

Research Article

THE EFFECTS OF TALKER VARIABILITY AND FREQUENCY OF EXPOSURE ON THE ACQUISITION OF SPOKEN WORD KNOWLEDGE

Takumi Uchihara *

Waseda University

Stuart Webb 

University of Western Ontario

Kazuya Saito 

University College London


Pavel Trofimovich 

Concordia University

Abstract

Eighty Japanese learners of English as a foreign language encountered 40 target words in one of four experimental conditions (three encounters, six encounters, three encounters with talker variability, and six encounters with talker variability). A picture-naming test was conducted three times (pretest, immediate posttest, and delayed posttest) and elicited speech samples were scored in terms of form-meaning connection (spoken form recall) and word stress accuracy (stress placement accuracy and vowel duration ratio). Results suggested that frequency of exposure consistently promoted the recall of spoken forms, whereas talker variability was more closely related to the enhancement of word stress accuracy. These findings shed light on how input

We would like to express thanks to the editor and anonymous *SSLA* reviewers for reading earlier versions of the manuscript and offering constructive feedback. We also thank Emi Iwaizumi, Shuhei Kudo, Dr. Akifumi Yanagisawa, and Dr. Tetsuo Harada for their help for data collection and analyses. This study was supported by the *Language Learning* Dissertation Grant Program (grant number: R5370A13).

 The experiment in this article earned an Open Data badge for transparent practices. The materials are available at osf.io/2rkpd

*Correspondence concerning this article should be addressed to Takumi Uchihara, Faculty of Science and Engineering, Waseda University, 3-4-1, Okubo, Shinjuku, Tokyo 169-8555, Japan. E-mail: tuchihar@aoni.waseda.jp

© The Author(s), 2021. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

quantity (frequency) and quality (variability) affect different stages of lexical development and provide implications for vocabulary teaching.

INTRODUCTION

Exposure to input through reading or listening is considered an important driving force for second language (L2) vocabulary acquisition (Ellis, 2002; Krashen, 1989; Nation, 2013; Webb & Nation, 2017). Learners develop word knowledge incrementally through repeatedly encountering new words within contexts (Uchihara et al., 2019) or in isolation (Rice & Tokowicz, 2020). However, the quantity of input (i.e., frequency of exposure) is not a sufficient condition for successful learning. The quality of input also matters for enriching and consolidating word knowledge through seeing and hearing words in different contexts (e.g., more vs. less informative contexts) and forms (e.g., derived and inflected forms) (Webb & Nation, 2017). Talker variability—characterized by differences in linguistic and nonlinguistic properties between and within talkers (e.g., voices, pitch height, speaking rate, speaking style, and loudness)—is one of the useful sources of variability that enhances input quality and facilitates vocabulary learning. Talker variability facilitates different aspects of L2 acquisition, including lexical knowledge of forms and form-meaning mappings (Barcroft & Sommers, 2014), recognition/perception accuracy of temporal and spectral features (Logan et al., 1991), and production accuracy (Bradlow et al., 1997). The benefits of talker variability can be attributed to talker-specific characteristics (i.e., indexical information) available in different voices, such that processing talker variability helps create 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 learning different aspects of word knowledge, little is known about how it affects learning the pronunciation of new words. Instead, researchers in this area have generally focused on pronunciation at a segmental level (e.g., vowel contrast /e/ vs. /ɛ/ in Kartushina & Martin, 2019). Previous studies of L2 word learning and acoustic variability have also tended to rely on measures of word form learning and form-meaning mapping accuracy (using picture-to-L2-recall test) without measuring the quality of the resulting spoken forms (e.g., pronunciation accuracy). 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 communication. Also, the extent to which different levels of frequency moderate the effects of variability 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 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 spoken word knowledge targeting Japanese learners studying English as a foreign language (EFL). Building upon Jiang’s (2000) psycholinguistic model of L2 lexical representation and frameworks of phonological refinement (Saito, 2018), spoken word knowledge was defined and measured in two ways: form-meaning connection (spoken form recall) and spoken form accuracy (pronunciation).

ACQUISITION OF SPOKEN WORD KNOWLEDGE

According to psycholinguistic models of bilingual lexicon (e.g., Jiang, 2000, 2002, 2004) and frameworks of sound and word learning (e.g., Saito, 2018; Tyler, 2019; Werker, 2018), the development of spoken word knowledge is assumed to take place at multiple stages. When they encounter new words, learners initially encode and retain novel word forms (e.g., composition and sequence of phonemes in a word) and then map the retained L2 forms onto meanings. One notable difference between initial word learning in learners' first language (L1) versus their L2 is that L2 learners can draw on their existing knowledge of L1 meanings (i.e., L1 translations) during the mapping process, while L1 learners have to develop semantic components of new words (Jiang, 2000). However, even if the initial mapping of word forms onto word meanings is successful, learners do not always develop targetlike meanings and spoken forms that are accurate in their linguistic detail. In the long run, through exposure to greater quantity and quality of input, learners can enhance not only the strength of form-meaning connections (i.e., mappings of word forms onto meanings) but also the accuracy of various aspects of word knowledge, which includes semantics (e.g., L1 translations → targetlike/accurate meaning; Jiang, 2000), morphosyntax (e.g., inaccurate or missing grammar → targetlike/accurate grammar; Jiang, 2004), and phonology (e.g., L1 influenced phonology → targetlike/accurate phonology; Saito, 2018).

