Accepted for publication in The Modern Language Journal (Wiley Blackwell)

Does Mode of Input Affect How Second Language Learners Create Form-Meaning Connections and Pronounce Second Language Words?

Takumi Uchihara (Waseda University)
Stuart Webb (University of Western Ontario)
Kazuya Saito (University of College London)
Pavel Trofimovich (Concordia University)

Abstract

This study examined how mode of input affects the learning of pronunciation and form-meaning connection of second language (L2) words. Seventy-five Japanese learners of English were randomly assigned to one of 3 conditions (reading-while-listening, reading-only, listening-only), studied 40 low-frequency words while viewing their corresponding pictures, and completed a picture-naming test 3 times (before, immediately, and about 6 days after treatment). The elicited speech samples were assessed for form-meaning connection (spoken form recall) and pronunciation accuracy (accentedness, comprehensibility). Results showed that the reading-while-listening group recalled a significantly greater number of spoken word forms than did the listening-only group. Learners in the reading-while-listening and listening-only modes were judged to be less accented and more comprehensible compared to learners in the reading-only mode. However, only learners receiving spoken input without orthographic support retained more targetlike (less accented) pronunciation compared to learners receiving only written input. Furthermore, sound-spelling consistency of words significantly moderated the degree to which different learning modes impacted pronunciation learning. Taken together, the findings suggest that simultaneous presentation of written and spoken forms is optimal for the development of form-meaning connection and comprehensibility of novel words, but provision of only spoken input may be beneficial for the attainment of targetlike accent.

Keywords: mode of input, vocabulary learning, pronunciation learning, orthographic influence, accentedness, comprehensibility

Accepted for publication in The Modern Language Journal (Wiley Blackwell)

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, 2012, 2020). 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 first language [L1] translations corresponding to L2 orthographic word forms provided) and targeting 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, despite a few studies measuring L2 pronunciation learning (e.g., Andreas Bürki, 2010), 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 input conditions (reading-while-listening, listening-only, and reading-only).

WRITTEN AND SPOKEN INPUT AND L2 VOCABULARY LEARNING

Research on incidental vocabulary learning has documented that learning occurs as a by-product of exposure to written input, such as reading short sentences (Webb, 2007) and reading graded readers (Brown et al., 2008). 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. Studies have demonstrated that vocabulary learning also occurs from exposure to spoken input (e.g., listening to TV interviews and lectures in 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, 2012, 2020). 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 were 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, 2020). Unlike previous studies focusing on contextualized learning, Bisson et al. (2014) had L2 beginners engage with a decontextualized activity (a letter-search task) where their attention was drawn to spellings of target words while receiving exposure to the spoken forms and pictorial information. In that study, as little as two exposures to the multimodal stimuli led to significant improvement of translation recognition accuracy (see also Krepel et al., 2021, for more evidence supporting multimodal effects on L2 vocabulary learning through a decontextualized activity using translation practice). The attested advantage of multimodal input over unimodal input 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).

As reviewed previously, most work on incidental vocabulary learning tends to measure receptive knowledge (e.g., Brown et al., 2008; Vidal, 2011; Webb & Chang, 2020), with few studies examining how mode of input affects productive knowledge of spoken forms (i.e., pronunciation). This is reasonable given that incidental learning research is most concerned with documenting the amount of

learning as a result of engagement with contextualized and meaning-focused activities where the main focus is on communicative content (Uchihara et al., 2019). Because the lack of explicit or deliberate focus on L2 words stands little chance of substantially improving productive knowledge, which is more challenging to acquire than receptive knowledge (Laufer & Goldstein, 2004), productive knowledge is rarely measured as a focus of incidental vocabulary research. As rare exceptions, two studies involving decontextualized and word-focused activities compared reading-while-listening with unimodal conditions (either reading-only or listening-only), measuring learning gains with production tests of spoken form and form-meaning connection (Andreas Bürki, 2010; Audrey Bürki et al., 2019). 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 (with productions of words elicited via a word-reading task 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 learners exposed to L2 orthographic input recalled significantly more spoken forms than those receiving only spoken input, but the 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), making it unclear whether learners could accurately pronounce L2 words spontaneously (without reading them). Also, neither study adopted global measures of L2 pronunciation, such as word intelligibility or comprehensibility. 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.

ORTHOGRAPHIC INFLUENCE IN L2 PRONUNCIATION LEARNING

Studies investigating the role of orthographic input in L2 phonological learning have produced inconsistent findings (for a review, see Bassetti, 2008; Hayes–Harb & Barrios, 2021), suggesting that orthography can have both 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 provided with orthographic support may have relied on their L1 orthographic system and substituted L1 vowels for their L2 counterparts, resulting in orthography-induced, nontargetlike pronunciations (Bassetti, 2008). Similar negative effects of orthography on L2 pronunciation due to L1-L2 incongruency in orthographic system have been documented in the studies of L1 Italian speakers learning L2 English words containing double letters such as *Finnish* versus *Finish* (Cerni et al., 2019), L1 English speakers learning L2 German-like words containing final voiced obstruents such as *Steid* versus *Steit* (Hayes-Harb et al., 2018), and L1 English speakers learning L2 Spanish words containing a range of incongruent segments such as <ll> pronounced as /j/ in English (Rafat, 2016).

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 the exception 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 L1 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.

