 40 unfamiliar English words while viewing their meanings conveyed through their pictures. Before, immediately after, and approximately one week after the treatment, participants completed a word production test, and the elicited speech was evaluated for pronunciation accuracy. The word learning format (i.e., paired-associate learning) and test format (i.e., picture-naming
test) were identical across the three studies but different in the input participants encountered during the treatment (i.e., input repetition, input variability, input modality). Study 1 investigated the extent to which learners can improve pronunciation of L2 words after hearing the spoken forms of target words repeatedly (1, 3, and 6 encounters). Study 2 investigated the extent to which learners benefit from hearing L2 words produced by different speakers (3 and 6 encounters with speaker variability) versus a single speaker (3 and 6 encounters without speaker variability). Study 3 investigated how mode of input (reading, listening, reading while listening) affects pronunciation of L2 words. Results showed that learners tended to be more accurate at pronouncing L2 words after hearing words repeatedly, and the effect of repetition was enhanced when learners heard the spoken forms produced by multiple speakers compared to a single speaker. Learners encountering spoken input were also better able to pronounce the words than learners encountering only written input. These findings suggest the importance of input repetition, input variability, and input modality for developing productive knowledge of spoken forms of L2 words. To conclude, I discuss several implications for researching and teaching the spoken forms of L2 words.

**Co-Authorship statement**

Three studies (Chapters 2, 3, & 4) are currently under review as co-authored papers. Study 1 was submitted to *Language Learning* on May 24th, 2020. Study 2 was submitted to *Studies in Second Language Acquisition* on August 3rd, 2020. Study 3 was submitted to *The Modern Language Journal* on August 21, 2020. For all studies, I am solely responsible for designing data collection materials (100% student contribution), recruiting participants (100%), collecting, processing, and analyzing the
data (100%), and preparing the manuscript (100%). The contribution made by the other authors was in the form of providing feedback and consultation throughout all stages of my research project, ranging from generating research ideas and questions, developing data collection materials (e.g., tests, visual stimuli, target items), conducting statistical analysis, and finalizing the manuscript. At the stage of manuscript preparation, each of the three papers was carefully read by the other authors multiple times and feedback was provided in many forms (e.g., general comments on conceptual, methodological, statistical, and stylistic issues, editing, suggestion for additional references), but I was responsible for deciding whether to accept and reflect their feedback in the finalized version of the manuscript.

Acknowledgements
This research was supported by the Language Learning Dissertation Grant Program (grant number: R5370A13). My PhD study at the University of Western Ontario was supported by the Japan Student Services Organization (JASSO). First, I would like to thank Dr. Stuart Webb for his great support as a hugely dedicated and supportive supervisor and mentor throughout my PhD journey. My special thanks also go to Dr. Kazuya Saito and Dr. Pavel Trofimovich for their immense help from the very beginning of my academic career. I would like to extend my gratitude for providing constructive feedback on my research design and data analysis, and helping me collect data and develop research materials to the following people: Shuhei Kudo, Dr. Michael Karas, Dr. Tetsuo Harada, Dr. Frank Boers, Dr. Paul Tremblay, Dr. Joan Mora, Dr. Ana Pellicer-Sánchez, Dr. Yuichi Suzuki, Dr. Mafuyu Kitahara, Dr. Yasuyo Sawaki, Dr. Masahisa Hirano, Dr. Yasuaki Shinohara, Dr. Toshinori Yasuda, Dr. Randy Appel, Takayuki Konishi, Misao Iwai, Martha Black, Emi Iwaizumi, Ciaran Bleak,
Akifumi Yanagisawa, Kaitlyn Yanagisawa, Su Kim, Masaki Eguchi, Shungo Suzuki, Yui Suzukida, Ryo Moriya, Takuma Fuchigami, Aoi Ichiyama, all students who participated in the main experiments and pilot studies, and vocabulary group members at Western. Finally, I am grateful to my family for supporting me all the time and encouraging me to pursue my dream.
Table of Contents

Abstract ........................................................................................................................................ ii
Summary for Lay Audience ........................................................................................................ iii
Co-Authorship statement ........................................................................................................... iv
Acknowledgements .................................................................................................................. v
List of Tables ............................................................................................................................... xii
List of Figures ............................................................................................................................... xiv
List of Appendices ...................................................................................................................... xvi
List of Abbreviations ................................................................................................................ xvii
Chapter 1: Introduction ............................................................................................................. 1
  1.2 Motivation for the Current Research ................................................................................... 8
  1.3 References .......................................................................................................................... 12
Chapter 2: To what extent does frequency of exposure affect accentedness and
comprehensibility in learners’ pronunciation of second language words? ..................... 16
  2.1 Introduction ......................................................................................................................... 16
  2.2 Background ........................................................................................................................ 17
    2.2.1 Accentedness and Comprehensibility ........................................................................... 17
    2.2.2 Repetition and Vocabulary Learning .......................................................................... 18
    2.2.3 Repetition and Pronunciation Learning ..................................................................... 20
    2.2.4 Word-Related Factors ............................................................................................... 21
      2.2.4.1 Cognateness ......................................................................................................... 22
      2.2.4.2 Word length ......................................................................................................... 22
      2.2.4.3 Phonotactic regularity ......................................................................................... 23
2.2.5 Motivation for the Current Study ................................................................. 23
2.3 Method ............................................................................................................. 25
  2.3.1 Overview of the Study .............................................................................. 25
  2.3.2 Participants ............................................................................................... 26
  2.3.3 Target Items .............................................................................................. 26
  2.3.4 Treatment and Testing ............................................................................. 29
  2.3.5 Procedure ................................................................................................ 30
  2.3.6 Vocabulary and Pronunciation Measures ................................................ 31
  2.3.7 Preliminary Analysis ................................................................................ 32
2.4 Results ............................................................................................................ 34
  2.4.1 Spoken Form Recall .................................................................................. 34
  2.4.2 Accentedness, Comprehensibility, and Processing Time ......................... 37
  2.4.3 Cognateness, Word Length, and Phonotactic Probability ......................... 39
    2.4.3.1 Cognateness ....................................................................................... 39
    2.4.3.2 Phoneme count ................................................................................. 42
    2.4.3.3 Phonotactic probability ................................................................. 43
2.5 Discussion ........................................................................................................ 45
  2.5.1 Repetition Effects on Spoken Form Recall and Pronunciation ................. 45
  2.5.2 Word-Related Factors and Repetition Effects on Pronunciation Learning 48
2.6 Implications and Future Directions .............................................................. 51
2.7 Notes ............................................................................................................... 54
2.8 References ...................................................................................................... 55

Chapter 3: To what extent do acoustic variability and frequency of exposure affect learning the spoken forms and form-meaning connections of second language words? ........................................................................................................... 61
3.1 Introduction ........................................................................................................ 61
3.2 Background .......................................................................................................... 62
   3.2.1 Word Knowledge, Processing, and Learning .................................................. 62
   3.2.2 Acoustic Variability and L2 Pronunciation Learning ...................................... 65
   3.2.3 Acoustic Variability and L2 Vocabulary Learning .......................................... 67
   3.2.4 Frequency of Exposure and L2 Word Learning .............................................. 68
   3.2.5 The Current Study ......................................................................................... 69
3.3 Method ................................................................................................................ 72
   3.3.1 Overview of the Study .................................................................................... 72
   3.3.2 Participants .................................................................................................... 72
   3.3.3 Target Items .................................................................................................. 73
   3.3.4 Treatment and Testing ................................................................................... 74
   3.3.5 Procedure ..................................................................................................... 76
   3.3.6 Dependent Measures ..................................................................................... 77
   3.3.7 Data Analysis ............................................................................................... 79
3.4 Results ................................................................................................................ 80
   3.4.1 Spoken Form Recall ...................................................................................... 80
   3.4.2 Stress Placement Accuracy ............................................................................ 82
   3.4.3 Vowel Duration Ratio ................................................................................... 84
3.5 Discussion ............................................................................................................ 86
3.6 Limitations and Future Directions ..................................................................... 92
3.7 Notes ................................................................................................................... 93
3.8 References .......................................................................................................... 94

Chapter 4: To what extent does mode of input affect form-meaning connection,
accentedness, and comprehensibility of second language words? ..................... 100
4.1 Introduction

4.2 Background

4.2.1 Written and Spoken Input and L2 Vocabulary Learning

4.2.2 Orthographic Influence in L2 Pronunciation Learning

4.2.3 Pronunciation Constructs: Comprehensibility and Accentedness

4.2.4 Motivation for the Current Study

4.3 Method

4.3.1 Overview of the Study

4.3.2 Participants

4.3.3 Target Items

4.3.4 Treatment and Testing

4.3.5 Procedure

4.3.6 Form-Meaning Connection and Pronunciation Measures

4.3.7 Data Analysis

4.4 Results

4.4.1 Spoken Form Recall

4.4.2 Accentedness and Comprehensibility

4.5 Discussion

4.6 Implications, Limitations, and Future Directions

4.7 Note

4.8 References

Chapter 5: Conclusion

5.1 Summary of Study 1 (repetition)

5.2 Summary of Study 2 (acoustic variability)

5.3 Summary of Study 3 (mode of input)
5.4 General Implications ................................................................. 142
5.5 General Limitations and Future Directions ................................... 144
5.6 References ............................................................................. 146
Appendices .................................................................................... 148
  Appendix A: Target Words ............................................................. 148
  Appendix B: Visual Prompts ............................................................ 150
  Appendix C: Ethics Approval .......................................................... 157
List of Tables

Chapter 1
Table 1 – What is involved in word knowledge (adapted from Nation, 2020, p. 16)

Chapter 2
Table 1 – Target words
Table 2 – Relationships between the three pronunciation measures
Table 3 – Means, standard deviations, and 95% confidence intervals of the pretest, immediate posttest, and delayed posttest scores on spoken form recall
Table 4 – Means, standard deviations, and 95% confidence intervals for pretest, immediate posttest, and delayed posttest pronunciation scores by group
Table 5 – Comparing effect sizes (Cohen’s $d$) for the immediate posttest between current study and earlier studies measuring productive knowledge of orthography
Table 6 – Pronunciation scores on cognates across three exposure groups at immediate posttest

Chapter 3
Table 1 – Sequence of talker presentations for four experimental groups
Table 2 – Means, standard deviations, and 95% confidence intervals for pretest, immediate posttest, and delayed posttest pronunciation scores by group
Table 3 – Summary of results

Chapter 4
Table 1 – 40 Target words with basic Item information and sound-spelling consistency scores

Table 2 – Means, standard deviations, and 95% confidence intervals for spoken form recall, accentedness, and comprehensibility by group

Chapter 5

No Tables
List of Figures

Chapter 1
No Figures

Chapter 2
Figure 1 – Exposure group means on spoken form recall over time
Figure 2 – Mean accentedness ratings by group and cognate status (immediate posttest)
Figure 3 – Mean comprehensibility ratings by group and cognate status (immediate posttest)
Figure 4 – Mean processing time by group and cognate status (immediate posttest)
Figure 5 – Mean comprehensibility ratings by group and cognate status (delayed posttest)
Figure 6 – Mean processing time by group and cognate status (delayed posttest)
Figure 7 – Relationship between the number of phonemes and processing time per group (immediate posttest)
Figure 8 – Relationship between phonotactic probability of target words and accentedness ratings per group (immediate posttest)
Figure 9 – Relationship between phonotactic probability of target words and comprehensibility ratings per group (immediate posttest)

Chapter 3
Figure 1 – Group means for spoken form recall over time. Error bars represent 95% confidence intervals around the mean
Figure 2 – Group means for stress placement accuracy over time. Error bars represent
95% confidence intervals around the mean

Figure 3 – Group means for vowel duration ratio over time. Error bars represent 95% confidence intervals around the mean

Chapter 4

Figure 1 – Group means for spoken form recall over time. Error bars represent standard errors (SE) of the means

Figure 2 – Group means for accentedness over time. Error bars represent standard errors (SE) of the means

Figure 3 – Group means for comprehensibility over time. Error bars represent standard errors (SE) of the means

Figure 4 – Relationship between sound-spelling consistency and accentedness per group (immediate posttest)

Figure 5 – Relationship between sound-spelling consistency and accentedness per group (delayed posttest)

Figure 6 – Relationship between sound-spelling consistency and comprehensibility per group (immediate posttest)

Figure 7 – Relationship between sound-spelling consistency and comprehensibility per group (delayed posttest)

Chapter 5

No Figures
List of Appendices

Appendix A – Target Words

Appendix B – Visual Prompts

Appendix C – Ethics Approval
### List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AV</td>
<td>Acoustic Variability</td>
</tr>
<tr>
<td>BNC</td>
<td>British National Corpus</td>
</tr>
<tr>
<td>COCA</td>
<td>Corpus of Contemporary American English</td>
</tr>
<tr>
<td>E1</td>
<td>One Encounter</td>
</tr>
<tr>
<td>E3</td>
<td>Three Encounters</td>
</tr>
<tr>
<td>E6</td>
<td>Six Encounters</td>
</tr>
<tr>
<td>EFL</td>
<td>English as a Foreign Language</td>
</tr>
<tr>
<td>ESL</td>
<td>English as a Second Language</td>
</tr>
<tr>
<td>HVPT</td>
<td>High Variability Phonetic Training</td>
</tr>
<tr>
<td>L1</td>
<td>First Language</td>
</tr>
<tr>
<td>L2</td>
<td>Second Language</td>
</tr>
<tr>
<td>LO</td>
<td>Listening Only</td>
</tr>
<tr>
<td>RO</td>
<td>Reading Only</td>
</tr>
<tr>
<td>RWL</td>
<td>Reading While Listening</td>
</tr>
<tr>
<td>TOPRA</td>
<td>Type of Processing – Resource Allocation</td>
</tr>
</tbody>
</table>
Chapter 1: Introduction

This chapter starts with a brief overview of theory relevant to this dissertation, followed by the rationale for the current research. After Chapter 1, three studies will be introduced: Chapter 2 (Study 1: Frequency of exposure), Chapter 3 (Study 2: Acoustic variability), and Chapter 4 (Study 3: Mode of input). Chapter 5 concludes the dissertation with a brief summary of the findings, discussion of the implications for researching and teaching L2 vocabulary, and suggestions for future studies.

1.1 Theoretical Background

This section first reviews how frequency plays a role in L2 vocabulary acquisition along with empirical evidence documented over the past few decades. Second, it discusses what constitutes L2 vocabulary acquisition from the perspective of limited cognitive capacity and language processing. Third, it extends the discussion of input quantity (i.e., frequency of exposure) to input quality (i.e., acoustic variability), and reviews how exposure to acoustically varied input affects the acquisition of the spoken forms of L2 words. Finally, it discusses pronunciation as a construct and introduces how it has been assessed in earlier L2 speech research.

1.1.1 Frequency Effects on L2 Word Learning

Usage-based theory posits that “the more times we experience something, the stronger our memory of it, and the more fluently it is accessed” (Ellis, 2014, p. 195). It is frequency of usage that explains all levels of language representation and acquisition including vocabulary, grammar, and pronunciation learning (Ellis, 2002). Empirical evidence supports the significant role of frequency in learning knowledge of form-meaning connections in both L1 and L2 (Webb, 2014; Webb & Nation, 2017).
Jenkins, Stein, and Wysocki (1984) demonstrated that school-age native speakers of English learn more words during reading as the number of encounters increases. Saragi, Nation, and Meister (1978) found a moderate relationship between frequency of encounters and learning gains \((r = .34)\) with native speakers of English. Elley’s (1989) study focusing on school-age native speakers of English indicates that learning tends to increase during listening as the number of exposures to target words increases \((r = .43 \text{ in Study 1, } r = .60 \text{ in Study 2})\). Similarly, in L2 research, frequency plays a crucial role in vocabulary learning. Hulstijn, Hollander, and Greidanus (1996) found that advanced Dutch learners of French acquired significantly more L2 words with three encounters than one encounter while reading a short story. Malone (2018) found that four encounters led to higher learning gains than two encounters in knowledge of form-meaning link in strictly controlled conditions. Rott (1999) found that six encounters resulted in greater learning gains than two or four encounters. Pellicer-Sánchez (2016) found that after eight encounters with nonwords, they were read in a similar manner to previously known real words, suggesting that durable learning may occur with eight encounters. Measuring ten aspects of word knowledge, Webb (2007) found that ten encounters yielded higher learning gains than one encounter for every vocabulary measure and seven encounters on four of the ten measures. Pellicer-Sánchez and Schmitt (2010) found that advanced learners reading a novel for one month received higher gain scores for words that had more than 10 encounters than those that had fewer than 10 encounters. Waring and Takaki (2003) confirmed positive frequency effects with 4–5 encounters yielding higher gains than one encounter and 8–10 encounters yielding higher gains than 4–5 encounters. A recent meta-analysis of 28 primary studies that included different modes of input (e.g., reading, listening, reading while listening, viewing) suggests a medium effect size of the relationship.
between frequency exposures and vocabulary learning, $r = .34$ (Uchihara et al., 2019).

Although findings of repetition effects are mostly based on written input (Chen & Truscott, 2010; Rott, 1999; Waring & Takaki, 2003; Webb, 2007), studies have started exploring spoken input, such as listening to academic lectures (Vidal, 2011), songs (Pavia et al., 2019), TV interviews (van Zeeland & Schmitt, 2013), and viewing full-length TV programs (Peters & Webb, 2018). It appears that repetition effects are diminished in spoken input compared to written input. For example, van Zeeland and Schmitt (2013) found that 15 encounters with spoken words did not lead to larger gains than seven or 11 encounters in knowledge of form, grammar, or meaning at either immediate or delayed posttests. Vidal (2011) reported a larger correlation between frequency of encounters and learning gains in reading ($r = .687$) than in listening ($r = .488$). One explanation for these results is that, during listening, learners experience difficulty segmenting continuous speech, which makes it harder for them to identify target words and notice them appearing multiple times (Vidal, 2011). However, this research area is still in its infancy; few studies have been conducted with limited aspects of word knowledge tested, and most results have focused on receptive knowledge (e.g., form/meaning recognition). To our knowledge, no studies have examined the effects of repetition on productive knowledge of spoken word forms (i.e., spoken form recall), let alone knowledge of pronunciation.

1.1.2 The Type of Processing – Resource Allocation (TOPRA) Model

In order to describe the process of L2 word learning, it is important to understand that learners have a limited capacity of attentional resources that they can allocate to different components of word knowledge. The Type of Processing –
Resource Allocation (TOPRA) Model (Barcroft, 2002, 2015; Kida, 2020; Kida & Barcroft, 2018) proposes that word knowledge can be divided into three components: form, meaning, and form-meaning mapping, and that word learning is constrained by limited cognitive capacity. Also, it is assumed that processing of form, processing of meaning, and processing of form-meaning mappings can operate largely independently. It is therefore hypothesized that learners whose attention is directed to a given component (e.g., meaning) may not have much attentional resources left for processing another component (e.g., form). Therefore, test takers who have better performance when the former is tested may have worse performance when the latter is tested. This hypothesis has been confirmed with accumulated evidence showing that learners engaging in a form-elaboration task (i.e., counting letters) outperformed learners completing a meaning-elaboration task (i.e., making pleasantness ratings about words) in recall of word forms, and the latter outperformed the former in recall of word meanings (Barcroft, 2002; see Kida, 2020; Kida & Barcroft, 2018 for recent research supporting the model). The TOPRA model helps explain why there is little cross-over between vocabulary and pronunciation fields in L2 research. Pronunciation researchers tend to use familiar or high-frequency words as target items with learners allocating their attentional resources exclusively to processing word forms (i.e., pronunciation). In contrast, little room is left for learning knowledge of form-meaning connection. The primary purpose of vocabulary researchers, on the other hand, is to explore ways to optimize the mapping of forms to meanings by using unfamiliar or low-frequency words. Consequently, limited attention is remaining for processing and enhancing knowledge of spoken forms.

1.1.3 Acoustic Variability and L2 Word Learning
Research has consistently shown a positive effect of acoustic variability on L2 vocabulary learning using measures of form-meaning connection (Barcroft & Sommers, 2014). Barcroft and Sommers (2005) used two recall tests—meaning recall (L2-to-L1 recall) and form recall (picture-to-L2 recall)—and compared three variability conditions. In their within-participants study, L1 English speakers with no prior formal instruction in Spanish completed a paired-associate word learning task in which they studied Spanish words while hearing the spoken forms of target items and viewing the pictures conveying their meanings. Participants learned 24 words, eight of which were presented in one of three conditions: high variability (6 occurrences produced by 6 different talkers), moderate variability (6 occurrences produced by 3 different talkers repeating each word twice), and low variability (6 occurrences of all words produced by a single talker). The results of meaning and form recall tests suggested that the words learned under high variability were recalled significantly more accurately compared to those learned under moderate variability, and both sets of words were recalled more accurately than those learned under low variability. Barcroft and Sommers concluded that acoustic variability is beneficial in developing knowledge of form-meaning connections of L2 words because it allows learners to process, encode, and store indexical information relevant to the L1 perceptual system, leading to a more distributed (robust) representation of the word form. A recent study (Sinkeviciute et al., 2019) investigated whether learner’s age moderates the positive effects of input variability on L2 vocabulary learning. In this study, English-speaking learners of different ages with no experience with a target language (Lithuanian) heard eight repetitions of six new words produced by a single talker (low-variability condition) or eight talkers (high-variability condition), and post-training performance was measured through meaning recognition (picture-to-word matching) and form
recall (picture-to-word recall) tests. The results were consistent with earlier studies showing beneficial effects of high variability for adult learners on form recall (but not on meaning recognition). However, no such benefit was observed for groups of children (7- to 8-year-olds and 10- to 11-year-olds), either in meaning recognition or form recall.

1.1.4 Pronunciation as a Construct

The construct of pronunciation is thought to be simply restricted to “accent improvement” or “accent modification” in general, but it is rather complex and multifarious in nature (Brinton, 2017, p. 259). The conceptual complexity involved in pronunciation is also evidenced by the absence of one-size-fits-all measurements, and the fact that there are numerous ways of operationalizing its construct. Broadly speaking, there are two ways to measure pronunciation, acoustic analysis and human rating. Acoustic analysis is characterized as an objective measurement involving quantification of given pronunciation features such as segmental (Lambacher et al., 2005; Piske et al., 2001), prosodic (Trofimovich & Baker, 2006), and temporal (Suzuki & Kormos, 2020) qualities of L2 speech. (Segmental features refer to vowels and consonants; prosodic or suprasegmental features refer to larger units of sounds beyond the segmental level, e.g., stress, intonation, rhythm, and tone). By way of illustration, word stress accuracy can be measured in terms of vowel duration ratio (i.e., duration ratio of unstressed to stressed vowels). Researchers often use speech-analysis software (e.g., Praat; Boersma & Weenink, 2014) to measure the duration (in milliseconds) of stressed and unstressed vowels (Lee et al., 2006). In English language, successful reduction of unstressed vowels in duration is one of the important characteristics determining acquisition of word stress (Beckman &
Pierrehumbert, 1986) and general oral proficiency (Trofimovich & Isaacs, 2012). In contrast, human rating is characterized as a subjective way of measuring pronunciation ability, since pronunciation scores are derived from trained or untrained listeners who rate speech samples in reference to given pronunciation criteria. The criteria that have been used in L2 speech research include global constructs such as accentedness (i.e., degree of nativelikeness) and comprehensibility (i.e., ease of understanding) (Derwing & Munro, 2015). Since Munro and Derwing’s (1995a) seminal study, accentedness and comprehensibility have been widely researched and measured in L2 pronunciation studies (Thomson & Derwing, 2015). Accentedness (or linguistic nativelikeness) is defined as listeners’ judgments of how different L2 speech sounds from the expected language variety, and comprehensibility refers to listeners’ perceived ease or difficulty of understanding L2 speech. These two constructs are measured through listeners’ ratings of speakers, using numerical point scales (e.g., 1 = no accent, 9 = heavily accented; 1 = easy to understand, 9 = hard to understand). Comprehensibility is often distinguished from intelligibility, which captures listeners’ actual understanding of L2 speech, measured through listener transcription of heard words or utterances (Field, 2005). However, conceptualized broadly, comprehensibility is an intuitive and easy-to-use measure used frequently as an alternative metric of listener understanding of words and utterances (Levis, 2005).

In this section, we looked at how frequency and input variability may enhance L2 vocabulary acquisition. The TOPRA model helps us understand the role of attention in facilitating (or preventing) the development of different aspects of word knowledge (i.e., form, meaning, and form-meaning mapping). It is also important to consider pronunciation as a multifaceted construct, rather than assuming that it is limited to accent reduction or elimination. The following section discusses why the
current research investigating the role of spoken input in learning productive knowledge of the spoken forms of L2 words is necessary and possible implications from this research for L2 vocabulary teaching and research.

1.2 Motivation for the Current Research

Input is fundamental for developing all linguistic aspects (e.g., grammar, lexis, phonology, and pragmatics) in first language (L1) and second language (L2) acquisition (Ellis, 2002). In L2 vocabulary acquisition, encountering the written forms of L2 words while reading texts (e.g., graded readers) helps learners to gradually and incrementally develop knowledge of the words encountered over time (Webb & Nation, 2017). Nation’s description of vocabulary knowledge (Nation, 2013, 2020), proposes that lexical knowledge consists of nine aspects under each of three categories (i.e., form, meaning, use) at receptive and productive dimensions (see Table 1 for the different aspects of word knowledge). The vast majority of earlier studies investigating L2 vocabulary learning have measured form-meaning connection using tests of meaning recognition (e.g., multiple-choice tests), form recognition (e.g., checklist tests), meaning recall (e.g., L2-to-L1 translation), and form recall (e.g., L1-to-L2 translation) (see Uchihara et al., 2019; Webb et al., in press; Yanagisawa et al., 2020 for a review). More recently, studies have started measuring aspects of word knowledge other than form-meaning connection, particularly with greater attention being paid to knowledge of collocation (Boers & Webb, 2018).

