




EMPIRICAL STUDY

Declarative and Automatized Phonological Vocabulary Knowledge: Recognition, Recall, Lexicosemantic Judgment, and Listening-Focused Employability of Second Language Words

Takumi Uchihara ^a, Kazuya Saito ^b, Satsuki Kurokawa,^a
Kotaro Takizawa ^c, and Yui Suzukida ^b

^aTohoku University ^bUniversity College London ^cWaseda University

Abstract: This study revisits the roles of different aspects of phonological vocabulary knowledge in second language (L2) listening. Japanese learners of English ($n = 114$) completed the TOEIC Listening test and three phonological vocabulary tests assessing (a) ability to recognize the meanings of aural forms (meaning recognition), (b) ability to recall the meanings of aural forms (meaning recall), and (c) ability to spontaneously judge the appropriate use of word meanings in sentential contexts (lexicose-

CRedit author statement – **Takumi Uchihara:** conceptualization (equal); funding acquisition (lead); methodology (equal); writing – original draft preparation (lead); formal analysis (equal); writing – review and editing (equal); project administration (lead). **Kazuya Saito:** conceptualization (equal); writing – original draft preparation (supporting); methodology (equal); writing – review and editing (equal). **Satsuki Kurokawa:** investigation (lead); formal analysis (equal); writing – review and editing (equal). **Kotaro Takizawa:** investigation (supporting); writing – review and editing (equal). **Yui Suzukida:** conceptualization (equal); investigation (supporting); writing – review and editing (equal); project administration (supporting).

A one-page Accessible Summary of this article in nontechnical language is freely available in the Supporting Information online and at <https://oasis-database.org>

This project was supported by the JSPS KAKENHI Grant Number 21K19995.

Correspondence concerning this article should be addressed to Takumi Uchihara, Tohoku University, Graduate School of International Cultural Studies, 41 Kawauchi, Aoba-ku, Sendai, 980–8576, Japan. Email: takumi@tohoku.ac.jp

The handling editor for this manuscript was Sarah Grey.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

mantic judgment task [LJT]). Among the three measures, the LJT best predicted the participants' ability to access the target words during real-life L2 listening comprehension of monologues and conversations (measured via TOEIC). Structural equation modeling demonstrated that the LJT was distinct from both meaning recognition and recall and revealed their different associations with listening comprehension scores. In line with the skill acquisition theory, we propose that the LJT reflects automatized knowledge, whereas meaning recognition and recall represent declarative knowledge.

Keywords vocabulary; listening; phonological vocabulary; automatization

Introduction

Research has provided converging evidence that second language (L2) learners with larger and richer lexicons are more skillful at comprehending spoken language (Stæhr, 2009; Vafae & Suzuki, 2020; Vandergrift & Baker, 2015, 2018; Wallace, 2022). The close vocabulary–listening link has a firm theoretical basis from the perspectives of psycholinguistics and speech perception (Cutler & Clifton, 1999; Field, 2009, 2013, 2019). Listening initially involves bottom-up processing where the incoming speech stream is interpreted by the listener as multiple units of language such as phonemes, syllables, words, phrases, and sentences. Successful processing at the bottom-up level depends on automaticity and efficiency in phonological decoding, lexical search, and syntactic processing. Developed L2 lexicon is an important driving force for the attainment of optimal comprehension because it facilitates accurate and rapid retrieval of words, determining the quality of formation of a literal meaning of the utterance. With robust and efficient bottom-up processing, lexically competent listeners do not need to rely too much on top-down processing through using world knowledge and cotextual information to aurally understand speakers' intended messages.

As a further step of inquiry in this area, there is an emerging paradigm that attempts to reconceptualize a multifaceted construct of vocabulary knowledge relevant to L2 listening (Cheng et al., 2023; McLean et al., 2015; Milton & Hopkins, 2006). This line of research highlights that vocabulary knowledge needs to be assessed aurally (Milton et al., 2010) in recall format (Cheng et al., 2023) in order for the vocabulary tasks to reflect real-life listening situations. Recent studies have also expanded the scope of word knowledge to examine the role of multiword items, confirming that knowledge of such items (e.g., phrasal verbs) is closely related to L2 listening success (Cheng et al., 2023; Matthews et al., 2024). However, earlier studies have relied on controlled and single-task formats, inducing learners' explicit attention to target items

presented in isolation (e.g., multiple choice, translation, and lexical decision tasks). Such controlled vocabulary measures serve as a useful proxy for learners' basic and declarative knowledge of the form–meaning mapping of individual words or multiword items. However, advanced listening requires more than explicit knowledge of a simple form–meaning link for words or phrases, extending to the ability to capitalize on multiple cues (e.g., collocational, grammatical, and contextual information) available in surrounding sentences in an integrated manner and retrieve appropriate meanings spontaneously (Nation, 2022; Schmitt, 2019). We argue that basic knowledge of form–meaning connections, measured separately through traditional controlled and decontextualized tasks, does not comprehensively capture advanced lexical knowledge relevant to L2 listening nor suffice to explain how the development of phonological vocabulary leads to improvement in listening ability.

Building on the notion of instructed skill acquisition theory (DeKeyser, 2020; Suzuki, 2023), the current study proposes that there are two fundamental aspects of phonological vocabulary knowledge: (a) the knowledge that allows listeners to retrieve L2 word meanings as they draw on declarative memories of form–meaning mapping (declarative knowledge) and (b) the knowledge that allows listeners to encode the semantic and collocational clues in immediate contexts to retrieve L2 word meanings spontaneously (automatized knowledge). According to Nation's (2022) componential model, declarative knowledge pertains to a component of the form–meaning connection of individual words (measured through meaning-recognition and -recall formats), whereas automatized knowledge additionally encompasses use-in-context aspects of vocabulary knowledge (indicated via a lexicosemantic judgment task, a newly developed lexical measure in this study). In essence, automatized phonological knowledge is distinct from declarative knowledge in that the multifaceted and integrative nature of automatized knowledge is hypothesized to allow for rapid and consistent retrieval of learned words, enabling the use of vocabulary in collocationally, grammatically, and contextually appropriate ways in L2 speech comprehension. Based on the conceptualization of the two types of lexical knowledge, we investigate the relative contribution of automatized and declarative phonological vocabulary knowledge to the actual usage of these words during L2 listening comprehension among 114 Japanese students studying English as a foreign language (EFL).