GLOBAL CONSTRUCTS OF PRONUNCIATION: 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 is defined as listeners' judgments of how different L2 speech sounds from the expected language variety. These two constructs are typically measured through listeners' ratings of speakers, using numerical point scales (e.g., 1 = *easy to understand*, 9 = *hard to understand*; 1 = *no accent*, 9 = *heavily accented*). A possible issue regarding the validity of global measures of pronunciation (particularly accentedness) is the question of who should be evaluating L2 speech. In research, the target variety of a given language as spoken by L1 speakers is often set as a reference point. However, "the native variety" frequently varies among speakers across different contexts, so researchers recruit listeners, with their specific experiential profiles, to represent a particular reference point (e.g., as trained L1-speaking raters assessing learners' speech or as naïve interlocutors interacting with learners). For instance, to provide a reference point for accentedness, researchers might sample listeners with homogeneous backgrounds, such as those who come from the same regional variety (e.g., British, American) and/or those who have knowledge of learners' L1. The extent to which various listener characteristics impact accentedness ratings has been researched but remains inconclusive (Saito, 2021). One way to confirm rating reliability is to examine interrater consistency to determine that it is sufficiently high for research purposes. Although listeners normally do not receive any detailed instruction as to the rating procedure except brief descriptions of the target constructs (e.g., accentedness, comprehensibility), high interrater consistency has been achieved regardless of various differences across listeners, for instance, in their L1 versus L2 status or prior training in linguistics and phonetics (e.g., see Saito, 2021, for a review).

Since Munro and Derwing (1995), researchers have measured L2 accentedness and comprehensibility with the aim of tracking the development of L2 pronunciation proficiency in naturalistic environments (e.g., Derwing & Munro, 2013; Saito, 2015) and instructional settings (e.g., Nagle, 2018; Saito & Hanzawa, 2018). The accumulated evidence suggests that comprehensibility improves with increased L2 exposure and intensive instruction, whereas attaining targetlike pronunciation remains a difficult task (Saito, 2021). L2 learners could improve in their production of some acoustic features determining segmental and suprasegmental accuracy, but other features are more difficult to improve (e.g., labial, alveolar, and pharyngeal constrictions) despite continued immersion or instruction (Trofimovich & Baker, 2006). For example, Japanese speakers initially aim to attain the comprehensible and intelligible form of a North American variety of English /r/ by using interlanguage strategies (tongue retraction, phonemic lengthening). After much immersion experience, some learners appear to approximate more targetlike English /r/ as they learn to produce novel acoustic (F3) and articulatory configurations, including labial, alveolar, and pharyngeal constrictions (Saito & Brajot, 2013). Reliance on a limited repertoire of acoustic features might be sufficient to render L2 speech comprehensible but not necessarily targetlike. While a small improvement of learners' accent can be expected with practice, estimated through a recent meta-analysis to be a relatively small effect (Cohen's

d = 0.28) following instruction (Saito, 2021), sounding non-accented might require a substantial amount of immersion experience (Flege & Fletcher, 1992; Trofimovich & Baker, 2006), an earlier age of onset (Flege et al., 2006), strong motivation (Moyer, 2014), and special language learning abilities (see Suzukida, 2021, for a review), including phonemic coding (e.g., Hu et al., 2013) and perceptual acuity (e.g., Saito et al., 2020).

In light of these empirical findings, many scholars have emphasized the importance of assessing L2 speech in terms of both L2 comprehensibility and accentedness as these measures might distinguish three different developmental stages of adult L2 speech learning (Derwing & Munro, 2013; Saito, 2021). At the onset of development, L2 speakers initially produce weakly specified (heavily accented) forms requiring considerable effort for listeners to understand (low comprehensibility). Through more exposure and ample conversational opportunities, L2 speakers might achieve communicatively adequate production (comprehensible *but* accented speech). Ultimately, some L2 speakers can master the phonological detail characteristic of advanced proficiency (comprehensible *and* targetlike speech). Using this developmental account of L2 speech learning, the current study set out to explore how training can help learners acquire not only comprehensible pronunciation forms of target L2 words (which would correspond to the initial-to-mid stage of L2 speech learning) but also more targetlike, refined, and advanced pronunciations of L2 words (corresponding to the mid-to-final stages of L2 speech learning).

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 largely based on the findings from research on incidental vocabulary learning which measured vocabulary learning in written form. In order to advance our insights into how and when spoken input facilitates (or inhibits) vocabulary learning, more research needs to employ different instructional approaches (e.g., explicit focus on L2 words vs. incidental learning) and measure other aspects of word knowledge beyond written mode of form-meaning connection.

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 targetlike accuracy (e.g., a forced-choice identification task by L1 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 and achievable goal) and accentedness (as a 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 adopted a decontextualized/deliberate learning procedure (i.e., paired associate word learning) and measured spoken forms of L2 words, was guided by the following research questions:

RQ1. To what extent does mode of input (reading-while-listening, reading-only, listening-only) affect the learning of form-meaning connections?
RQ2. 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)?
RQ3. 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). Last, different effects of mode of input might arise for different pronunciation measures. Learners receiving written input might 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 influence 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.

METHOD

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.

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_{age} = 19.5, range = 18–24) had studied English for a minimum of 6 years in instructional settings. All participants except one had scored 90% or higher on the 1,000 word level of the Vocabulary Levels Test (Webb et al., 2017), and all except one had scored 80% 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.