Surprisingly, few studies of L2 vocabulary learning have focused on productive knowledge of spoken forms or pronunciation. Researchers tend to measure word knowledge in written form, asking learners to see visual prompts (e.g., meanings) and write responses (e.g., spellings). Although some researchers have
highlighted the important role of test modality (written vs. spoken) and assessed spoken word knowledge (Jelani & Boers, 2018; Peters, 2019; Peters & Webb, 2018), almost all studies have measured receptive knowledge of spoken forms (e.g., form recognition). The lack of research examining productive knowledge of the spoken forms of words is surprising given that it is crucial for L2 speakers success in oral communication. Even if learners are able to produce the spoken forms of L2 words, it is important to further ensure that the produced forms are sufficiently accurate and ultimately intelligible/comprehensible to listeners so that L2 speakers are successful in oral communication (Derwing & Munro, 2015). One possible reason for the lack of research investigating productive knowledge of spoken form might be due to a tendency in vocabulary research to examine learning from written input (e.g., reading). Spoken input receives less research attention possibly because researchers hold the view that listening is less effective for vocabulary learning than reading (Brown et al., 2008; Bürki, 2010; Vidal, 2011). However, such a view might be biased by results based on tests measuring form-meaning connection in written form. The value of spoken input for vocabulary learning may be underestimated due to the limited aspects of word knowledge tested so far, and little attention being paid to the pronunciation of L2 words. It is therefore important to measure productive knowledge of the spoken forms of words in order to examine the true value of spoken input. It is also pedagogically useful to explore different ways to optimize spoken input for learning the pronunciation of words.

Table 1. What is involved in word knowledge (adapted from Nation, 2020, p. 16)

<table>
<thead>
<tr>
<th>Form</th>
<th>Spoken</th>
<th>Written</th>
<th>Word parts</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>What does the word sound like?</td>
<td>What does the word look like?</td>
<td>What parts are recognizable in this word?</td>
</tr>
<tr>
<td>P</td>
<td>How is the word pronounced?</td>
<td>How is the word written and spelled?</td>
<td></td>
</tr>
<tr>
<td>Meaning</td>
<td>Form and meaning</td>
<td>P</td>
<td>What word parts are needed to express the meaning?</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------------------------------------------------------</td>
<td>---</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>Concept and referents</td>
<td>R What meaning does this word form signal?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P What word form can be used to express this meaning?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Associations</td>
<td>R What is included in the concept?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P What items can the concept refer to?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use</td>
<td>Grammatical functions</td>
<td>R</td>
<td>In what patterns does the word occur?</td>
</tr>
<tr>
<td></td>
<td>P In what patterns must we use this word?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collocations</td>
<td>R What other words does this make us think of?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P What other words could we use instead of this one?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constraints on use</td>
<td>R Where, when, and how often would we expect to meet</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>this word?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P Where, when, and how often can we use this word?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note. R = receptive; P = productive.*

The present study was designed to respond to these research gaps with the aim of examining the value of spoken input for improving not only form-meaning connection but also pronunciation. This research involves three studies, each of which explored different input factors to optimize the effect of spoken input on L2 vocabulary learning. All data used for three studies were elicited through a large-scale, single data collection from the same pool of participants (i.e., Japanese L1 speakers from the same university). Study 1 examined how repeated exposure to the spoken forms of unfamiliar words affects the learning of form-meaning connection (i.e., spoken form recall) and pronunciation (i.e., accentedness, comprehensibility, processing time). Study 2 investigated how acoustic variability and frequency of exposure affect the learning of form-meaning connection (i.e., spoken form recall) and pronunciation (i.e., word stress placement accuracy, vowel duration ratio). Study 3 examined how mode of input affects the learning of form-meaning connection (i.e., spoken form recall) and pronunciation (i.e., accentedness, comprehensibility).
There are several reasons why the current research is needed. First, it advances our understanding of an unexplored dimension of vocabulary learning. Earlier studies have tended to rely on written measures of form-meaning connection, and it is likely that our understanding of L2 vocabulary acquisition is heavily biased by such results. Second, spoken input has received little attention in vocabulary research. A few studies investigating spoken input have suggested the limited value of spoken input for L2 vocabulary acquisition. However, as highlighted above, these studies have tended to measure learning of limited aspects of word knowledge, and virtually no research has looked into pronunciation. It is important to consider test modality (spoken or written) in order to evaluate the true value of spoken input in L2 vocabulary learning. Third, there has been little discussion of the pronunciation construct in the vocabulary literature. In particular, vocabulary scholars have not discussed how input exposure or learning activities influence comprehensibility of L2 words. In L2 pronunciation research, it is agreed among scholars that language teachers should make the most of their class time to ensure that L2 speakers are sufficiently comprehensible to listeners instead of (or before) aiming to reduce foreign accent. Finally, this research can help bridge the gap between vocabulary and pronunciation research. In the field of second language research, pronunciation studies tend to measure the learning of words whose form–meaning connections are already established through targeting high-frequency items. Exploration of new words and lexical acquisition in parallel therefore will inform extant L2 pronunciation research, which will in turn update L2 vocabulary literature with more attention being paid to spoken modality.
1.3 References


Ellis, N. C. (2002). Frequency effects in language processing: A review with


Chapter 2: To what extent does frequency of exposure affect accentedness and comprehensibility in learners’ pronunciation of second language words?

2.1 Introduction

Frequency of encounters is a key determinant of first language (L1) and second language (L2) acquisition and processing (Ellis, 2002). One of the most extensively researched areas exploring frequency effects is incidental vocabulary acquisition (Uchihara et al., 2019). Earlier studies suggested varying numbers of encounters necessary for significant vocabulary learning to occur, spanning six (Rott, 1999), eight (Horst et al., 1998), 10 (Webb, 2007), and more than 20 encounters (Waring & Takaki, 2003). This line of research has advanced our understanding of frequency effects in vocabulary acquisition by measuring not only knowledge of form–meaning connection but also various aspects of word knowledge, including knowledge of collocation (Webb et al., 2013), grammar (van Zeeland & Schmitt, 2013), association (Horst et al., 1998), and spelling (Webb, 2007). However, findings are predominantly based on testing word knowledge in written form, and knowledge of pronunciation remains unexplored. The lack of attention to pronunciation in vocabulary research is surprising in view of the prominence placed on it as one of the fundamental aspects of word knowledge (Nation, 2013) and speaking proficiency (de Jong et al., 2012). The current study therefore aimed to bridge this research gap by exploring the effects of repeated encounters with the spoken forms of words on word pronunciation learning. This research should shed further light on our understanding of frequency effects in L2 acquisition and provide important implications for L2 pronunciation instruction.
2.2 Background

2.2.1 Accentedness and Comprehensibility

Since Munro and Derwing’s (1995a) seminal study, several global constructs, including accentedness and comprehensibility, have been widely researched in L2 pronunciation studies (Derwing & Munro, 2015). Accentedness (or linguistic nativelikeness) is defined as listeners’ judgments of how different L2 speech sounds from the expected language variety, and comprehensibility refers to listeners’ perceived ease or difficulty of understanding L2 speech. These two constructs are measured through listeners’ ratings of speakers, using numerical point scales (e.g., 1 = no accent, 9 = heavily accented; 1 = easy to understand, 9 = hard to understand). Comprehensibility is often distinguished from intelligibility, which captures listeners’ actual understanding of L2 speech, measured through listener transcription of heard words or utterances (Field, 2005). However, conceptualized broadly, comprehensibility is an intuitive and easy-to-use measure used frequently as an alternative metric of listener understanding of words and utterances (Levis, 2005).

Accentedness and comprehensibility appear to be at least partially independent. For example, L2 speakers with a stronger foreign accent do not necessarily sound less comprehensible or intelligible (Munro & Derwing, 1995a). Similarly, when listeners rate L2 utterances for comprehensibility and accentedness, the processing cost indicated by response latency data significantly predicts raters’ comprehensibility but not accentedness judgments (Ludwig & Mora, 2017; Munro & Derwing, 1995b), implying that the two constructs can be distinguished through a reaction-time measure. According to cross-sectional and longitudinal investigations (Derwing & Munro, 2013; Saito, 2015), L2 learners appear to continue to improve various dimensions of language relevant to comprehensibility, as long as they use the
target language daily. In contrast, although the degree of foreign accent greatly diminishes within early phases of L2 immersion (Derwing & Munro, 2013), its further development is likely followed by a plateau and may be limited to learners with greater phonetic aptitude (Granena & Long, 2013) and memory (Darcy et al., 2015). In light of prior work highlighting accentedness and comprehensibility as separate constructs, the current study targets these two constructs through three measures—scalar ratings of accentendness and comprehensibility and a reaction-time measure of processing time—to capture several dimensions of L2 pronunciation knowledge.

### 2.2.2 Repetition and Vocabulary Learning

Repetition is an important condition for L2 vocabulary learning (Webb & Nation, 2017). Although the positive effect of repetition was confirmed in deliberate vocabulary learning (Nakata, 2017), the prominence given to the effect has mostly stemmed from incidental vocabulary learning research (Uchihara et al., 2019). This line of research involves looking for the optimal number of encounters with words necessary for significant learning to occur while learners engage in a meaning-focused activity such as reading a short story (Horst et al., 1998), listening to songs (Pavia et al., 2019), and viewing television (Peters, 2019). Webb (2007) conducted an experimental study with Japanese English as a foreign language (EFL) learners who read sets of sentences including 10 target words. Participants were randomly assigned to one of four treatment groups that encountered the target words one, three, seven, and 10 times. After the treatment, learning was assessed in tests measuring receptive and production knowledge of five aspects of vocabulary knowledge (orthography, association, syntax, form–meaning link, and grammar). Webb found that repeated encounters promoted vocabulary learning, and also highlighted considerable variation
of the repetition effect across different knowledge aspects. At one encounter, sizable gains in both receptive and productive knowledge of orthography were found, such that five out of 10 target words were learned (50% gain). However, participants were less successful at learning form–meaning connections, as measured through a meaning recall test, where they demonstrated only a 29% gain. Building on Webb’s (2007) study, Chen and Truscott (2010) conducted a replication study in which participants encountered target words one, three, and seven times, and their learning was measured for receptive and productive knowledge of four aspects (orthography, form–meaning connection, grammar, association). The results showed the largest repetition effect between one and three encounters for productive knowledge of orthography ($d = 1.13$) compared to other knowledge aspects ($d = −0.07$ to 1.02). These findings suggest that formal aspects of word knowledge may be sensitive to repetition effects and that measurable learning gains might arise after a small number of encounters (e.g., one to three).

Although findings of repetition effects are mostly based on written input (Chen & Truscott, 2010; Rott, 1999; Waring & Takaki, 2003; Webb, 2007), studies have started exploring spoken input, such as listening to academic lectures (Vidal, 2011), songs (Pavia et al., 2019), TV interviews (van Zeeland & Schmitt, 2013), and viewing full-length TV programs (Peters & Webb, 2018). It appears that repetition effects are diminished in spoken input compared to written input. For example, van Zeeland and Schmitt (2013) found that 15 encounters with spoken words did not lead to larger gains than seven or 11 encounters in knowledge of form, grammar, or meaning at either immediate or delayed posttests. Vidal (2011) reported a larger correlation between frequency of encounters and learning gains in reading ($r = .687$) than in listening ($r = .488$). One explanation for these results is that, during listening,
learners experience difficulty segmenting continuous speech, which makes it harder for them to identify target words and notice them appearing multiple times (Vidal, 2011). However, this research area is still in its infancy; few studies have been conducted with limited aspects of word knowledge tested, and most results have focused on receptive knowledge (e.g., form/meaning recognition). To our knowledge, no studies have examined the effects of repetition on productive knowledge of spoken word forms (i.e., spoken form recall), let alone knowledge of pronunciation.

2.2.3 Repetition and Pronunciation Learning

The lack of research investigating the effects of repetition on pronunciation learning may be due to the discrepancy in modality between learning (i.e., listening) and testing (i.e., speaking). However, there are theoretical underpinnings that account for a close interface between L2 perception and production, supporting the hypothesis that repeated encounters with L2 words would first help establish new phonological representations in the brain, which will in turn result in their improved perception and production abilities. According to Flege’s (1995) speech learning model, difficulties in perception are responsible for difficulties in production. Once an adequate perceptual specification of an L2 sound is established, such that it is not confused with an L1 sound, production will become more accurate with continued exposure over time. This view that perception precedes production has been empirically tested via a considerable body of perceptual training studies (for a review, see Sakai & Moorman, 2018). In the seminal work conducted by Bradlow et al. (1997), Japanese learners who completed three to four weeks of input-only perception training (i.e., focusing on identification of English /r/ and /l/) showed improvement not only in their perception but also in their production accuracy for these target sounds.
Prior work on auditory word priming also provides support for the view that repetition promotes pronunciation learning. Auditory word priming refers to the phenomenon in which prior exposure to spoken words leads to more rapid processing of the same words at subsequent encounters (Church & Fisher, 1998). This processing advantage that repeated words have over unrepeated words is characterized by unconscious and unintentional facilitation, supporting the learning of spoken word forms. Such repetition-driven processing advantage for words, observed in L1 speakers, also appears to be available to L2 learners (Trofimovich, 2005; Trofimovich & Gatbonton, 2006). If L2 learners (as auditory word priming research suggests) are indeed more sensitive to spoken words they had recently encountered than words that they had not, then manipulating the frequency of encounters with spoken word forms has great pedagogical value for improving L2 learners’ pronunciation through classroom instruction.

2.2.4 Word-Related Factors

The number of encounters necessary to learn words varies because words have varying degrees of difficulty (Chen & Truscott, 2010; Ellis & Beaton, 1993; Lotto & de Groot, 1998; Peters, 2020; Webb, 2014; Webb & Nation, 2017). Words are different in many characteristics, including frequency, length, imageability, part of speech, cognateness, pronounceability, and concreteness, making some words easier to learn than others. Among the many factors that may affect the learning of L2 word pronunciation, three (cognateness, word length, and phonotactic regularity) were considered most relevant.
2.2.4.1 Cognateness

Cognates are typically defined as words that are phonologically or orthographically, semantically, and etymologically related across languages (Peters, 2020). However, this definition has been extended to word pairs that are shared across languages in form and meaning regardless of the presence or absence of an etymological relationship (Rogers et al., 2015). An example falling under this definition is loanwords in Japanese such as *cable* /ケーブル (keeburu) and *cup* /カップ (kappu). Research has consistently indicated that cognates are easier to learn than noncognates regardless of learning conditions. For instance, in paired-associate learning, learners were more accurate and faster at recalling the forms of cognates than noncognates with fewer encounters (Lotto & de Groot, 1998). Similarly, in incidental learning research, participants were likely to acquire cognates before noncognates (Peters, 2019; Peters & Webb, 2018), and the positive effect of cognateness might be larger for learning through spoken input than written input (Vidal, 2011). However, research is yet to compare the learning of pronunciation for cognates versus noncognates.

2.2.4.2 Word length.

Ellis and Beaton (1993) found a negative relationship between number of letters and word learning, and the effect was more salient in productive learning (written form recall) than in receptive learning (written meaning recall). One explanation why longer words might be more difficult to learn relates to the possibility that shorter words, compared to longer words, could be subvocally repeated more frequently before the auditory percept decays in short-term memory (Dahlen & Caldwell–Harris, 2013). However, longer words might not necessarily
increase the learning burden because they tend to be morphologically transparent and phonologically familiar to learners with words composed of recognizable parts (Service, 1998). Influence of word length therefore needs to be considered in combination with learners’ familiarity with allowable phonological sequences or phonotactic regularity of words.

2.2.4.3 Phonotactic regularity.

Language learners become gradually familiar with phonological sequence patterns or phonotactic rules specific to a language, for example, learning that no English word ends in /h/. Knowledge of phonotactics is implicit as it simply develops as a result of using language rather than explicitly analyzing the phonology of that language (Ellis, 2002). In examining how word-related factors affect English-speaking students learning German as a target language, Ellis and Beaton (1993) measured phonotactic regularity by calculating the summed probabilities of all bi-phonemes (i.e., two consecutive phonemes) in a German word occurring in an English language corpus. This probability measure indicated the degree of similarity in phonotactic patterns between English (L1) and German (L2). Their results showed that phonotactic probability was significantly correlated with the measure of productive learning, supporting the view that words that contain phonologically familiar sequences tend to be learned more easily.

2.2.5 Motivation for the Current Study

There are several reasons why research on pronunciation learning as a function of frequency of encounters is needed. First, it advances our understanding of an unexplored dimension of vocabulary learning. Second, it may indicate the extent to
which pronunciation needs to be taught explicitly. For the past 10 years, the number of pronunciation teaching studies has dramatically increased with evidence supporting the effect of explicit instruction (Saito & Plonsky, 2019). Despite a great deal of variation in teaching methods and approaches (Celce-Murcia et al., 2010), one feature in common across studies is that they all draw learners’ attention to target language features (e.g., Saito, 2011), including segmental and prosodic features (e.g., intonation, word stress, rhythm) by means of providing corrective feedback or metalinguistic information (e.g., place and manner of articulation). Despite the well-attested role of explicit instruction in pronunciation teaching (Saito & Plonsky, 2019; Sakai & Moorman, 2018), recent studies suggest that many instructors still lack training to teach pronunciation (Foote et al., 2016) and share the time constraints on L2 curricula leaving limited space for pronunciation instruction (Martin, 2020). The present study sheds light on this issue by revealing the extent to which pronunciation of L2 words can be learned as a by-product of input exposure alone without explicit attention drawn to specific phonological properties. Findings of this research might help relieve in-class time pressure arising as a result of increasing demands for integrating explicit pronunciation instruction into L2 curricula. Third, determining the number of encounters necessary to learn the pronunciation of L2 words might provide a useful guide to how new words can be introduced in the classroom. Such guidance will also help teachers reexamine the importance of providing spoken input when teaching new words. Given the growing primacy of producing comprehensible speech in international communication (Levis, 2005), it is crucial for pronunciation of newly learned words to be available for immediate use in oral exchange. Finally, this research can help bridge the gap between vocabulary and pronunciation research. In the field of second language research, pronunciation studies tend to measure the
learning of words whose form–meaning connections are already established through targeting high-frequency items. Exploration of new words and lexical acquisition in parallel therefore will inform extant L2 pronunciation research, which will in turn update L2 vocabulary literature with more attention being paid to spoken modality.

The present study was guided by the following research questions:

1. How does frequency of exposure (1, 3, and 6 exposures) influence learners’ recall of the spoken forms of previously unknown L2 words?

2. How does frequency of exposure (1, 3, and 6 exposures) influence three pronunciation aspects (accentedness, comprehensibility, and processing time) of learners’ production of previously unknown L2 words?

3. To what extent do cognateness, word length, and phonotactic regularity moderate the relationship between repetition and learners’ performance on measures of pronunciation learning?

2.3 Method

2.3.1 Overview of the Study

The study adopted a pre–post design with three experimental groups (one, three, and six encounters) and three testing trials (pretest, immediate posttest, and delayed posttest). Participants were randomly assigned to the three experimental groups and received different frequencies of exposures to target words: one encounter, three encounters, and six encounters. During the treatment, participants learned 40 English words through listening to the words and viewing their corresponding pictures. A picture-naming test was administered at the three testing times, and the elicited samples were evaluated for vocabulary and pronunciation measures.
2.3.2 Participants

Seventy-nine Japanese university EFL students in Japan participated in this experiment. Four participants were excluded from the subsequent analysis because they had lived abroad for an extended period of time (2–12 years). The remaining 75 participants had studied English for a minimum of six years in instructional settings. They had scored 90 percent or higher on the 1,000 word level of the Vocabulary Levels Test (Webb et al., 2017), and all except two had scored 80 percent or higher on the 2,000 word level of the test. Their mean score at the 2,000 level was 28.44, indicating that they had mastered that level, and they had receptive knowledge of almost all of the 2,000 most frequent words. The 75 participants were randomly assigned to three experimental groups: one encounter (E1), three encounters (E3), and six encounters (E6). There was no between-group difference in vocabulary test scores, $F(2, 72) = 1.70, p = .191$. All participants reported normal hearing.

2.3.3 Target Items

Forty target words were quasi-randomly selected from a pool of candidate words collected according to the following three criteria (Table 1). First, because the purpose of this study was to examine the learning of “unknown” or “new” words instead of already known words, a pool of low-frequency words was created by collecting English word items that were beyond the most frequent 5,000 word families in Nation’s BNC/COCA word lists (Nation, 2012). Second, because the treatment involved learning spoken forms attached to meanings conveyed in visual images (pictures), only concrete nouns were selected as candidate target items. Third, words that could be replaced with high-frequency synonyms were avoided to reduce the
possibility that high-frequency synonyms of the target items would be produced in the picture-naming test. Candidate target items were coded for three word characteristics: (a) cognateness, (b) number of phonemes, and (c) phonotactic probability.

Cognateness was determined by having four L1 Japanese-speaking raters judge whether the target word was a loanword (Rogers et al., 2015). If items were considered cognates by all raters, they were labelled as a cognate item. There was 90% agreement among the four raters. For phonotactic probability, positional segment frequency was measured using a web-based calculator of English phonotactic probability accessible at https://calculator.ku.edu/phonotactic/about (Vitevitch & Luce, 2004). Higher phonotactic probability scores indicate that the word contains phonemes that appear frequently in a given position in English language (e.g., probability of /str-/ occurring in a word-initial position is higher than that of the initial /dw-/ in English). Each of the 40 target words was recorded twice by a female native speaker of English using a TASCAM DR-05 audio recorder and digitized into a wav format (44.1 kHz sampling rate with 16-bit quantization). The better of the two productions was selected in terms of clarity, naturalness, and lack of background noise and then stored as an individual sound file, with peak intensity normalized using digital speech-analysis software (Praat). The stimuli were clear and comprehensible based on the judgement of another native English speaker.

Table 1. Target words

<table>
<thead>
<tr>
<th></th>
<th>Items</th>
<th>Cognate</th>
<th>Phoneme count</th>
<th>Phonotactic probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>abalone</td>
<td>No</td>
<td>7</td>
<td>0.040</td>
</tr>
<tr>
<td>2</td>
<td>acorn</td>
<td>No</td>
<td>5</td>
<td>0.058</td>
</tr>
<tr>
<td>3</td>
<td>armadillo</td>
<td>Yes</td>
<td>8</td>
<td>0.045</td>
</tr>
<tr>
<td>4</td>
<td>binoculars</td>
<td>No</td>
<td>10</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>Item</td>
<td>Choice</td>
<td>Location</td>
<td>Score</td>
</tr>
<tr>
<td>---</td>
<td>--------------</td>
<td>--------</td>
<td>----------</td>
<td>-------</td>
</tr>
<tr>
<td>5</td>
<td>caramel</td>
<td>Yes</td>
<td>6</td>
<td>0.077</td>
</tr>
<tr>
<td>6</td>
<td>carousel</td>
<td>No</td>
<td>7</td>
<td>0.070</td>
</tr>
<tr>
<td>7</td>
<td>catapult</td>
<td>No</td>
<td>8</td>
<td>0.064</td>
</tr>
<tr>
<td>8</td>
<td>celery</td>
<td>Yes</td>
<td>6</td>
<td>0.064</td>
</tr>
<tr>
<td>9</td>
<td>chameleon</td>
<td>Yes</td>
<td>8</td>
<td>0.062</td>
</tr>
<tr>
<td>10</td>
<td>chandelier</td>
<td>Yes</td>
<td>9</td>
<td>0.046</td>
</tr>
<tr>
<td>11</td>
<td>chisel</td>
<td>No</td>
<td>4</td>
<td>0.046</td>
</tr>
<tr>
<td>12</td>
<td>cicada</td>
<td>No</td>
<td>6</td>
<td>0.070</td>
</tr>
<tr>
<td>13</td>
<td>clover</td>
<td>Yes</td>
<td>6</td>
<td>0.032</td>
</tr>
<tr>
<td>14</td>
<td>crayon</td>
<td>Yes</td>
<td>5</td>
<td>0.069</td>
</tr>
<tr>
<td>15</td>
<td>croissant</td>
<td>Yes</td>
<td>7</td>
<td>0.066</td>
</tr>
<tr>
<td>16</td>
<td>escalator</td>
<td>Yes</td>
<td>9</td>
<td>0.048</td>
</tr>
<tr>
<td>17</td>
<td>ladle</td>
<td>No</td>
<td>4</td>
<td>0.046</td>
</tr>
<tr>
<td>18</td>
<td>loquat</td>
<td>No</td>
<td>6</td>
<td>0.043</td>
</tr>
<tr>
<td>19</td>
<td>lotus</td>
<td>No</td>
<td>5</td>
<td>0.063</td>
</tr>
<tr>
<td>20</td>
<td>maracas</td>
<td>Yes</td>
<td>7</td>
<td>0.051</td>
</tr>
<tr>
<td>21</td>
<td>marshmallow</td>
<td>Yes</td>
<td>8</td>
<td>0.039</td>
</tr>
<tr>
<td>22</td>
<td>mermaid</td>
<td>Yes</td>
<td>6</td>
<td>0.035</td>
</tr>
<tr>
<td>23</td>
<td>pacifier</td>
<td>No</td>
<td>8</td>
<td>0.053</td>
</tr>
<tr>
<td>24</td>
<td>parakeet</td>
<td>No</td>
<td>8</td>
<td>0.065</td>
</tr>
<tr>
<td>25</td>
<td>persimmon</td>
<td>No</td>
<td>8</td>
<td>0.051</td>
</tr>
<tr>
<td>26</td>
<td>podium</td>
<td>No</td>
<td>6</td>
<td>0.052</td>
</tr>
<tr>
<td>27</td>
<td>porcupine</td>
<td>No</td>
<td>9</td>
<td>0.049</td>
</tr>
<tr>
<td>28</td>
<td>protractor</td>
<td>No</td>
<td>10</td>
<td>0.056</td>
</tr>
<tr>
<td>29</td>
<td>raccoon</td>
<td>No</td>
<td>5</td>
<td>0.054</td>
</tr>
<tr>
<td>30</td>
<td>raisin</td>
<td>Yes</td>
<td>4</td>
<td>0.050</td>
</tr>
<tr>
<td>31</td>
<td>razor</td>
<td>No</td>
<td>5</td>
<td>0.034</td>
</tr>
<tr>
<td>32</td>
<td>spatula</td>
<td>No</td>
<td>7</td>
<td>0.051</td>
</tr>
<tr>
<td>33</td>
<td>strainer</td>
<td>No</td>
<td>7</td>
<td>0.054</td>
</tr>
<tr>
<td>34</td>
<td>syringe</td>
<td>No</td>
<td>6</td>
<td>0.060</td>
</tr>
<tr>
<td>35</td>
<td>tadpole</td>
<td>No</td>
<td>6</td>
<td>0.043</td>
</tr>
<tr>
<td>36</td>
<td>toboggan</td>
<td>No</td>
<td>7</td>
<td>0.043</td>
</tr>
<tr>
<td>37</td>
<td>toupee</td>
<td>No</td>
<td>4</td>
<td>0.041</td>
</tr>
<tr>
<td>38</td>
<td>treadmill</td>
<td>No</td>
<td>7</td>
<td>0.039</td>
</tr>
<tr>
<td>39</td>
<td>walrus</td>
<td>No</td>
<td>6</td>
<td>0.048</td>
</tr>
<tr>
<td>40</td>
<td>xylophone</td>
<td>No</td>
<td>7</td>
<td>0.048</td>
</tr>
</tbody>
</table>
2.3.4 Treatment and Testing

Paired-associate vocabulary learning was implemented as a learning intervention. The learning and testing schedule was programmed with PsychoPy (Peirce, 2007). Before the treatment began, participants put on headphones equipped with a microphone (AT810 Cardioid Headset Microphone) and familiarized themselves with the vocabulary learning task by working through three practice examples. During the treatment, participants saw the meanings of the target words conveyed in visual images (i.e., copyright-free pictures retrieved from the Internet, standardized to a size of 400 × 400 pixels, see Appendix B for visual prompts) while hearing the spoken forms of the words. For each target item, the picture was displayed on the computer screen for 4 seconds, with the auditory presentation of the target word beginning 750 milliseconds (ms) after the picture appeared. The picture remained visible for the entire 4 seconds. A 2-second blank interval was inserted between trials. During the treatment, the 40 target items were presented in a sequence of eight blocks of five items. The different experimental groups (E1, E3, E6) received different numbers of encounters with the 40 target items. Thus, the total number of encounters with target items varied between the groups: 40 (= 40 × 1 encounter), 120 (= 40 × 3 encounters), and 240 (= 40 × 6 encounters) items in E1, E3, and E6, respectively. For all groups, the order of item presentation was randomized across participants. For E3 and E6, the interval between the first encounter and the next encounter with the same target word remained constant to control for spacing effects.