Background Literature

Vocabulary Knowledge in L2 Listening

Vocabulary knowledge is essential for successful L2 listening performance (e.g., Vafae & Suzuki, 2020; Vandergrift & Baker, 2015, 2018; Wallace, 2022). Earlier studies often demonstrate a varying yet robust and positive relationship between vocabulary and L2 listening, with values of the correlation coefficient r ranging from .38 (Mecarty, 2000) to .94 (Matthews et al., 2024). Meta-analytic findings of a mean correlation value of .56 confirm the moderate and positive relationship between vocabulary and listening measures (In'nami et al., 2022; Zhang & Zhang, 2022). Multivariate analyses of the vocabulary–listening link, while accounting for cognitive and affective individual differences (Vafae & Suzuki, 2020; Vandergrift & Baker, 2015, 2018; Wallace, 2022), also reveal that L2 vocabulary is the most reliable contributor to general L2 listening proficiency over and above participants' first language (L1) vocabulary, metacognition, auditory discrimination, L2 grammatical knowledge, memory capacity, anxiety, and topical knowledge.

The literature of vocabulary and listening has advanced to start exploring this issue further with the goal of determining the specific aspects of vocabulary knowledge relevant to L2 listening (Cheng et al., 2023; Cheng & Matthews, 2018; Matthews et al., 2024; McLean et al., 2015; Milton et al., 2010; Milton & Hopkins, 2006). One important issue concerns the test modality (spoken vs. written) in which vocabulary tasks are delivered. This issue is particularly relevant to the case of learners in instructional contexts where L2 spoken input outside the classroom is severely limited and a notable gap is observed in the knowledge of aural and written vocabulary (Hamada & Yanagawa, 2023; Uchihara, 2023; Uchihara & Harada, 2018).

Another emerging issue concerns the dimension of lexical knowledge (recognition vs. recall) that is tested. From a theoretical perspective of lexical development, meaning recognition is considered to indicate partial knowledge of the form–meaning connection of a word (Nagy et al., 1985), manifesting itself in the early stages of vocabulary development (Bordag et al., 2021). With increased exposure and practice, the representation of the word becomes more robust with a stronger form–meaning link, enabling learners to demonstrate the ability to not only recognize but also produce the meaning of the word (González-Fernández & Schmitt, 2020; Jiang, 2000; Laufer & Goldstein, 2004; Webb, 2007).

The majority of earlier studies have used meaning-recognition tasks in multiple-choice formats to measure knowledge of the form–meaning connection of L2 words (e.g., Stæhr, 2009; Vafae & Suzuki, 2020; Wallace,

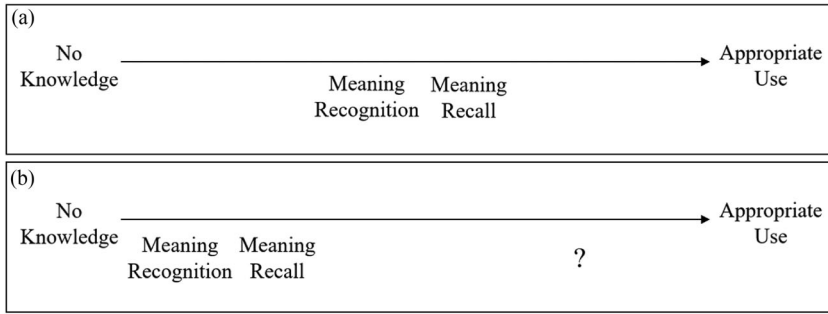


Figure 1 Hypothetical developmental trajectory of L2 vocabulary knowledge from no knowledge to the complete mastery required for appropriate use of words in L2 comprehension.

2022). From a standpoint of vocabulary assessment, researchers have pointed out potential issues with recognition tasks, including the confounding effect of guessing strategies (Gyllstad et al., 2015) and an incongruence between the cognitive processes elicited by meaning-recognition and listening tasks (Cheng et al., 2023). Alternatively, meaning-recall tasks such as L2 cued meaning production are considered more appropriate as this elicitation format likely reflects the psychological process involved in retrieving word meanings during speech comprehension. However, the meta-analysis by Zhang and Zhang (2022) indicated a comparable size of vocabulary–listening correlations between meaning recall, $r = .63$, 95% CI [.53, .72], and meaning recognition, $r = .58$, 95% CI [.54, .62]. Perhaps, despite some differences in the quality of mapping knowledge, both tasks may tap into a very similar aspect of word knowledge (i.e., declarative knowledge of form–meaning connections).

Despite mounting evidence for the important role of meaning-recognition and meaning-recall knowledge in L2 listening proficiency, building the declarative knowledge of form–meaning mapping is not sufficient for learners to achieve advanced L2 listening ability (Nation, 2022; Schmitt, 2019). Based on the framework of the strength of form–meaning knowledge (Laufer & Goldstein, 2004) and a developmental perspective on vocabulary acquisition (Schmitt, 2017, 2019), the hypothetical developmental sequence of phonological vocabulary knowledge relevant to L2 listening is charted. As illustrated in Figure 1(a), increasing the precision of lexical knowledge from a recognition to recall level can be regarded as a step forward along a developmental continuum, helping the word to be slightly more employable in speech comprehension. However, the reality may be better reflected in Figure 1(b) rather

than 1(a), with much larger space remaining before the attainment of complete mastery where the word is made readily available for use in authentic listening activities. Nation's (2022) comprehensive view of lexical development may give a useful insight into this issue. According to this framework, advanced lexicons can be described by the increase in the number of words for which form–meaning connections are developed (*size*), the enhancement of various knowledge aspects for an individual word (*depth*), the refinement of precision of each knowledge aspect (*strength*), and the enrichment of lexical networks within and between words (*organization*). Thus, it is logical to presume that advanced L2 listeners should have sizable, multilayered, precise, and rich lexical knowledge. Yet, this framework is rather generic, and little remains known about the specific characteristics of advanced lexical knowledge relevant to successful speech comprehension.

Automatized Versus Declarative Phonological Vocabulary Knowledge

To move forward the understanding of the relationship between vocabulary knowledge and listening, it is essential to understand how phonological vocabulary develops into advanced knowledge after the basic knowledge of form–meaning mapping is acquired. The usage-based account of language comprehension states that with continued exposure to L2 input, a learner becomes “an optimal word processor” (Ellis, 2006, p. 2). Efficiency in retrieving L2 word meanings improves as a function of frequency (how often a word has been encountered in the past), recency (how long ago the word was last accessed), and context (what word it often occurs with in immediate contexts). From this perspective, advanced knowledge of a word can be characterized not only by enhancement of declarative or episodic memories of its form–meaning mapping mostly due to explicit training in classroom settings (Jiang, 2000; Nation, 2022), but also the further refinement of the knowledge as a result of encoding the probabilities of co-occurrence of the word with other words through context-driven implicit learning (Ellis, 2022).