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*). To illustrate using the consistency scores from this dataset, *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 Canadian speaker of English from Ontario 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 L1 English speaker. Pilot testing showed that two L1 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

No.	Target word	Phoneme counts	Number of syllables	Consistency scores
1	abalone	7	4	Off-List
2	acorn	5	2	0.511
3	armadillo	8	4	0.469
4	binoculars	10	4	0.414
5	caramel	6	3	0.918
6	carousel	7	3	0.297
7	catapult	8	3	0.647
8	celery	6	3	0.586
9	chameleon	8	4	0.481
10	chandelier	9	4	0.632
11	chisel	4	2	0.702
12	cicada	6	3	0.185
13	clover	6	2	0.658
14	crayon	5	2	0.820
15	croissant	7	2	0.055
16	escalator	9	4	0.556
17	ladle	4	2	0.573
18	loquat	6	2	Off-List

No.	Target word	Phoneme counts	Number of syllables	Consistency scores
19	lotus	5	2	0.623
20	maracas	7	3	Off-List
21	marshmallow	8	3	0.490
22	mermaid	6	2	0.949
23	pacifier	8	4	0.435
24	parakeet	8	3	0.525
25	persimmon	8	3	0.619
26	podium	6	3	0.488
27	porcupine	9	3	0.575
28	protractor	10	3	0.643
29	raccoon	5	2	0.716
30	raisin	4	2	0.544
31	razor	5	2	0.279
32	spatula	7	3	0.476
33	strainer	7	2	0.879
34	syringe	6	2	0.536
35	tadpole	6	2	0.796
36	toboggan	7	3	0.370
37	toupee	4	2	0.128
38	treadmill	7	2	0.934
39	walrus	6	2	0.637
40	xylophone	7	3	0.658

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.

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. Immediately after the final exposure to each block of five items, a picture-naming test was administered. 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). The same test procedure (except for exposure treatment) was adopted for both pretest and delayed posttest. 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 is available in the Online Supporting Information).

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)[1] 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 asked 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.

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 L1 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), 6 L1 speakers of Canadian English from Ontario (3 females, 3 males) were recruited to participate in a series of rating sessions. Speakers of Canadian English were chosen because they represented the variety of English that was readily accessible to the researchers and that provided a reasonable L1-speaker benchmark for assessing the pronunciation of L2 learners. Three of 6 speakers had language teaching experiences in EFL and English-as-a-second language (ESL) contexts. All 6 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). Raters were first presented with 40 target words produced by the model speaker as a reference point and asked to evaluate whether L2 speech samples were intelligible to an average speaker of their variety of English. For some of the words having multiple variant pronunciations (e.g., /baɪ̲nάkjʊlərz/ and /bɪ̲nάkjʊlərz/ for *binoculars*), the one which was produced by the model speaker was considered as the expected target sound. 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 (e.g., through a 9-point scale in Munro & Derwing, 1995), 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 L1 speakers' samples (included as distracter items), totaling 4,141 items (4,061 from Japanese speakers + 80 from English speakers). The inclusion of the English 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.

Data Analysis

Preliminary analysis of raters' responses to English speaker samples showed that raters consistently judged the English-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*). The interrater reliability for accentedness (α = .75) and comprehensibility (α = .72) was not perfect yet satisfactory for research purposes (α > .70; Larson–Hall, 2010). These preliminary results confirmed 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). 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.

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. Full results of mixed-effects modelling conducted to answer research questions (i.e., variance components, R-squared, random parameter correlations, and model fit indices) can be found in the Online Supporting Information.

TABLE 2

Means, Standard Deviations, and 95% Confidence Intervals for Spoken Form Recall, Accentedness, and Comprehensibility by Group

Subgroups	Spoken form recall			Accentedness		Comprehensibility	
	Pretest	Immediate posttest	Delayed posttest	Immediate posttest	Delayed posttest	Immediate posttest	Delayed posttest
LO	0.25 (0.05)	0.48 (0.10)	0.33 (0.07)	3.78 (0.27)	3.73 (0.41)	3.18 (0.35)	2.82 (0.50)
	[0.22, 0.27]	[0.43, 0.52]	[0.30, 0.36]	[3.67, 3.89]	[3.56, 3.90]	[3.04, 3.33]	[2.61, 3.03]
RO	0.25 (0.07)	0.49 (0.10)	0.34 (0.07)	4.08 (0.24)	3.98 (0.31)	3.43 (0.30)	3.15 (0.35)
	[0.22, 0.27]	[0.45, 0.53]	[0.31, 0.37]	[3.97, 4.18]	[3.85, 4.11]	[3.30, 3.55]	[3.01, 3.29]
RWL	0.27 (0.05)	0.53 (0.10)	0.36 (0.07)	3.80 (0.26)	3.86 (0.38)	3.20 (0.30)	2.98 (0.41)
	[0.25, 0.29]	[0.49, 0.57]	[0.33, 0.39]	[3.69, 3.90]	[3.70, 4.02]	[3.07, 3.33]	[2.81, 3.15]