Immediately after the final exposure to each block of five items, a picture-naming test was administered. In this test, participants were presented with the same pictures that were presented during the learning trial and asked to twice orally produce the words corresponding to the pictures shown on the computer screen. If participants
did not remember a word, they were instructed to move to the next item. Their speech was recorded with a TASCAM DR-05 audio recorder and digitized into a wav format (44.1 kHz sampling rate with 16-bit quantization). One out of two productions per word (i.e., a speech sample without fillers or self-corrections during articulation) was selected and stored in an individual sound file, with peak intensity normalized using Praat. Prior to data collection, issues with clarity of visual stimuli, trial procedures, and testing procedures were resolved through a pilot study with 20 university students with a similar learning background. Data for pilot study participants were not included in the main data analysis.

2.3.5 Procedure

The experiment was conducted over two sessions on two different days. On Day 1, participants took a pretest, completed the treatment, an immediate posttest, and the Vocabulary Levels Test. For participants listening to words multiple times (E3 and E6), a 5-minute break was provided halfway through the treatment to reduce participant fatigue. On Day 2, approximately one week ($M = 6.05$ days) after the first session, participants completed a delayed posttest and filled out language background questionnaires. The test format (i.e., picture naming) across three time points was the same except that 10 high-frequency items were added to the pretest to boost motivation, but these items were not counted for any measures. Participants were told to learn the English words, but were not forewarned that their pronunciation would be assessed. The treatment and tests were conducted individually with the researcher or a research assistant. All speech samples were recorded in a sound-attenuated booth at the university. A total of 4,443 speech samples were elicited from 75 speakers on three test trials and evaluated for vocabulary and pronunciation measures.
2.3.6 Vocabulary and Pronunciation Measures

To assess vocabulary knowledge, spoken form recall (e.g., production of accurate forms of words in a picture-naming test) was measured. Form recall is considered the most difficult measure of form–meaning knowledge compared to three other measures: form recognition, meaning recognition, and meaning recall (Laufer & Goldstein, 2004). For pronunciation measures, following Derwing and Munro (2015) and Munro and Derwing (1995b), three constructs were measured: accentedness (i.e., listener rating of the extent to which learners’ word productions deviated from a native variety of the target language), comprehensibility (i.e., listener rating of the degree of effort needed to comprehend learners’ word productions), and processing time (i.e., listener’s response latencies derived from a timed dictation task). All vocabulary and pronunciation measures derived from a timed dictation task completed by raters.

Two native English-speaking teachers, both speakers of North American English (one female, one male), were recruited to participate in a series of rating sessions. They had extensive language teaching and speaking assessment experiences targeting learners with different L1 backgrounds in different countries (e.g., Korea, China, Canada). Their familiarity with Japanese-accented English was moderate (2 and 3 in response to 1 = not familiar at all, 6 = very familiar). They had no hearing problems. Raters completed a timed dictation task programmed using PsychoPy (Peirce, 2007). In this task, raters listened to each of the 4,443 speech samples and typed the spelling of the word they heard as fast as possible. Raters were presented with 44 blocks of 100 samples and a block of 43 samples that contained random selection of pretest, immediate posttest, and delayed posttest items. Listeners rated accentedness (1 = not accented at all, 5 = heavily accented) and comprehensibility (1
= easy to understand, 5 = hard to understand). Recordings were played only once. Form recall scores per rater were derived from transcription accuracy with minor misspellings considered accurate (e.g., chisle, camelieon, ladel). Coded dichotomously (1 = accurate, 0 = inaccurate), the form recall scores captured the accuracy with which participants recalled each word form, as judged by listeners. Processing time was defined as the time lapse between the onset of audio recording to the first key press on the computer keyboard. Before completing the rating task, raters completed a practice set of 15 samples representing varying pronunciation qualities (not included in the main dataset). Due to the large sample size and task demand, the two raters completed the listening task in multiple sessions (i.e., 14 to 16 1-hour sessions). All rating sessions were implemented individually in the researcher’s office.

2.3.7 Preliminary Analysis

Before addressing research questions, preliminary analysis was conducted on vocabulary and pronunciation measures. First, interrater agreement in the dictation task was checked (98% agreement), and the average score was calculated to yield a single score per participant for each target word. Second, moderate interrater reliability between the two raters were obtained for accentedness ($\alpha = .562$), comprehensibility ($\alpha = .663$), and processing time ($\alpha = .615$), and similarly average scores were calculated.¹ The latter three pronunciation scores were calculated based on 4,443 responses, comprising 264 cases from pretest E1, 272 cases from pretest E3, 274 cases from pretest E6, 627 cases from immediate posttest E1, 900 cases from immediate posttest E3, 961 cases from immediate posttest E6, 366 cases from delayed posttest E1, 381 from delayed posttest E3, and 398 cases from delayed posttest E6. All measures except processing time were considered normally distributed through a
Shapiro-Wilk test of normality, visual inspection of histograms, and examination of skewness statistics. Processing time values were log-normalized before statistical analysis was computed.

Pearson correlation analysis was conducted between the four measures (see Table 2). A correlation of .760 between comprehensibility and accentedness was comparable to earlier findings (e.g., $r = .74$ to .80 in Crowther et al., 2018; $r = .624$ in Munro & Derwing, 1995b), confirming that these two global constructs were related but partially independent of each other. A paired-samples $t$ test showed that comprehensibility ratings were significantly lower than accentedness ratings, $t(4442) = 131.42, p < .001$, the finding compatible with earlier research suggesting that comprehensibility was judged more favorably than accentedness (Munro & Derwing, 1995a). The strong correlation between processing time and comprehensibility rating in particular also supports Munro and Derwing’s (1995b) view that processing cost relates to native listeners’ judgements of comprehensibility, with less comprehensible speech requiring longer processing times. Thus, the obtained relationships between the four target pronunciation measures were similar to those reported in earlier L2 speech research, supporting the validity of each measured construct.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Accentedness $r$</th>
<th>Comprehensibility $r$</th>
<th>Processing time $r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accentedness</td>
<td>.760</td>
<td>.538</td>
<td>.538</td>
</tr>
<tr>
<td>Comprehensibility</td>
<td>.776</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.4 Results

2.4.1 Spoken Form Recall

The descriptive statistics for spoken form recall and three pronunciation measures (accentedness rating, comprehensibility rating, and processing time) are presented in Tables 3 and 4. In response to the first research question regarding the effect of repetition on spoken form recall, a mixed-design analysis of variance (ANOVA) was conducted using the JASP version 0.11 with exposure as a between-groups independent variable (E1, E3, and E6) and time as a within-groups independent variable (pretest, immediate posttest, and delayed posttest). Before conducting the analysis, statistical assumptions—Levene’s test for homogeneity of between-groups variances and Mauchly’s test for sphericity of within-groups variances—were confirmed.

The analysis showed significant effects for exposure, $F(2, 72) = 30.40, p < .001, \eta_p^2 = 0.458$, and for time, $F(2, 148) = 576.22, p < .001, \eta_p^2 = 0.886$, as well as a significant Time $\times$ Exposure interaction, $F(4, 144) = 57.31, p < .001, \eta_p^2 = 0.614$ (see Figure 1). Post hoc comparisons revealed no significant differences among the three groups at the pretest, $F(2, 72) = 0.47, p = .629, \eta_p^2 = 0.013$, where participants recalled on average 10 words, or the delayed posttest, $F(2, 72) = 1.12, p = .333, \eta_p^2 = 0.030$, where the mean recall was approximately 15 words. A significant difference was found among the three groups at immediate posttest, $F(2, 72) = 85.38, p < .001, \eta_p^2 = 0.703$. Post hoc pairwise test with Bonferroni correction revealed that the E6 group outperformed the E1 (mean difference = 13.28, $p < .001, d = 3.735$) and E3 groups (mean difference = 3.06, $p = .016, d = 0.845$), and the E3 group outperformed the E1 group (mean difference = 10.22, $p < .001, d = 2.498$).
Table 3. Means, standard deviations, and 95% confidence intervals of the pretest, immediate posttest, and delayed posttest scores on spoken form recall

<table>
<thead>
<tr>
<th>Participant subgroups</th>
<th>Spoken form recall</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pretest</td>
<td>Immediate posttest</td>
<td>Delayed posttest</td>
</tr>
<tr>
<td>E1 (n = 25)</td>
<td>10.16 (2.31)</td>
<td>23.30 (4.03)</td>
<td>14.08 (2.96)</td>
</tr>
<tr>
<td></td>
<td>[9.21, 11.11]</td>
<td>[21.64, 24.97]</td>
<td>[12.86, 15.30]</td>
</tr>
<tr>
<td>E3 (n = 25)</td>
<td>10.56 (2.13)</td>
<td>33.52 (4.15)</td>
<td>14.62 (3.20)</td>
</tr>
<tr>
<td></td>
<td>[9.68, 11.44]</td>
<td>[31.81, 35.23]</td>
<td>[13.30, 15.94]</td>
</tr>
<tr>
<td>E6 (n = 25)</td>
<td>10.74 (2.08)</td>
<td>36.58 (3.00)</td>
<td>15.50 (3.55)</td>
</tr>
<tr>
<td></td>
<td>[9.88, 11.60]</td>
<td>[35.34, 37.82]</td>
<td>[14.04, 16.96]</td>
</tr>
</tbody>
</table>

Note. Maximum score was 40. Standard deviations are in parentheses and upper and lower confidence intervals are in square brackets.
Table 4. Means, standard deviations, and 95% confidence intervals for pretest, immediate posttest, and delayed posttest pronunciation scores by group

<table>
<thead>
<tr>
<th>Item subsets</th>
<th>Accentedness</th>
<th>Comprehensibility</th>
<th>Processing time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pretest</td>
<td>Immediate posttest</td>
<td>Delayed posttest</td>
</tr>
<tr>
<td>E1</td>
<td>3.55 (0.47)</td>
<td>3.59 (0.35)</td>
<td>3.42 (0.61)</td>
</tr>
<tr>
<td></td>
<td>[3.34, 3.75]</td>
<td>[3.48, 3.70]</td>
<td>[3.21, 3.64]</td>
</tr>
<tr>
<td>E3</td>
<td>3.52 (0.42)</td>
<td>3.46 (0.25)</td>
<td>3.56 (0.60)</td>
</tr>
<tr>
<td></td>
<td>[3.32, 3.71]</td>
<td>[3.38, 3.54]</td>
<td>[3.35, 3.76]</td>
</tr>
<tr>
<td>E6</td>
<td>3.54 (0.38)</td>
<td>3.37 (0.26)</td>
<td>3.41 (0.50)</td>
</tr>
<tr>
<td></td>
<td>[3.37, 3.72]</td>
<td>[3.28, 3.45]</td>
<td>[3.25, 3.57]</td>
</tr>
</tbody>
</table>

*Note. Accentedness (1 = not accented at all to 5 = heavily accented); comprehensibility (1 = easy to understand to 5 = hard to understand). Standard deviations are in parentheses and upper and lower confidence intervals are in square brackets.*
Figure 1. Exposure group means on spoken form recall over time.

2.4.2 Accentedness, Comprehensibility, and Processing Time

In response to the second research question regarding the effect of repetition on measures of L2 word pronunciation, we conducted a crossed-random effects multilevel analysis (Locker et al., 2007) using one participant-level variable (exposure group) as the predictor and three item-level variables (cognateness, phoneme count, phonotactic probability) as covariates. The two variables (exposure, cognateness) were dummy coded. The two variables (phoneme count, phonotactic probability) were grand-mean centered. There was no multicollinearity among all item variables. Pretest scores were not included in the model given that there were no significant differences between groups for accentedness, $F(2, 72) = 0.67, p = .513$; comprehensibility, $F(2, 72) = 1.49, p = .231$; or processing time, $F(2, 72) = 1.40, p = .253$. A separate analysis was conducted for each measure with immediate and delayed posttest results. All statistical analyses were computed using IBM SPSS version 20.

The results of immediate posttests showed that E6 had significantly lower scores for accentedness (implying less accented production) than E1 ($\beta = -0.18$, SE =
0.07, \( t = -2.52, p = .014 \), which indicated that participants listening six times sounded more nativelike than those listening once. However, the differences between E6 and E3 (\( \beta = -0.09, \text{SE} = 0.07, t = -1.32, p = .191 \)) and E3 and E1 (\( \beta = -0.09, \text{SE} = 0.07, t = -1.23, p = .222 \)) did not reach statistical significance. For comprehensibility and processing time, E6 had significantly lower scores (implying more comprehensible production that was processed faster by raters) than E1 (comprehensibility: \( \beta = -0.34, \text{SE} = 0.09, t = -3.95, p < .001 \); processing time: \( \beta = -0.06, \text{SE} = 0.01, t = -4.34, p < .001 \)), and E3 had significantly lower scores than E1 (comprehensibility: \( \beta = -0.19, \text{SE} = 0.09, t = -2.26, p = .026 \); processing time: \( \beta = -0.03, \text{SE} = 0.01, t = -2.48, p = .015 \)). The differences between E6 and E3 were approaching significance (comprehensibility: \( \beta = -0.14, \text{SE} = 0.08, t = -1.73, p = .087 \); processing time: \( \beta = -0.03, \text{SE} = 0.01, t = -1.92, p = .059 \)). These findings indicate a linear relationship between frequency of exposures and comprehensibility and processing time gains; the more exposures the participants received, the more comprehensible their pronunciation of the words became, and the less effort was necessary for listeners to recognize the words in the dictation task. Unlike the results of the immediate posttest, no statistical differences were found for all three pronunciation measures between the three groups at the delayed posttest.

For the main effects of item-level covariates, cognateness and phoneme count significantly predicted pronunciation scores. Cognateness significantly predicted all pronunciation measures at all test times except accentedness at the delayed posttest (\( \beta = -0.14, \text{SE} = 0.11, t = -1.21, p = .234 \)), indicating that cognates were judged to be more nativelike, comprehensible, and faster to recognize than noncognates. Phoneme count significantly predicted accentedness scores at the delayed posttest, indicating that words with more phonemes were perceived to be more heavily accented (\( \beta = \).
0.12, SE = 0.04, t = 3.23, p = .003). No significant main effects of phonotactic probability were found for any pronunciation measures.

2.4.3 Cognateness, Word Length, and Phonotactic Probability

In response to the third research question targeting the extent to which word properties moderate the relationship between repetition and L2 word pronunciation, three interaction terms between the three item-level variables and exposure were created and added to the crossed-random effects model previously built in answer to the second research question.

2.4.3.1 Cognateness.

There were significant interactions between cognates and exposure group for all pronunciation measures at the immediate posttest. For accentedness, two significant interactions were detected: first between E6 and E1 (β = 0.27, SE = 0.07, t = 4.17, p < .001), and second between E3 and E1 (β = 0.18, SE = 0.07, t = 2.70, p = .007), indicating that accentedness was reduced as a function of repeated encounters particularly when participants learned noncognates in comparison to cognates (Figure 2). For comprehensibility, three interactions were found: first between E6 and E1 (β = 0.51, SE = 0.09, t = 5.60, p < .001), second between E3 and E1 (β = 0.24, SE = 0.09, t = 2.61, p = .009), and third between E6 and E3 (β = 0.27, SE = 0.08, t = 3.27, p = .001), indicating that comprehensibility improved as a function of repeated exposures when participants learned noncognates in comparison to cognates (Figure 3). For processing time, three interactions were found between exposure group and cognates: first between E6 and E1 (β = 0.09, SE = 0.02, t = 5.33, p < .001), second between E3 and E1 (β = 0.05, SE = 0.02, t = 2.90, p = .004), and third between E6 and
E3 ($\beta = 0.04$, $SE = 0.02$, $t = 2.64$, $p = .008$), indicating that processing time became shorter with increased exposures when participants learned noncognates in comparison to cognates (Figure 4).

Figure 2. Mean accentedness ratings by group and cognate status (immediate posttest).

Figure 3. Mean comprehensibility ratings by group and cognate status (immediate posttest).
On the delayed posttest, a significant interaction between cognates and exposure was found for comprehensibility between E3 and E1 ($\beta = -0.41$, $SE = 0.13$, $t = -3.14$, $p = .002$), indicating that comprehensibility was reduced with increased exposures (E1→E3) around a week after participants learned noncognates in comparison to cognates (Figure 5). For processing time, a significant interaction was found between E6 and E3 ($\beta = 0.06$, $SE = 0.02$, $t = 2.61$, $p = .009$), indicating that processing time became shorter from three exposures to six exposures around a week after participants learned noncognates in comparison to cognates (Figure 6).
2.4.3.2 Phoneme count.

A significant interaction was found for processing time at the immediate posttest between E6 and E1 ($\beta = -0.01$, $SE = 0.01$, $t = -2.06$, $p = .039$), revealing a
tendency for longer words to be processed more rapidly by raters while listening to those who had learned L2 words from six exposures in comparison to only one (Figure 7).

Figure 7. Relationship between the number of phonemes and processing time per group (immediate posttest).

2.4.3.3 Phonotactic probability.

Significant interactions were found for accentedness and comprehensibility on the immediate posttests. For accentedness, two significant interactions were detected: first between E6 and E1 ($\beta = -9.56, SE = 2.70, t = -3.54, p < .001$), and second between E6 and E3 ($\beta = -768, SE = 2.49, t = -3.09, p = .002$) (Figure 8). For comprehensibility, two interactions were also found: first between E6 and E1 ($\beta = -10.25, SE = 3.75, t = -2.73, p = .006$) and second between E6 and E3 ($\beta = -8.57, SE = 3.46, t = -2.48, p = .013$) (Figure 9). These results indicate that pronunciation of words that contain more frequently occurring sequences of segments in a given
position became more comprehensible and nativelike after six exposures. However, the direction of the effect appeared to be opposite in groups receiving one or three exposures, indicating that phonotactically more regular words tended to become less comprehensible and more heavily accented after being heard once or three times.

Figure 8. Relationship between phonotactic probability of target words and accentedness ratings per group (immediate posttest).
2.5 Discussion

2.5.1 Repetition Effects on Spoken Form Recall and Pronunciation

In response to the first research question regarding repetition and spoken form recall, the results showed that learners receiving six exposures recalled a larger number of spoken word forms than learners receiving one or three exposures, and learners receiving three exposures outperformed those receiving one exposure. These findings support earlier studies highlighting the important role of repetition in developing learners’ knowledge of form–meaning connection (Nakata, 2017; Uchihara et al., 2019; Webb, 2007). This result adds to the growing body of evidence from research targeting word learning in a written modality, by showing that positive effects of repetition can be extended to improving form recall in an aural modality. However, the absence of significant repetition effects on the delayed posttest suggests that the effect was likely not long lasting. This might be expected given that form recall is the
most difficult measure of form–meaning connection (Laufer & Goldstein, 2004).

In answer to the second research question regarding repetition and pronunciation of previously unknown L2 words, the results showed that learners receiving six exposures produced words that were more comprehensible, more nativelike, and less effortful to process than learners receiving one exposure. Similarly, learners receiving three exposures outperformed those receiving one exposure in comprehensibility and processing time. These findings indicate that repetition enhanced the quality of spoken forms for unfamiliar words, while also likely contributing to the development of form–meaning connections. Put differently, learners’ production of spoken word forms became more comprehensible and less accented after learners had encountered these spoken forms multiple times while attempting to remember word meanings. Learners might benefit from multiple auditory encounters with novel word forms because repetition might help learners refine the phonetic detail they perceive and subsequently store for these words in their lexicons. A more refined lexical representation may then guide learners’ production, resulting in listeners recognizing the intended word faster (Ludwig & Mora, 2017; Munro & Derwing, 1995b). A more refined lexical representation not only would correspond to more nativelike (less accented) production but also might decrease listeners’ processing cost and increase word comprehensibility. A repetition-based processing benefit is also consistent with prior work on auditory word priming, which suggests that repeated encounters with spoken words facilitate the processing of the same words at a subsequent encounter (Trofimovich, 2005).

However, accentedness appeared to be less impacted by repetition effects, compared to comprehensibility and processing time, as evidenced by the absence of a significant difference between learners with one and three exposures. These findings
indicate a different learning trajectory of accentedness and comprehensibility, with the
development of the former being slow and gradual to the extent that at least six
encounters were necessary compared to that of the latter improving relatively quickly
after three encounters (Derwing & Munro, 2013; Saito, 2015). This differential
learning curve might be attributed to the different rater behaviors during L2
comprehensibility and accentedness judgements. For L2 comprehensibility, raters
attend to as much phonological information available in accented L2 speech as
possible, including segmentals (e.g., vowels), prosody (e.g., word stress), and fluency
(e.g., articulation speed), such that they can arrive at speakers’ intended message
(words) quickly and efficiently (Crowther et al., 2018; Saito et al., 2016). For L2
accentedness, raters rely on a restricted range of phonological information especially
at a segmental level, the improvement of which involves much learning experience
(Munro, 1993) and is less likely to develop with a few encounters.

The current findings for spoken word forms contrast with the findings
reported for written word forms reported by earlier studies (Chen & Truscott, 2010;
Webb, 2007). Table 5 suggests that effect sizes for repeated encounters in this study ($d$
= 0.498 to 0.504) were smaller than Webb’s (2007) and Chen and Truscott’s (2010)
studies measuring productive knowledge of orthography ($d$ = 0.863 and 0.990). On the
one hand, this result is surprising, given that the learners in this study experienced new
words under relatively favorable conditions (i.e., they heard each word spoken clearly,
in isolation, while seeing the image depicting the word’s meaning), whereas in Webb
(2007) and Chen and Truscott (2010), learners had to infer the meaning of a word
embedded in a sentence or passage. On the other hand, smaller repetition-driven
learning effects in the aural modality would be expected, considering the transient,
memory-dependent nature of speech, compared to the relative permanence of
orthographic representations embedded in meaningful contexts. To clarify potential modality differences in future work, it would be important to compare repetition effects across aural and written modalities directly, using similar learning and assessment procedures.

Table 5. Comparing effect sizes (Cohen’s $d$) for the immediate posttest between current study and earlier studies measuring productive knowledge of orthography

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>PO</td>
<td>PO</td>
<td>Accentedness</td>
</tr>
<tr>
<td>E1–E3</td>
<td>0.523</td>
<td>1.130</td>
<td>E1–E3</td>
</tr>
<tr>
<td>E3–E7</td>
<td>0.738</td>
<td>0.434</td>
<td>E3–E6</td>
</tr>
<tr>
<td>E1–E7</td>
<td>1.328</td>
<td>1.405</td>
<td>E1–E6</td>
</tr>
<tr>
<td>Mean</td>
<td>0.863</td>
<td>0.990</td>
<td>Mean</td>
</tr>
</tbody>
</table>

Note. PO = productive knowledge of orthography; PT = processing time.

2.5.2 Word-Related Factors and Repetition Effects on Pronunciation Learning

In response to the third research question, the results revealed that the effect of repetition was significantly moderated by different word characteristics (cognateness, word length, phonotactic probability). In this study, cognates were perceived to be more comprehensible, less accented, and easier to process than noncognates. However, cognateness interacted with exposure frequency at the immediate posttest, such that only noncognates sounded more comprehensible, less accented, and easier to process with increased exposure (E1→E3→E6). In fact, Table 6 shows that none of the pronunciation measures for cognates changed meaningfully with increased exposure, consistent with the idea that cognates are generally easier to learn than noncognates (Lotto & de Groot, 1998; Rogers et al., 2015). The L1/L2 form
and meaning overlap for cognates ostensibly provides learners with sufficient learning
benefit immediately after exposure, enabling them to produce cognates in a nativelike
and comprehensible way and allowing listeners to process them rapidly, regardless of
the number of times cognates occur in the input. In contrast, for noncognates, the
learning benefits of repeated exposure appear uneven and non-linear, particularly in
the long run. As shown by interactions between cognateness and exposure frequency
at the delayed posttest, noncognates became less comprehensible when exposure
increased from one to three repetitions (E1→E3) yet listeners processed them faster
when exposure increased further to six repetitions (E3→E6). Thus, although increased
replication might benefit noncognates in the short run, long-term benefits of repetition
for noncognates might take time to accrue and might extend to only some measures of
L2 pronunciation (e.g., processing time) rather than others (e.g., accentedness).