In L2 speech literature, it has been argued that L2 learners' sound and word learning takes place sequentially, although such developmental stages overlap to some degree (e.g., Tyler, 2019, for PAM-L2). Bundgaard-Nielsen et al. (2011) found that L2 learners beyond a certain threshold of vocabulary size (6,000–7,000 word families) demonstrated relatively superior L2 vowel perception, which suggests that development of form-meaning connection precedes refinement of L2 phonology (for the link between vocabulary growth and vowel production, see Bundgaard-Nielsen et al., 2012). L2 learners first need to develop large vocabularies before confronting many confusing minimal pairs (e.g., “beat” vs. “bit,” “adapt” vs. “adopt,” “cap” vs. “cab”), shifting their attention to fine-grained phonemic discrimination (e.g., [i] vs. [ɪ], [æ] vs. [a], [p] vs. [b]), and acquiring phonologically refined representations of L2 words (Saito, 2018). Although most studies have concerned segmental features (vowels and consonants), suprasegmental features, such as fluency and stress, are also subject to refinement processes (Saito, 2018). For word stress, for instance, Japanese learners of English initially rely on acoustic cues available in their L1, including pitch height (and vowel duration to some degree), and they might gradually become more able to use other cues absent in their L1 inventory (vowel quality) for perceiving and producing word stress (Lee et al., 2006).

To provide a nuanced understanding of how L2 learners acquire spoken vocabulary, it is important to distinguish the processes by which learners establish a form-meaning connection (e.g., through form specification, meaning specification, and form-meaning mapping) from those by which learners engage in form refinement. Learners with fully specified and established representations of L2 forms (i.e., learners who have established form-meaning connections) would be able to retrieve word forms accurately and efficiently (Elgort et al., 2018; Pellicer-Sánchez, 2016; Rice & Tokowicz, 2020). Comparatively, learners with partially specified representations might fail to recall some phonemes or might retrieve them in a wrong sequence even if all phonemes are present. However, regardless of whether L2 forms are represented fully or partially, such

representations may not be sufficiently refined to the extent that they are considered targetlike. In real-life situations, for instance, it is not rare to find learners who can produce all phonemes of a word in a correct sequence but whose word productions are heavily accented and hard to understand. In this regard, phonological refinement is not equivalent to form specification (or novel word form learning). Refinement concerns an approximation toward targetlike accuracy in word pronunciation, whereas form specification involves committing L2 formal/structural information to memory regardless of the degree to which the phonological representations of the words are targetlike.

Although the learning of novel word forms is achievable (Pavia et al., 2019), phonological refinement is less likely to occur especially for EFL learners whose learning experiences are largely shaped by interaction with nonnative speakers with low proficiency (vs. native or near-native speakers) and by a disproportionately large amount of exposure to the orthographic forms of words (vs. phonological word forms) (see Tyler, 2019, for a review of the conditions promoting and inhibiting L2 phonological refinement). Critically, as reviewed in the preceding text, the scope of lexical development is not restricted to building form-meaning connections but also includes formal (and linguistic) refinement/enhancement with continued L2 exposure (see Jiang, 2002, for semantic refinement; Jiang, 2004, for morphosyntactic refinement). Therefore, the current study focused on two aspects of L2 vocabulary learning—(a) establishment of an initial form-meaning connection (acquiring L2 meaning but using L1 phonology) and (b) phonological refinement (mastering both L2 meaning and L2 phonology)—and examined the effects of input quantity (i.e., frequency of exposure) and quality (i.e., talker variability) on these two aspects of L2 word knowledge.

ACOUSTIC VARIABILITY AND L2 PRONUNCIATION LEARNING

In the field of L2 speech research, acoustic variability has been considered crucial for the recognition and production of individual L2 phonetic categories (sounds). Learners need ample exposure to a wide range of exemplars in different phonetic, talker, and task contexts so they can gradually become 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, a procedure that 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 high-variability training show significant improvement (10–20% gain) in identifying various segmental targets spanning vowels (Lambacher et al., 2005), liquids (Bradlow et al., 1997), stops (Flege, 1995), and Japanese geminate consonants (Hirata, 2004). 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/) following high-variability training led to improvement in their production accuracy and intelligibility for the same sounds.

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 in their production of 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 the 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 beginner-level L1 English learners studying L1 Mandarin tones after they had received 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.

ACOUSTIC VARIABILITY AND L2 VOCABULARY LEARNING

Research has consistently shown a positive effect of acoustic variability on L2 vocabulary learning using cued recall measures (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 (six occurrences produced by six different talkers), moderate variability (six occurrences produced by three different talkers repeating each word twice), and no variability (six occurrences of all words produced by a single talker). The results of cued recall tests suggested that the words learned with high variability were recalled significantly more accurately compared to those learned with moderate variability, and the words learned with moderate variability were recalled significantly more accurately compared to those learned with no variability. Barcroft and Sommers concluded that acoustic variability is beneficial in developing knowledge of form-meaning connections (i.e., novel form learning and mapping) 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 the target language (Lithuanian) heard eight repetitions of six new words produced by a single talker (no-variability condition) or eight talkers (high-variability condition), with posttraining performance measured through picture recognition (picture-to-L2 matching) and form recall (picture-to-L2 recall) tests. The results were consistent with those reported by Barcroft and Sommers (2005) showing beneficial effects of high variability for adult

learners' form recall (but not on picture recognition). However, no such benefit was observed for groups of children (7- to 8-year-olds and 10- to 11-year-olds), either in picture recognition or form recall.

FREQUENCY OF EXPOSURE AND L2 WORD LEARNING

Frequency of exposure is a key factor contributing to L2 vocabulary learning (e.g., Horst et al., 1998; Nakata, 2017; Peters, 2019; Rott, 1999; Uchihara et al., 2019; Vidal, 2011; Webb, 2007; for review, see Rice & Tokowicz, 2020; 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. In contextualized learning activities (e.g., learning through reading graded readers), greater numbers of encounters with target words seem necessary, with 6 (Rott, 1999), 8 (Horst et al., 1998; Pellicer-Sánchez, 2016), 10 (Webb, 2007), or even more than 20 encounters (Waring & Takaki, 2003) required for learning. 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 three aspects of knowledge (spoken form recognition, part of speech, meaning recognition) for words encountered 3, 7, 11, and 15 times in oral passages. The obtained learning gains were moderate, and frequency appeared to have less impact on vocabulary learning gains compared to the findings of reading studies. Other listening studies further suggest that frequency has a positive but moderate effect on word learning through listening to songs (Pavia et al., 2019) and listening to academic lectures (Vidal, 2011). To the best of our knowledge, no prior research has examined the effects of frequency on learners' productive knowledge of spoken word forms.