The skill acquisition model for instructed L2 acquisition also provides further insights into the developmental trajectories of phonological vocabulary knowledge (DeKeyser, 2020; Suzuki, 2023). Under this view, learners first build declarative knowledge of form–meaning mapping for a novel word through explicit vocabulary training (Jiang, 2000). At this stage of building declarative knowledge, learners know about what words sound like (i.e., knowledge of spoken forms) and what kind of semantic information they signify (i.e., form–meaning knowledge). With such declarative knowledge alone, learners can recognize or recall the meanings of the word when cued

by L2 forms via, for instance, multiple-choice or translation tasks. At the subsequent stage, knowledge of how words are used in global contexts (i.e., use-in-context knowledge) develops with increased exposure to L2 input (proceduralization). Sustained contextual exposure continues to refine and enhance lexical representations while learners encode information about a word's usage and its co-occurrence with other words (Landauer et al., 1998; Webb, 2007). By employing various lexical cues including semantic, grammatical, and collocational information, learners can eventually acquire the ability to retrieve learned vocabulary knowledge with high levels of automaticity and stability (automatization). A key characteristic of automatized lexical knowledge is the ability to retrieve L2 word meanings fluently and consistently under varying processing conditions (DeKeyser, 2020). Declarative knowledge may be sufficient to retrieve accurate meanings for spoken words presented in isolation through explicit analyses of individual words. However, automatized knowledge is required when multiple words are spoken in more taxing and communicatively authentic contexts, given that lexical processing needs to be executed while attentional resources are directed for simultaneous processing of other aspects of language in context (e.g., morphology, syntax, and discourse; Ellis et al., 2008).

In the L2 morphosyntax literature, automatized L2 knowledge (i.e., accurate and fluent use of acquired knowledge) is often assessed using acceptability judgment tasks (Spinner & Gass, 2019). In experiments aiming to measure automatized morphosyntactic knowledge, L2 learners hear or read a set of short sentences within a restricted time frame, some of which are grammatically incorrect, and judge whether they are accurate (for a review of grammaticality judgment tasks, see Plonsky et al., 2020). It has been shown that the timed grammaticality judgment results are substantially different from controlled measures of grammatical knowledge (e.g., fill-in-the-blanks; Gutiérrez, 2013; for a discussion of the construct validity of grammar tests, see Vafae & Kachinske, 2019). Such automatized explicit knowledge may underpin the development of implicit knowledge (Suzuki & DeKeyser, 2017). In a comprehensive review of existing measures of automatized L2 knowledge, Suzuki and Elgort (2023) pointed out the lack of attention to measuring auditory lexical processing in the L2 literature (the main focus of the present study).

In the context of L2 spoken word recognition, when L2 phonological vocabulary knowledge is automatized, learners are considered to store it together with strongly collocated words as a chunk. Such advanced lexicons should promote fluent and accurate lexical retrieval under varying processing conditions (Ellis et al., 2008; Fitzpatrick & Izura, 2011; Tavakoli & Uchihara, 2020).

Building upon the literature of measuring automatized grammatical knowledge (Plonsky et al., 2020), we propose that acceptability judgments be applied to assessing automatized knowledge of phonological vocabulary. Upon hearing sentences containing contextually correct and incorrect word usage, advanced listeners should be able to instantly judge whether the target words are contextually appropriate or not. Such tasks require listeners to employ knowledge of semantic and collocational properties of target words in relation to surrounding contexts and evaluate the appropriateness of the word usage in a spontaneous manner.

To date, very few studies have used acceptability judgment tasks to assess L2 phonological vocabulary knowledge. The vast majority of earlier studies (e.g., Elgort, 2011; Sonbul & Schmitt, 2013) adopted lexicality judgment tasks in the written modality to measure accuracy and fluency in recognizing individual words or judging the idiomaticity of phrases presented out of context (for a review, see Suzuki & Elgort, 2023). The incongruence of the test modality between vocabulary measures (written) and listening comprehension (spoken) is problematic as it may misleadingly attenuate the important link that could be revealed if the congruence of the test modality was rigorously considered (Cheng & Matthews, 2018; Milton et al., 2010). As one exception, Saito et al. (2023) utilized an acceptability judgment task within the framework of L2 speech recognition to measure automatized phonological vocabulary knowledge. Their task, a lexicosemantic judgment task (LJT), was purported to measure spontaneous and contextualized recognition of the meanings of 80 target words (e.g., *publish*) under time pressure. Japanese learners of EFL took the LJT, auditory multiple-choice tests as a measure of meaning recognition, and a listening proficiency test. For the LJT procedure, test takers heard a set of short sentences, half of which were contextually appropriate (e.g., *He has published many books*) and the other half inappropriate (e.g., *Mary published her left hand*), and judged the appropriateness of each sentence as quickly as possible. Their results showed that although aural meaning recognition moderately correlated with the LJT ($r = .50$), the LJT was a more reliable predictor of general listening proficiency ($r = .66$) than was meaning recognition ($r = .43$).

Nation's (2022) componential framework of word knowledge clarifies the distinctiveness of the construct measured via the LJT from declarative knowledge measured through meaning-recognition and -recall tasks. Aural meaning-recognition and meaning-recall tests primarily involve recognition of acoustic input as meaningful words (knowledge of spoken form) and retrieval of the meanings cued by the identification of the spoken forms (knowledge of form–meaning connection). In contrast, the LJT taps additional components of

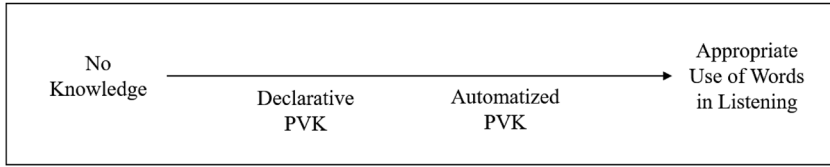


Figure 2 Hypothetical developmental trajectory of phonological vocabulary knowledge (PVK) towards development of the ability to use L2 words appropriately in real-life L2 listening.