Table 6. Pronunciation scores on cognates across three exposure groups at immediate
posttest

<table>
<thead>
<tr>
<th>Measure</th>
<th>E1</th>
<th>E3</th>
<th>E6</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>SD</td>
<td>SD</td>
<td>SD</td>
<td>SD</td>
</tr>
<tr>
<td>Accentedness</td>
<td>3.35</td>
<td>3.34</td>
<td>3.30</td>
</tr>
<tr>
<td>0.29</td>
<td>0.24</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>Comprehensibility</td>
<td>2.08</td>
<td>1.99</td>
<td>2.00</td>
</tr>
<tr>
<td>0.37</td>
<td>0.29</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>Processing time</td>
<td>0.017</td>
<td>0.001</td>
<td>0.004</td>
</tr>
<tr>
<td>0.072</td>
<td>0.053</td>
<td>0.054</td>
<td></td>
</tr>
</tbody>
</table>

Note. Accentedness rating score ranges from 1 = not accented at all to 5 = heavily
accented. Comprehensibility rating score ranges from 1 = easy to understand to 5 =
hard to understand.

At the delayed posttest, longer words were perceived to be more heavily
accented than shorter words, implying that longer words present learners with more
opportunity to produce segment- or syllable-level content that might be perceived by
listeners as less nativelike (more accented) but not necessarily harder to understand (less comprehensible). However, phoneme count interacted with exposure frequency at the immediate posttest, such that longer words elicited shorter response times from listeners as frequency of exposure increased between one and six repetitions. At first glance, this finding seems to run counter to the general view that longer words are harder to learn (Ellis & Beaton, 1993; Peters, 2020). In this study, however, this might not be altogether surprising, given that longer words are more informative than shorter words in that they provide more cues for listeners to use for word recognition. If response time offers a satisfactory metric of processing effort for the listener, then longer words might be challenging for learners to produce comprehensibly, particularly when a word’s exposure frequency is low. However, with repeated exposure, learners can produce longer words to the extent that listeners recognize them with less effort.

Words with high phonotactic probability comprise frequently occurring sequential patterns in English, and such phonotactically regular words are easier to learn (Ellis & Beaton, 1993). At the immediate posttest, up to three exposures were insufficient to enable learners to produce phonotactically regular words as less accented and more comprehensible, compared to irregular words. It was not until after six exposures that phonotactic probability had a positive effect on L2 pronunciation measures. This might seem counterintuitive, because spoken forms of phonotactically more regular words would be expected to improve at a faster rate than those of less regular words. One possible reason for this finding might relate to the difference in the way that learning gains have been tested. In Ellis and Beaton’s study, learning was measured in the written form (via L1–L2 translation), whereas in this study, it was assessed in speaking (through picture naming). It may be that the effects of
phonotactic regularity are less pronounced or perhaps less salient for learners in the aural modality, compared to written input, where regular patterns are likely more perceptible. However, a more likely explanation is that the learners who experienced words six times had more opportunity to (subvocally) rehearse spoken words (Dahlen & Caldwell–Harris, 2013) such that the facilitative effect of repetition might have started to come into play. In any case, the impact of phonotactic probability on L2 word pronunciation is likely cumulative so that it might be detected only after a certain minimum threshold of repeated exposure has been reached (i.e., six repetitions).

2.6 Implications and Future Directions

A primary implication from this study comes from the finding that learners could improve pronunciation through repeated exposure to spoken words, implying that pronunciation of individual L2 words can be enhanced without explicit attention drawn to their specific pronunciation features. In view of growing concerns expressed among researchers and teachers regarding lack of teacher training and time constraints on L2 curricula (Foote et al., 2016), recent research has attempted to explore ways to improve pronunciation without taking up valuable class time (e.g., homework-based approach, Martin, 2020). The current research is compatible with this increasing trend, suggesting that providing repeated spoken input could supplement explicit instruction with a minimal burden on the teacher. Although pronunciation gains were not consistently retained in this study, the value of repeated input should not be underestimated since provision of repeated exposure over an extended period of time might result in sizable gains in speech comprehensibility (Trofimovich et al., 2009).

Increased exposure to spoken word forms would likely benefit learners’
pronunciation of some words over others, with the amount of benefit depending on specific word characteristics. Because longer and more phonotactically regular words were amenable to repeated exposure (≥ 6 encounters), comprehension-based activities might lend themselves best to improving pronunciation of longer words particularly containing phonotactically common patterns. In contrast, shorter words should be taught explicitly by drawing learners’ attention to their specific phonetic features, for instance, through providing metalinguistic information and corrective feedback (Saito, 2011). Even if some segments or syllables in a longer word are mispronounced, listeners may still be able to reconstruct the whole word from the remaining intact fragments. However, if a shorter word is mispronounced, it might be difficult for listeners to infer the intended word from limited information available in the word. Furthermore, although cognates generally enjoy a learning advantage over noncognates (Lotto & de Groot, 1998; Nation, 2013; Peters & Webb, 2018; Vidal, 2011), cognates may need to be taught explicitly as their pronunciation is less likely to be improved with repeated exposure. When cognates are introduced, special attention may need to be paid to their spoken forms, with the view to improving learners’ pronunciation of such deceptively easy words, for example, through raising awareness of the difference between the spoken forms of Japanese katakana words and their English cognates.

Second, it is important to ensure that spoken forms of words should be presented when teachers introduce new vocabulary. Particularly when focusing on high-frequency vocabulary, spoken word forms should be introduced using audio materials or teacher’s own pronunciation. Given that 2,000 word families provide around 90 percent of lexical coverage in spoken discourse (Milton, 2009), acquiring comprehensible pronunciation of these words is crucial for learners to become
functional L2 speakers. One notable issue in L2 pedagogy concerns a lack of crossover between the domains of vocabulary and pronunciation research. Vocabulary researchers are primarily interested in learning defined as the development of form–meaning connections, while pronunciation researchers mainly examine learning operationalized through various measures of spoken word form. However, given limited instruction time (Martin, 2020) and a strong call for an interdisciplinary approach to teaching pronunciation (Field, 2005), we suggest that vocabulary teachers should be responsible for improving learners’ knowledge of pronunciation, besides focusing on how quickly and how well learners establish the link between a word’s form and meaning.

Lastly, findings of this study suggest that vocabulary researchers should test word knowledge in different ways and that they cannot generalize findings for written forms to spoken forms. The extent to which repetition affects acquisition might depend on whether learning is tested in the written or spoken modality. Moreover, learners’ word knowledge should be tested in a way that reflects their ability to use words in real language communication (Kremmel & Schmitt, 2016). In this sense, measuring L2 pronunciation through global, listener-based constructs (comprehensibility, accentedness) offers a useful way to capture learners’ ability to use words in spontaneous oral communication.

Several elements of the current design might be modified in future research to provide further insight into how repetition impacts L2 pronunciation. First, this study was conducted in a laboratory setting, and the facilitative effects of repeated exposure found in this particular context should not be generalized to language classrooms. Given that vocabulary can be learned through listening to teacher talk (Jin & Webb, 2020), it is worth exploring whether teacher talk can be a useful source of input for
improved pronunciation. Second, the target words were presented in isolation, so it remains to be seen whether repeated exposure to contextually embedded spoken words (e.g., Webb, 2007) will help improve pronunciation. One promising area worth investigating is television viewing as an ecologically valid source of spoken input (Peters, 2019; Peters & Webb, 2018). Wisniewska and Mora (2020) have recently reported that watching television increased pronunciation accuracy. It would be intriguing to explore to what extent repeated encounters with spoken words with/without orthographic support (captions) help improve pronunciation. Finally, different aspects of pronunciation should be measured (Saito & Plonsky, 2019). Pronunciation involves not only global constructs but also specific features such as segments and prosody. Using a diverse toolkit of pronunciation measures would provide further insight into our understanding of the role of repetition in L2 pronunciation development.

2.7 Notes

1. Interrater reliability for accentedness, comprehensibility, and processing time was not satisfactory ($\alpha < .70$). This was probably due to the limited number of raters ($N = 2$). Given the high task demand on raters (i.e., listening to 4,443 samples and typing words as fast as possible), recruitment of multiple raters was not feasible in this study. However, with evidence of correlation analysis supporting the concurrent validity of the three pronunciation measures, we used these measures as dependent variables for subsequent analysis. Future research can take an alternative approach, for example, by distributing a subset of raters to different blocks of speech samples (see Martin, 2020 for the rating procedure) instead of having the same raters evaluate all samples.
2. For accentedness, comprehensibility, and processing time, descriptive statistics were calculated for each item (i.e., 40 target items in the column of the spreadsheet) due to considerable variation in the number of items produced across groups and in order to minimize the influence of item-related factors on descriptive findings.

2.8 References


Chapter 3: To what extent do acoustic variability and frequency of exposure affect learning the spoken forms and form-meaning connections of second language words?

3.1 Introduction

Acoustic variability or input variation in acoustic cues can be characterized by differences in linguistic and non-linguistic properties between and within talkers (e.g., voices, pitch height, speaking rate, speaking style, and loudness). Research has shown that talker variability facilitates different aspects of second language (L2) acquisition including form-meaning mapping of new words (Barcroft & Sommers, 2014a), recognition/perception accuracy of temporal and spectral features (Logan et al., 1991), and their production accuracy (Bradlow et al., 1997). The benefits of talker variability can be attributed to two factors, repeated exposure and talker-specific characteristics (i.e., indexical information) available in different voices. The positive effects of repetition have long been documented in L2 vocabulary research with evidence suggesting that words encountered a greater number of times are more likely to be learned and retained (Uchihara et al., 2019; Webb, 2007, 2014). Indexical information available in multiple voices helps form more “associative hooks” with which learners can retrieve and recall spoken forms of words efficiently and fluently (Barcroft & Sommers, 2005, p. 410).

Despite the potential benefits of talker variability for different aspects of word knowledge, little is known about how it affects learning the pronunciation of new words (for segmental features, see Brosseau-Lapré et al., 2013; Kartushina & Martin, 2019). Previous studies of L2 word learning and acoustic variability have also tended to rely on measures of form-meaning connection without measuring how
accurately or fluently spoken forms are produced. This is surprising given that pronunciation is considered one of the fundamental aspects of word knowledge (Nation, 2013). Even if learners are able to produce the spoken forms of L2 words, it is important to further ensure that the produced forms are sufficiently accurate and ultimately intelligible to the listener so that L2 speakers are successful in oral communication. Also, the extent to which repeated exposure and input variability contribute to L2 word learning (in terms of the knowledge of meaning and form) remains underexplored. This is mainly because the effects of acoustic variability tend to be examined at one frequency, and no studies have attempted to tease apart the contributions of the two factors (input frequency and variability) to L2 word learning. In response to these research gaps, the current study aimed to examine the effects of talker variability and frequency of exposure on knowledge of form-meaning connection (spoken form recall) and spoken form accuracy (pronunciation) targeting Japanese learners studying English as a foreign language (EFL).

3.2 Background

3.2.1 Word Knowledge, Processing, and Learning

Among numerous aspects of word knowledge, form-meaning connection is regarded as perhaps the most important aspect and has been extensively researched in L2 vocabulary studies (Nakata, 2017; Rott, 1999; Waring & Takaki, 2003). Form-meaning connection is conceptualized as having different degrees of strength, and is measured using different test formats such as meaning recognition, form recognition, meaning recall, and form recall. According to Nation’s (2013) framework of word knowledge, knowledge of spoken form is conceptualized at both receptive and productive levels, suggesting that lexically proficient learners should know “what the
word sounds like” (i.e., speech perception or recognition) and “how the word is pronounced” (i.e., speech production or pronunciation). Recent studies have measured recognition of spoken forms using a checklist test (Feng & Webb, 2019) and a multiple-choice test (Pavia et al., 2019) to examine gains in word knowledge through vocabulary learning activities (e.g., viewing television programs, listening to academic lectures, listening to songs). However, no research has explored how productive knowledge of the spoken forms of new words (i.e., pronunciation) develops as a result of hearing spoken input or completing vocabulary learning activities.

A major reason for the lack of studies investigating pronunciation in the field of vocabulary research may be due to the difference between learning the pronunciation of new/unfamiliar versus old/familiar words. According to the type of processing – resource allocation (TOPRA) model (Barcroft, 2002, 2015; Kida & Barcroft, 2018), word learning is constrained by limited cognitive capacity, and processing of form, processing of meaning, and processing of form-meaning mappings can operate largely independently. It is therefore hypothesized that learners whose attention is directed to a given component (e.g., meaning) may not devote sufficient attentional resources to processing another component (e.g., form). Therefore, when tested on word meanings, test takers with better performance may show impaired performance for word forms. The TOPRA model helps explain why there is little crossover between vocabulary and pronunciation fields in L2 research. Pronunciation researchers typically use familiar or high-frequency words as target items with learners allocating their attentional resources exclusively to processing word forms (i.e., pronunciation). In this line of research, the focus on development of spoken form leaves little room for gaining knowledge of form-meaning connection
(cf. Flege et al., 1998, for the roles of lexical frequency and familiarity in L2 segmental accuracy). The primary purpose of vocabulary research, on the other hand, has been to explore ways to optimize the mapping of forms to meanings by using unfamiliar or low-frequency words. Consequently, limited attention has been paid to the processing and enhancing of knowledge of spoken forms.

Another reason might relate to the limited scope of L2 lexical development considered and addressed by earlier research. According to the psycholinguistic model of L2 lexical representation and development (Jiang, 2000), learners initially go through a process of mapping L2 forms to first language (L1) meanings (i.e., L1 translations) of new words. At this stage of learning, form-meaning connections of words are considered relatively weak, causing production of L2 words to be laborious and effortful. With continued exposure to unfamiliar words, form-meaning connections become stronger and more robust, allowing for more efficient and fluent retrieval of L2 words. However, learners are still unlikely to have developed target-like, accurate meanings and spoken forms that are accurate in the phonetic details. Finally, a greater quantity and quality of input further enhances not only the strength of form-meaning connections but also semantic (e.g., L1 translations→L2 specific/accurate meaning) and formal (e.g., L1 influenced phonology→L2 specific/accurate phonology) aspects of word knowledge. The phenomenon of phonological re-finetuning during rapid vocabulary growth has been observed in the context of L1 and L2 acquisition processes (Bundgaard-Nielsen et al., 2011; Werker, 2018).

Critically, the scope of lexical development is not restricted to the mapping process but also formal (and semantic) enhancement with continued L2 exposure. As shown in the L1 and L2 acquisition literature, while newly learned words might be
initially colored by L1 phonology, learners who are induced to notice, attend to, and fill in the phonetic details learn to attain target-like production (Saito, 2013). Therefore, the current study defined the acquisition of spoken word knowledge as a two-step process—(a) establishment of an initial form-meaning connection (L2 meaning and L1 phonology) and (b) phonological refinement (L2 meaning and L2 phonology). In the current investigation, we aimed to explore the effects of input quantity (i.e., frequency of exposure) and quality (i.e., acoustic variability) on these two knowledge aspects.

3.2.2 Acoustic Variability and L2 Pronunciation Learning

Acoustic variability is crucial for the recognition and production of individual L2 sounds. Learners need ample exposure to a wide range of exemplars in different phonetic, talker, and task contexts so they can gradually get attuned to the acoustic information that distinguishes new phonetic categories (e.g., third formant of 1,500-2,000 Hz as a threshold for English /r/ vs. /l/). One especially promising intervention maximizing learners’ access to variability involves high variability phonetic training (HVPT), a procedure which exposes learners to the target L2 sounds produced by different talkers and spoken in varied phonetic contexts (Thomson, 2018). Since seminal work by Logan et al. (1991), numerous studies have observed that learners completing HVPT show significant improvement (10–20% gain) in identifying various segmental targets spanning vowels (Lambacher et al., 2005), liquids (Lively et al., 1993), stops (Flege, 1995b), and Japanese geminate consonants (Hirata, 2004). Although most earlier studies focused on the benefits of HVPT for perceptual (or recognition) performance, perception training with high-variability input also promotes production accuracy of L2 sounds (Bradlow et al., 1997; Lambacher et al.,...
2005). For example, Bradlow et al. (1997) found that Japanese learners’ increased accuracy in correctly recognizing phonemic contrasts (i.e., English /r/ vs. /l/) led to improvement of production accuracy of the same sounds. Flege’s (1995a) speech learning model provides a theoretical account for a close interface between receptive and productive ability. According to this model, difficulties in production arise from difficulties in recognition. Once an adequate perceptual specification of an L2 sound is established, production will become more accurate with continued exposure over time, a view that has been extensively tested through prior research (for a meta-analytic review, see Sakai & Moorman, 2018).

However, recent studies looking in greater depth into the effects of talker variability suggest mixed findings regarding the extent to which high-variability training brings about larger learning gains in production accuracy compared to low-variability training (Brosseau-Lapré et al., 2013; Kartushina & Martin, 2019; Wiener et al., 2020). Brosseau-Lapré et al. (2013) investigated whether English speakers with limited knowledge of French improve accuracy at producing the French unrounded and rounded mid-vowels. There were no beneficial effects of a multiple-talker condition (three talkers) over a single-talker condition after completion of two 1-hour perception training sessions over two days. In contrast, Kartushina and Martin (2019) found that Spanish speakers with no experience with French improved production accuracy of the French mid-open and mid-close front unrounded vowels to a greater degree when they listened to target sounds produced by five talkers than by a single talker. Wiener et al. (2020) confirmed the superiority of high-variability training (four talkers) over low-variability training (single talker) for L1 English beginners studying L1 Mandarin tones after they were given explicit instruction and perception training sessions over four consecutive days. In summary, listening to multiple talkers appears
more effective in improving perception accuracy than listening to a single talker, but
the degree to which a high-variability advantage extends to production accuracy
remains unclear.

3.2.3 Acoustic Variability and L2 Vocabulary Learning

Research has consistently shown a positive effect of acoustic variability on
L2 vocabulary learning using measures of form-meaning connection (Barcroft &
Sommers, 2014a). Barcroft and Sommers (2005) used two recall tests—meaning
recall (L2-to-L1 recall) and form recall (picture-to-L2 recall)—and compared three
variability conditions. In their within-participants study, L1 English speakers with no
prior formal instruction in Spanish completed a paired-associate word learning task in
which they studied Spanish words while hearing the spoken forms of target items and
viewing the pictures conveying their meanings. Participants learned 24 words, eight of
which were presented in one of three conditions: high variability (6 occurrences
produced by 6 different talkers), moderate variability (6 occurrences produced by 3
different talkers repeating each word twice), and low variability (6 occurrences of all
words produced by a single talker). The results of meaning and form recall tests
suggested that the words learned under high variability were recalled significantly
more accurately compared to those learned under moderate variability, and both sets
of words were recalled more accurately than those learned under low variability.
Barcroft and Sommers concluded that acoustic variability is beneficial in developing
knowledge of form-meaning connections of L2 words because it allows learners to
process, encode, and store indexical information relevant to the L1 perceptual system,
leading to a more distributed (robust) representation of the word form. A recent study
(Sinkeviciute et al., 2019) investigated whether learner’s age moderates the positive
effects of input variability on L2 vocabulary learning. In this study, English-speaking
learners of different ages with no experience with a target language (Lithuanian) heard
eight repetitions of six new words produced by a single talker (low-variability
condition) or eight talkers (high-variability condition), and post-training performance
was measured through meaning recognition (picture-to-word matching) and form
recall (picture-to-word recall) tests. The results were consistent with earlier studies
showing beneficial effects of high variability for adult learners on form recall (but not
on meaning recognition). However, no such benefit was observed for groups of
children (7- to 8-year-olds and 10- to 11-year-olds), either in meaning recognition or
form recall.

3.2.4 Frequency of Exposure and L2 Word Learning

Frequency of exposure is a key factor contributing to L2 vocabulary learning
(Webb, 2014; Webb & Nation, 2017). In decontextualized learning activities (e.g.,
paired-associate learning), Nakata (2017) found that five and seven retrievals of target
words produced significantly larger gains than one and three retrievals regardless of
different test timings. For contextualized leaning activities (e.g., learning through
reading graded readers), greater numbers of encounters with target words seem
necessary ranging from six (Rott, 1999), eight (Horst et al., 1998), 10 (Webb, 2007),
to more than 20 encounters (Waring & Takaki, 2003). A recent meta-analysis of 26
studies (Uchihara et al., 2019) showed a significant mean correlation of .34 between
frequency of exposure and contextualized vocabulary learning. However, the majority
of earlier studies focused on reading activities, with few studies investigating
vocabulary learning through spoken input. van Zeeland and Schmitt (2013) measured
learning gains in knowledge of spoken form recognition, grammar (part of speech),
and meaning recognition for words encountered three, seven, 11, and 15 times in oral passages. They found that gains in spoken form and grammar occurred between three and seven encounters, whereas at least 11 encounters were needed for learners to recognize the form-meaning connections of target items. At delayed posttest, only learning gains of words encountered 11 times remained significantly higher than the gains for words presented in other frequency ranges. Other listening studies which are similar to van Zeeland and Schmitt (2013) in research design and findings further suggest that frequency has a positive effect on word learning through listening to songs (Pavia et al., 2019), viewing full-length television programs (Peters & Webb, 2018), and listening to academic lectures (Vidal, 2011). However, to the best of our knowledge, no prior research has examined the effects of frequency on learners’ productive knowledge of spoken word forms.

3.2.5 The Current Study

There are several reasons why it is important to investigate the effects of acoustic variability and frequency of exposure on learning the spoken forms and form-meaning connections of unknown words. First, no studies have examined the effects of acoustic variability and frequency of exposure on productive knowledge of spoken form (i.e., pronunciation) of unknown words, with most previous research relying on receptive measures of form-meaning connection. This is an important gap in research that needs to be filled as learners’ ability to pronounce words accurately is essential for successful oral communication. Second, the relative contributions of frequency of exposure and acoustic variability to L2 lexical acquisition remain underexplored. For instance, prior research has not explored the minimum number of encounters necessary for the positive effects of acoustic variability to emerge in L2 word
learning, because variability effects have been examined at one frequency in each study (6 encounters in Barcroft & Sommers, 2005; 8 encounters in Sinkeviciute et al., 2019). Determining the minimum number of encounters needed for a variability benefit to arise should be useful for L2 instructors as it might help them introduce input variability effectively to optimize variability benefits for L2 word learning. Third, this research, conceptualized within the framework of lexical knowledge and development defined through multi-componential perspectives (Jiang, 2000; Webb & Nation, 2017), examines how the quantity (i.e., frequency) and quality (i.e., variability) of input promotes different stages of word learning (form-meaning connection and phonological refinement), thus promising to shed further insight into the role of input in L2 lexical acquisition. Last but not least, evidence of variability benefits in L2 pronunciation learning has predominantly come from work focusing on segmental (vowels and consonants) rather than suprasegmental aspects of language such as rhythm, intonation, word stress, and fluency (Thomson, 2018). This is surprising because the important role of suprasegmentals has been increasingly recognized in L2 pronunciation teaching (Zhang & Yuan, 2020), and a growing number of studies have suggested that L2 production (e.g., measured through comprehensibility and intelligibility) is associated with a range of suprasegmental features including word stress (J. Field, 2005), sentence stress (Hahn, 2004), and temporal fluency (Suzuki & Kormos, 2020). The present study therefore focused on word stress accuracy as an essential aspect of L2 word knowledge (J. Field, 2005), particularly in light of previous calls for vocabulary instruction to be responsible for helping learners improve word stress accuracy (J. Field, 2005; Murphy & Kandil, 2004). Therefore, this study was designed to examine to what extent each factor—frequency of exposure (3 vs. 6 encounters, without talker variability) or acoustic
variability (3 vs. 6 encounters, with talker variability)—enhances knowledge of pronunciation (word stress accuracy) and knowledge of form-meaning connection (spoken form recall), for L1 Japanese participants learning novel L2 words. This study was guided by the following research questions and predictions:

1. To what extent does talker variability (3 vs. 6 voices) affect learning two aspects of word knowledge: initial form-meaning connection (measured through spoken form recall) and subsequent phonological refinement (assessed through production of word stress)?

2. To what extent does frequency of exposure (3 vs. 6 encounters) affect two aspects of word knowledge: form-meaning connection (measured through spoken form recall) and phonological refinement (assessed through word stress accuracy)?

On the basis of vocabulary literature (Nakata, 2017; Rott, 1999; Webb, 2014; Webb & Nation, 2017), it was predicted that L2 learners would require six encounters to trigger the establishment of form and meaning mappings. At this initial learning stage, however, learners would most likely rely on L1 phonology (producing word stress inaccurately), especially if the treatment lacks any acoustic enhancement (1 or 3 voices). Thus, it might be necessary for learners to experience not only ample word encounters (six encounters), but also get exposed to various acoustic models (6 voices) so that they can attain target-like spoken word knowledge.
3.3 Method

3.3.1 Overview of the Study

The present study involved four experimental groups and three testing trials (pretest, immediate posttest, delayed posttest). Participants were randomly assigned to the four experimental groups and received different frequencies of exposure with or without acoustic variability to target words: three encounters with acoustic variability (E3+AV), six encounters with acoustic variability (E6+AV), three encounters without acoustic variability (E3), and six encounters without acoustic variability (E6). During the treatment, participants were instructed to learn 40 low-frequency English words through listening to the words and viewing their corresponding pictures. A picture-naming test was administered at the three testing times, and the elicited samples were evaluated for form recall and word stress measures.

3.3.2 Participants

Eighty Japanese university EFL students (age = 18–23) in Japan participated in this experiment. All participants had never lived in English-speaking countries longer than one month. All participants scored 90 percent or higher on the 1,000 word level of the Vocabulary Levels Test (Webb et al., 2017), and all except one participant scored 80 percent or higher on the 2,000 word level of the test. Their mean score at the 2,000 level was 28.31, indicating that they had receptive knowledge of almost all of the 2,000 most frequent words. The 80 participants were randomly assigned to four experimental groups (E3, E6, E3+AV, and E6+AV). There were no between-group differences in overall vocabulary test scores, $F(3, 76) = 1.31, p = .278$. All participants reported normal hearing.
3.3.3 Target Items

Forty target words were selected according to the following three criteria. First, a pool of low-frequency words was created by collecting English words that were beyond the most frequent 5,000 word families in Nation’s BNC/COCA word lists (Nation, 2012). Second, because the treatment involved learning spoken forms attached to meanings conveyed in visual images (pictures), only concrete nouns were selected as target items. Third, words that could be replaced with high-frequency synonyms were avoided to reduce the possibility that high-frequency synonyms of the target items would be produced in the picture-naming test (see Appendix A for target items).