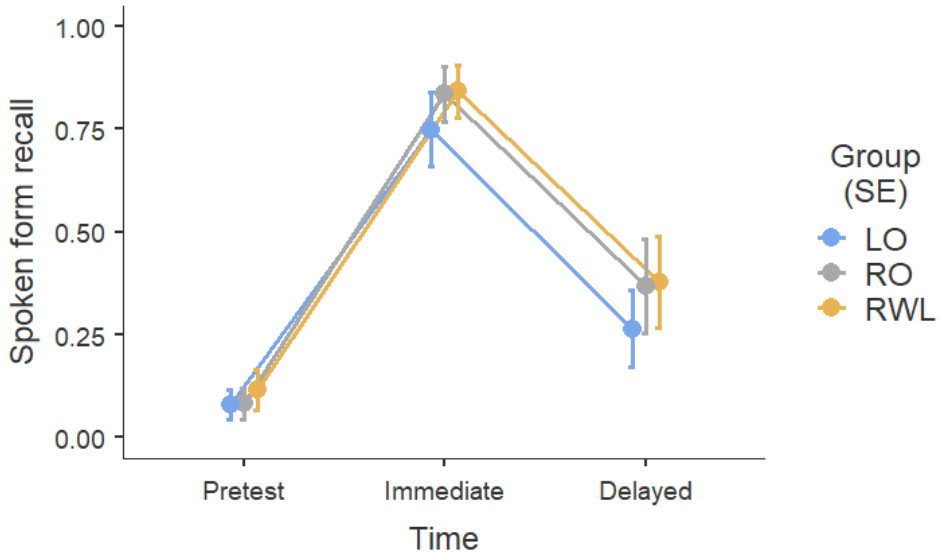
Note. Accentedness (1 = *not accented at all* to 5 = *heavily accented*); comprehensibility (1 = *easy to understand* to 5 = *hard to understand*). Standard deviations are in parentheses and upper and lower confidence intervals are in square brackets.

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 × 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: $\beta = 0.47$, $SE = 0.33$, $z = 1.41$, $p = .159$).[2] At immediate posttest, learners receiving both written and spoken input significantly outperformed those receiving only spoken input (RWL, $M = 53\%$, $SD = 10\%$; 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 Means for Spoken Form Recall Over Time.



Note. Error bars represent standard errors (SE) of the means.

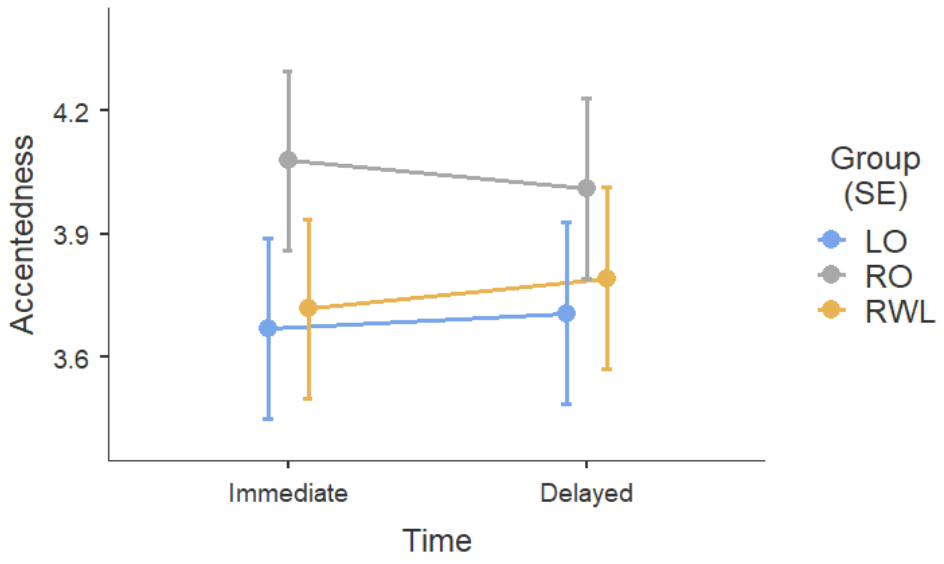
Accentedness and Comprehensibility

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 vs. LO: $\beta = 0.42$, $SE = 0.10$, $t = 4.21$, $d = 1.19$, $p < .001$; RO vs. RWL, $\beta = 0.37$, $SE = 0.09$, $t = 4.04$, $d = 1.14$, $p < .001$). No significant difference was found between the LO and RWL groups ($\beta = 0.05$, $SE = 0.09$, $t = 0.57$, $d = 0.16$, $p = .570$). At delayed posttest, the superiority of spoken input over the written-input-only condition remained present for LO versus RO ($\beta = 0.35$, $SE = 0.13$, $t = 2.68$, $d = 0.76$, $p = .010$), and also for RWL versus RO to a lesser degree ($\beta = 0.20$, $SE = 0.11$, $t = 1.84$, $d = 0.52$, $p = .071$). No significant difference was found between the LO and RWL groups ($\beta = 0.16$, $SE = 0.11$, $t = 1.39$, $d = 0.39$, $p = .174$).

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 vs. RO: $\beta = 0.36$, $SE = 0.12$, $t = 3.05$, $d = 0.86$, $p = .003$; RWL vs. RO: $\beta = 0.33$, $SE = 0.10$, $t = 3.27$, $d = 0.92$, $p = .002$). There was no significant difference between the LO and RWL groups ($\beta = 0.03$, $SE = 0.10$, $t = 0.35$, $d = 0.10$, $p = 0.73$). At delayed posttest, no significant differences were observed between the three groups with the trend that RWL was perceived to be more comprehensible than RO (LO vs. RO: $\beta = 0.27$, $SE = 0.17$, $t = 1.63$, $d = 0.46$, $p = .109$; LO vs. RWL: $\beta = 0.04$, $SE = 0.14$, $t = 0.27$, $d = 0.08$, $p = .788$; RWL vs. RO: $\beta = 0.23$, $SE = 0.13$, $t = 1.75$, $d = 0.49$, $p = .087$).