word knowledge pertaining to the use-in-context aspects (collocations, grammatical functions, and constraints on use). It induces evaluating the degree of collocational associations between the target and surrounding words and retrieving the most contextually plausible combinations (e.g., *publish* [verb] + *books* [object] vs. *publish* [verb] + *hand* [object]). This seemingly simplistic definition of the two constructs in light of the number of aspects involved can be reconceptualized on a developmental cline in terms of the lexical retrieval efficiency required for advanced listening performance (see Figure 2). In this regard, automatized knowledge (proxied by the LJT), enabling efficient retrieval of meanings cued by multiple knowledge sources (spoken form, form–meaning link, use-in-context aspects), can be placed somewhere closer towards the endpoint indicating the appropriate use of L2 words in real-world speech comprehension. This knowledge-integrated approach to conceptualizing advanced lexical competence is distinct from the componential approach widely utilized to date in L2 vocabulary research (Nation, 2022). Studies taking the componential view assess different lexical aspects separately with multiple discrete tests (e.g., translation tests to assess knowledge of single-word items and multiword units) and relate them to L2 listening proficiency (Cheng et al., 2023; Matthews et al., 2024). These studies have provided novel insights into the vocabulary–listening relationship, underscoring the multifaceted nature of phonological vocabulary knowledge. Building on this line of emerging research, however, we now adopt a novel approach, shifting our focus to L2 word employability and taking the integrative (rather than componential) perspective in order to define word knowledge closely relevant to L2 listening success.

In essence, we define declarative phonological vocabulary knowledge as the construct that concerns understanding the connections between forms and meanings at the individual word level. Automatized phonological vocabulary knowledge, in contrast, spans a wider array of capabilities. It allows for rapid

and consistent retrieval of declarative knowledge, enabling the use of vocabulary in collocationally, grammatically, and contextually appropriate ways across entire sentences. This multifaceted and integrated nature distinguishes automatized phonological vocabulary knowledge from its declarative counterpart. Accordingly, the LJT is specifically designed to assess these varied skills that are critical to the process of automatization.

The Current Study

Rationale

Adopting a componential view of vocabulary knowledge (Nation, 2022), recent studies have documented the close relationship between vocabulary knowledge and L2 listening proficiency by using a range of independent vocabulary tests (e.g., multiple choice, translation, lexical decision). Since the call for attending to the modality of vocabulary tests was made in Stæhr (2009), the field has witnessed an increase in the number of studies testing phonological vocabulary and establishing its vital role in L2 listening success (e.g., Milton et al., 2010; Zhang & Zhang, 2022). Scholars in this paradigm have also expanded their scope to demonstrate the unique contribution to L2 listening proficiency made by knowledge of multiword expressions (e.g., phrasal verbs; Cheng et al., 2023; Matthews et al., 2024). These studies have made a novel contribution to our understanding of the link between vocabulary and listening proficiency as they aim to measure multiple aspects of word knowledge beyond that of single-word items. Building on this line of work, the present study, adopting a more integrated and comprehensive approach, aims to measure the ability to access L2 words in a collocationally, grammatically, and contextually appropriate manner at sentence level. A distinctive aspect in our approach lies in the focus on sentence-level lexical retrieval, which allows us to assess the real-time use of vocabulary in a way that more accurately reflects the cognitive processes involved in the fluent retrieval of appropriate word meanings that is required for L2 listening success.

In this line of reasoning, Saito et al. (2023) drew on the L2 skill acquisition theory to conceptualize automatized phonological vocabulary knowledge and attempted to measure this knowledge through an acceptability judgment task (a LJT). Although Saito et al. implied the important role of automatized phonological vocabulary in advanced L2 listening proficiency, their findings were limited to the results from two vocabulary tasks (namely, aural meaning recognition and the LJT); thus, the construct of declarative knowledge was only represented by the recognition-level or partial knowledge developed at the initial stage of L2 lexical development. We aim to address this limitation by

measuring meaning recall as advanced knowledge of form–meaning connection, considered to more accurately mirror a cognitive process of L2 listening (Cheng et al., 2023; Matthews et al., 2024).

Another key difference from Saito et al. (2023) is that we measure L2 listening comprehension using the test materials from which target words for vocabulary tests are sampled. In other words, participants encounter target vocabulary that appears in listening comprehension tests. This direct approach allows us to gauge the extent to which learners can employ declarative and automatized knowledge of L2 target vocabulary in real-life L2 listening comprehension (i.e., word employability; Schmitt, 2019). Hence, this research informs the developmental perspective of phonological vocabulary acquisition by evaluating the theoretical distance between declarative knowledge (represented by meaning recognition and recall), automatized knowledge (by the LJT), and actual usage of target words during L2 comprehension (the rightmost end of the developmental continuum in Figure 2).

Therefore, the current study is designed to examine how 114 Japanese EFL listeners recognize and recall the meanings of 80 target words when they are aurally presented on two tests of declarative phonological vocabulary knowledge (a multiple-choice test and a translation test) and one test of automatized phonological vocabulary knowledge (a LJT), and how their test scores can differentially predict their actual usage of these words during L2 listening comprehension of monologues and conversations. The findings of this study are expected to provide additional insights into a multifaceted construct of phonological vocabulary knowledge (automatized and declarative knowledge) and update the understanding of how lexical knowledge is related to the employability of L2 words in speech comprehension.

Research Questions and Predictions

Our research questions (RQs) were as follows:

1. How are different aspects of phonological vocabulary knowledge associated with the employability of L2 words in listening comprehension?
2. Is automatized phonological vocabulary knowledge, as an independent construct, more accurately reflective of the employability of L2 words in listening comprehension compared to declarative phonological vocabulary knowledge?

Regarding RQ 1, based on prior work (Saito et al., 2023) and the developmental framework of L2 vocabulary acquisition (Schmitt, 2017, 2019), we

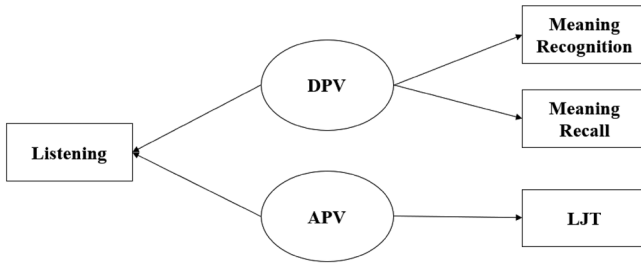


Figure 3 Model 1 (automatized–declarative model). DPV = declarative phonological vocabulary; APV = automatized phonological vocabulary; LJT = lexicosemantic judgment task.

predicted that automatized phonological vocabulary knowledge (measured by the LJT) would be more closely associated with participants’ actual usage of the target words during L2 listening comprehension than would declarative phonological vocabulary knowledge (represented by meaning recognition and meaning recall). Regarding the relative contribution of meaning recognition and recall to L2 listening comprehension, we predicted that meaning-recall knowledge would be more strongly associated with L2 listening than meaning-recognition knowledge, given that recall tests are regarded as a more cognitively valid and reliable measure of form–meaning knowledge (Matthews et al., 2024; Zhang & Zhang, 2022).