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, earlier studies have examined the effects of talker variability on how accurately novel word forms are encoded, retained, and mapped onto meanings (Barcroft & Sommers, 2005; Sinkeviciute et al., 2019) but not on how accurately the resulting spoken forms are pronounced. This is an important gap in research that needs to be filled as learners' ability to pronounce words accurately is essential for successful communication (Saito et al., 2017). Second, the relative contributions of frequency of exposure and acoustic variability to L2 lexical acquisition remain under-explored. 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 (six encounters in Barcroft & Sommers, 2005; eight 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 is conceptualized within the two-step process of the acquisition of spoken word knowledge (Jiang, 2000; Saito, 2013, 2018)—form-meaning connection (i.e., novel form learning and mapping) followed by phonological refinement (i.e., approximation toward targetlike accuracy). Therefore, this framework allows for examining 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 (Field, 2005), sentence stress (Hahn, 2004), and temporal fluency (Suzuki & Kormos, 2020).

The present study therefore focused on word stress accuracy in L2 English, defined as the ability to pronounce stressed syllables with higher pitch, greater amplitude, and longer duration while deemphasizing unstressed syllables (Lee et al., 2006). Although word stress is one of numerous prosodic features predicting overall pronunciation proficiency, this research focuses on this given aspect for empirical and pedagogical reasons. In English, because stressed syllables serve as vital cues to word segmentation (Cutler, 1990) and word identification (Grosjean & Gee, 1987), lexical stress errors (e.g., misplacement, missing stress) are likely to significantly reduce L2 speech comprehensibility (Isaacs & Trofimovich, 2012) and intelligibility (Field, 2005). Furthermore, lexical stress is one of few pronunciation aspects receiving attention among L2 vocabulary scholars (Nation, 2013), suggesting that it might have a central place in vocabulary teaching (Field, 2005; Murphy & Kandil, 2004).

Hence, this study was designed to examine to what extent each factor—frequency of exposure (three vs. six encounters without talker variability) or talker variability (three vs. six encounters with talker variability)—enhances knowledge of pronunciation (word stress accuracy) and knowledge of form-meaning connection (spoken form recall)¹ for L1 Japanese participants learning novel L2 words. This study was guided by the following research questions and predictions:

1. To what extent does frequency of exposure (three vs. six encounters) and talker variability (single vs. multiple voices [three or six voices]) affect form-meaning connection of L2 words (measured through spoken form recall)?
2. To what extent does frequency of exposure (three vs. six encounters) and talker variability (single vs. multiple voices [three or six voices]) affect phonological refinement of L2 words (assessed through word stress accuracy)?

It was predicted that L2 learners would be more likely to learn novel word forms and establish form-meaning mappings after L2 words are encountered six times than three

times. At this initial learning stage, however, learners would most likely rely on L1 phonology (producing word stress inaccurately), especially if the exposure is low in acoustic variability (one or three 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 (six voices) so that they can attain targetlike spoken word knowledge.

METHOD

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 talker variability (E3 + TV), six encounters with talker variability (E6 + TV), three encounters without talker variability (E3), and six encounters without talker variability (E6). During the treatment, participants were instructed to learn 40 low-frequency English words by 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. Although originally motivated by Barcroft and Sommers (2005), this study adopted several different methodological procedures, such as controlling talker intelligibility (cf. talker rotation method used by Barcroft and Sommers).

PARTICIPANTS

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

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 (three females, three 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 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, 2014), we chose to optimize the effectiveness of learning procedures by sequencing presentations of six speakers in the order of intelligibility (see Webb, 2008, for a similar approach to contextual informativeness in vocabulary learning). First, 10 out of 40 items produced by each of the six speakers were selected randomly (60 samples = 10 items × 6 speakers). An additional group of native English listeners ($n = 8$) were recruited to listen to these 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).

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, 2014; Sinkeviciute et al., 2019). 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) 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.

TABLE 1. Sequence of talker presentations for four experimental groups

Group	Repetition					
	1	2	3	4	5	6
E3	Talker 1	Talker 1	Talker 1			
E3 + TV	Talker 1	Talker 2	Talker 3			
E6	Talker 1	Talker 1	Talker 1	Talker 1	Talker 1	Talker 1
E6 + TV	Talker 1	Talker 2	Talker 3	Talker 4	Talker 5	Talker 6

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 + TV, E6 + TV) 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 versus six times: E3 and E3 + TV listened to 120 items (40 items \times 3 encounters), and E6 and E6 + TV listened to 240 items (40 items \times 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 + TV and E6 + TV, 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 produce the words corresponding twice orally 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 (visual stimuli are available through the Open Science Framework at <https://osf.io/2rkpd/>).

PROCEDURE

The experiment was conducted over two sessions on two different days. In the first session, 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. In the second session, approximately one week ($M = 6.6$ days) after the first session, participants took a surprise delayed posttest and filled out language background questionnaires.² 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 + TV and E6 + TV 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.

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 /streɪnər/ → /sUtɹeɪnər/ 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; Saito et al., 2017), performance of word stress was categorized in terms of placement accuracy: (a) primary stress is correctly placed on the right syllable (e.g., *TREAD*mill), (b) primary stress is misplaced (e.g., *tread*MILL), 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 intercoder agreements for spoken form recall ($\kappa = .963$) and stress placement accuracy ($\kappa = .967$). 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 structure, 20 samples were not analyzed for stress placement accuracy.