FIGURE 2

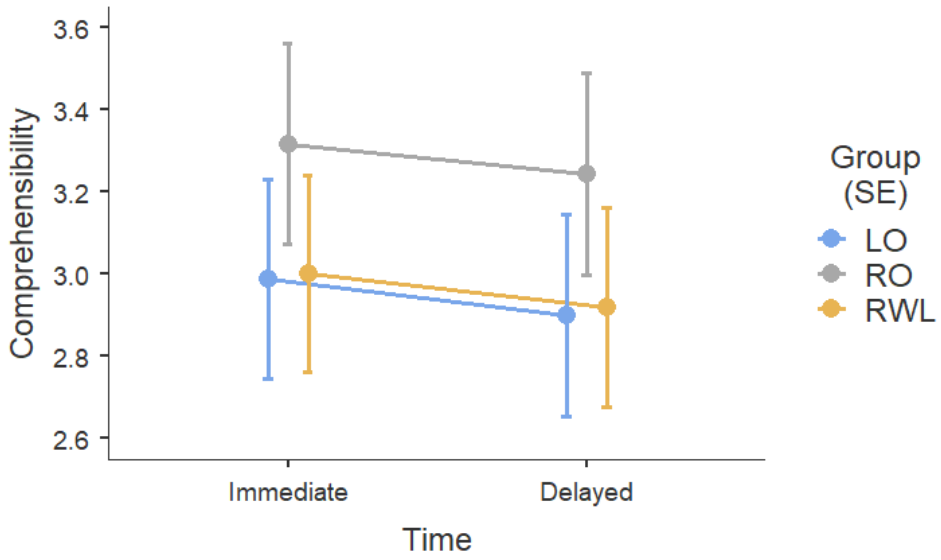
Group Means for Accentedness Over Time.



Note. Error bars represent standard errors (SE) of the means.

FIGURE 3

Group Means For Comprehensibility Over Time.



Note. 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 targetlike 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 targetlike 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: β = 0.41, SE = 0.59, t = 0.70, p = .488).

FIGURE 4

Relationship Between Sound-Spelling Consistency and Accentedness per Group (Immediate Posttest).

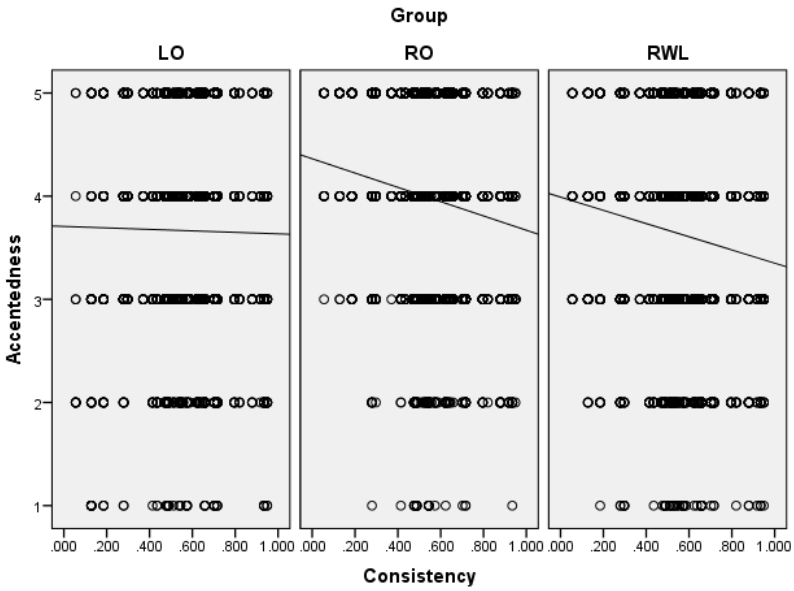


FIGURE 5

Relationship Between Sound-Spelling Consistency and Accentedness per Group (Delayed Posttest)

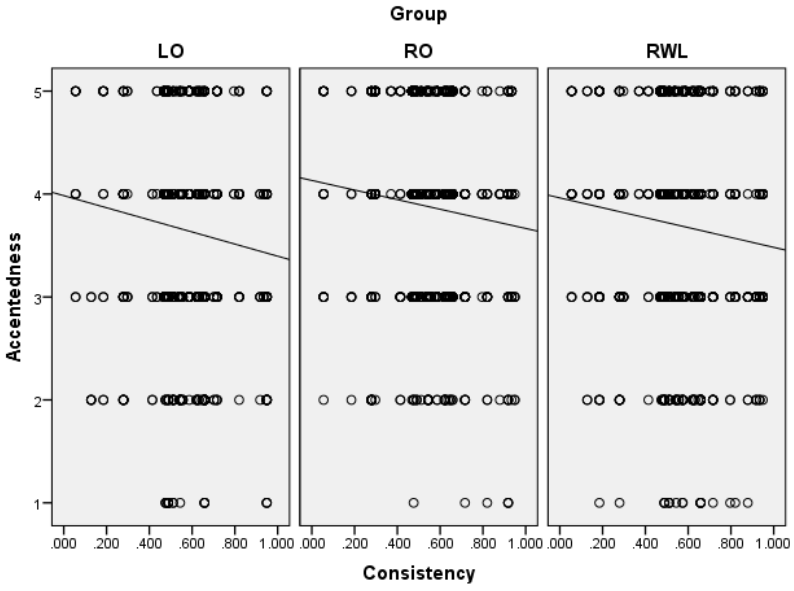


FIGURE 6

Relationship Between Sound-Spelling Consistency and Comprehensibility per Group (Immediate Posttest)