Regarding RQ 2, we employed structural equation modeling analysis to test the hypothesis that automatized phonological vocabulary knowledge would be an independent and stronger predictor of participants’ actual usage of the target words during L2 comprehension of monologues and conversations compared to declarative phonological vocabulary knowledge. Our hypothesized model (the automatized–declarative model) consisted of two factors distinguishing automatized knowledge (LJT) from declarative knowledge (meaning recognition, meaning recall), both of which would contribute to L2 listening test scores (Figure 3). We also built an alternative, equally plausible model (the recall–recognition model), contrasting recall (meaning recall) with recognition (meaning recognition, LJT), for comparison against our original two-factor model (Figure 4). The comparison of the two models was motivated by the notion of dimensions of word knowledge (recall vs. recognition: González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004). Given that the LJT is essentially a recognition task in which listeners judge whether the utterance of a short sentence makes sense (accept or reject), it is theoretically

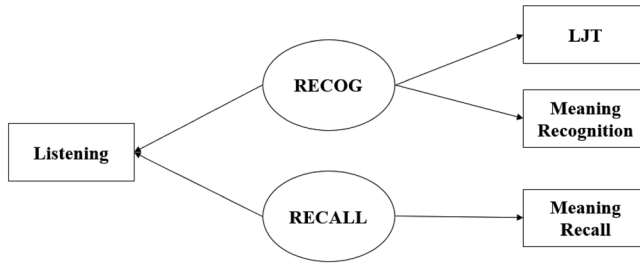


Figure 4 Model 2 (recall–recognition model). RECOG = recognition; LJT = lexicosemantic judgment task.

plausible to hypothesize that a meaning-recognition multiple-choice task and the LJT would tap into a similar construct (recognition knowledge). Based on the developmental perspective of phonological vocabulary acquisition, we predicted that the automatized–declarative model would be retained against the recall–recognition model and automatized knowledge would be a stronger predictor of participants’ vocabulary use during L2 listening comprehension than declarative knowledge.

Method

Participants

A total of 114 Japanese university students studying EFL in Japan participated in this study (52 males and 62 females, $M_{\text{age}} = 20.3$ years, range = 18–26). All students had received English education in junior high and high school in Japan; they were studying in various departments at the time of this experiment. The majority of participants started learning English in primary school (77%). Participants’ L2 English proficiency varied considerably, ranging from levels A2 to C1 of the Common European Framework of Reference for Languages (CEFR) based on their Test of English for International Communication (TOEIC) listening scores. The mean score of English vocabulary knowledge measured with the Lexical Test for Advanced Learners of English (LexTALE; Lemhöfer & Broersma, 2012) was 63.5% ($SD = 8.5\%$), indicating that participants’ vocabulary size was estimated to be around 8,000 word families based on a previous study exploring Japanese learners’ English vocabulary size by using multiple vocabulary tests (Nakata et al., 2020). No participants reported hearing difficulties.

Listening Test

In this study, we assessed different facets of participants' knowledge (both declarative and automatized) of the 80 target words by using meaning-recognition, meaning-recall, and LJT methods (for details, see below). To gauge how successfully participants actively accessed these words during holistic L2 listening comprehension, we employed a retired version of the TOEIC Listening test. This test encompassed a variety of real-life monologues and conversations in which the 80 target words were embedded. We posit that participants' comprehension of this aural content can indicate their actual usage of the target words during L2 listening.

In Japan, the TOEIC exam is commonly used to evaluate the listening skills of Japanese adult learners (as mentioned in studies by Cheng et al., 2023; Hamada & Yanagawa, 2023; Matthews et al., 2024; McLean et al., 2015). The materials used in this study were adapted from the New Official Workbook provided by the Educational Testing Service in Volume 4. Participants recorded their answers on a scoring sheet via Google Forms. This test was divided into three sections: Part 2, Part 3, and Part 4. In Part 2 ($k = 30$), participants had to choose the best response from three options for single-sentence questions (with 5–10 words). In Part 3 ($k = 30$), participants listened to a conversation between a male and a female speaker (consisting of 80–100 words) and answered three comprehension questions by selecting the most appropriate response from four options. In Part 4 ($k = 30$), participants heard a business announcement delivered by a single speaker (80–100 words) and answered three comprehension questions by choosing the best response from four options.

Vocabulary Measures

We developed three computerized vocabulary tests to assess different aspects of phonological vocabulary knowledge. We adopted an aural multiple-choice test to measure meaning recognition, an aural L1-cued translation test to measure meaning recall, and a LJT to assess contextualized meaning recognition. The example test formats and materials, including target words, answer keys, distractor items, and test sentences, can be found in the Supporting Information online (Appendix S1 for target items and distractors, Appendix S2 for test sentences, and Appendix S3 for sample test formats) and via the OSF (<https://osf.io/jgudx/>).

Target Vocabulary

To investigate the direct relationship between vocabulary knowledge and the ability to aurally process L2 words in global contexts, we selected 80 target words from the TOEIC listening passages used in this study. We initially compiled a speech corpus based on scripts from the TOEIC listening test. In total, 2,731 tokens used in the passages were evaluated based on the following three criteria, and the top 80 most phonologically and lexically challenging words were extracted:

1. **Word frequency:** We prioritized the selection of less frequent words using Nation's (2012) BNC/COCA word family lists (where related forms such as *happy*, *happier*, and *unhappy* are treated as part of the same family). These lists (based on corpus data) divide word families into frequency bands: the first 1,000 most frequent ("1K"), the second 1,000 most frequent ("2K"), and so on.
2. **Cognateness:** We excluded cognates as they might aid in L2 comprehension (Uchihara, 2023). Three experienced L1 Japanese teachers evaluated and determined the cognate status of target words.
3. **Phonological difficulty:** We prioritized L2 words that contain difficult segmentals (English [r] and [l]), consonant clusters, and multiple syllables with primary stress not on the first syllable (Field, 2005; Saito, 2014).

As a result, the 80 words selected for this research were distributed as follows: 22 words from the 2K frequency band, 35 words from the 3K band, 13 words from the 4K band, and 10 words from the 5K–8K range. The decision to focus more on the 2K and 3K words, which accounted for 57 out of the 80 words, was based on the rationale that previous findings, as demonstrated by Matthews (2018), indicated that lexical knowledge of frequently occurring vocabulary items, especially those in the high- and mid-frequency categories, had a substantial influence on the listening test scores of EFL learners. The selection process did not consider the relevance or importance of the words for answering the comprehension questions; instead, we took an inclusive approach and selected all potentially unfamiliar and phonologically challenging words.