In addition to the measure of stress placement accuracy, vowel duration ratio (i.e., duration ratio of unstressed to stressed vowels) was also measured as one of the acoustic properties important to the perception of lexical stress (Beckman, 1986; Lee et al., 2006; Parlak & Ziegler, 2017; Trofimovich & Baker, 2006). 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 (e.g., vowel quality reduction) 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).³ Therefore, 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 structure, 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. Lower scores for vowel duration ratio indicate the ability to successfully reduce the duration of unstressed vowels relative to the duration of stressed vowels. Finally, five English native speakers (three females, two 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 ($r = -.275, p = .014$), immediate posttest ($r = -.625, p < .001$), and delayed posttest ($r = -.396, p < .001$), such that more accurate stress placement was associated with a smaller vowel duration ratio (i.e., more English-like duration of unstress vowels), which supported the validity of the two pronunciation measures.

DATA ANALYSIS

To answer the first and second research questions regarding the effects of input quantity (i.e., frequency of exposure) and quality (i.e., talker variability) on the two stages of spoken word acquisition (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 + TV, E6 + TV) and time as a within-participants variable (pretest, immediate posttest, delayed posttest) for each of the three dependent measures (spoken form recall, stress placement accuracy, vowel duration ratio). A follow-up analysis (Bonferroni-corrected pairwise post hoc comparisons) was conducted to confirm where the differences occurred between groups. 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 the 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). Descriptive analyses for the three dependent measures and detailed results of post hoc comparison tests can be found in the Supplementary Material.

RESULTS

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 \times 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

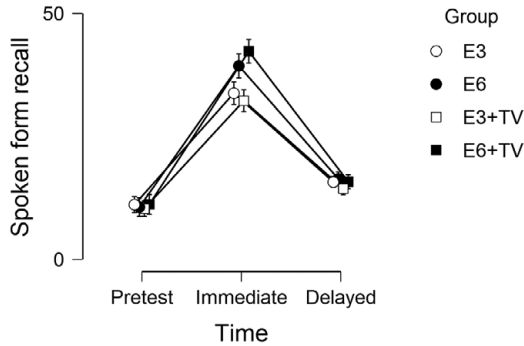


FIGURE 1. Group estimated marginal means for spoken form recall over time. Error bars represent 95% confidence intervals around the mean.

immediate posttest, $F(3, 41.4) = 11.94, p < .001, \eta_p^2 = 0.46$. Bonferroni-corrected pairwise post hoc comparisons showed that E6 + TV significantly outperformed E3 ($d = 1.29$) and E3 + TV ($d = 1.69$) but did not significantly outperform E6 ($d = 0.32$). E6 had significantly higher scores than E3 ($d = 0.91$) and E3 + TV ($d = 1.20$). No significant difference was found between E3 and E3 + TV ($d = 0.18$). 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.

STRESS PLACEMENT ACCURACY

The analysis showed significant effects for Time, $F(1.90, 150.28) = 563, p < .001, \eta_p^2 = 0.88$, but not for Group, $F(3, 76) = 0.23, p = .878, \eta_p^2 = 0.01$; however, there was a significant Time \times Group interaction, $F(5.78, 146.44) = 2.58, p = .022, \eta_p^2 = 0.09$ (see Figure 2). Post hoc comparisons showed no significant effects for Group at the pretest, $F(3, 76) = 0.93, p = .430, \eta_p^2 = 0.04$, or the delayed posttest, $F(3, 76) = 0.34, p = .797, \eta_p^2 = 0.01$. A significant effect was found for Group at the immediate posttest, $F(3, 41.5) = 3.54, p = .023, \eta_p^2 = 0.20$. Bonferroni-corrected post hoc tests revealed that E6 + TV significantly outperformed E3 ($d = 0.97$), but that there were no significant differences between E6 + TV and the remaining two groups: E6 ($d = 0.48$) or E3 + TV ($d = 0.10$). E3 + TV had significantly higher scores than E3 ($d = 0.91$). No significant differences were found between E6 and E3 ($d = 0.65$) or E3 + TV ($d = 0.39$). 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.

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 \times Group interaction, $F(5.48, 138.88) = 4.19, p < .001, \eta_p^2 = 0.14$ (see Figure 3). Post hoc

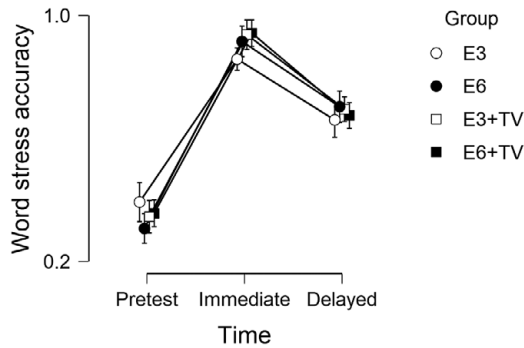


FIGURE 2. Group estimated marginal means for stress placement accuracy over time. Error bars represent 95% confidence intervals around the mean.

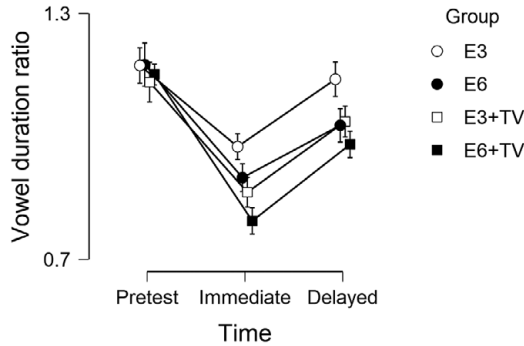


FIGURE 3. Group estimated marginal means for vowel duration ratio over time. Error bars represent 95% confidence intervals around the mean.