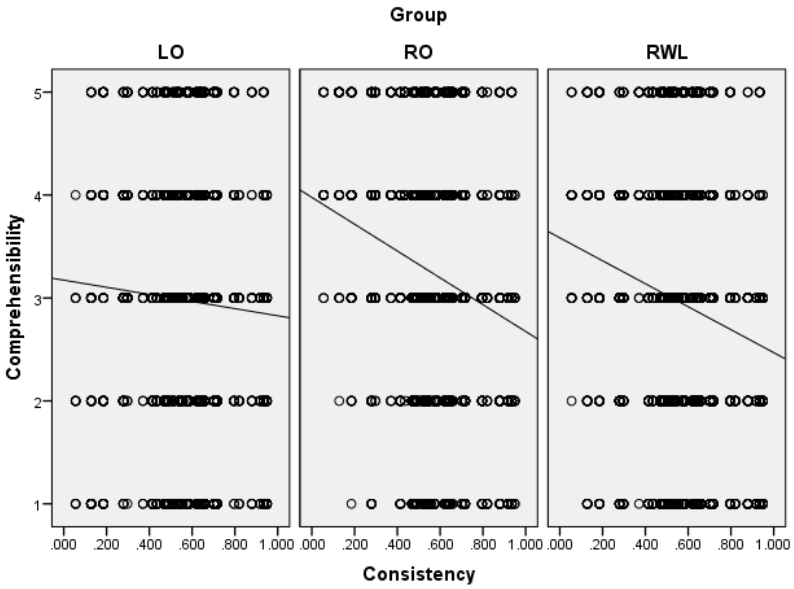
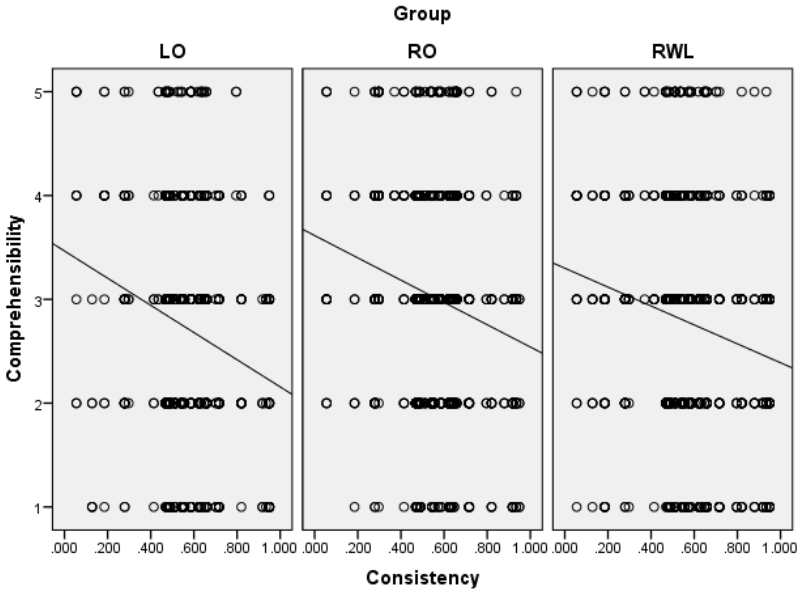


FIGURE 7

Relationship Between Sound-Spelling Consistency and Comprehensibility per Group (Delayed Posttest)



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.

DISCUSSION

Overall results showed that learners in the RWL group recalled a significantly larger number of spoken word forms than learners in the LO group, 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; Krepel et al., 2021; Malone, 2018; Webb & Chang, 2012). Also, learners in the RWL and LO groups 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 impacted by a potential increase in cognitive workload. 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; Krepel et al., 2021; Malone, 2018; Webb & Chang, 2012). Despite the statistically significant results, these bimodal effects were relatively moderate in magnitude when considered from a practical standpoint (RWL group learners gained 5% or 1.6 words more than LO group on the immediate posttest). 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. Furthermore, the lack of significant differences between RO and LO was not expected and merits further discussion. It was predicted that compared to the LO condition, learners in the RO condition would show smaller gains due to the absence of congruency between the input and test modes (Jelani & Boers, 2018). These findings suggest that learners encountering the written forms of words might perform as well as learners receiving spoken input in building form-meaning connections of L2 words *and* in achieving a minimum-threshold accuracy of the phonological forms according to the criteria set in this study (i.e., judged as acceptable by all six L1 listeners).

In answer to the second research question, the findings that RWL and LO outperformed RO for accentedness and comprehensibility with medium-to-large effects (d = 0.86–1.19) 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 targetlike 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 L1 speakers use the L1 romanization system (i.e., *Roomaji*) 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 /ə/ (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., *toboggan*: /t**ə**b**ɑ**gən/→*/t**o**b**o**gən/) and devoicing L2 consonants (e.g., *chisel*: /tʃizəl/→*/tʃ**ɪ**səl/). 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.