Each of the 40 target words was recorded twice by six native speakers of English (3 females, 3 males) using a TASCAM DR-05 audio recorder and digitized into a wav format (44.1 kHz sampling rate with 16-bit quantization). The better of the two productions was selected in terms of clarity, naturalness, and lack of background noise and then stored as an individual sound file, with peak intensity normalized using digital speech-analysis software (Praat; Boersma & Weenink, 2014). Pilot testing showed that two native English speakers successfully identified all 240 productions recorded by the six speakers. Instead of presenting different voices randomly as in earlier studies (Barcroft & Sommers, 2005, 2014b), this study took an approach to optimizing the effectiveness of learning procedures by sequencing presentations of six speakers in the order of intelligibility (see Webb, 2008, for a similar approach in contextual informativeness and vocabulary learning). First, 10 out of 40 items produced by each of the six speakers were randomly selected (60 samples = 10 items × 6 speakers). An additional panel of native English listeners (n = 8) were recruited to listen to 60 speech samples embedded in cafeteria noise (signal-to-noise ratio = 8 dB).
and write down the words they heard in an answer sheet. A point was awarded for correctly spelled words with minor misspellings accepted (e.g., *chameleon* → *cameleon*). Although intelligibility scores were not significantly different across native listeners, $F(5, 35) = 0.57, p = .725$, average scores indicated a slight variation, and these scores were used to sequence the intelligibility of the speakers from higher to lower scores: Talkers 1 ($M = 0.80$), 2 ($M = 0.79$), 3 ($M = 0.78$), 4 ($M = 0.75$), 5 ($M = 0.74$), and 6 ($M = 0.71$) (see Table 1).

Table 1. Sequence of talker presentations for four experimental groups

<table>
<thead>
<tr>
<th>Group</th>
<th>Repetition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>E3</td>
<td>Talker 1</td>
</tr>
<tr>
<td>E3+AV</td>
<td>Talker 1</td>
</tr>
<tr>
<td>E6</td>
<td>Talker 1</td>
</tr>
<tr>
<td>E6+AV</td>
<td>Talker 1</td>
</tr>
</tbody>
</table>

3.3.4 Treatment and Testing

A paired-associate vocabulary learning procedure was implemented as the learning intervention, following earlier studies of acoustic variability and L2 word learning (Barcroft & Sommers, 2005, 2014a, 2014b; Sinkeviciute et al., 2019; Sommers & Barcroft, 2007). The learning and testing schedule was programmed with PsychoPy (Peirce, 2007). Before the treatment began, participants put on headphones equipped with a microphone (AT810 Cardioid Headset Microphone) and familiarized themselves with the vocabulary learning task by working through three practice examples. During the treatment, participants saw the meanings of the target words conveyed in visual images (i.e., copyright-free pictures retrieved from the Internet, standardized to a size of $400 \times 400$ pixels) while hearing the spoken forms of the
words. For each target item, the picture was displayed on the computer screen for 4 seconds, with the auditory presentation of the target word beginning 750 milliseconds (ms) after the picture appeared. The picture remained visible for the entire 4 seconds. A 2-second blank interval was inserted between trials.

During the treatment, the 40 target items were presented in a sequence of eight blocks of five items. The different experimental groups (E3, E6, E3+AV, E6+AV) received different numbers of encounters with the 40 target items with or without talker variability. Thus, the total number of encounters with target items was different between groups listening to spoken words three times and six times: E3 and E3+AV listened to 120 items (40 items × 3 encounters), and E6 and E6+AV listened to 240 items (40 items × 6 encounters). For all groups, the order of item presentation was randomized across participants, and the interval (or the number of items) between the first encounter and the subsequent encounter with the same word remained constant to control for spacing effects. For E3+AV and E6+AV, the order of presentations of talkers was fixed within all blocks so that participants always encountered new words produced by more intelligible talkers first and then gradually less intelligible talkers subsequently.

Immediately after the final exposure to each block of five items, a picture-naming test was administered. The assessment of knowledge after each block provided participants with a greater chance to recall the items than if the test was administered after the final exposure to a single block of 40 items. In the picture-naming test, participants were presented with the same pictures that were presented during the learning trial and asked to twice orally produce the words corresponding to the pictures shown on the computer screen. If participants did not remember a word, they were instructed to move to the next item. Their speech was recorded with a
TASCAM DR-05 audio recorder and digitized into a wav format (44.1 kHz sampling rate with 16-bit quantization). One out of two productions per word (i.e., a speech sample without fillers or self-corrections during articulation) was selected and stored in an individual sound file, with peak intensity normalized using Praat. Prior to data collection, issues with clarity of visual stimuli, trial procedures, and testing procedures were resolved through a pilot study with 20 university students with a similar learning background. Data for pilot study participants were not included in the main data analysis.

### 3.3.5 Procedure

The experiment was conducted over two sessions on two different days. On Day 1, participants completed a pretest, the treatment, an immediate posttest, and the Vocabulary Levels Test. For all participants, a 5-minute break was provided halfway through the treatment to reduce participant fatigue. On Day 2, approximately one week ($M = 6.6$ days) after the first session, participants took a surprise delayed posttest and filled out language background questionnaires. The test format (i.e., picture naming) across the three time points was the same except that 10 high-frequency items were added to the pretest to boost motivation. The 10 high-frequency items were not included in the analyses. Participants were told to learn the English words and were forewarned that they would be asked to produce words in response to pictures immediately after learning trials. Participants in the E3+AV and E6+AV were told that they would hear different voices. The treatment and tests were conducted individually with the researcher or a research assistant. All speech samples were recorded in a sound-attenuated booth. A total of 5,056 speech samples were elicited from 80 speakers on the pretest, immediate posttest, and delayed posttest.
3.3.6 Dependent Measures

Speech samples of words produced by Japanese learners were assessed for spoken form recall and word stress accuracy. The former measure was used to capture the process of form-meaning mappings, and the latter measure was meant to document the degree of phonological refinement. For spoken form recall, a binary coding scheme was adopted (correct = 1 point, incorrect = 0 points). Cases in which words were intelligible but influenced by L1 phonological system (e.g., substituting Japanese lateral flap for /r/ in *razor*, inserting vowels between consonant clusters such as /strənər/ → /sUstrənər/ in *strainer*) were counted as correct because the purpose of this test was to determine whether participants could link spoken form to meaning (see Sinkeviciute et al., 2019, for a similar approach). Word stress accuracy was measured in two ways. First, following L2 speech research (e.g., Isaacs & Trofimovich, 2012), performance of word stress was categorized in terms of placement accuracy: (a) primary stress is correctly placed on the right location (e.g., *TREADmill*), (b) primary stress is misplaced (e.g., *treadMILL*), and (c) primary stress is missing. One point was awarded to cases of accurate production and zero points to cases of misplacement or missing stress errors. The researcher and a native Japanese-speaking teacher who had extensive English language teaching experience in EFL and English-as-a-second-language (ESL) programs independently coded 100 speech samples (not included in the main dataset) for spoken form recall and stress placement measures. The results of Cohen’s kappa analyses confirmed high inter-coder agreements for spoken form recall ($k = .963$) and stress placement accuracy ($k = .967$). After disagreements were resolved through discussion, the remaining speech samples were coded by the researcher. Due to some instances of deletion of target vowels and significant changes
to syllable structures, 20 samples were not analyzed for stress placement accuracy.

In addition to the measure of stress placement accuracy, I measured vowel duration ratio (i.e., duration ratio of unstressed to stressed vowels) as one of the acoustic features important to the perception of lexical stress. In English, successful reduction of unstressed vowels in duration is one of the key characteristics determining acquisition of word stress (Beckman & Pierrehumbert, 1986) and more advanced L2 pronunciation proficiency (Trofimovich & Baker, 2006). Focusing on vowel duration instead of other acoustic correlates of stress such as vowel quality reduction (for discussion of the important role of vowel quality, see Zhang & Francis, 2010) was considered suitable given that L1 Japanese speakers were found to be able to acquire this feature over time with continued L2 exposure (Lee et al., 2006). It was reasonable to expect that this prosodic feature would improve to some degree after completion of the short-term training procedure adopted in this study. Using Praat (Boersma & Weenink, 2014), the duration (in milliseconds) of stressed and unstressed vowels was measured manually between two cursors placed at the onset and offset of voicing in each vowel (see Appendix A for target vowels). The ratio of unstressed to stressed vowels was calculated by dividing the duration of unstressed vowels by that of stressed vowels (when multiple unstressed vowels were available, average duration was calculated). Due to some instances of deletion of target vowels, significant changes to syllable structures, or poor sound quality, 221 speech samples were excluded from this analysis. The ratios for each word were averaged to yield a single score per participant. Finally, five English native speakers (3 females, 2 males) were recruited to read aloud 40 target words, and their vowel duration ratio was measured, which served as baseline data. A preliminary analysis showed that vowel duration ratio was significantly correlated with stress placement accuracy: pretest ($r = – .275, p$
such that more accurate stress placement was associated with a smaller vowel duration ratio (i.e., more English-like duration of unstressed vowels), which supported the validity of the two pronunciation measures.

3.3.7 Data Analysis

The descriptive statistics for spoken form recall, stress placement accuracy, and vowel duration ratio are presented in Table 2. The mean of vowel duration ratio for the native speaker baseline ($M = 0.70$, $SD = 0.07$, 95% CI [0.61, 0.79]) was lower than the means of learner groups. Lower scores for vowel duration ratio indicate the ability to successfully reduce the duration of unstressed vowels relative to the duration of stressed vowels.

In order to explore the effects of acoustic variability and frequency of exposure on form-meaning connection and phonological refinement, a series of mixed-design analysis of variance (ANOVA) with an alpha level of .05 were conducted with group as a between-participants variable (E3, E6, E3+AV, E6+AV) and time as a within-participants variable (pretest, immediate posttest, delayed posttest) for the three dependent measures (spoken form recall, stress placement accuracy, vowel duration ratio). Prior to conducting the analysis, normality of distribution was confirmed according to Shapiro-Wilk’s test, skewness statistics, and visual inspection of histograms for each group at the three test times. Mauchly’s test for sphericity of within-participants variances was significant for spoken form recall, placement accuracy, and vowel duration ratio, and therefore Greenhouse-Geisser correction for degrees of freedom was applied. Levene’s test for homogeneity of between-participants variances was significant for spoken form recall and stress
placement accuracy at the immediate posttest; therefore, Welch’s tests were employed to analyze group-mean differences for these two measures. To report the effect sizes of the group effects (frequency and talker variability), Cohen’s $d$ was calculated and was interpreted as small ($0.40 \leq d < 0.70$), medium ($0.70 \leq d < 1.00$), and large ($1.00 \leq d$) for between-participants contrasts (Plonsky & Oswald, 2014).

3.4 Results

3.4.1 Spoken Form Recall

The analysis showed significant effects for Time, $F(1.63, 128.97) = 1306, p < .001$, $\eta_p^2 = 0.94$, and Group, $F(3, 76) = 5.89, p = .001$, $\eta_p^2 = 0.19$, as well as a significant Time × Group interaction, $F(5.36, 135.85) = 7.26, p < .001$, $\eta_p^2 = 0.22$ (see Figure 1). Post hoc comparisons revealed no significant differences among the four groups at the pretest, $F(3, 76) = 0.58, p = .628$, $\eta_p^2 = 0.02$ or the delayed posttest, $F(3, 76) = 1.35, p = .263$, $\eta_p^2 = 0.05$. However, a significant difference was found among the four groups at the immediate posttest, $F(3, 41.4) = 11.94, p < .001$, $\eta_p^2 = 0.46$. Bonferroni-corrected pairwise post hoc comparisons showed that E6+AV outperformed E3 ($M_{\text{diff}} = 5.20, p < .001, d = 1.29$) and E3+AV ($M_{\text{diff}} = 6.00, p < .001, d = 1.69$) but did not outperform E6 ($M_{\text{diff}} = 1.10, p = 1.00, d = 0.32$). E6 had significantly higher scores than E3 ($M_{\text{diff}} = 4.10, p = .013, d = 0.91$) and E3+AV ($M_{\text{diff}} = 4.90, p = .002, d = 1.20$). No significant difference was found between E3 and E3+AV ($M_{\text{diff}} = 0.80, p = 1.00, d = 0.18$). In sum, at the immediate posttest, high frequency was especially useful for spoken form recall, whether or not high variability was present in the learning input.
Table 2. Means, standard deviations, and 95% confidence intervals for pretest, immediate posttest, and delayed posttest pronunciation scores by group

<table>
<thead>
<tr>
<th>Participant subgroups</th>
<th>Spoken form recall</th>
<th></th>
<th></th>
<th>Stress placement accuracy</th>
<th></th>
<th></th>
<th>Vowel duration ratio</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pretest</td>
<td>Immediate Posttest</td>
<td>Delayed Posttest</td>
<td>Pretest</td>
<td>Immediate Posttest</td>
<td>Delayed Posttest</td>
<td>Pretest</td>
<td>Immediate Posttest</td>
<td>Delayed Posttest</td>
</tr>
<tr>
<td>E3</td>
<td>10.9 (2.1)</td>
<td>30.4 (4.9)</td>
<td>14.9 (2.7)</td>
<td>0.39 (0.18)</td>
<td>0.86 (0.11)</td>
<td>0.66 (0.16)</td>
<td>1.17 (0.10)</td>
<td>0.98 (0.09)</td>
<td>1.14 (0.12)</td>
</tr>
<tr>
<td></td>
<td>[9.9, 11.9]</td>
<td>[28.0, 32.7]</td>
<td>[13.7, 16.1]</td>
<td>[0.31, 0.48]</td>
<td>[0.81, 0.91]</td>
<td>[0.58, 0.74]</td>
<td>[1.13, 1.22]</td>
<td>[0.93, 1.02]</td>
<td>[1.08, 1.20]</td>
</tr>
<tr>
<td>E6</td>
<td>10.6 (2.3)</td>
<td>34.5 (4.0)</td>
<td>15.4 (3.3)</td>
<td>0.31 (0.13)</td>
<td>0.92 (0.06)</td>
<td>0.70 (0.16)</td>
<td>1.18 (0.15)</td>
<td>0.90 (0.11)</td>
<td>1.03 (0.12)</td>
</tr>
<tr>
<td></td>
<td>[9.5, 11.6]</td>
<td>[32.6, 36.3]</td>
<td>[13.9, 16.9]</td>
<td>[0.24, 0.37]</td>
<td>[0.89, 0.94]</td>
<td>[0.63, 0.78]</td>
<td>[1.11, 1.25]</td>
<td>[0.85, 0.95]</td>
<td>[0.97, 1.08]</td>
</tr>
<tr>
<td>E3+AV</td>
<td>10.3 (2.3)</td>
<td>29.6 (4.2)</td>
<td>13.7 (2.2)</td>
<td>0.35 (0.18)</td>
<td>0.94 (0.05)</td>
<td>0.70 (0.14)</td>
<td>1.13 (0.13)</td>
<td>0.86 (0.14)</td>
<td>1.04 (0.13)</td>
</tr>
<tr>
<td></td>
<td>[9.2, 11.4]</td>
<td>[27.6, 31.5]</td>
<td>[12.7, 14.7]</td>
<td>[0.26, 0.43]</td>
<td>[0.91, 0.96]</td>
<td>[0.63, 0.76]</td>
<td>[1.07, 1.19]</td>
<td>[0.80, 0.93]</td>
<td>[0.98, 1.10]</td>
</tr>
<tr>
<td>E6+AV</td>
<td>11.2 (2.1)</td>
<td>35.6 (2.9)</td>
<td>14.9 (2.9)</td>
<td>0.36 (0.16)</td>
<td>0.94 (0.05)</td>
<td>0.68 (0.15)</td>
<td>1.15 (0.10)</td>
<td>0.80 (0.10)</td>
<td>0.98 (0.12)</td>
</tr>
<tr>
<td></td>
<td>[10.2, 12.1]</td>
<td>[34.2, 36.9]</td>
<td>[13.5, 16.3]</td>
<td>[0.28, 0.43]</td>
<td>[0.92, 0.97]</td>
<td>[0.61, 0.74]</td>
<td>[1.11, 1.20]</td>
<td>[0.75, 0.84]</td>
<td>[0.93, 1.04]</td>
</tr>
</tbody>
</table>

*Note.* Maximum score for spoken form recall was 40, and the maximum score for stress placement accuracy was 1. Standard deviations are in parentheses and upper and lower confidence intervals are in square brackets.
Figure 1. Group means for spoken form recall over time. Error bars represent 95% confidence intervals around the mean.

3.4.2 Stress Placement Accuracy

The analysis showed significant effects for Time, $F(1.90, 150.28) = 563$, $p < .001$, $\eta^2_p = 0.88$, but not for Group, $F(3, 76) = 0.23$, $p = .878$, $\eta^2_p = 0.01$; however, there was a significant Time × Group interaction, $F(5.78, 146.44) = 2.58$, $p = .022$, $\eta^2_p = 0.09$. Post hoc comparisons showed no significant effects for group at the pretest, $F(3, 76) = 0.93$, $p = .430$, $\eta^2_p = 0.04$, or the delayed posttest, $F(3, 76) = 0.34$, $p = .797$, $\eta^2_p = 0.01$. A significant effect was found for group at the immediate posttest, $F(3,
Bonferroni-corrected post hoc tests revealed that E6+AV outperformed E3 ($M_{\text{diff}} = 0.09$, $p = .003$ $d = 0.97$), but that there were no significant differences between E6+AV and the remaining two groups: E6 ($M_{\text{diff}} = 0.03$, $p = 1.00$, $d = 0.48$) or E3+AV ($M_{\text{diff}} = 0.01$, $p = 1.00$, $d = 0.10$). E3+AV had significantly higher scores than E3 ($M_{\text{diff}} = 0.08$, $p = .006$, $d = 0.91$). No significant differences were found between E6 and E3 ($M_{\text{diff}} = 0.06$, $p = .092$, $d = 0.65$) or E3+AV ($M_{\text{diff}} = 0.02$, $p = 1.00$, $d = 0.39$). In sum, at the immediate posttest, a combination of high frequency and high variability was most helpful for stress placement accuracy, and low frequency with variability was more helpful than low frequency alone.

Figure 2. Group means for stress placement accuracy over time. Error bars represent
95% confidence intervals around the mean.

### 3.4.3 Vowel Duration Ratio

The analysis showed significant effects for Time, $F(1.77, 139.71) = 199, p < .001, \eta_p^2 = 0.72$, and Group, $F(3, 76) = 5.48, p = .002, \eta_p^2 = 0.18$, as well as a significant Time × Group interaction, $F(5.48, 138.88) = 4.19, p < .001, \eta_p^2 = 0.14$.

Post hoc comparisons revealed no significant group effect at the pretest, $F(3, 76) = 0.56, p = .646, \eta_p^2 = 0.02$, but significant effects at both the immediate, $F(3, 76) = 9.40, p < .001, \eta_p^2 = 0.27$, and delayed posttests, $F(3, 76) = 5.96, p = .001, \eta_p^2 = 0.19$.

At the immediate posttest, Bonferroni-corrected post hoc tests revealed that E6+AV had significantly lower (more native-like) scores than E6 ($M_{\text{diff}} = 0.11, p = .020, d = 1.03$) and E3 ($M_{\text{diff}} = 0.18, p < .001, d = 1.92$) but the difference between E6+AV and E3+AV did not reach statistical significance ($M_{\text{diff}} = 0.07, p = .289, d = 0.59$). E3+AV also had significantly lower scores than E3 ($M_{\text{diff}} = 0.11, p = .012, d = 0.95$) but the difference between E3+AV and E6 was not significant ($M_{\text{diff}} = 0.04, p = 1.00, d = 0.29$). There was no significant difference between E3 and E6 ($M_{\text{diff}} = 0.08, p = .195, d = 0.75$). At the immediate posttest, it appears that high variability alone, with or without high frequency, was generally sufficient to encourage at least some change in vowel duration ratios.

At the delayed posttest, Bonferroni-corrected post hoc tests revealed that E6+AV showed lower (more native-like) vowel duration ratios than E3 ($M_{\text{diff}} = 0.16, p < .001, d = 1.32$) but did not show any significant difference when compared with E6 ($M_{\text{diff}} = 0.05, p = 1.00, d = 0.39$) and E3+AV ($M_{\text{diff}} = 0.06, p = .926, d = 0.45$). E6 had significantly lower scores than E3 ($M_{\text{diff}} = 0.11, p = .030, d = 0.92$) but no significant difference was found between E6 and E3+AV. Lastly, E3+AV appeared to show lower
vowel duration ratios than E3 as the difference approached statistical significance ($M_{diff} = 0.10, p = .058, d = 0.82$). Unlike the immediate posttest, where high variability seemed to undergird the production of unstressed vowel durations, the delayed posttest results appeared to be generally driven by high frequency, with a diminished contribution from high variability. Comparison of learner performance with a native speaker baseline at both the immediate and delayed posttests showed that all differences between the learner and baseline performances, except E6+AV at the immediate posttest, $t(38) = 1.71, p = .909$, were statistically significant. This result indicates that almost all learners’ performance, whether tested immediately or approximately one week after treatment, did not reach the level of native speakers’ performance.¹

Figure 3. Group means for vowel duration ratio over time. Error bars represent 95 %
confidence intervals around the mean.

### 3.5 Discussion

The present study was conducted to examine the effects of frequency of exposure and talker variability on L2 learners’ developing knowledge of form-meaning connection (spoken form recall) and phonological refinement (word stress accuracy). According to the two-step process of L2 lexical acquisition, the results (summarized in Table 3) overall supported our prediction. Frequency of exposure appeared to have a larger impact on the first stage of learning (form-meaning mapping) as supported by the findings that six encounters (E6 and E6+AV) produced significantly larger gains than three encounters (E3 and E3+AV) on a form recall test with relatively large effects ($d = 0.91–1.69$). In contrast, talker variability was more closely related to the second stage of learning (phonological refinement) given the findings that exposure to multiple voices (E3+AV and E6+AV) consistently led to larger gains than exposure to a single voice (E3 and E6) on most of the word stress measures with moderate-to-large effects ($d = 0.91–1.03$).

**Table 3. Summary of results**

<table>
<thead>
<tr>
<th></th>
<th>Immediate posttest</th>
<th>Delayed posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spoken form recall</td>
<td>E6+AV &gt; E3, E3+AV</td>
<td>No difference</td>
</tr>
<tr>
<td></td>
<td>E6 &gt; E3, E3+AV</td>
<td></td>
</tr>
<tr>
<td>Stress placement accuracy</td>
<td>E6+AV &gt; E3</td>
<td>No difference</td>
</tr>
<tr>
<td></td>
<td>E3+AV &gt; E3</td>
<td></td>
</tr>
<tr>
<td>Vowel duration ratio</td>
<td>E6+AV &lt; E3, E6</td>
<td>E6+AV &lt; E3, E6 &lt; E3</td>
</tr>
<tr>
<td></td>
<td>E3+AV &lt; E3</td>
<td>E3+AV &lt; E3*</td>
</tr>
<tr>
<td></td>
<td>NS &lt; E3, E6, E3+AV</td>
<td>NS &lt; E3, E6, E3+AV, E6+AV</td>
</tr>
</tbody>
</table>

*Note.* An asterisk indicates that the difference between E3+AV and E3 was marginally significant ($p = .058$). NS = native speaker.
Results of word stress accuracy showed a general pattern supporting a stronger effect of talker variability for both stress placement and duration measures compared to frequency of exposure. This finding adds to the value of a high-variability input for improving L2 pronunciation (Thomson, 2018), revealing that acoustic variability helps enhance pronunciation accuracy not only at a segmental (e.g., Kartushina & Martin, 2019) but also at a word level. Specifically, regarding stress placement accuracy, the E3+AV condition produced significantly larger gains than the E3 condition at immediate posttest. The absence of the expected advantage of high-variability input for E6+AV over E6 might be due to a ceiling effect for E6 ($M = 94\%$ accuracy). For vowel duration ratio, the two variability conditions (E6+AV and E3+AV) outperformed corresponding conditions without talker variability (E6 and E3) at immediate posttest. No significant difference between E6+AV and the native speakers’ baseline indicated that the performance of L2 learners approximated nativelike performance.

The advantage of E3+AV over E3 for stress placement accuracy and vowel duration ratio at the immediate posttest contrasts with the results of spoken form recall, where no such advantage emerged ($M_{E3+AV} = 29.6$ vs. $M_{E3} = 30.4$). In fact, the variability benefit was not expected to emerge for the E3+AV condition, given that three encounters are generally insufficient for a form-meaning connection to be established (Rott, 1999; van Zeeland & Schmitt, 2013). We reasoned that learners’ attention would be directed to the mapping of form to the word’s meaning, leaving few attentional resources for processing spoken forms (Barcroft, 2015). A possible explanation for this result might be that talker-specific voice information tends to be retained incidentally, even at low frequencies of word occurrence and with no explicit
instructions for learners to pay attention to talkers’ voice characteristics (Geiselman & Bellezza, 1976). Learners might thus have encoded talker-specific cues unintentionally and automatically while attempting to remember new words. As a result, the spoken forms that learners managed to recall by listening to three talkers were enhanced to a greater extent than those that they produced in a single-talker learning situation. However, because learners’ attention was focused on spoken forms, with limited cognitive resources remaining to process form-meaning mapping (Barcroft, 2015), no variability benefit was observed for establishing form-meaning connections (hence, E3+AV did not outperform E3 on form recall). Although acoustic variability appeared to be chiefly responsible for the learning of word stress, frequency effects also seemed to play an increasingly positive role in word stress production, as evidenced by the finding that the E6 condition outperformed the E3 condition at delayed posttests. Nevertheless, the effect of talker variability appears to have remained large ($d = 1.32$ for E6+AV vs. E3, $d = 0.92$ for E6 vs. E3).