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 + TV had significantly lower (more targetlike) scores than E6 ($d = 1.03$) and E3 ($d = 1.92$) but the difference between E6 + TV and E3 + TV did not reach statistical significance ($d = 0.59$). E3 + TV also had significantly lower scores than E3 ($d = 0.95$) but the difference between E3 + TV and E6 was not significant ($d = 0.29$). There was no significant difference between E3 and E6 ($d = 0.75$). 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 + TV showed significantly lower (more targetlike) vowel duration ratios than E3 ($d = 1.32$) but did not show any significant difference when compared with E6 ($d = 0.39$) and E3 + TV ($d = 0.45$). E6 had significantly lower scores than E3 ($d = 0.92$) but no significant difference was found between E6 and E3 + TV. Lastly, E3 + TV appeared to show lower vowel duration ratios than E3 as the difference approached statistical significance

TABLE 2. Summary of pairwise comparisons between groups showing significant differences

	Immediate posttest	Delayed posttest
Spoken form recall	E6 + TV > E3, E3 + TV E6 > E3, E3 + TV	No difference
Stress placement accuracy	E6 + TV > E3 E3 + TV > E3	No difference
Vowel duration ratio	E6 + TV < E3, E6 E3 + TV < E3 NS < E3, E6, E3 + TV	E6 + TV < E3, E6 < E3 E3 + TV < E3* NS < E3, E6, E3 + TV, E6 + TV

Note: An asterisk indicates that the difference between E3 + TV and E3 was marginally significant ($p = .058$). NS = native speaker. Lower scores for vowel duration ratio indicate more targetlike production of word stress.

($p = .058$, $d = 0.82$). Unlike the immediate posttest, where high variability seemed to support 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 + TV at the immediate posttest, $t(38) = 1.71$, $p = .009$, 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.⁴

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 (i.e., a composite of novel form learning and form-meaning mapping, measured using spoken form recall) and phonological refinement (i.e., approximating targetlike pronunciation, measured using word stress placement accuracy and vowel duration ratio). According to the two-step process of L2 lexical acquisition (Jiang, 2000; Saito, 2018), the results (summarized in Table 2) 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 + TV) produced significantly larger gains than three encounters (E3 and E3 + TV) on the 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 + TV and E6 + TV) 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).

EFFECTS OF FREQUENCY AND VARIABILITY ON SPOKEN L2 WORD ACQUISITION

The findings of this study expand on earlier studies of L2 lexical acquisition. Both input quantity (frequency of exposure) and quality (talker variability) improve word learning,

but the degree of such facilitative effects differs across different stages of lexical development (Jiang, 2000).⁵ First, learners hearing target words six times (E6 and E6 + TV) recalled a greater number of spoken forms than those hearing the words three times (E3 and E3 + TV). In contrast, no clear advantage of E6 over either E3 or E3 + TV was observed for pronunciation measures (except for E6 outperforming E3 for vowel duration ratio at the delayed posttest). These findings suggest that frequency of encounters with spoken word forms exerts a larger impact at the initial stage of learning, promoting the development of form-meaning connections (i.e., encoding and retaining novel word forms and mapping them onto 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 /sUtreinær/). It is therefore possible to suggest that although repeated exposure facilitates the development of form-meaning connections, 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 process of developing formal and mapping components is completed, additional repetitions may not motivate learners to attend to specific form details, unless they receive explicit phonetic instruction to do so (Saito, 2013) or encounter communication breakdowns due to their misunderstanding and mispronunciation of phonologically similar words (Saito et al., 2020). However, benefits of talker variability for pronunciation measures reveal that talker 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 phonological representations of L2 words.

SPOKEN FORM RECALL

Results of spoken form recall showed that six encounters with spoken word forms (E6 and E6 + TV) produced larger learning gains than three encounters (E3 and E3 + TV). These findings are consistent with the previous L2 vocabulary literature revealing positive effects of frequency on learning gains (e.g., Nakata, 2017; Rice & Tokowicz, 2020; Rott, 1999; Uchihara et al., 2019; Webb & Nation, 2017). Despite the slightly higher recall rate for E6 + TV ($M = 35.6$) 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 make a simple comparison difficult (e.g., between-participants vs. within-participants design; experienced vs. inexperienced learners). However, a key difference was that the current study did not adopt the approach of earlier studies (e.g., Barcroft & Sommers,

2005) for controlling the influence of talkers' characteristics (i.e., Barcroft and Sommers rotated different talkers used in no-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.

Another notable finding was related to the absence of a variability advantage for learners who encountered target words three times (E3 vs. E3 + TV). As highlighted in the preceding text, this finding might be attributed to the methodological difference (i.e., lack of rotation of talkers across conditions), indicating that listening to the most intelligible talker three times is as effective as listening to different talkers who may vary in their intelligibility (however slightly). Yet, it should be noted that the mean score of the E3 + TV condition was the lowest of the four conditions ($M = 29.6$), in contrast to the findings for E6 + TV and E6, where E6 + TV ($M = 35.6$) appeared to outperform E6 ($M = 34.5$), although the difference did not reach statistical significance. The tendency toward a negative influence of talker variability in E3 + TV indicates the possibility that L2 learners may need sufficient encounters without talker variability to create initial form-meaning mappings. Only after novel word forms are learned and 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 talker variability on perceptually ready vs. unready learners; cf. Saito et al., 2020, for learners' readiness for L2 phonological refinement).