IMPLICATIONS, LIMITATIONS, AND FUTURE DIRECTIONS

The current study provides methodological and pedagogical implications for assessing and teaching L2 vocabulary. First, learners encountering the written forms of L2 words without the availability of spoken input could achieve a minimum-threshold level of production accuracy. However, the role of spoken input comes into play in further enhancing the form accuracy, such that newly learned spoken words might be perceived by listeners as less accented and easier to understand. 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 targetlike pronunciation forms (Derwing & Munro, 2015), we argue that RWL may be an optimal method for 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 accent reduction) 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, and the inefficiency of learning was evident from a large drop between immediate and delayed posttest scores for spoken form recall. To increase the ecological validity of the research and improve learning, 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). Third, this study did not explore the influence of individual differences, such as L2 proficiency and auditory processing skills (Saito et al., 2020) on vocabulary learning. Although exploration of learner-related variables was beyond the scope of this study, individual differences play an important role in L2 pronunciation learning (for a review, see Suzukida, 2021). How learner-internal variables interact with mode of input needs to be investigated further. Fourth, findings of this study based on the learning of concrete words may not generalize to a situation where learners study abstract words because concrete and imageable meanings are easier to learn than abstract meanings (Ellis & Beaton, 1993). The extent to which varying degrees of imageability affect word pronunciation learning remains unknown, which is also worth investigating in the future. Finally, testing word knowledge via phonological form for the RO group, who received only written input, might not be ecologically valid. Essentially, participants in the RO condition were required to draw on grapheme-to-phoneme translations to articulate target words. In this study, the RO condition served as a baseline group, and the data were employed to explore the extent to which exposure to written forms, especially when words of higher consistency were targeted, could lead to comprehensible word pronunciation. This methodological decision was also pedagogically motivated given the fact that vocabulary learning tends to be implemented via written mode, especially in EFL contexts (Nation, 2013). Moreover, understanding the benefits and limitations of learning L2 words from written input alone provides useful pedagogical implications. Yet, to increase the ecological validity in light of learning-test congruency, future studies should also measure knowledge of written forms of words, which will allow researchers to evaluate both advantages and disadvantages of written input for multifaceted aspects of vocabulary learning more comprehensively.

ACKNOWLEDGMENT

This research was supported by the *Language Learning* Dissertation Grant Program (grant number: R5370A13). We would like to thank Shuhei Kudo for his help with data collection.

NOTE

[1]The interval means for LO, RO, and RWL were 5.4 days (*SD* = 3.3), 6.2 days (*SD* = 3.5), and 6.6 days (*SD* = 4.0), and there was no statistical difference across groups, F (2, 72) = 0.68, p = .51.

[2]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.

REFERENCES

Baddeley, J. (1986). *Working memory*. Oxford University Press.

Bassetti, B. (2008). Orthographic input and second language phonology. In P. Thorsten & M. Young-Scholten, *Input matters in SLA* (pp. 191–206). Multilingual Matters.

Bassetti, B., & Atkinson, N. (2015). Effects of orthographic forms on pronunciation in experienced instructed second language learners. *Applied Psycholinguistics*, *36*, 67–91. https://doi.org/10.1017/S0142716414000435

Bisson, M.-J., van Heuven, W. J. B., Conklin, K., & Tunney, R. J. (2014). The role of repeated exposure to multimodal input in incidental acquisition of foreign language vocabulary. *Language learning*, *64*, 855–877. https://doi.org/10.1111/lang.12085

Boersma, P., & Weenink, D. (2014). *Praat: Doing phonetics by computer* (5.3.80) [Computer software]. https://www.fon.hum.uva.nl/praat/

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, *20*, 136–163.

Bürki, Andreas. (2010). Lexis that rings a bell: On the influence of auditory support in vocabulary acquisition. *International Journal of Applied Linguistics*, *20*, 206–231. https://doi.org/10.1111/j.1473-4192.2009.00246.x

Bürki, Audrey, Welby, P., Clément, M., & Spinelli, E. (2019). Orthography and second language word learning: Moving beyond "friend or foe?" *The Journal of the Acoustical Society of America*, *145*, EL265–EL271. https://doi.org/10.1121/1.5094923

Cerni, T., Bassetti, B., & Masterson, J. (2019). Effects of orthographic forms on the acquisition of novel spoken words in a second language. *Frontiers in Communication*, *4*, 31. https://doi.org/10.3389/fcomm.2019.00031

Chee, Q. W., Chow, K. J., Yap, M. J., & Goh, W. D. (2020). Consistency norms for 37,677 english words. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01391-7

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, *63*, 163–185. https://doi.org/10.1111/lang.12000

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives*

for L2 teaching and research. John Benjamins.

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language learning*, *43*, 559–617. https://doi.org/10.1111/j.1467-1770.1993.tb00627.x

Erdener, V. D., & Burnham, D. K. (2005). The role of audiovisual speech and orthographic information in nonnative speech production. *Language Learning*, *55*, 191–228. https://doi.org/10.1111/j.0023-8333.2005.00303.x

Flege, J. E., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Tsukada, K. (2006). Degree of foreign accent in English sentences produced by Korean children and adults. *Journal of Phonetics*, *34*, 153–175. https://doi.org/10.1016/j.wocn.2005.05.001

Flege, J. E., & Fletcher, K. L. (1992). Talker and listener effects on degree of perceived foreign accent. *The Journal of the Acoustical Society of America*, *91*, 370–389. https://doi.org/10.1121/1.402780

Hatami, S. (2017). The differential impact of reading and listening on L2 incidental acquisition of different dimensions of word knowledge. *Reading in a Foreign Language*, *29*, 61–85.