Meaning Recognition

A multiple-choice test was adopted to elicit learners' recognition of target word meanings from phonological forms (see Figure S3.1 in

Appendix S3 of the Supporting Information online for a sample test format). In this test, the participants heard a target word once and chose as quickly as possible the right L1 meaning out of four options (L1 Japanese meanings) orthographically presented on a computer screen. A female native speaker of American English recorded the pronunciation of the target words. The parts of speech for the distractor items matched those in the answer key, and all the distractor items were chosen from a list of words commonly found in TOEIC test materials. Before we conducted this study, three Japanese speakers with EFL teaching experience reviewed the test materials. They addressed concerns related to translations of answers and distractors in the participants' native language. We replaced answer key meanings that did not align with the context of the passage and made revisions to distractor items that could potentially be misinterpreted as correct answers.

Meaning Recall

A L2-to-L1 translation test was employed to elicit learners' recall of target word meanings prompted by the aural forms of words (see Figure S3.2 in Appendix S3 of the Supporting Information online for a sample test format). In this test, the participants heard a target word once and typed a L1 translation for the word within 15 s. The time restriction was added for the purpose of maintaining participants' focus on the task and preventing them from pausing for a long time in the middle of the test. The appropriateness of the time for answering questions was determined based on a pilot study. Target stimuli were recorded by the same native speaker who recorded the stimuli for the meaning-recognition test. In scoring the elicited responses, two Japanese speakers first coded 474 responses collected from a pilot study with participants whose data were not included in the main study. After the intercoder agreement reached 100%, one coder proceeded to score responses from all participants in this study.

Lexicosemantic Judgment Task

The LJT was adopted to measure learners' ability to spontaneously judge the contextual and semantic appropriateness of target word meanings (see Figure S3.3 in Appendix S3 of the Supporting Information online for a sample test item). According to Nation's (2022) model of word knowledge, the LJT is purported to tap into aspects of form (spoken forms), meaning (form–meaning connection), and use-in-context (collocations, grammatical functions, and constraints on use). These knowledge sources are meant to be employed in

an integrated manner to feed into optimization in retrieving target word meanings in sentential contexts.

In this judgment task, participants encountered short sentences once, and upon hearing each sentence, they had to decide whether it was “semantically appropriate” or “semantically inappropriate” as quickly as possible. Each sentence featured a specific target word, and to ensure that listeners paid attention to the entire sentence, these target words were not placed at the beginning. The majority of the words in stimuli sentences ($k = 160$) were chosen from the 1K frequency band (the 1,000 most frequent word families) or were familiar proper names, altogether constituting 93% of all the words used in the sentences. Although a small fraction (7%) came from the 2K frequency band (the second 1,000 most frequent word families), these were mainly words present in Japanese as loanwords. In half of the sentences (80 out of 160), the target words were used in a way that made sense within the context (semantically appropriate). In the other 80 sentences, the target words were used in a manner that did not fit the context (semantically inappropriate). For example, if the target word was *estate*, participants heard the semantically appropriate sentence *My grandfather bought an estate* and the semantically inappropriate sentence *My friend's estate was very kind*. Learners can quickly and intuitively accept the former, appropriate sentence if they recognize the auditory form of the word as *estate*, know the meaning mapped to the spoken form, and have an intuition of the probability of its occurrence that it is frequently used with *bought*. On the other hand, learners can reject the latter, inappropriate sentence if they know the word's form and meaning and draw on the knowledge that *estate* does not match with an adjective describing personality (semantically incongruent) and rarely occurs with *kind* (low co-occurrence probability).

The sentences were kept short, ranging from 3 to 8 words, to avoid making excessive demands on participants' working memory while they took the test. To mitigate the possibility that the test would primarily assess knowledge of syntactic structures and speech perception skills, we maintained syntactic simplicity, without using any complex subordination, and ensured that a native speaker recorded careful productions of each stimulus sentence rather than natural-speed utterances (see Appendix S5 in the Supporting Information online for preliminary analyses of the influence of item-level characteristics on the LJT scores). Thus, in these sentences, aside from the target word, everything else remained syntactically accurate and comprehensible, making the appropriateness of the target word the sole determinant of semantic correctness.

It is important to note that the judgment of semantic appropriateness can be valid only so long as it does not involve individual variations in learners' prior or world knowledge. To ensure that participants' prior knowledge would not confound the test results, researchers first drafted an initial pool of prompt sentences, and the candidate sentences were rated by two native speakers of English for contextual appropriateness (1 = *definitely true*, 2 = *probably true*, 3 = *not sure*, 4 = *probably false*, 5 = *definitely false*). Revision and rewriting of the test sentences continued until all sentences were rated as unambiguously appropriate (1 = *definitely true*) or inappropriate (5 = *definitely false*). The 160 sentences of the finalized list were recorded by the same native speaker who recorded the meaning-recognition and meaning-recall stimuli.

Procedure

Because of the COVID-19 pandemic restrictions on in-person testing, we recruited participants through online advertisements, and the data collection process spanned three separate days conducted in a virtual environment. Students who visited the recruitment webpage with interest filled out an eligibility questionnaire online. We excluded students who had not received formal education in Japan (e.g., returnees and international students) and those who did not speak Japanese as their L1.

Eligible participants were given the choice of three time slots for completing tests. On the first day of the experiment, a group of around 10 participants assigned to the same time slot met with one or two research assistants via Zoom. Following a brief explanation of the experiment by the research assistants, participants consented to take part. Before beginning the listening test, all participants were instructed to wear headsets, check their sound volume, and turn on their cameras. This allowed the research assistants to monitor how participants took the tests. Participants took the TOEIC listening test (lasting 40 min) in the presence of the research assistants. After finishing the TOEIC test, participants signed out from the online meeting platform and took the LexTALE (lasting 5 min). Throughout these tasks, the research assistants kept track of task progress and were available online to offer real-time support to those facing technical issues. On the second day of the experiment (within a few days after the first day of the experiment), participants began with the LJT (lasting 20 min) and then engaged in distractor memory tasks, which involved memorizing sequences of random numbers. Finally, they completed the meaning-recognition test (lasting 10 min). The LJT was administered first because completing the decontextualized test of meaning recognition before the LJT might draw participants' attention to target words, potentially inducing

their selective attention to the words during the judgment task. On the last day of the experiment (approximately 1 month after the second day of the experiment), participants took the meaning-recall test (lasting 15 min).