STRESS PLACEMENT ACCURACY AND VOWEL DURATION RATIO

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 high-variability input for improving L2 pronunciation (Thomson, 2018), revealing that acoustic variability helps enhance pronunciation accuracy not only at the level of segments (e.g., Kartushina & Martin, 2019) but also at the level of individual words. Specifically, regarding stress placement accuracy, the E3 + TV condition produced significantly larger gains than the E3 condition at the immediate posttest. The absence of the expected advantage of high-variability input for E6 + TV over E6 might be due to a ceiling effect for E6 ($M = 94\%$ accuracy). For vowel duration ratio, the two variability conditions (E6 + TV and E3 + TV) outperformed corresponding conditions without talker variability (E6 and E3) at the immediate posttest. No significant difference between E6 + TV and the native speakers' baseline indicated that the performance of L2 learners approximated target-like performance.

The advantage of E3 + TV 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. $M_{E3} = 30.4$). This result suggests that three encounters might not be sufficient for the benefit of talker variability to manifest itself in the ability to recall newly learned spoken forms. During the first few experiences with spoken words, learners' attention is likely directed to the process of creating an initial form-meaning mapping, and additional variability is generally irrelevant to this attention-

demanding process (Barcroft, 2015). Therefore, increased variability was not particularly beneficial for the creation of a basic form-meaning connection (hence, E3 + TV did not outperform E3 on form recall). Nevertheless, 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 refined to a greater extent than those that they produced in a single-talker learning situation, leading to the benefit for pronunciation-specific measures. 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 + TV vs. E3, $d = 0.92$ for E6 vs. E3).

IMPLICATIONS, LIMITATIONS, AND FUTURE WORK

In the L2 vocabulary acquisition literature, input quantity and quality have been extensively researched and suggested as a driving force for vocabulary learning (Webb & Nation, 2017). Adopting the models of sound and word learning (Jiang, 2000; Tyler, 2019), this study sheds light on the value of input quantity (defined here as frequency of word occurrence) and input quality (operationalized through talker variability) in the two different stages of L2 spoken word knowledge. That is, frequency of exposure had a robust effect on the first stage of lexical development (i.e., encoding and retaining novel forms, and mapping forms onto meanings; Jiang, 2000), while talker variability further enhanced the subsequent stage of learning (i.e., phonological refinement toward targetlike production of L2 words; Saito, 2018).

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. Critically, listening to different talkers significantly improved the use of lexical stress compared to listening to the most intelligible talker multiple times. This highlights the value of integrating talker variability into L2 vocabulary learning. 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 (see Nation, 2013; Schmitt, 2010; Webb, 2007), yet 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.

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, in the current study, we followed a pedagogically oriented approach (cf. Webb, 2008) to distributing talkers across experimental conditions (see Table 1), prioritizing talker intelligibility (measured through listeners' word transcription) as a way of enhancing potential learning benefits for L2 learners. Rather than rotating all talker voices across the high- and low-variability conditions (for more details about this procedure, see Barcroft & Sommers, 2005, pp. 400–401), we assigned L2 learners in the no-variability conditions (E3 and E6) to target words spoken by the most intelligible talker (Talker 1), contrasting these learning situations with the variability conditions (E3 + TV and E6 + TV) where talkers varied in their intelligibility. In this sense, we ostensibly made it harder for us to detect potential effects of talker variability. Against this backdrop, our findings in favor of enhanced variability, particularly as it pertains to the refinement of phonological forms, is noteworthy. Similarly, the absence of a talker-variability advantage for spoken form recall should not be interpreted as evidence against previous research findings suggesting a positive role for variability (Barcroft & Sommers, 2005; Sinkeviciute et al., 2019). Second, this study measured word stress as a target pronunciation feature, and findings might not be easily generalized to other pronunciation features. Phonological refinement (as defined in this study) was therefore limited to word stress placement accuracy and reduction of vowel duration in unstressed syllables, which was assumed to be relatively easy for L1 Japanese speaking participants to acquire. There is a need for more studies looking at changes in different aspects of word stress such as vowel quality (Zhang & Francis, 2010) and measuring L2 speech using different approaches, including listener judgement (Bradlow et al., 1997). Finally, 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) because encountering words in context is more common than in isolation across instructional 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 the 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.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit <http://doi.org/10.1017/S0272263121000218>.

NOTES

¹As a reviewer pointed out, spoken form recall taps into not only form-meaning mapping but also encoding and retention of novel word forms. To be consistent with the previous L2 vocabulary literature (see Schmitt, 2010, pp. 84–88), the present study used the term “form-meaning connection” as a composite construct indicating learning novel word forms and form-meaning mapping.

²Participants were not forewarned that the same target words would be tested again approximately a week after the first treatment session. After the delayed posttest, all participants received a debriefing to clarify the real purpose of this study.

³The acoustic correlates of lexical stress in English include intensity, pitch, duration, and vowel quality. Pitch was not measured because this acoustic feature which is available as a primary cue in the Japanese language system was considered easy for L1 Japanese speakers to acquire (Ueyama, 2000). Although vowel quality is perhaps the most important cue to the perception of English lexical stress (Zhang & Francis, 2010), it was not examined given that vowel-quality reduction is extremely difficult for Japanese adult learners to acquire (Lee et al., 2006), thus not appropriate to expect them to improve within a limited training session (three or six exposures without feedback). Intensity was not selected as this acoustic feature was less important than other features in stress perception (Beckman, 1986).

⁴Significantly 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 + TV, $t(38) = 3.00, p = .036$ (immediate posttest) and $t(38) = 5.55, p < .001$ (delayed posttest); and E6 + TV, $t(38) = 4.62, p < .001$ (delayed posttest).

⁵A reviewer pointed out that this study was not designed to distinguish between different stages of vocabulary learning as compared to piecemeal vocabulary learning over time. While we agree with this point, we would like to note that we did not intend to provide evidence for or against a particular vocabulary learning model. The two-step model was used for the purpose of conceptualizing the acquisition of spoken word knowledge.