Hayes–Harb, R., & Barrios, S. (2021). The influence of orthography in second language phonological acquisition. *Language Teaching*, 1–30. doi:10.1017/S0261444820000658

Hayes–Harb, R., Brown, K., Smith, B. L. (2018). Orthographic input and the acquisition of German final devoicing by native speakers of English. *Language and Speech*, *61*, 547–564. https://doi.org/10.1177/0023830917710048

Hu, X., Ackermann, H., Martin, J. A., Erb, M., Winkler, S., & Reiterer, S. M. (2013). Language aptitude for pronunciation in advanced second language (L2) learners: Behavioural predictors and neural substrates. *Brain and Language*, *127*, 366–376. https://doi.org/10.1016/j.bandl.2012.11.006

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

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Jelani, N. A. M., & Boers, F. (2018). Examining incidental vocabulary acquisition from captioned video: Does test modality matter? *ITL - International Journal of Applied Linguistics*, *169*, 169–190. https://doi.org/10.1075/itl.00011.jel

Kang, S. H. K., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin & Review*, *20*, 1259–1265. https://doi.org/10.3758/s13423-013-0450-z

Krepel, A., de Bree, E.H. & de Jong, P.F. (2021). Does the availability of orthography support L2 word learning? *Reading and Writing*, *34*, 467–496. https://doi.org/10.1007/s11145-020-

10078-6

Larson–Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. Routledge.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*, 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*, 369–377. https://doi.org/10.2307/3588485

Malone, J. (2018). Incidental vocabulary learning in SLA: Effects of frequency, aural enhancement, and working memory. *Studies in Second Language Acquisition*, *40*, 651–675. https://doi.org/10.1017/S0272263117000341

Mayer, R. E. (2014). Introduction to multimedia learning. In R. E. Mayer, *The Cambridge handbook of multimedia learning* (pp. 1–24). Cambridge University Press.

Montero Perez, M., Van Den Noortgate, W., & Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System*, *41*, 720–739. https://doi.org/10.1016/j.system.2013.07.013

Moyer, A. (2014). Exceptional outcomes in L2 phonology: The critical factors of learner engagement and self-regulation. *Applied Linguistics*, *35*, 418–440. https://doi.org/10.1093/applin/amu012

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*, 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Nagle, C. (2018). Motivation, comprehensibility, and accentedness in L2 Spanish: Investigating motivation as a time-varying predictor of pronunciation development. *The Modern Language Journal*, *102*, 199–217. https://doi.org/10.1111/modl.12461

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.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Heinle Cengage Learning.

Niegeman, H., & Heidig, S. (2012). Multimedia learning. In N. M. Seel, *Encyclopedia of sciences of learning* (pp. 2372–2375). Springer.

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

Rafat, Y. (2016). Orthography-induced transfer in the production of English-speaking learners of Spanish. *The Language Learning Journal*, *44*, 197–213. https://doi.org/10.1080/09571736.2013.784346

Saito, K. (2015). Experience effects on the development of late second language learners' oral

proficiency. *Language Learning*, *65*, 563–595. https://doi.org/10.1111/lang.12120

Saito, K. (2021). What characterizes comprehensible and native-like pronunciation among English-as-a-second-language speakers? Meta-analyses of phonological, rater, and instructional factors. *TESOL Quarterly*. https://doi.org/10.1002/tesq.3027

Saito, K., & Brajot, F-X. (2013). Scrutinizing the role of length of residence and age of acquisition in the interlanguage pronunciation development of English /ɹ/ by late Japanese bilinguals. *Bilingualism: Language and Cognition*, *16*, 847–863. https://doi.org/10.1017/S1366728912000703

Saito. K., & Hanzawa, K. (2018). The role of input in second language oral ability development in foreign language classrooms: A longitudinal study. *Language Teaching Research*, *22*, 398–417. https://doi.org/10.1177/1362168816679030

Saito, K., Kachlicka, M., Sun, H., & Tierney, A. (2020). Domain-general auditory processing as an anchor of post-pubertal second language pronunciation learning: Behavioural and neurophysiological investigations of perceptual acuity, age, experience, development, and attainment. *Journal of Memory and Language*. https://doi.org/10.1016/j.jml.2020.104168

Solier, C., Perret, C., Baqué, L., & Soum-Favaro, C. (2019). Written training tasks are better than oral training tasks at improving L2 learners' speech production. *Applied Psycholinguistics*, *40*, 1455–1480. https://doi.org/10.1017/S014271641900033X

Suzukida, Y. (2021). The contribution of individual differences to L2 pronunciation learning: Insights from research and pedagogical implications. *RELC Journal*, *52*, 48–61. https://doi.org/10.1177/0033688220987655

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

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

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

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*, 46–65. https://doi.org/10.1093/applin/aml048

Webb, S., & Chang, A. C.-S. (2020). How does mode of input affect the incidental learning of collocations? *Studies in Second Language Acquisition*. https://doi.org/10.1017/S0272263120000297

Webb, S., & Chang, A. C.-S. (2012). Vocabulary learning through assisted and unassisted repeated

reading. *The Canadian Modern Language Review*, *68*, 267–290. https://doi.org/10.3138/cmlr.1204.1

Webb, S., Sasao, Y., & Oliver, B. (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