Results of spoken form recall showed that six encounters with spoken word forms (E6 and E6+AV) produced larger learning gains than three encounters (E3 and E3+AV). These findings suggest that frequency effects are stronger for learning knowledge of form-meaning connection than variability effects. Despite the slightly higher recall rate for E6+AV ($M = 35.6$) than E6 ($M = 34.5$), the absence of a significant high-variability benefit appears to run counter to the findings by Barcroft and Sommers (2005), who found that exposure to words spoken by six talkers yielded significantly larger gains in recall of forms and meanings than exposure to words spoken by a single talker. A number of methodological differences between this study and Barcroft and Sommers’s research makes a simple comparison difficult (e.g., between-participants vs. within-participants design; experienced vs. inexperienced
learners). One possible reason for the difference in findings might be due to a ceiling effect for the E6 condition ($M = 86\%$ accuracy). Sinkeviciute et al. (2019) also suggested that the absence of a high-variability benefit for their L2 participants may have been due to ceiling effects. Another notable finding was related to the absence of a variability advantage for learners who encountered target words three times (E3 vs. E3+AV). In fact, the mean score of the E3+AV condition was lowest of the four conditions ($M = 29.6$), indicating that more than three encounters were necessary for a positive effect of talker variability to come into play. Or it could be that L2 learners may need enough encounters without acoustic variability to acquire the semantic information of new words. Only after form-meaning mappings are adequately established, learners may be ready to take advantage of acoustic variability enhancement (see Perrachione et al., 2011, for the facilitative vs. detrimental effects of HVPT on perceptually ready vs. unready learners). This result does not seem to align with Barcroft and Sommers’s (2005) finding of an advantage of the moderate variability condition (3 talkers) over the low-variability condition (1 talker). However, the number of encounters was different between the two studies: participants in the current study listened to each word produced once by each of three talkers (3 repetitions) while participants in Barcroft and Sommers’ study listened to each word produced twice by each of three talkers (6 repetitions). Goldinger et al. (1991) suggested that hearing different talkers’ voices requires more time for listeners to rehearse spoken words and transfer knowledge to long-term memory than dealing with a single-talker voice because more cognitive resources are needed to process “extra” information (i.e., indexical features) at every encounter with different voices. Learners in Barcroft and Sommers’ study might have been able to make significant learning gains by familiarizing themselves with different voices through repeated
encounters, reducing the cognitive load required to process words at the second encounter with the same talker. In contrast, learners in this study might not have been able to encode and rehearse words as efficiently due to a higher processing demand required at every encounter with a new talker, causing a negative or null effect on overall learning.

Taken together the findings of this study expand on earlier studies of L2 lexical acquisition. Both input quantity (frequency of exposure) and quality (acoustic variability) overall improve word learning, but the degree of such facilitative effects differs across different stages of lexical development (Jiang, 2000). First, learners hearing target words six times (E6 and E6+AV) recalled a greater number of spoken forms than those hearing three times (E3 and E3+AV). In contrast, no clear advantage of E6 over either E3 or E3+AV was observed on pronunciation measures (except for E6 outperforming E3 on vowel duration ratio at the delayed posttest). These findings suggest that frequency of encounters with spoken word forms exerts a larger impact on the initial stage of learning, promoting the mapping of forms to meanings and allowing learners to recall phonological forms when prompted by corresponding meanings. However, the quality of spoken forms produced may not be accurate yet at this stage (although these forms were sufficiently intelligible even in the presence of L1 Japanese accent or Japanese-specific minor pronunciation errors such as *strainer* produced as /sUtremər/). It is therefore possible to suggest that although repeated exposure promotes form-meaning mapping significantly, it may not be a sufficient condition for further finetuning of formal aspects (or phonetic details) of L2 words, at least in the context of a short-term learning procedure as adopted here. Once the mapping process is completed, additional repetitions may not motivate learners to attend to specific formal details unless instructed to do so (Saito, 2013). On the other
hand, benefits of talker variability for pronunciation measures reveal that acoustic variability may facilitate not only the mapping process (Barcroft & Sommers, 2005) but also further refinement of phonological forms. Exposure to acoustically varied speech triggers attention to and processing of spoken forms with indexical information encoded simultaneously (Geiselman & Bellezza, 1976; Goldinger et al., 1991), encouraging learners to discard irrelevant talker-specific information, extract common phonetic patterns across talkers (e.g., duration of stressed vowels is longer and duration of unstressed vowels is shorter), and develop accurate pronunciation of L2 words.

Findings of this study suggest that pronunciation of L2 words can be learned through exposure to the spoken forms of new words during vocabulary instruction. One way to do this is to utilize vocabulary learning apps and give opportunities for learners to encounter spoken forms of words recorded by multiple talkers. Apps that have the function to let users add audio information would allow learners to study new words while being exposed to the spoken word forms produced by different talkers multiple times. The YouGlish website (https://youglish.com) may serve this purpose as it provides multiple instances of searched words spoken by different English speakers. Classroom teachers are also encouraged to make use of audio materials that include a variety of talker voices (including teacher and learner voices). In introducing talker variability, teachers should remember that more than three repetitions might be needed to create the best learning conditions for both form-meaning mapping and spoken form enhancement. Lastly, it is important for teachers and researchers to be aware that word knowledge is a multifaceted construct involving various aspects other than form-meaning connection. This idea is not new, as we see the increase in the number of studies looking at different knowledge aspects such as collocations (Boers
& Webb, 2018). However, virtually no research has directly or systematically investigated development of pronunciation of L2 words within the framework of L2 lexical knowledge and acquisition. Exploration of this topic has not only theoretical but also pedagogical value given that the amount and type of exposure to new words may determine whether learners improve their pronunciation and form-meaning mapping of those words.

3.6 Limitations and Future Directions

The present study has several limitations and suggestions for future research with the view of enhancing our understanding of the effects of acoustic variability on L2 word learning. First, this study did not adopt the approach of earlier studies (e.g., Barcroft & Sommers, 2005) in controlling the influence of talkers’ characteristics (i.e., Barcroft and Sommers rotated different talkers used in low-variability conditions). Although a potentially confounding influence of talker intelligibility was minimized through conducting preliminary analysis and pilot testing, it is possible that this methodological difference might have influenced the current results. It is also worth investigating intelligibility (e.g., measured through word transcription) as a variable, exploring how learning gains through listening to multiple talkers with varied intelligibility levels compares to learning through listening to talkers with similar intelligibility levels. This would make this line of research more reflective of real-life learning because not all instances of speech that learners encounter are perfectly intelligible between and within talkers. Second, some data might have been close to ceiling (E6 for spoken form recall and stress placement accuracy). This may be attributed to the small number of words presented per block (i.e., 5 items) for experienced L2 learners, compared to earlier studies (i.e., 8 items for novice learners.
in Barcroft & Sommers, 2005). Although the large number of target items (i.e., 40 items) allowed sufficient variations in all test scores for the purpose of statistical analysis (absolute skewness statistics < 2; Field, 2009), the potential for ceiling effects highlights the methodological difficulty of eliciting a sufficiently large number of responses for pronunciation analysis and collecting nicely spread scores on vocabulary tests. Future research could perhaps consider adding an implicit measure (e.g., response latency) for assessing robustness of form-meaning connection. Third, this study measured word stress as a target pronunciation feature, and findings may not be generalized to improvement of other phonological features. There is a need for more studies looking at changes in different aspects of L2 speech using listener judgements (Bradlow et al., 1997) and acoustic analyses (Lambacher et al., 2005). Fourth, given that this study was conducted in a laboratory setting, findings are not immediately applicable to practical L2 learning contexts. One way to make this line of research more relevant to practical situations is to explore how talker variability affects learning when spoken forms of words are presented within sentences (Hirata, 2004), since encountering words in context is more common than in isolation across classroom settings. Another way is to use nonnative speakers as sources of talker variability and explore whether the variability benefit can be replicated. Such work would increase ecological validity of research as many language courses and programs today are taught by not only native speakers but also proficient L2 speakers and perhaps the most commonly heard spoken input within classrooms may be that of other L2 learners.

3.7 Notes

1. Significantly lower vowel duration ratios were found for a native speaker baseline
than for Japanese learners in E3, \( t(38) = 5.05, p < .001 \) (immediate posttest) and \( t(38) = 7.26, p < .001 \) (delayed posttest); E6, \( t(38) = 3.66, p = .005 \) (immediate posttest) and \( t(38) = 5.39, p < .001 \) (delayed posttest); E3+AV, \( t(38) = 3.00, p = .036 \) (immediate posttest) and \( t(38) = 5.55, p < .001 \) (delayed posttest); and E6+AV, \( t(38) = 4.62, p < .001 \) (delayed posttest).

### 3.8 References


Hirata, Y. (2004). Training native English speakers to perceive Japanese length


Chapter 4: To what extent does mode of input affect form-meaning connection, accentedness, and comprehensibility of second language words?

4.1 Introduction

Learners encountering the written forms of second language (L2) words tend to acquire more vocabulary than learners encountering their spoken forms (Brown et al., 2008; Vidal, 2011). However, mounting evidence reveals the value of spoken input when it is used as an additional mode of input to support reading (Brown et al., 2008; Andreas Bürki, 2010; Webb & Chang, in press, 2012). Learners pick up more words from reading texts with auditory support than reading without such support (Malone, 2018; Webb & Chang, 2012). The benefit of bimodal input has also been corroborated by findings that learners tend to acquire more words through watching L2 television with captions than without captions (Montero Perez et al., 2013). However, earlier studies have not focused on L2 learners’ productive knowledge (i.e., pronunciation), for the most part using written measures of form-meaning connection (e.g., choosing the first language [L1] translation corresponding to L2 orthographic word form provided) and their receptive knowledge of spoken forms (i.e., recognition). This is surprising because pronunciation is considered an important aspect of word knowledge (Nation, 2013) and essential for successful oral communication (Derwing & Munro, 2015). Lack of attention to pronunciation and the overuse of written measures may underestimate the value of encountering words in speech. Critically, no research has examined how mode of input affects how comprehensibly (easy for listeners to understand) L2 learners produce novel words.
Because L2 speakers can be sufficiently comprehensible despite having a noticeable foreign accent (Munro & Derwing, 1995) and because increasing comprehensibility is an appropriate goal of pronunciation teaching in globalized contexts (Levis, 2005), it is important to ensure that learners can produce the spoken forms of L2 words, and that the produced forms are sufficiently comprehensible to the listener so that L2 speakers are successful in oral communication. Therefore, the present study aimed to examine the value of spoken input for developing two aspects of L2 learners’ vocabulary knowledge: pronunciation (measured through comprehensibility and accentedness) and form-meaning connection (measured through spoken form recall) by comparing three conditions (reading-while-listening, listening-only, and reading-only).

4.2 Background

4.2.1 Written and Spoken Input and L2 Vocabulary Learning

Numerous research has documented that vocabulary learning occurs from exposure to written input, such as reading short sentences (Webb, 2007), reading graded readers (Brown et al., 2008), and studying word lists (Andreas Bürki, 2010). Researchers often measure learning gains in terms of form-meaning connection by asking learners whether they can recognize and recall word meanings and forms in written format. A growing number of studies have demonstrated that vocabulary learning also occurs through exposure to spoken input (e.g., van Zeeland & Schmitt, 2013), yet learners appeared to acquire more words from written input than from spoken input (Brown et al., 2008; Hatami, 2017; Vidal, 2011). For example, Vidal (2011) assigned first-year university students studying English as a foreign language (EFL) to two groups, either
listening to academic lectures or reading the transcribed texts. Participants were tested in written format before, immediately after, and a month after the treatment using a lexical developmental scale assessing knowledge ranging from partial (i.e., form recognition) to full competence (i.e., the ability to use the word in a sentence). Vidal concluded that reading was a more efficient source of input for vocabulary learning than listening, particularly for low-proficiency learners, who might have experienced difficulty segmenting connected speech for text processing and comprehension. However, the limited benefit of spoken input has been documented primarily through research focusing on word knowledge in terms of form-meaning connection in written format (Brown et al., 2008; Vidal, 2011) or recognition of spoken forms, for instance, via a multiple-choice test (Hatami, 2017).

Spoken input is considered useful for vocabulary learning as an additional mode to support reading (Brown et al., 2008; Andreas Bürki, 2010; Audrey Bürki et al., 2019; Malone, 2018; Webb & Chang, in press, 2012). Brown et al. (2008) compared vocabulary learning in three modes (reading-while-listening, reading-only, listening-only), with Japanese university students studying three graded readers. Different levels of form-meaning connection of target words was measured using tests of meaning recognition (via a multiple-choice test) and meaning recall (via an L2-to-L1 translation test) in written format. Participants showed the greatest learning gains in all test formats in the reading-while-listening condition, followed by reading-only and then listening-only condition. Studies following up on Brown et al. (2008) support the advantage of reading-while-listening over reading-only, targeting participants of different L2 proficiency (e.g., L2 beginners; Webb & Chang, 2012) and using different test formats (e.g., form recognition:
Malone, 2018; collocation recognition: Webb & Chang, in press). The attested advantage of bimodal input (reading and listening) over unimodal input (reading or listening) aligns with the cognitive theory of multimedia learning (Mayer, 2014). According to this model, presenting simultaneous modalities (e.g., written and spoken modes) leads to greater learning outcomes, such that success in learning depends on how multiple sensory systems are employed to integrate both verbal and non-verbal information into coherent mental representations (Niegeman & Heidig, 2012).

Few studies have examined how mode of input affects productive knowledge of spoken forms (i.e., pronunciation). To the best of our knowledge, two studies (Andreas Bürki, 2010; Audrey Bürki et al., 2019) have compared reading-while-listening with unimodal conditions (either reading-only or listening-only) and measured learning gains with tests of pronunciation and form-meaning connection. Andreas Bürki (2010) compared the effectiveness of an audio-supported paired-associate learning approach (i.e., combination of written and spoken input) with that of a traditional paired-associate learning approach (i.e., only written input) in the learning of multiple aspects of word knowledge including form-meaning connection (L1-to-L2 form recall in written mode) and pronunciation (productions of words elicited via a word-reading task were assessed for lexical stress and segmental accuracy). L1 Korean participants studying L2 English words in the audio-supported condition showed a significantly higher rate of written form recall and pronunciation accuracy in comparison to those in the reading-only condition.

On the other hand, Audrey Bürki et al. (2019) compared reading-while-listening with listening-only conditions, in which L1 French learners studied English-like pseudowords in a paired-associate format while viewing the meanings conveyed through corresponding
pictorial information. Spoken form recall was measured through a picture-naming test, and accuracy of vowel production was assessed with acoustic analysis and listener judgement. The results indicated that L2 orthographic input helped learners recall significantly more spoken forms than those receiving only spoken input, while learners with orthographic support substituted significantly more L1 sounds erroneously in vowel production. Audrey Bürki et al. concluded that exposure to written forms facilitates form-meaning mapping, but leads to non-targetlike pronunciation. It should be noted, however, that Andreas Bürki (2010) elicited production of L2 words using a controlled task (i.e., word reading) limiting the generalizability of the findings to the extent to which learners could accurately pronounce L2 words spontaneously (without reading the spellings of the words). Also, neither of the studies adopted measures reflecting listener’s holistic understanding of the word. Given that the first hurdle that learners need to overcome is to become understandable to listeners (Levis, 2005), assessing the degree of listener comprehension of L2 speech would increase the ecological validity of the pronunciation measures.

4.2.2 Orthographic Influence in L2 Pronunciation Learning

Studies investigating the role of orthographic input in L2 phonological acquisition have produced inconsistent findings (see Bassetti, 2008 for a review), suggesting that orthography can have positive (Erdener & Burnham, 2005; Solier et al., 2019) and negative (Bassetti & Atkinson, 2015; Audrey Bürki et al., 2019) effects. These mixed findings could be due to the degree to which L1 and L2 orthographic systems overlap with or deviate from each other. For example, Audrey Bürki et al. (2019)
attributed their finding of the negative influence of orthography to the incongruencies of
the grapheme-to-phoneme conversion rules between L1 French and L2 English.
Participants saw orthographies involving ⟨i⟩ and ⟨o⟩ which can be pronounced in
English as /ɪ/ (e.g., pick, kick, sick) and /ɑ/ (e.g., log, hot, cod) respectively; however, /ɪ/
is absent and ⟨o⟩ never corresponds to /ɑ/ in French. Thus, learners with orthographic
support tended to rely on their L1 orthographic system and substitute L1 vowels for L2
counterparts, resulting in orthography-induced, non-targetlike pronunciations (Bassetti,
2008). Another factor concerns the extent to which an orthographic system deviates from
one-to-one grapheme-to-phoneme correspondences or orthographic depth, which is
conceptualized on a transparent-to-opaque continuum. Spanish is a good example of a
transparent language with exceptions of a few letters (i.e., v, b, c, and ll) that can
correspond to two phonemes. In contrast, English has a rather opaque language system
with many instances of graphemes corresponding to two or more phonemes such as ⟨i⟩
as /ɪ/ (e.g., pick), /i/ (e.g., taxi), and /aɪ/ (e.g., kite). It is hypothesized that L1 users of
phonologically transparent writing systems rely on L2 orthographic input more than L1
users of phonologically opaque writing systems (Bassetti, 2008). Erdener and Burnham
(2005) tested and supported this hypothesis, investigating whether two groups of
participants speaking L1 Turkish (transparent) and L1 English (opaque) can accurately
repeat L2 words in two target languages: Spanish (transparent) and Irish (opaque). All
learners pronounced target words more accurately when they viewed a written
representation of the words. However, the benefit of orthographic input was greater for
L1 Turkish speakers than L1 English speakers in production of L2 Spanish words,
probably because native users of the phonologically transparent language could make
better use of L2 orthographic input in processing L2 auditory input. In contrast, when repeating L2 Irish words, L1 Turkish speakers were negatively impacted by the orthographic representation, while L1 English speakers were not.

**4.2.3 Pronunciation Constructs: Comprehensibility and Accentedness**

Since Munro and Derwing’s (1995) seminal study, several global constructs, including comprehensibility and accentedness, have been widely researched in L2 pronunciation studies (Derwing & Munro, 2015). Comprehensibility refers to listeners’ perceived ease or difficulty in understanding L2 speech, and accentedness (or linguistic nativelikeness) is defined as listeners’ judgments of how different L2 speech sounds from the expected language variety. These two constructs are measured through listeners’ ratings of speakers, using numerical point scales (e.g., $1 = \text{easy to understand}$, $9 = \text{hard to understand}$; $1 = \text{no accent}$, $9 = \text{heavily accented}$). Comprehensibility and accentedness are related but different constructs that need to be measured separately, given that L2 speakers with a stronger foreign accent do not necessarily sound less comprehensible (Munro & Derwing, 1995). Moreover, comprehensibility and accentedness proxy two different dimensions of L2 pronunciation learning (meeting minimum threshold for successful communication vs. mastering phonological detail characteristic of advanced proficiency). The first priority in L2 pronunciation instruction for learners, particularly in English, is to become understandable to listeners regardless of the severity of their foreign accent (Levis, 2005). This idea reflects the fact that speakers using L2 English far outnumber native speakers and that oral communication often takes place between non-native speakers in international contexts (Pennycook, 2017). The development of
comprehensible pronunciation is attainable for many adult L2 learners, as long as they practice the target language. Studies exploring L2 adult learners immersed into naturalistic settings (Derwing & Munro, 2013; Saito, 2015) suggest that comprehensibility appears to improve with increased L2 exposure. By contrast, attaining nativelike pronunciation is a difficult task. While learners’ accent often improves with practice, sounding non-accented requires a substantial amount of immersion experience (Trofimovich & Baker, 2006), earlier age of onset (Flege et al., 2006), strong motivation (Moyer, 2014), and special language learning abilities such as phonemic coding (Granena & Long, 2013). Although the majority of L2 pronunciation studies have focused on nativelike accuracy (captured through accentedness ratings), scholars have emphasized the importance of approaching the dynamic nature of instructed L2 speech learning from multiple angles including comprehensibility and accentedness (Thomson & Derwing, 2015). These two measures were thus adopted in this study to capture two distinct yet complementary aspects of L2 pronunciation learning.

4.2.4 Motivation for the Current Study

There are several reasons why more research is needed to investigate the effects of mode of input on L2 vocabulary acquisition. First, our understanding of the value of spoken input is biased toward the predominant use of written vocabulary measures. A recent study conducted by Webb and Chang (in press) has suggested that the choice of test format determines the effectiveness of input modality for vocabulary learning. In their study, participants were assigned to three experimental groups (reading, listening, reading-while-listening) and encountered collocations in the same graded reader.
Although the superiority of reading-while-listening over reading or listening was found as expected, the listening group learned as many collocations as the reading group. This finding contrasted with those of earlier studies testing knowledge of single-word items, concluding that spoken input had limited benefits (Brown et al., 2008; Hatami, 2017; Vidal, 2011). Just as spoken exposure favors developing knowledge of collocations, there may be other aspects of word knowledge that benefit from listening, which needs further exploration before determining the true value of spoken input for vocabulary acquisition.

Second, little is known about how mode of input affects productive knowledge of spoken forms (i.e., pronunciation). Although two studies investigated input modality using measures of pronunciation and form-meaning connection, neither of them compared reading versus listening (reading-while-listening vs. reading in Andreas Bürki, 2010; reading-while-listening vs. listening in Audrey Bürki et al., 2019). In order to determine the true value of spoken input, it is necessary to compare all three modality types at one time and examine the relative contribution of the three modes to vocabulary learning.

Third, previous studies used pronunciation measures focusing on nativelike accuracy (e.g., a forced-choice identification task by native listeners) and provided little insight into the degree to which listeners understand L2 speech. Given that instructed L2 speech learning is a multifaceted phenomenon that needs to be examined from multiple angles, it is important to include both comprehensibility (as a fundamental, realistic, and achievable goal) and accentedness (as an ideal, specialized, and advanced-level goal) (Derwing & Munro, 2015).

Finally, investigation of the extent to which (in)congruencies between spellings and sounds affect pronunciation acquisition was limited to segmental features (e.g.,
vowels and consonants), and studies have yet to examine the effects of sound-spelling consistency at the word level. Therefore, the present study, which was designed to respond to these research gaps, was guided by the following research questions:

1. To what extent does mode of input (reading-while-listening, reading-only, listening-only) affect the learning of form-meaning connections?

2. To what extent does mode of input (reading-while-listening, reading-only, listening-only) affect the learning of pronunciation (in terms of accentedness and comprehensibility)?

3. To what extent does sound-spelling consistency of words influence the effects of input mode on pronunciation learning?

Based on the cognitive theory of multimedia learning (Mayer, 2014) and findings of earlier studies (Andreas Bürki, 2010; Audrey Bürki et al., 2019; Malone, 2018; Webb & Chang, 2012), the reading-while-listening mode was predicted to facilitate learners’ development of form-meaning connections to a greater degree than the reading-only or listening-only mode. This is because exposure to multimodal input (audio and orthographic) can help L2 learners access greater linguistic resources from different angles, resulting in deeper processing and greater acquisition of new words. For pronunciation measures, learners receiving spoken input (reading-while-listening and listening-only) were expected to perform better than those receiving only written input (reading-only). However, the predicted superiority of the spoken input modes over the reading-only mode might be gradually reduced as sound-spelling consistency of target
words increases. Conversely, an additional mode of input might place demands on learners’ limited cognitive capacity and result in a negative impact for multimodal input on learning, particularly in the present study where participants received only one exposure to each target word (Baddeley, 1986). Lastly, different effects of mode of input might arise for different pronunciation measures. Learners receiving written input might tend to sound more heavily accented than learners receiving only spoken input because the availability of orthographic information triggers grapheme-to-phoneme recoding applying L2 and L1 conversion rules, so that learners’ production of L2 words is influenced by their L1 (Audrey Bürki et al., 2019). Because L1 influences might be more detrimental for listener judgments of accentedness than comprehensibility, a negative effect of orthography might be reduced for comprehensibility compared to accentedness, particularly as sound-spelling consistency of words increases.

4.3 Method

4.3.1 Overview of the Study

The study adopted a pretest–posttest design with three experimental groups (reading-only, listening-only, and reading-while-listening) and three testing trials (pretest, immediate posttest, and delayed posttest). Participants were randomly assigned to three experimental groups which encountered target words in different modes of input: reading-only (RO), listening-only (LO), and reading-while-listening (RWL). During the treatment, participants learned 40 English words through seeing and/or hearing the words while viewing their corresponding pictures. A picture-naming test was administered at the three testing times, and the elicited samples were evaluated for form-meaning connection and
pronunciation measures.

4.3.2 Participants

Seventy-nine Japanese university EFL students in Japan participated in this experiment. Four participants were excluded from the analysis because three had lived abroad for an extended period of time (5–12 years) and one did not complete a delayed posttest. The remaining 75 participants ($M_{\text{age}} = 19.5$, range = 18–24) had studied English for a minimum of six years in instructional settings. All participants except one had scored 90 percent or higher on the 1,000 word level of the Vocabulary Levels Test (Webb et al., 2017), and all except one had scored 80 percent or higher on the 2,000 word level of the test. Their mean score at the 2,000 level was 28.76, indicating that they had receptive knowledge of almost all of the most frequent 2,000 words. The 75 participants were randomly assigned to three experimental groups: RO ($n = 25$), LO ($n = 25$), and RWL ($n = 25$). There was no between-group difference in vocabulary test scores, $F(2, 72) = 0.70, p = .503, \eta_p^2 = 0.02$. All participants reported normal hearing.