REFERENCES

- Barcroft, J. (2015). *Lexical input processing and vocabulary learning*. John Benjamins.
- Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition*, 27, 387–414. <https://doi.org/10.1017/S0272263105050175>
- Barcroft, J., & Sommers, M. S. (2014). A theoretical account of the effects of acoustic variability on word learning and speech processing. In V. Torrens & L. Escobar (Eds.), *The processing of lexicon and morphosyntax* (pp. 7–14). Cambridge Scholars Publishing.
- Beckman, M. E. (1986). *Stress and non-stress accent*. Foris.
- Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology*, 3, 255–309. <https://doi.org/10.1017/S095267570000066X>
- Boersma, P., & Weenink, D. (2014). Praat: *Doing phonetics by computer (5.3.80) [Computer software]*. <https://www.fon.hum.uva.nl/praat>
- Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /t/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, 101, 2299–2310. <https://doi.org/10.1121/1.418276>
- Brosseau-Lapr e, F., Rvachew, S., Clayards, M., & Dickson, D. (2013). Stimulus variability and perceptual learning of nonnative vowel categories. *Applied Psycholinguistics*, 34, 419–441. <https://doi.org/10.1017/S0142716411000750>
- Bundgaard-Nielsen, R. L., Best, C. T., Kroos, C., & Tyler, M. D. (2012). Second language learners' vocabulary expansion is associated with improved second language vowel intelligibility. *Applied Psycholinguistics*, 33, 643–664. <https://doi.org/10.1017/S0142716411000518>
- Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2011). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in Second Language Acquisition*, 33, 433–461. <https://doi.org/10.1017/S027226311000040>

- Cutler, A. (1990). Exploring prosodic possibilities in speech segmentation. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 105–121). MIT Press.
- Elgort, I., Candry, S., Boutorwick, T. J., Eyckmans, J., & Brysbaert, M. (2018). Contextual word learning with form-focused and meaning-focused elaboration. *Applied Linguistics*, 39, 646–667. <http://doi.org/10.1093/applin/amw029>
- Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24, 143–188. <https://doi.org/10.1017/S0272263102000204>
- Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, 39, 399–423. <https://doi.org/10.2307/3588487>
- Flege, J. E. (1995). Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguistics*, 16, 425–442. <https://doi.org/10.1017/S0142716400066029>
- Geiselman, R. E., & Bellezza, F. S. (1976). Long-term memory for speaker's voice and source location. *Memory & Cognition*, 4, 483–489. <https://doi.org/10.3758/BF03213208>
- Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 152–162. <https://doi.org/10.1037/0278-7393.17.1.152>
- Grosjean, F., & Gee, J. (1987). Prosody structure and spoken word recognition. *Cognition*, 25, 135–155.
- Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, 38, 201–223. <https://doi.org/10.2307/3588378>
- Hirata, Y. (2004). Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts. *The Journal of the Acoustical Society of America*, 116, 2384–2394. <https://doi.org/10.1121/1.1783351>
- Horst, M., Cobb, T., & Meara, P. (1998). Beyond *A Clockwork Orange*: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, 11, 207–223.
- Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, 34, 475–505. <https://doi.org/10.1017/S0272263112000150>
- Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics*, 21, 47–77. <https://doi.org/10.1093/applin/21.1.47>
- Jiang, N. (2002). Form-meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, 24, 617–637. <http://doi.org/10.1017/S0272263102000407>
- Jiang, N. (2004). Morphological insensitivity in second language processing. *Applied Psycholinguistics*, 25, 603–634. <http://doi.org/10.1017/S0142716404001298>
- Kartushina, N., & Martin, C. D. (2019). Talker and acoustic variability in learning to produce nonnative sounds: Evidence from articulatory training. *Language Learning*, 69, 71–105. <https://doi.org/10.1111/lang.12315>
- Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal*, 73, 440–464. <http://doi.org/10.1111/j.1540-4781.1989.tb05325.x>
- Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, 26, 227–247. <https://doi.org/10.1017/S0142716405050150>
- Lee, B., Guion, S. G., & Harada, T. (2006). Acoustic analysis of the production of unstressed English vowels by early and late Korean and Japanese bilinguals. *Studies in Second Language Acquisition*, 28, 487–513. <https://doi.org/10.1017/S0272263106060207>
- Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America*, 89, 874–886. <https://doi.org/10.1121/1.1894649>
- Murphy, J., & Kandil, M. (2004). Word-level stress patterns in the academic word list. *System*, 32, 61–74. <https://doi.org/10.1016/j.system.2003.06.001>
- Nakata, T. (2017). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition*, 39, 653–679. <https://doi.org/10.1017/S0272263116000280>
- Nation, I. S. P. (2012). The BNC/COCA word family lists. <https://www.wgtn.ac.nz/lals/about/staff/paul-nation>
- Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