The order of the test administration for meaning-recognition and meaning-recall tasks did not follow the recommended procedure for avoiding practice effects (recall → recognition) for logistical reasons. However, such potential effects were considered minimal, given that (a) all participants took the meaning-recall test 1 month after the completion of the meaning-recognition test, and (b) the transitory nature of the auditory stimuli in the recognition test (played only once) made it unrealistic for the participants to check the meanings of the target words afterwards and study them before taking the recall test.

All vocabulary tests were computerized, and participants' responses were recorded using Gorilla, an online experiment builder (Anwyl-Irvine et al., 2020). The research assistants and researchers monitored the completion of tasks online. The order in which target items were presented in all vocabulary tests was randomized for each participant. Additionally, participants were given practice questions before each of the three vocabulary tests.

Data Analysis

For the meaning-recognition and meaning-recall tests, 1 point was awarded for each correct response (max. = 80). For the LJT, responses of correctly accepting appropriate sentences ($k = 80$) and correctly rejecting inappropriate sentences ($k = 80$) were transformed to a binary score with responses to both appropriate and inappropriate sentences defined as correct (max. = 80). This means that 0 points were awarded if participants did not correctly respond to either an appropriate or an inappropriate sentence (or both). Prior to conducting main analyses, we confirmed that all the vocabulary-task and listening-test scores showed adequate internal consistency: LJT, $\alpha = .90$, 95% CI [.88, .92]; meaning recognition, $\alpha = .88$, 95% CI [.86, .91]; meaning recall, $\alpha = .91$, 95% CI [.89, .93]; and TOEIC, $\alpha = .93$, 95% CI [.91, .95]. The descriptive statistics showed that all the scores were normally distributed (the absolute values of skewness statistics for all test scores were greater than 2.0).

To answer the first research question, regarding the relationship between phonological vocabulary knowledge and employability of target words in listening comprehension, we conducted standard multiple regression analysis, using the `lm` function of R (R Core Team, 2023). We created a regression model with the TOEIC test scores as the dependent variable and the LJT, meaning-recognition, and meaning-recall test scores as independent variables.

We used the *car* package (Fox & Weisberg, 2019) to produce the quantile–quantile plot and variance inflation factors (VIFs) for each predictor variable (obtaining VIFs of 1.6 for the LJT, 2.3 for meaning recall, and 2.3 for meaning recognition, whereas $VIF > 5.0$ indicates problematic collinearity; Heiberger & Holland, 2004) and confirmed that the current data set met the assumptions of multiple linear regression (i.e., normality of residuals and absence of multicollinearity). To determine the importance of each predictor variable in the model while accounting for correlations between all predictor variables (Mizumoto, 2022), we used dominance analysis with the *domir* package (Luchman, 2023).

To answer the second research question, asking whether automatized vocabulary knowledge (as opposed to declarative vocabulary knowledge) uniquely contributes to listening performance, we implemented structural equation modeling analyses in Mplus, Version 8.4 (Muthén & Muthén, 1998–2017). We specified the original model with automatized and declarative vocabulary factors as the predictors of listening performance (Figure 3) and the alternative model with recall and recognition knowledge factors as the predictors of listening performance (Figure 4). For each of the models, automatized and recall knowledge factors were established as the single indicator while measurement error was accounted for in the structural model (Kline, 2016). Compared to the traditional approach (i.e., adding an observed variable to a structural model), explicitly representing measurement error using the single-indicator method is considered favorable as it enhances the precision of estimated coefficients while not changing overall model fit (Kline, 2016). Following the recommended procedure, we calculated the estimated proportion of total variance due to random error based on the reliability coefficient (Cronbach’s alpha) for each of the lexical measures (the LJT and meaning recall), using the following formula: $\text{observed variance} \times (1 - \alpha)$. To scale the single-indicator latent variable, the unstandardized pattern coefficients for the LJT and meaning recall were fixed to 1.0. The maximum likelihood technique was used as the method of model parameter estimation. Following suggestions made regarding model-fit evaluation (Hu & Bentler, 1999; Kline, 2016; Vafae & Kachinske, 2019), we adopted the following model indices: (a) comparative fit index ($CFI \geq .95$), (b) Tucker–Lewis fit index ($TLI \geq .95$), (c) root-mean-square error of approximation ($RMSEA \leq .06$), (d) standardized root-mean-square residual ($SRMR \leq .08$), (e) Akaike information criterion (AIC), and (f) Bayesian information criterion (BIC). A full summary of statistical analyses, including descriptive and inferential statistics, data, and analysis codes,

can be found in the Supporting Information online (Appendix 4 for descriptive statistics) and via the OSF (<https://osf.io/jgudx/>).

Results

Relationships Between the Lexicosemantic Judgment Task, Meaning Recognition, and Meaning Recall

The results of Pearson correlation analyses showed that meaning recognition was strongly correlated with meaning recall, $r = .72$, 95% CI [.62, .80], $p < .001$. The LJT was moderately correlated with meaning recognition, $r = .58$, 95% CI [.44, .69], $p < .001$, and with meaning recall, $r = .58$, 95% CI [.44, .69], $p < .001$. Although all test scores were positively correlated, the relatively larger correlation between meaning recognition and recall (shared variance = 52%) indicated that the two tests tapped similar aspects of word knowledge (recognition of spoken form and form–meaning mapping). A paired-sample t test demonstrated that the meaning-recognition scores were higher than the meaning-recall scores, $t = 25.27$, $p < .001$, $d = 2.37$, 95% CI [2.01, 2.72], and than the LJT scores, $t = 25.31$, $p < .001$, $d = 2.37$, 95% CI [2.01, 2.73]. The meaning-recall scores were higher than the LJT scores, $t = 4.97$, $p < .001$, $d = 0.47$, 95% CI [0.27, 0.66]. These results indicated varying degrees of difficulty across three vocabulary measures (LJT > meaning recall > meaning recognition) with a relatively large gap between meaning recognition and the other two measures.