4.3.3 Target Items

Forty target words were selected according to three criteria. First, a pool of low-frequency words was created by collecting English words that were beyond the most frequent 5,000 word families in Nation’s BNC/COCA word lists (Nation, 2012). Second, because the treatment involved learning written and spoken forms attached to meanings conveyed in visual images (pictures), only concrete nouns were selected as target items. Third, words that could be replaced with high-frequency synonyms were avoided to
reduce the possibility that high-frequency synonyms of the target items would be produced in the picture-naming test. The selected items were measured in terms of sound-spelling consistency (i.e., the degree to which the pronunciation of a word matches its spelling). Using consistency norms for English words developed by Chee et al. (2020), a feedforward (i.e., spelling-to-sound) rime consistency score was calculated for each target word. This score accounts for the frequencies of similarly spelled words for a given pronunciation (e.g., “-oar” can be regarded as consistent due to many instances of words which contain the rime and are pronounced similarly among the words such as soar, boar, and hoar). As an example, the consistency score in our data indicates that toupee (0.128) is less consistent than spatula (0.476) or parakeet (0.525). The score for three words (i.e., abalone, loquat, maracas) was not available and hence not analyzed in this study (see Table 1 for target items and consistency scores).

Each of the 40 target words was recorded twice by a female native speaker of English using a TASCAM DR-05 audio recorder and digitized into a wav format at a sampling rate of 44.1 kHz (16-bit resolution). The better of the two productions was selected according to clarity, naturalness, and lack of background noise and then stored as an individual sound file, with peak intensity normalized using digital speech-analysis software (Praat) (Boersma & Weenink, 2014). The stimuli were clear and comprehensible based on the judgement of another native English speaker. Pilot testing showed that two native English speakers successfully identified all 40 productions recorded by the model speaker.
Table 1. 40 Target words with basic Item information and sound-spelling consistency scores

<table>
<thead>
<tr>
<th>No.</th>
<th>Target word</th>
<th>Phoneme counts</th>
<th>Number of syllables</th>
<th>Consistency scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>abalone</td>
<td>7</td>
<td>4</td>
<td>Off-List</td>
</tr>
<tr>
<td>2</td>
<td>acorn</td>
<td>5</td>
<td>2</td>
<td>0.511</td>
</tr>
<tr>
<td>3</td>
<td>armadillo</td>
<td>8</td>
<td>4</td>
<td>0.469</td>
</tr>
<tr>
<td>4</td>
<td>binoculars</td>
<td>10</td>
<td>4</td>
<td>0.414</td>
</tr>
<tr>
<td>5</td>
<td>caramel</td>
<td>6</td>
<td>3</td>
<td>0.918</td>
</tr>
<tr>
<td>6</td>
<td>carousel</td>
<td>7</td>
<td>3</td>
<td>0.297</td>
</tr>
<tr>
<td>7</td>
<td>catapult</td>
<td>8</td>
<td>3</td>
<td>0.647</td>
</tr>
<tr>
<td>8</td>
<td>celery</td>
<td>6</td>
<td>3</td>
<td>0.586</td>
</tr>
<tr>
<td>9</td>
<td>chameleon</td>
<td>8</td>
<td>4</td>
<td>0.481</td>
</tr>
<tr>
<td>10</td>
<td>chandelier</td>
<td>9</td>
<td>4</td>
<td>0.632</td>
</tr>
<tr>
<td>11</td>
<td>chisel</td>
<td>4</td>
<td>2</td>
<td>0.702</td>
</tr>
<tr>
<td>12</td>
<td>cicada</td>
<td>6</td>
<td>3</td>
<td>0.185</td>
</tr>
<tr>
<td>13</td>
<td>clover</td>
<td>6</td>
<td>2</td>
<td>0.658</td>
</tr>
<tr>
<td>14</td>
<td>crayon</td>
<td>5</td>
<td>2</td>
<td>0.820</td>
</tr>
<tr>
<td>15</td>
<td>croissant</td>
<td>7</td>
<td>2</td>
<td>0.055</td>
</tr>
<tr>
<td>16</td>
<td>escalator</td>
<td>9</td>
<td>4</td>
<td>0.556</td>
</tr>
<tr>
<td>17</td>
<td>ladle</td>
<td>4</td>
<td>2</td>
<td>0.573</td>
</tr>
<tr>
<td>18</td>
<td>loquat</td>
<td>6</td>
<td>2</td>
<td>Off-List</td>
</tr>
<tr>
<td>19</td>
<td>lotus</td>
<td>5</td>
<td>2</td>
<td>0.623</td>
</tr>
<tr>
<td>20</td>
<td>maracas</td>
<td>7</td>
<td>3</td>
<td>Off-List</td>
</tr>
<tr>
<td>21</td>
<td>marshmallow</td>
<td>8</td>
<td>3</td>
<td>0.490</td>
</tr>
<tr>
<td>22</td>
<td>mermaid</td>
<td>6</td>
<td>2</td>
<td>0.949</td>
</tr>
<tr>
<td>23</td>
<td>pacifier</td>
<td>8</td>
<td>4</td>
<td>0.435</td>
</tr>
<tr>
<td>24</td>
<td>parakeet</td>
<td>8</td>
<td>3</td>
<td>0.525</td>
</tr>
<tr>
<td>25</td>
<td>persimmon</td>
<td>8</td>
<td>3</td>
<td>0.619</td>
</tr>
<tr>
<td>26</td>
<td>podium</td>
<td>6</td>
<td>3</td>
<td>0.488</td>
</tr>
<tr>
<td>27</td>
<td>porcupine</td>
<td>9</td>
<td>3</td>
<td>0.575</td>
</tr>
<tr>
<td>28</td>
<td>protractor</td>
<td>10</td>
<td>3</td>
<td>0.643</td>
</tr>
<tr>
<td>29</td>
<td>raccoon</td>
<td>5</td>
<td>2</td>
<td>0.716</td>
</tr>
<tr>
<td>30</td>
<td>raisin</td>
<td>4</td>
<td>2</td>
<td>0.544</td>
</tr>
<tr>
<td>31</td>
<td>razor</td>
<td>5</td>
<td>2</td>
<td>0.279</td>
</tr>
</tbody>
</table>
Note. Sound-spelling consistency scores were calculated using Chee et al.’s (2020) consistency norms for 37,677 English words with type selected as a counting unit. Items labelled as “Off-List” were not available in the norm list and hence not analyzed in this study.

### 4.3.4 Treatment and Testing

Paired-associate vocabulary learning was implemented as the learning intervention for three reasons. It allowed for careful control of the presentation of the target items, it has been found to positively contribute to learning the written forms of words, and it has been used frequently in studies of vocabulary learning (Nation & Webb, 2011). The learning and testing schedule was programmed with PsychoPy (Peirce, 2007). Before the treatment began, participants put on headphones equipped with a microphone (AT810 Cardioid Headset Microphone) and familiarized themselves with the vocabulary learning task by working through three practice examples. During the treatment, participants encountered the meanings of the target words conveyed in visual images (i.e., copyright-free pictures retrieved from the Internet, standardized to a size of 400 × 400 pixels) while seeing and/or hearing the target word forms. For each target item, the
picture was displayed on the computer screen for 4 seconds. For the conditions involving spoken input (LO and RWL), the auditory presentation of the target word began 750 milliseconds after the picture appeared. For the conditions receiving written input (RO and RWL), the orthographic presentation of the target word appeared under the corresponding picture for 4 seconds. A 2-second blank interval was inserted between trials.

During the treatment, the 40 target items were presented in a sequence of eight blocks of five items. The experimental groups received exposure to each of the 40 target items once in one of three different modes of input (LO, RO, and RWL). For all groups, the order of item presentation was randomized across participants, and the interval (or the number of items) between the first encounter and the next encounter with the same word remained constant to control for spacing effects. Immediately after the final exposure to each block of five items, a picture-naming test was administered. Measuring participants’ knowledge after each block of five items was meant to enhance the learning, in the sense that participants were encouraged to recall newly learned words to a greater extent than if the test was administered after the final exposure to a single block of 40 items. In the picture-naming test, participants were presented with the same pictures that were presented during the learning trial and asked to twice orally produce the words corresponding to the pictures shown on the computer screen. If participants did not remember a word, they were instructed to move to the next item. Their speech was recorded with a TASCAM DR-05 audio recorder and digitized into a wav format at a sampling rate of 44.1 kHz with 16-bit resolution. One out of two productions per word (i.e., a speech sample without fillers or self-corrections during articulation) was selected.
and stored in an individual sound file, with peak intensity normalized using Praat (Boersma & Weenink, 2014). Prior to data collection, issues with clarity of visual stimuli, trial procedures, and testing procedures were resolved through a pilot study with 20 university students with a similar learning background. Data for pilot study participants were not included in the main data analysis (see Appendix B for visual prompts).

4.3.5 Procedure

The experiment was conducted over two sessions on two different days. On Day 1, participants took the pretest, completed the treatment, an immediate posttest, and the Vocabulary Levels Test. On Day 2, approximately 6 days \((M = 6.1, SD = 3.6)\) after the first session, participants completed a surprise delayed posttest and filled out language background questionnaires. The test format (i.e., picture naming) across three time points was the same except that 10 high-frequency items were added to the pretest to boost motivation. The 10 high-frequency items were not included in the analyses. Participants were told to learn the English words, and forewarned that they would be asked to produce words in response to pictures immediately after learning trials. Participants in the RO condition were told that they would see the spellings of words without any auditory information presented. Participants in the RWL condition were told that they would see and hear target words simultaneously. The treatment and tests were conducted individually with the researcher or a research assistant. All speech samples were recorded in a sound-attenuated booth at a university. A total of 4,061 speech samples were elicited from 75 speakers on the pretest, immediate posttest, and delayed posttest and evaluated for form-meaning connection and pronunciation measures.
4.3.6 Form-Meaning Connection and Pronunciation Measures

To assess form-meaning connection, spoken form recall (e.g., production of accurate forms of words in a picture-naming test) was measured. Form recall is considered the most difficult measure of form–meaning knowledge compared to three other measures: form recognition, meaning recognition, and meaning recall (Laufer & Goldstein, 2004). For pronunciation measures, following (Derwing & Munro, 2015), two constructs were measured: accentedness (i.e., listener rating of the extent to which learners’ word productions deviated from a native variety of the target language) and comprehensibility (i.e., listener rating of the degree of effort needed to comprehend learners’ word productions).

To measure three aspects of word knowledge (spoken form recall, accentedness, comprehensibility), six native speakers of North American English (three females, three males) were recruited to participate in a series of rating sessions. Three of six speakers had language teaching experiences in EFL and English-as-a-second language (ESL) contexts. All six speakers had no hearing problems and were highly familiar with Japanese-accented English ($M = 5.1$, range = 4–6 in response to 1 = not familiar at all, 6 = very familiar). Raters completed a word listening task programmed using PsychoPy (Peirce, 2007). In this task, raters first listened to each of the speech samples and pressed an “f” key for correct and a “j” key for incorrect word pronunciation. Pronunciation was considered correct if it was sufficiently intelligible with minor errors or foreign accents present (Kang et al., 2013). Form recall was coded dichotomously with 1 point assigned to responses judged as correct by all six raters and 0 points to responses judged as
incorrect by one or more raters or missing responses (i.e., failure to name pictures).

Second, for samples judged correct, listeners rated accentedness (1 = not accented at all, 5 = heavily accented) and comprehensibility (1 = easy to understand, 5 = hard to understand). The 5-point numerical scale was adopted because in contrast to earlier studies measuring L2 speech at sentence- or discourse-levels (cf. Munro & Derwing, 1995 using a 9-point scale), this study focused on words as a unit of speech samples. Given the relatively limited amount of linguistic information available at the word level, using a large number of scale points might make the rating task excessively challenging or even confusing. Also, for intuitive L2 speech judgments of this kind, rating performance using a 5-point scale could be as reliable as when a 9-point scale is used (Isaacs & Thomson, 2013). A pilot study also confirmed that the choice of a 5-point rating scale was appropriate for rating word pronunciation in this study. Prior to main rating sessions, raters first familiarized themselves with 40 target words and a rating procedure through completing a practice listening task with 50 items (not included for analysis in this study). They then listened to each of the speech samples from the main dataset, completed a binary rating task (correct vs. incorrect), and rated accentedness and comprehensibility for items they had judged as correct. Raters were presented with 41 blocks of 100 samples and a final block of 41 samples. These samples consisted of random selection of pretest, immediate posttest, and delayed posttest items, as well as native speakers’ samples (included as distracter items), totaling 4,141 items (4,061 from Japanese speakers + 80 from native speakers). The inclusion of the native speaker samples also allowed us to confirm the reliability of raters’ performance. Recordings were played only once. In the first meeting with the researcher, the raters first practiced rating
50 samples and then rated the first block of 100 samples. Raters subsequently evaluated the remaining samples in their own time.

4.3.7 Data Analysis

Preliminary analysis of raters’ responses to native speaker samples showed that raters consistently judged the native-speaker baseline as correct (100% accuracy), least accented (94% of samples were rated as 1 = not accented at all), and easiest to understand (99% of samples were rated as 1 = easy to understand), confirming the reliability of the raters’ performance and their understanding of the tasks. In response to the first research question, data of form recall (1 = correct, 0 = incorrect) were analyzed in a generalized linear mixed-effects model with a binomial distribution (Jaeger, 2008). The fixed factors included (dummy-coded) mode of input (LO, RO, RWL), (dummy-coded) time (pretest, immediate posttest, delayed posttest), and the interaction term. We included random intercepts for participant (75 levels), word (40 levels), and rater (6 levels), a by-word random slope for the mode-of-input factor, and the correlation between the slope and the intercept. Before conducting analysis to answer the second and third research questions, accentedness and comprehensibility ratings were calculated only for responses to the target items that learners did not recall at pretest but recalled after treatment, such that pronunciation scores reflected the development of the spoken forms of unfamiliar words. The resulting datapoints (or observations) for accentedness and comprehensibility were 10,434 cases (= 1,739 × 6 raters). The interrater reliability for accentedness (α = .754) and comprehensibility (α = .722) was not perfect yet satisfactory for research purposes (α > .70) (Larson-Hall, 2010). Data of accentedness and comprehensibility were analyzed in
a mixed-effects model. The fixed factors included (dummy-coded) mode of input (LO, RO, RWL), (dummy-coded) time (immediate and delayed posttests), (grand-mean centered) sound-spelling consistency, and all of the interactions between them (stepwise model comparison was not adopted here). We included random intercepts for participant (75 levels), word (40 levels), and rater (6 levels), a by-word random slope for the mode-of-input factor, a by-participant random slope for the consistency factor, and the correlations between the slopes and the intercepts.

4.4 Results

Descriptive statistics of spoken form recall, accentedness, and comprehensibility are presented in Table 2. Changes in scores for spoken form recall, accentedness, and comprehensibility between different test timepoints are illustrated in Figures 1, 2, and 3.
Table 2. Means, standard deviations, and 95% confidence intervals for spoken form recall, accentedness, and comprehensibility by group

<table>
<thead>
<tr>
<th>Subgroups</th>
<th>Spoken form recall</th>
<th>Accentedness</th>
<th>Comprehensibility</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pretest</td>
<td>Immediate posttest</td>
<td>Delayed posttest</td>
</tr>
<tr>
<td>LO</td>
<td>0.26 (0.06)</td>
<td>0.56 (0.10)</td>
<td>0.35 (0.07)</td>
</tr>
<tr>
<td></td>
<td>[0.24, 0.29]</td>
<td>[0.52, 0.60]</td>
<td>[0.33, 0.38]</td>
</tr>
<tr>
<td>RO</td>
<td>0.26 (0.07)</td>
<td>0.63 (0.11)</td>
<td>0.39 (0.08)</td>
</tr>
<tr>
<td></td>
<td>[0.23, 0.29]</td>
<td>[0.58, 0.67]</td>
<td>[0.35, 0.42]</td>
</tr>
<tr>
<td>RWL</td>
<td>0.28 (0.05)</td>
<td>0.64 (0.10)</td>
<td>0.40 (0.08)</td>
</tr>
<tr>
<td></td>
<td>[0.26, 0.30]</td>
<td>[0.60, 0.68]</td>
<td>[0.36, 0.43]</td>
</tr>
</tbody>
</table>

Note. Accentedness (1 = not accented at all to 5 = heavily accented); comprehensibility (1 = easy to understand to 5 = hard to understand). Standard deviations are in parentheses and upper and lower confidence intervals are in square brackets.
4.4.1 Spoken Form Recall

Logistic mixed-effects modeling revealed significant effects for Time, $\chi^2(2) = 6502.68, p < .001$, but not for Group (mode of input), $\chi^2(2) = 5.23, p = .073$. However, there was a significant Time $\times$ Group interaction, $\chi^2(4) = 22.12, p < .001$ (see Figure 1). There were no statistically significant differences between the three groups at pretest (LO, $M = 25\%$, $SD = 5\%$; RO, $M = 25\%$, $SD = 7\%$; RWL, $M = 27\%$, $SD = 5\%$; LO vs. RO: $\beta = 0.03, SE = 0.33, z = 0.11, p = .915$; LO vs. RWL: $\beta = 0.50, SE = 0.33, z = 1.51, p = .130$; RO vs. RWL: $\beta = 0.47, SE = 0.33, z = 1.41, p = .159$).\(^1\) At immediate posttest, learners receiving both written and spoken input significantly outperformed those receiving only spoken input (R WL, $M = 53\%$, $SD = 10\%$; LO, $M = 48\%$, $SD = 10\%$; $\beta = 0.51, SE = 0.41, z = 1.99, p = .047$), but learners in the RWL group did not outperform those receiving only written input (RO, $M = 49\%$, $SD = 10\%$; $\beta = 0.01, SE = 0.25, z = 0.05, p = .961$). There was a tendency for the RO group to yield a larger gain than for the LO group ($\beta = 0.50, SE = 0.27, z = 1.82, p = .068$). At delayed posttest, no significant differences were found between the three groups with the difference between RWL and LO approaching statistical significance (LO, $M = 33\%$, $SD = 7\%$; RO = 34\%, $SD = 7\%$; RWL, $M = 36\%$, $SD = 7\%$; LO vs. RO: $\beta = 0.63, SE = 0.40, z = 1.58, p = .114$; RO vs. RWL: $\beta = 0.05, SE = 0.38, z = 0.13, p = .899$; RWL vs. LO: $\beta = 0.68, SE = 0.39, z = 1.76, p = .078$).
Figure 1. Group estimated marginal means for spoken form recall over time. The estimated means are different from the raw means in that the former were extracted from a statistical model (i.e., mixed-effects model); the latter was a product of averaging raw data points. Error bars represent standard errors (SE) of the means.

4.4.2 Accentedness and Comprehensibility

Mixed effects modelling was used to explore a three-way interaction effect (Time × Group × Consistency), separately for accentedness and comprehensibility. The interaction was statistically significant for accentedness, $F(2, 6770.9) = 3.51, p = .030$, and comprehensibility, $F(2, 7373.7) = 5.71, p = .003$. For accentedness ratings at immediate posttest, learners receiving only written input were perceived significantly more heavily accented than those receiving only spoken input or both spoken and written input together (RO, $M = 4.08, SD = 0.24$; LO, $M = 3.78, SD = 0.27$; RWL, $M = 3.80, SD = 0.26$; RO vs. LO: $\beta = 0.42, SE = 0.10, t = 4.21, p < .001$; RO vs. RWL, $\beta = 0.37, SE = 0.09, t = 4.04, p < .001$). No significant difference was found between the LO and RWL groups ($\beta = 0.05, SE = 0.09, t = 0.57, p = .570$). At
delayed posttest, the superiority of spoken input over the written-input-only condition remained present for LO versus RO (LO, $M = 3.73$, $SD = 0.41$; RO, $M = 3.98$, $SD = 0.31$; $\beta = 0.35$, $SE = 0.13$, $t = 2.68, p = .010$), and also for RWL versus RO to a lesser degree (RWL, $M = 3.86$, $SD = 0.26$; $\beta = 0.20$, $SE = 0.11$, $t = 1.84, p = .071$). No significant difference was found between the LO and RWL groups ($\beta = 0.16$, $SE = 0.11$, $t = 1.39, p = .174$).

For comprehensibility ratings at immediate posttest, learners receiving spoken input or both spoken and written input simultaneously were perceived significantly more comprehensible than those receiving written input only (LO, $M = 3.18$, $SD = 0.35$; RWL, $M = 3.20$, $SD = 0.30$; RO, $M = 3.43$, $SD = 0.30$; LO vs. RO: $\beta = 0.36$, $SE = 0.12$, $t = 3.05, p = .003$; RWL vs. RO: $\beta = 0.33$, $SE = 0.10$, $t = 3.27, p = .002$). There was no significant difference between the LO and RWL groups ($\beta = 0.03$, $SE = 0.10$, $t = 0.35, p = .73$). At delayed posttest, no significant differences were observed between the three groups with the difference between RWL and RO approaching significance (LO, $M = 2.82$, $SD = 0.50$; RO, $M = 3.15$, $SD = 0.35$; RWL, $M = 2.98$, $SD = 0.41$; LO vs. RO: $\beta = 0.27$, $SE = 0.17$, $t = 1.63, p = .109$; LO vs. RWL: $\beta = 0.04$, $SE = 0.14$, $t = 0.27, p = .788$; RWL vs. RO: $\beta = 0.23$, $SE = 0.13$, $t = 1.75, p = .087$).
Figure 2. Group estimated marginal means for accentedness over time. Error bars represent standard errors (SE) of the means.

Figure 3. Group estimated marginal means for comprehensibility over time. Error bars represent standard errors (SE) of the means.
Sound-spelling consistency of words was in general negatively associated with pronunciation measures (accentedness: $\beta = -0.64$, $SE = 0.24$, $t = -2.70$, $p = .001$; comprehensibility: $\beta = -1.11$, $SE = 0.32$, $t = -3.46$, $p = .001$), indicating that productions of consistent words tended to be perceived more nativelike and comprehensible than those of inconsistent words. However, Figures 4 to 7 illustrate that the effect of consistency appeared to vary across experimental groups and testing times. At immediate posttest for both accentedness and comprehensibility, the strength of the relationship between consistency and pronunciation measures was significantly different between the LO and RO groups as well as between the LO and RWL groups (accentedness: RO vs. LO: $\beta = -0.88$, $SE = 0.38$, $t = -2.33$, $p = .024$; RWL vs. LO: $\beta = -0.81$, $SE = 0.31$, $t = -2.57$, $p = .014$; comprehensibility: RO vs. LO: $\beta = -1.43$, $SE = 0.52$, $t = -2.73$, $p = .009$; RWL vs. LO: $\beta = -1.08$, $SE = 0.41$, $t = -2.64$, $p = .011$), indicating that the extent to which productions of consistent words become more nativelike and comprehensible was greater for learners receiving written input (RO and RWL) than those receiving spoken input only (LO). Such an effect was not found when two groups (RWL and RO) receiving written input were compared either for accentedness ($\beta = 0.08$, $SE = 0.32$, $t = 0.24$, $p = .814$) or comprehensibility ($\beta = 0.35$, $SE = 0.42$, $t = 83$, $p = .410$). At delayed posttest, no significant variations in the effect of consistency on pronunciation measures were found between the three groups for either accentedness (RO vs. LO: $\beta = -0.07$, $SE = 0.61$, $t = -0.12$, $p = .908$; RWL vs. LO: $\beta = 0.02$, $SE = 0.51$, $t = 0.04$, $p = .968$; RWL vs. RO: $\beta = 0.09$, $SE = 0.47$, $t = 0.20$, $p = .846$) or comprehensibility (RO vs. LO: $\beta = -0.43$, $SE = 0.77$, $t = -0.56$, $p = .581$; RWL vs. LO: $\beta = -0.01$, $SE = 0.63$, $t = -0.02$, $p = .985$; RWL vs. RO: $\beta = 0.41$, $SE = 0.59$, $t = 0.70$, $p = .488$).
Figure 4. Relationship between sound-spelling consistency and accentedness per group (immediate posttest).

Figure 5. Relationship between sound-spelling consistency and accentedness per group (delayed posttest).
Finally, a follow-up analysis was conducted to examine whether learners in the written-input conditions (RO and RWL) could perform better than those in the
spoken-input-only condition (LO) when learning words that are highly consistent in sound-spelling correspondence. The target words with consistency scores (i.e., 37 items) were sequenced in the order of consistency and organized into three categories (12 low-consistency words, 13 mid-consistency words, and 12 high-consistency words). A mixed-effects modelling analysis was conducted on the set of 12 high-consistency words. The analysis showed a similar pattern of the results found in the original analysis with all target words included (see above). Although no significant between-groups differences were found for the delayed-posttest result, the LO and RWL groups significantly outperformed RO at immediate posttest for accentedness (RO vs. LO: $\beta = 0.40$, $SE = 0.12$, $t = 3.37$, $p = .002$; RO vs. RWL: $\beta = 0.44$, $SE = 0.15$, $t = 3.03$, $p = .006$) and comprehensibility (RO vs. LO: $\beta = 0.36$, $SE = 0.15$, $t = 2.44$, $p = .022$; RO vs. RWL: $\beta = 0.38$, $SE = 0.17$, $t = 2.19$, $p = .041$). No significant differences were found between the LO and RWL groups for accentedness ($\beta = 0.04$, $SE = 0.10$, $t = 0.38$, $p = .709$) or comprehensibility ($\beta = 0.02$, $SE = 0.10$, $t = 0.16$, $p = .872$). These findings together indicated that the effect of sound-spelling consistency was larger for RO and RWL compared to LO, yet learning gains in accentedness and comprehensibility were larger for RWL and LO in comparison to RO, as well as gains for RWL were comparable to gains for LO, regardless of the degree to which target words were consistent in sound-spelling correspondence.

4.5 Discussion

Overall results showed that learners in the RWL recalled a significantly larger number of spoken word forms than learners in the LO mode, which aligned with earlier research findings showing that reading with auditory support is an effective way to build form-meaning connection for L2 words (Brown et al., 2008; Audrey
Bürki et al., 2019; Malone, 2018; Webb & Chang, 2012). Also, learners in the RWL and LO modes produced L2 words in a manner that listeners perceived to be less accented and more comprehensible compared to words produced by learners assigned to the RO mode. Although our understanding of the value of spoken input might have been biased by the findings of earlier studies measuring vocabulary gains in written format, the current study confirmed the important role of spoken input when pronunciation of novel words was measured. Furthermore, sound-spelling consistency of words had a significantly larger impact on accentedness and comprehensibility in the RWL and RO conditions compared to the LO condition. However, no difference was found between RWL and RO, indicating that learners in the RWL mode processed orthographic information to the same extent as did learners in the RO mode. The superiority of RWL in (a) processing orthographic input (vs. LO) and (b) enhancing form-meaning connection (vs. LO) and pronunciation (vs. RO) reveals that learners could process and benefit from two modes of input presented simultaneously at one exposure without being excessively demanded by their limited cognitive capacity, supporting the efficacy of multimodal (or multi-sensory) input for L2 word learning (Mayer, 2014). These findings help re-affirm the pedagogical value of reading-while-listening with the goal of enhancing multifaceted aspects of word knowledge including form-meaning connection and pronunciation. Further interpretation and discussion of the results follow in response to each of the three research questions.