- Parlak, Ö., & Ziegler, N. (2017). The impact of recasts on the development of primary stress in a synchronous computer-mediated environment. *Studies in Second Language Acquisition*, *39*, 257–285. <http://doi.org/10.1017/S0272263116000310>
- Pavia, N., Webb, S., & Faez, F. (2019). Incidental vocabulary learning through listening to songs. *Studies in Second Language Acquisition*, *41*, 745–768. <https://doi.org/10.1017/S0272263119000020>
- Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8–13. <https://doi.org/10.1016/j.jneumeth.2006.11.017>
- Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition from and while reading: An eye-tracking study. *Studies in Second Language Acquisition*, *38*, 97–130. <http://doi.org/10.1017/S0272263115000224>
- Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130*, 461–472. <https://doi.org/10.1121/1.3593366>
- Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *TESOL Quarterly*, *53*, 1008–1032. <http://doi.org/10.1002/tesq.531>
- Ploonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912. <https://doi.org/10.1111/lang.12079>
- Rice, C. A., & Tokowicz, N. (2020). A review of laboratory studies of adult second language vocabulary training. *Studies in Second Language Acquisition*, *42*, 439–470. <http://doi.org/10.1017/S0272263119000500>
- Rott, S. (1999). The effect of exposure frequency on intermediate language learners’ incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, *21*, 589–619. <https://doi.org/10.1017/S0272263199004039>
- Saito, K. (2013). Reexamining effects of form-focused instruction on L2 pronunciation development: The role of explicit phonetic information. *Studies in Second Language Acquisition*, *35*, 1–29. <https://doi.org/10.1017/S0272263112000666>
- Saito, K. (2018). Advanced segmental and suprasegmental acquisition. In P. Malovrh & A. Benati (Eds.), *The handbook of advanced proficiency in second language acquisition* (pp. 282–303). Wiley Blackwell.
- Saito, K., Suzuki, S., Oyama, T., & Akiyama, Y. (2020). How does longitudinal interaction differentially promote experienced vs. inexperienced learners’ L2 speech learning? *Second Language Research*. Advance online publication. <https://doi.org/10.1177/0267658319884981>
- Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*, 439–462. <http://doi.org/10.1093/applin/amv047>
- Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.
- Sinkeviciute, R., Brown, H., Brekelmans, G., & Wonnacott, E. (2019). The role of input variability and learner age in second language vocabulary learning. *Studies in Second Language Acquisition*, *41*, 795–820. <https://doi.org/10.1017/S0272263119000263>
- Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*, 143–167. <https://doi.org/10.1017/S0272263119000421>
- Thomson, R. I. (2018). High Variability [Pronunciation] Training (HVPT): A proven technique about which every language teacher and learner ought to know. *Journal of Second Language Pronunciation*, *4*, 208–231. <https://doi.org/10.1075/jslp.17038.tho>
- Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, *28*, 1–30. <https://doi.org/10.1017/S0272263106060013>
- Tyler, M. D. (2019). PAM-L2 and phonological category acquisition in the foreign language classroom. In A. M. Nyvad, M. Hejná, A. Højen, A. B. Jespersen, & M. Hjortshøj Sørensen (Eds.), *A sound approach to language matters—In honor of Ocke-Schwen Bohn* (pp. 607–630). Department of English, School of Communication & Culture, Aarhus University.
- Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, *69*, 559–599. <https://doi.org/10.1111/lang.12343>
- Ueyama, M. (2000). *Prosodic transfer: An acoustic study of L2 English vs. L2 Japanese*. (Unpublished doctoral dissertation). University of California, Los Angeles.

- van Zeeland, H., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, 41, 609–624. <https://doi.org/10.1016/j.system.2013.07.012>
- Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, 61, 219–258. <https://doi.org/10.1111/j.1467-9922.2010.00593.x>
- Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, 15, 130–163. <https://doi.org/10.1177/003368828501600214>
- Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, 28, 46–65. <https://doi.org/10.1093/applin/aml048>
- Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language*, 20, 232–245.
- Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford University Press.
- Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL—International Journal of Applied Linguistics*, 168, 33–69. <https://doi.org/10.1075/itl.168.1.02web>
- Werker, J. F. (2018). Speech perception, word learning, and language acquisition in infancy: The voyage continues. *Applied Psycholinguistics*, 39, 769–777. <https://doi.org/10.1017/S0142716418000243>
- Wiener, S., Chan, M. K. M., & Ito, K. (2020). Do explicit instruction and high variability phonetic training improve nonnative speakers' Mandarin tone productions? *The Modern Language Journal*, 104, 152–168. <https://doi.org/10.1111/modl.12619>
- Zhang, R., & Yuan, Z. (2020). Examining the effects of explicit pronunciation instruction on the development of L2 pronunciation. *Studies in Second Language Acquisition*. Advance online publication. <https://doi.org/10.1017/S0272263120000121>
- Zhang, Y., & Francis, A. (2010). The weighting of vowel quality in native and non-native listeners' perception of English lexical stress. *Journal of Phonetics*, 38, 260–271. <http://doi.org/10.1016/j.wocn.2009.11.002>

APPENDIX A

LIST OF FORTY TARGET WORDS

No.	Target word	Phonetic symbol	Number of syllables	Location of main stress
1	abalone	æbəlóʊni	4	3
2	acorn	éikɔːm	2	1
3	armadillo	ɑːrmədílloʊ	4	3
4	binoculars	bainákjʊlərz	4	2
5	caramel	kæərəməɪl	3	1
6	carousel	kæərəseɪl	3	1
7	catapult	kætəpʌlt	3	1
8	celery	séləri	3	1
9	chameleon	kəmflíən	4	2
10	chandelier	ʃændəlíər	4	3
11	chisel	tʃízəl	2	1
12	cicada	sikédə	3	2
13	clover	klóʊvər	2	1
14	crayon	kréɪən	2	1
15	croissant	krəsáːnt	2	2
16	escalator	éskæleɪtər	4	1
17	ladle	léɪdəl	2	1
18	loquat	lókwət	2	1
19	lotus	lóʊtəs	2	1
20	maracas	mərəːkəs	3	2
21	marshmallow	mɑːrʃmeloʊ	3	1
22	mermaid	məːrmeɪd	2	1
23	pacifier	pæsəfáɪər	4	1
24	parakeet	pærəkiːt	3	1
25	persimmon	pərsímən	3	2
26	podium	póʊdiəm	3	1
27	porcupine	pɔːrkjʊpam	3	1
28	protractor	prətræktər	3	2
29	raccoon	rækúːn	2	2
30	raisin	réɪzən	2	1
31	razor	réɪzər	2	1
32	spatula	spætʃələ	3	1
33	strainer	stréɪnər	2	1
34	syringe	səɪrɪndʒ	2	2
35	tadpole	tædpóʊl	2	1
36	toboggan	təbógən	3	2
37	toupee	tuːpéi	2	2
38	treadmill	trédmɪl	2	1
39	walrus	wɔːlɹəs	2	1
40	xylophone	záɪləfóʊn	3	1

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.