In answer to the first research question, at the immediate posttest, RWL (53%) led to significantly larger gains in recall of spoken forms than LO (48%) but no significant difference was found between RWL and RO (48%) or between LO and RO. At the delayed posttest, no significant differences were found across the three groups, but a similar pattern emerged with RWL (36%) leading to the greatest gains,
while LO (33%) led to the smallest gains. The larger gains for RWL aligned with earlier research findings (Brown et al., 2008; Andreas Bürki, 2010; Audrey Bürki et al., 2019; Malone, 2018; Webb & Chang, 2012), but no clear advantage of RWL over RO or RO over LO appeared to contrast with previous findings of contextualized vocabulary learning: RWL outperforms RO (Webb & Chang, 2012) and RO outperforms LO (Vidal, 2011). Because the target items were presented in isolation in this study, auditory support most likely did not help participants either divide written texts into meaningful chunks of language (Webb & Chang, 2012) or segment connected speech (Vidal, 2011) for improving text processing and comprehension.

In answer to the second research question, the findings that RWL and LO outperformed RO for accentedness and comprehensibility at the immediate posttest suggested that encountering spoken input is beneficial for the development of productive knowledge of spoken forms. The absence of a significant difference between RWL and LO at the immediate and delayed posttests indicated that the orthographic representation did not help learners produce L2 words in a more nativelike or comprehensible manner. One possible reason for this is the crosslinguistic influence of orthographic depth in participants’ L1 (Japanese) and L2 (English) (Erdener & Burnham, 2005). Although Japanese is not an alphabetic language, Japanese native speakers use the L1 romanization system (i.e., *Romaji*) to represent L2 English, which is considered phonologically transparent. For example, <o> corresponds to a single phoneme /o/ in Japanese, can be pronounced differently in English such as /ʌ/ (e.g., *computer*), /ɑ/ (e.g., *hot*), and /oʊ/ (e.g., *token*). In this example, the orthographic information presented in RWL and RO might negatively affect L2 pronunciation accuracy because learners tend to apply L1 grapheme-phoneme conversion rules to recoding L2 orthographic forms into L2 sounds. This
may result in spoken production involving segmental errors such as substituting L1 sounds (e.g., *tobogan*: /təbɑɡən/ → */təbəɡən/) and devoicing L2 consonants (e.g., *chisel*: /tʃɪzəl/ → */tʃɪsəl/). However, unlike previous studies focusing on segments (Audrey Bürki et al., 2019), this study did not show any significantly negative effect of L2 orthography on L2 pronunciation. Perhaps segmental errors resulting from erroneous recoding of L2 written to spoken forms were compensated by accurate pronunciation of the remaining parts so that the errors might not have a significantly negative impact on the listener judgment of the whole word. Approximately one week after the treatment, the advantage of spoken input over written input retained for LO versus RO ($\beta = 0.35$, $p = .010$) but not for RWL versus RO ($\beta = 0.20$, $p = .071$) for accentedness rating, suggesting that orthographic input might prevent learners from reducing the degree of foreign accent in the long term. Given that speech is transient and orthography is permanent, learners in the RWL mode might still have access to targetlike phonological forms of L2 words in their memory immediately after the treatment, allowing them to produce the spoken forms more accurately than learners in the RO mode. However, at the delayed posttest, since the visual orthographic trace of the word remains accessible longer than the phonological information (Solier et al., 2019), learners might have relied on the orthographic representation to recode the written forms into L2 sounds using L1 (and L2) grapheme-phoneme conversion rules. As a result, the recalled spoken forms might have been as heavily accented as the forms elicited from learners receiving only written input. For comprehensibility rating, the benefit of spoken input was not durable either for RWL versus RO ($\beta = 0.23$, $p = .087$) or LO versus RO ($\beta = 0.27$, $p = .109$). Crosslinguistic influence from L1 grapheme-phoneme conversion might have less of an impact on how easily or effortlessly spoken words are understood regardless of the presence or absence of
foreign accent.

In answer to the third research question, words of higher sound-spelling consistency were in general perceived to be less accented and more comprehensible than words of lower consistency, but the extent to which consistency of words impacted listener judgements differed across groups and test times. The results of the immediate posttest showed stronger consistency effects for RWL and RO compared to LO for accentedness and comprehensibility, aligning with our prediction because RWL and RO were the only conditions where learners were exposed to the spellings of words, which likely triggered orthographic recoding. However, we did not expect the consistency effect to become stronger in the LO condition from the immediate to delayed posttest, as evidenced by the finding that the significant between-group differences in the effect of consistency initially observed at the immediate posttest disappeared at the delayed posttest. This result likely occurred because participants in the LO mode, immediately after the treatment, could produce spoken word forms, whether consistent or inconsistent, because the phonological representation of the words remained available in their working memory (considering that the knowledge of words was tested after each block of five items). However, given the transient nature of auditory information, success in recall of L2 forms at the delayed posttest might have been largely dependent on the orthographic representations of the words, developed as a result of the phonology-to-orthography recoding at the exposure phase. The recoding process might have been executed more easily and successfully for words of higher consistency, therefore enabling learners to be more accurate at pronouncing consistent words than inconsistent words. This explanation is speculative and the role of sound-spelling consistency in L2 pronunciation acquisition needs to be further investigated in future studies. Finally, a follow-up analysis of the high-
consistency words confirmed that RO yielded the least learning gains for pronunciation accuracy of the three modes regardless of the degree to which target items were consistent. This finding suggests that exposure to the written form alone may not be sufficient in order for pronunciation of L2 words to be improved, even though these words are highly consistent in their sound-spelling correspondence.

4.6 Implications, Limitations, and Future Directions

The current study provides methodological and pedagogical implications for assessing and teaching L2 vocabulary. First, it is important to remember that test modality determines the efficacy of input modality for vocabulary acquisition. Although this idea is not new (Jelani & Boers, 2018), we provided additional evidence suggesting that when there is a mismatch between input and test modalities, learning gains are likely to be smaller than when there is test-modality congruence. Second, ideally learners should be presented both the written and spoken forms of L2 words together so that knowledge of form-meaning connection and pronunciation can develop simultaneously. In many instructional contexts where spoken input is limited outside the classroom, learners tend to devote most of their time to studying the written forms of words, for example, through reading written texts intensively, using flashcards and word lists, and writing the spellings of words repeatedly. It is important for language teachers to ensure that learners are exposed to the spoken forms of words by teaching strategies such as encouraging learners to listen to the pronunciation of unfamiliar words when looking them up in online dictionaries, choosing vocabulary exercise books or textbooks that include audio support, using vocabulary learning apps that have the function to present the spoken forms of words, watching L2 television, movies, and video clips with captioning options available (e.g., YouTube),
and listening to other audio materials (e.g., songs, podcasts, radio). Second, the superiority of spoken input over written input persisted for even words that are highly consistent in their sound-spelling correspondence, suggesting that if the primary goal of L2 instruction is to enhance the spoken forms of words, spoken input always needs to be introduced even when the pronunciations of new words are easily inferred from the spellings of words.

Finally, exposure to spoken input without orthographic support helps L2 speakers develop targetlike pronunciation of words in the long term, as evidenced by the finding that it was only the LO condition that maintained its advantage over the RO condition for accentedness. Learners’ full attention may need to be drawn to phonological details without being distracted by the presence of orthographic representation if the pedagogical focus is on accent reduction. Given that many scholars have emphasized the importance of setting a realistic goal, such as the development of comprehensible rather than nativelike pronunciation forms (Derwing & Munro, 2015), we argue that RWL may be an optimal method of developing L2 oral skills relative to RO (typical of foreign language education) and LO (characteristic of naturalistic immersion). While written modality enables students to develop and reinforce stronger form-meaning mappings for new words (Vidal, 2011), audio modality can help students reach the minimum threshold for successful understanding (comprehensibility rather than nativelikeness) in the efficient and effective manner (Derwing & Munro, 2015).

The present study has several limitations which should be considered in future studies investigating how mode of input affects L2 word learning. First, participants received only one exposure to each of the target items in the treatment. To increase the ecological validity of the research, different numbers of repetitions should
be explored in future studies. The findings of the current study might then be used as a baseline for comparison. Second, we urge caution with generalization of the findings because they are restricted to a specific population of learners (L1 Japanese, an orthographically transparent language) and target language (L2 English, an orthographically opaque language), and may not apply to other situations where, for example, learners’ L1 is opaque (e.g., English) and their L2 is transparent (e.g., Spanish). Finally, because this study focused on decontextualized learning (i.e., a paired-associate procedure), the extent to which the combination of written and spoken input leads to gains in knowledge of the spoken forms of words in contextualized learning (e.g., reading graded readers with auditory support) remains unknown. More studies are needed to examine how mode of input promotes (or hinders) the leaning of different aspects of word knowledge across various vocabulary learning activities including contextualized and decontextualized learning tasks, in order to determine the effectiveness of multimodal input for L2 vocabulary learning.

4.7 Note

1. Because the initial maximal model did not converge, we reduced the model complexity by removing the by-word intercept-slope correlation and the random slope for the group factor.

4.8 References


Chapter 5: Conclusion

This section first provides a brief summary of the results from the three studies, followed by general implications for L2 vocabulary research and instruction, as well as a discussion of the limitations of the current research along with suggestions for future research.

5.1 Summary of Study 1 (repetition)

The goal of the first study was to examine how repeated exposure to spoken input (1, 3, and 6 exposures) would affect the learning of form-meaning connection (i.e., spoken form recall) and pronunciation (i.e., accentedness, comprehensibility, processing time) of novel words. The study also explored how word-related factors (cognacy, word length, phonotactic probability) would moderate the impact of repetition on L2 word learning. Regarding spoken form recall, results showed that learners receiving six exposures recalled a larger number of spoken word forms than learners receiving one or three exposures, and learners receiving three exposures outperformed those receiving one exposure. Regarding pronunciation measures, the results showed that learners receiving six exposures produced words that were more comprehensible, more nativelike, and less effortful to process than learners receiving one exposure. Similarly, learners receiving three exposures outperformed those receiving one exposure in comprehensibility and processing time. However, instead of three encounters, six encounters were necessary for pronunciation of words to sound more nativelike than one encounter. Regarding word-related factors, the results revealed that the effect of repetition was significantly moderated by different word characteristics (cognateness, word length, phonotactic probability).
5.2 Summary of Study 2 (acoustic variability)

The goal of the second study was to explore how acoustic variability and frequency of exposure (3 exposures with talker variability, 6 exposures with talker variability, 3 exposures without talker variability, 6 exposures without talker variability) would affect the learning of form-meaning connection (i.e., spoken form recall) and pronunciation (i.e., word stress placement accuracy, vowel duration ratio) of novel words. Frequency of exposure appeared to have a larger impact on form-meaning connection as supported by the finding that six exposures with and without talker variability produced significantly larger gains than three exposures with and without talker variability. In contrast, talker variability was more closely related to word stress acquisition given the finding that exposure to multiple voices (3 and 6 talkers) consistently led to larger gains than exposure to a single voice (1 talker) on most of the word stress measures.

5.3 Summary of Study 3 (mode of input)

The goal of the third study was to explore how mode of input (reading-while-listening, reading-only, listening-only) would affect the learning of form-meaning connection (i.e., spoken form recall) and pronunciation (i.e., accentedness, comprehensibility) of novel words. The study also investigated how sound-spelling consistency of words would moderate the impact of mode of input on pronunciation learning. Results showed that learners in the reading-while-listening mode recalled a significantly larger number of spoken word forms than learners in the listening-only mode, and learners in the reading-while-listening and listening-only modes reduced accentedness as well as increased comprehensibility of the recalled words to a greater
extent than learners in the reading-only mode. However, the listening-only mode was
the only condition that outperformed the reading-only mode for accentedness rating at
the delayed posttest. Furthermore, sound-spelling consistency of words had a
significantly larger impact on accentedness and comprehensibility in the conditions
receiving written input compared to the listening-only condition.

5.4 General Implications

The present research provides several implications for L2 vocabulary
assessment and instruction. First, it is important for researchers to be aware that the
extent to which learners show vocabulary gains may be largely dependent on the
choice of test modality (Jelani & Boers, 2018). Earlier studies tended to assess L2
vocabulary learning through written measures of form-meaning connection, asking
participants to complete multiple-choice, word-matching, translation, and yes/no
checklist tasks in written format. Overuse of written measures might have
underestimated the role that spoken input has to play in vocabulary acquisition
(Brown et al., 2008; Vidal, 2011). Findings of the current research suggested that
exposure to spoken input has great potential to facilitate vocabulary learning,
particularly when knowledge of pronunciation is measured. It is important for
researchers to consider test modality in selection of vocabulary tests to evaluate the
efficacy of learning activities or interpret the results of studies appropriately. Second,
findings of this research inform the existing framework of conditions contributing to
vocabulary learning (Webb & Nation, 2017). Webb and Nation suggested that five
conditions had the greatest impact on L2 vocabulary learning, three of which—
repetition, varied encounters, and elaboration—are relevant here. Words encountered
frequently are more likely to be acquired than words encountered once or twice. The
quality of each encounter is enhanced and learning is facilitated when words encountered previously occur subsequently in a new form or context, rather than words being encountered repeatedly in exactly the same form and context. Elaboration arises and improves learning when enrichment of knowledge of a word occurs through encountering more aspects of its form, meaning, and use. The present research adds to Webb and Nation’s framework by revealing that the benefit of repetition that has been observed for learning the written forms of words can be extended to the learning of pronunciation. This research also informs the framework as it reveals that manipulating spoken forms by using different talkers’ voices is one way to create the condition of varied encounters. Finally, exposing learners to both written and spoken forms helps elaborate and enrich knowledge of forms, strengthening form-meaning connection and enhancing pronunciation accuracy. Third, it is important to ensure that learners are exposed to spoken input. This is particularly important for learners in foreign language contexts where the amount of spoken input is limited outside the classroom. One way to increase exposure to spoken input is to introduce extensive viewing (Webb, 2015). Extensive viewing of L2 television programs serves as a useful source of spoken input which optimizes L2 vocabulary learning because it meets the three conditions mentioned above that contribute to vocabulary learning (i.e., repetition, varied encounters, and elaboration). For example, watching related television programs (i.e., narrow viewing) makes it more likely for learners to encounter the spoken forms of words repeatedly than watching unrelated television programs (Rodgers & Webb, 2011). Viewers can also hear the spoken forms produced by different speakers as more than one person (e.g., actors and actresses) typically appear and talk in television programs, serving as a source of varied encounters. Nowadays captioning options are available for television shows, movies, and video
clips (e.g., YouTube), which provides the written forms of words along with presentations of the spoken forms of the same words, leading to elaboration of knowledge of unfamiliar words.

5.5 General Limitations and Future Directions

It is important to note several limitations of the present research. Future research can address the limitations of the current studies with the view of further exploring the role of spoken input and looking for methods to optimize spoken input for L2 vocabulary acquisition. First, it is important to note that the significant effects of input factors (e.g., frequency) observed in this research at the immediate posttest were not often retained until the delayed posttest. This might limit the value of input factors such as frequency of exposure for vocabulary learning in the long term. Further research should consider other conditions that contribute to learning, including distributed learning (e.g., spaced exposures), corrective feedback, and retrieval opportunities (Webb & Nation, 2017). First, findings of this research cannot be directly generalized to contextualized learning (e.g., reading books, watching television). The results suggested the benefits of spoken input but were based on the results of learners studying words out of context (i.e., paired-associate learning). In the paired-associate learning conditions learners might have been able to pay greater attention to the spoken forms than they could do while engaging in meaning-focused input activities. It would be useful for future studies to examine whether the positive effect of spoken input can be replicated when learning words in meaningful contexts, such as viewing television (Webb, 2015) and reading sentences (Webb, 2007). For instance, future studies could compare repeated reading of graded readers with or without talker variability in order to test the effectiveness of acoustic variability in
contextualized learning. Second, because this research was conducted in laboratory settings with many learner-related or external factors controlled, findings may not reflect practical L2 learning situations. One factor worth noting is time on task. The amount of time allowed for reading and hearing each of target words and the intervals between item presentations was the same across all participants. It is possible, for example, that if participants had been allowed more time to memorize the words, they might have subvocally rehearsed the spoken forms repeatedly, which likely improves performance of form recall and may have an influence on pronunciation accuracy. Third, relatively low interrater reliability was found for accentedness and comprehensibility ratings, particularly for Study 1. There are several possible reasons for this issue. First, I adopted a 5-point scale rather than 9 points (Munro & Derwing, 1995; Saito et al., 2016; Suzuki & Kormos, 2020), which might have reduced the data spread in rating performance and yielded restricted correlation between raters. However, the 5-point numerical scale was adopted because in contrast to earlier studies measuring L2 speech at sentence- or discourse-levels (cf. Munro & Derwing, 1995 using a 9-point scale), this study focused on words as a unit of speech samples. Given the relatively limited amount of linguistic information available at the word level, using a large number of scale points might make the rating task excessively challenging. Also, rating performance using a 5-point scale was reported to be as reliable as when a 9-point scale is used (Isaacs & Thomson, 2013). A pilot study also confirmed that the choice of a 5-point rating scale was appropriate for rating word pronunciation in this study. Second, sample size was too large for listeners to rate entire speech samples consistently. Third, the number of raters was small (i.e., 2 raters in Study 1), as evidenced by the finding that when 6 listeners were involved in Study 3, the interrater reliability was considered acceptable (alpha > .70; Larson-Hall, 2010).
A possible solution to this issue for future research might, instead of asking the same listener to rate entire speech samples, divide speech samples into several blocks and assign multiple raters to each of the blocks, which helps lessen the burden on each rater. By allowing a portion of the same speech samples to appear multiple times between blocks, researchers can calculate interrater reliability for overlapping speech samples. A final limitation that needs to be addressed by future studies is that all raters in this research were L1 English speakers. The extent to which pronunciation of words was judged to be comprehensible here may not be generalized to the case of L2 English speakers as listeners. Now that oral communication between non-native speakers is more common than communication between non-native and native speakers (J. Jenkins, 2000; Levis, 2005), future studies need to take non-native speaking raters into consideration in order to increase the ecological validity of the pronunciation measures.

5.6 References


Appendices

Appendix A: Target Words

<table>
<thead>
<tr>
<th>No.</th>
<th>Target word</th>
<th>Phonetic symbol</th>
<th>Number of syllables</th>
<th>Location of main stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>abalone</td>
<td>æbəlóoni</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>acorn</td>
<td>éikɔːrn</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>armadillo</td>
<td>aːrmadiloo</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>binoculars</td>
<td>bainúkjolærz</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>caramel</td>
<td>kærəmal</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>carousel</td>
<td>kærəsel</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>catapult</td>
<td>kætəpalt</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>celery</td>
<td>séləri</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>chameleon</td>
<td>kæmɪliən</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>chandelier</td>
<td>fændəliər</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>chisel</td>
<td>tjɪzɔl</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>cicada</td>
<td>stɪkɛdə</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>clover</td>
<td>klóʊvɔr</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>crayon</td>
<td>krέən</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>croissant</td>
<td>kræsənt</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>escalator</td>
<td>ēskəleɪtɔr</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>ladle</td>
<td>lɛɪdəl</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>loquat</td>
<td>lóokwʌt</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>lotus</td>
<td>lʊötəs</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>maracas</td>
<td>mərəkəs</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>21</td>
<td>marshmallow</td>
<td>mʊːrʃmeloʊ</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>mermaid</td>
<td>mɔːrmɛrd</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>pacifier</td>
<td>pæsəfaɪər</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>parakeet</td>
<td>pærakiːt</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>persimmon</td>
<td>pərsɪmən</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>26</td>
<td>podium</td>
<td>póʊdʊm</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>porcupine</td>
<td>póːrkjʊpɪn</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>protractor</td>
<td>prətrɛktər</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>29</td>
<td>raccoon</td>
<td>rækʊən</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>30</td>
<td>raisin</td>
<td>rɛɪzən</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>razor</td>
<td>rɛɪzər</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>No.</td>
<td>Word</td>
<td>Phonetic Symbol</td>
<td>Stressed Syllable(s)</td>
<td>Unstressed Vowels</td>
</tr>
<tr>
<td>-----</td>
<td>---------</td>
<td>-----------------</td>
<td>----------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>32</td>
<td>spatula</td>
<td>spætʃəla</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>33</td>
<td>strainer</td>
<td>stréɪnər</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>34</td>
<td>syringe</td>
<td>sɑːrɪndʒ</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>35</td>
<td>tadpole</td>
<td>tædpoʊl</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>36</td>
<td>toboggan</td>
<td>təbʊɡən</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>37</td>
<td>toupee</td>
<td>tuːpɛi</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>38</td>
<td>treadmill</td>
<td>tredmɪl</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>39</td>
<td>walrus</td>
<td>wɔːlras</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>40</td>
<td>xylophone</td>
<td>zæləfʊn</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

*Note. Target words are presented with phonetic symbols where stressed syllables are marked with an acute accent and unstressed vowels are marked in bold. When two-syllable words contained vowels with secondary stress (e.g., *mermaid, tadpole*), they were considered unstressed and the duration of such vowels were compared to that of vowels with primary stress. When tense vowels appeared at the end of words (e.g., *celery*), they tended to be substantially lengthened and were not measured as unstressed vowels.*
Appendix B: Visual Prompts

- Acorn
- Ladle
- Glade
- Mermaid
- Walrus
- Tadpole
Appendix C: Ethics Approval

Dear Dr. Stuart Webb,

The Western University Non-Medical Research Ethics Board (NMREB) has reviewed and approved the WREM application form for the above mentioned study, as of the date noted above. NMREB approval for this study remains valid until the expiry date noted above, conditional to timely submission and acceptance of NMREB Continuing Ethics Review.

This research study is to be conducted by the investigator noted above. All other required institutional approvals must also be obtained prior to the conduct of the study.

Documents Approved:

<table>
<thead>
<tr>
<th>Document Name</th>
<th>Document Type</th>
<th>Document Date</th>
<th>Document Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detaching/Form (English speakers baseline)</td>
<td>Detaching document</td>
<td>28/Mar/2019</td>
<td>3</td>
</tr>
<tr>
<td>Detaching/Form (Japanese learners of English)</td>
<td>Detaching document</td>
<td>28/Mar/2019</td>
<td>3</td>
</tr>
<tr>
<td>Detaching/Form (Japanese learners of English) (translated in Japanese)</td>
<td>Detaching document</td>
<td>28/Mar/2019</td>
<td>3</td>
</tr>
<tr>
<td>Flyer for Japanese Participants</td>
<td>Recruitment Materials</td>
<td>19/Mar/2019</td>
<td>2</td>
</tr>
<tr>
<td>In Class Recruitment</td>
<td>Oral Script</td>
<td>19/Mar/2019</td>
<td>1</td>
</tr>
<tr>
<td>In Class Recruitment (translated in Japanese)</td>
<td>Oral Script</td>
<td>19/Mar/2019</td>
<td>1</td>
</tr>
<tr>
<td>Language Background Questionnaire (in English)</td>
<td>Paper Survey</td>
<td>19/Mar/2019</td>
<td>2</td>
</tr>
<tr>
<td>Language Background Questionnaire (in Japanese)</td>
<td>Paper Survey</td>
<td>19/Mar/2019</td>
<td>2</td>
</tr>
<tr>
<td>Letter of information and consent (English language teachers)</td>
<td>Written Consent/Assent</td>
<td>28/Mar/2019</td>
<td>3</td>
</tr>
<tr>
<td>Letter of information and consent (English speakers baseline)</td>
<td>Written Consent/Assent</td>
<td>20/Mar/2019</td>
<td>3</td>
</tr>
<tr>
<td>Letter of information and consent (English speakers registered)</td>
<td>Written Consent/Assent</td>
<td>20/Mar/2019</td>
<td>3</td>
</tr>
<tr>
<td>Letter of information and consent (Japanese learners of English)</td>
<td>Written Consent/Assent</td>
<td>20/Mar/2019</td>
<td>3</td>
</tr>
<tr>
<td>Letter of information and consent (Japanese learners of English) (translated in Japanese)</td>
<td>Written Consent/Assent</td>
<td>20/Mar/2019</td>
<td>3</td>
</tr>
<tr>
<td>List of target words</td>
<td>Other Data Collection</td>
<td>15/Feb/2019</td>
<td>1</td>
</tr>
<tr>
<td>Recruitment_Brief (English language teachers)</td>
<td>Recruitment Materials</td>
<td>15/Feb/2019</td>
<td>1</td>
</tr>
<tr>
<td>Recruitment_Brief (English speakers baseline)</td>
<td>Recruitment Materials</td>
<td>15/Feb/2019</td>
<td>1</td>
</tr>
<tr>
<td>Recruitment_Brief (English speakers registered)</td>
<td>Recruitment Materials</td>
<td>15/Feb/2019</td>
<td>1</td>
</tr>
<tr>
<td>Vocabulary Levels Test</td>
<td>Paper Survey</td>
<td>15/Feb/2019</td>
<td>1</td>
</tr>
</tbody>
</table>
No deviations from, or changes to the protocol should be initiated without prior written approval from the NMPREB, except when necessary to eliminate immediate hazard(s) to study participants or when the change(s) involves only administrative or logistical aspects of the trial.

The Western University NMPREB operates in compliance with the Tri-Council Policy Statement Ethical Conduct for Research Involving Humans (TCPS2), the Ontario Personal Health Information Protection Act (PHIPA, 2004), and the applicable laws and regulations of Ontario. Members of the NMPREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB. The NMPREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000941.

Please do not hesitate to contact us if you have any questions.

Sincerely,

Kelly Paterson, Research Ethics Officer on behalf of Dr. Randall Graham, NMPREB Chair

Note: This correspondence includes an electronic signature (validation and approval via an online system that is compliant with all regulations).