The Relationship Between Declarative and Automatized Phonological Vocabulary and Actual Usage of Target Vocabulary in Listening Comprehension

The results of Pearson correlation analyses showed that all vocabulary tests significantly correlated with TOEIC listening: LJT, $r = .71$, 95% CI [.61, .79], $p < .001$; meaning recall, $r = .62$, 95% CI [.50, .73], $p < .001$; and meaning recognition, $r = .57$, 95% CI [.43, .68], $p < .001$. As summarized in Table 1, the results of multiple regression analysis with dominance analysis showed that the three vocabulary measures together explained 58% of the total variance in TOEIC listening scores ($R^2 = .58$). Regarding the unique contribution of each predictor variable, the LJT explained the most variance ($R^2 = .29$), followed by the meaning-recall test ($R^2 = .17$). The meaning-recognition test was not a significant predictor ($p = .368$) when the LJT and meaning-recall scores were controlled for. Of the three predictor variables, the LJT accounted for 49.95% of the total variance explained by the three vocabulary measures and can thus

Table 1 Summary of multiple regression of three vocabulary measures predicting listening comprehension scores

Variable	<i>b</i>	<i>SE</i>	95% CI (<i>b</i>)		<i>t</i>	<i>p</i>	Dominance weight (%)	Ranks
			<i>LL</i>	<i>UL</i>				
Intercept	21.64	8.41	4.97	38.31	2.57	.011		
Meaning recognition	0.15	0.17	-0.18	0.49	0.90	.368	.12 (21.09)	3
Meaning recall	0.30	0.10	0.09	0.51	2.88	.005	.17 (28.95)	2
LJT	0.52	0.08	0.36	0.68	6.41	< .001	.29 (49.95)	1
Total							.58 (100)	

Note. *LL* = lower limit; *UL* = upper limit; LJT = lexicosemantic judgment task. $R^2 = .58$; adjusted $R^2 = .57$.

be considered a more influential predictor than meaning recall (28.95%) or meaning recognition (21.09%).

Automatized Versus Declarative Phonological Vocabulary in Relation to Actual Usage of L2 Words in Listening Comprehension

As summarized in Table 2, the results of structural equation modeling analysis showed that the automatized–declarative model fitted data well, whereas the goodness-of-fit indices for the recall–recognition model indicated poor model fit (only the SRMR was within the range of acceptable fit). The results of the AIC and BIC confirmed that the automatized–declarative model explained data more accurately than the recall–recognition model ($\Delta AIC = 14.645$, $\Delta BIC = 14.645$). A close inspection of the model parameters for the recall–recognition model (Figure 6 and Table 3) revealed that the factor correlation (RECOG–RECALL) was considered extremely high ($r = .901$), potentially compromising the discriminant validity of the constructs (Rönkkö & Cho, 2022). Thus, a one-factor model with three vocabulary measures loaded onto a single latent factor of general phonological vocabulary knowledge was built (Figure 7) and compared with the automatized–declarative model. The chi-square difference test showed that the automatized–declarative model fitted data significantly better than the one-factor model, $\chi^2_{diff}(1, N = 114) = 17.18$, $p < .001$ ($\Delta AIC = 44.198$, $\Delta BIC = 41.566$). The results of parameter estimates for the automatized–declarative model (Figure 5 and Table 3) showed that despite the relatively high correlation between automatized and declarative phonological

Table 2 Model fit indices for the hypothesized models

Criterion	$\chi^2(df)$	CFI	TLI	RMSEA	SRMR	AIC	BIC
		$\geq .95$	$\geq .95$	$\leq .06$	$\leq .08$		
Automatized-declarative model	0.871 (1), $p = .351$	1.000	1.000	.000	.007	3248.282	3283.853
Recall-recognition model	15.517 (1), $p < .001$.938	.626	.357	.038	3262.927	3298.498
One-factor model	18.054 (2), $p < .001$.934	.803	.257	.040	3292.480	3325.419

Note. CFI = comparative fit index; TLI = Tucker–Lewis fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Table 3 Structural equation model results

Path	<i>b</i> (<i>SE</i>)	95% CI	<i>p</i>
Model 1: Automatized and declarative knowledge in predicting L2 listening comprehension			
APV → Listening	0.566 (0.270)	[0.218, 0.841]	.036
DPV → Listening	0.287 (0.271)	[−0.057, 0.568]	.291
APV ↔ DPV	0.740 (0.095)	[0.529, 0.898]	< .001
Model 2: Recall and recognition knowledge in predicting L2 listening comprehension			
RECALL → Listening	−0.555 (2.253)	[−10.872, 0.274]	.805
RECOG → Listening	1.343 (2.244)	[0.542, 11.565]	.549
RECALL ↔ RECOG	0.901 (0.087)	[0.669, 0.988]	< .001
Model 3: Phonological vocabulary as a single factor in predicting L2 listening comprehension			
PVK → Listening	0.796 (0.058)	[0.665, 0.893]	< .001

Note. APV = automatized phonological vocabulary; DPV = declarative phonological vocabulary; PVK = phonological vocabulary knowledge; RECOG = recognition.

vocabulary knowledge ($r = .740$), automatized knowledge (APV: $b = 0.566$, $SE = 0.270$, 95% CI [0.218, 0.841], $p = .036$) was regarded as a more reliable and stronger predictor of listening comprehension than declarative knowledge (DPV: $b = 0.287$, $SE = 0.271$, 95% CI [−0.057, 0.568], $p = .291$).

Additionally, we further examined the indirect effect of the declarative-knowledge factor on listening performance via the automatized-knowledge factor (DVK → AVK → Listening) using bias-corrected bootstrapped 95% confidence intervals (MacKinnon et al., 2007). The estimated indirect effect with its confidence interval above zero ($b = 0.419$, $SE = 0.419$, 95% CI [0.220, 0.802]) confirmed the presence of mediation, indicating that declarative knowledge (DVK) not directly but indirectly affected L2 listening performance via the development of automatized knowledge (AVK). These findings suggest that automatized and declarative phonological vocabulary knowledge contributed to actual usage of L2 words in real-life listening comprehension in a different way.

Discussion

Drawing on the skill acquisition account of instructed second language acquisition (Suzuki, 2023), the current study reconceptualized how two different

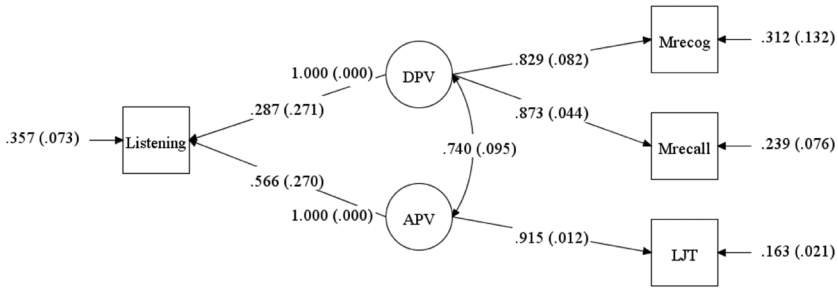


Figure 5 Automated–declarative model. DPV = declarative phonological vocabulary; APV = automatized phonological vocabulary; Mrecog = meaning recognition; Mrecall = meaning recall; LJT = lexicosemantic judgment task.

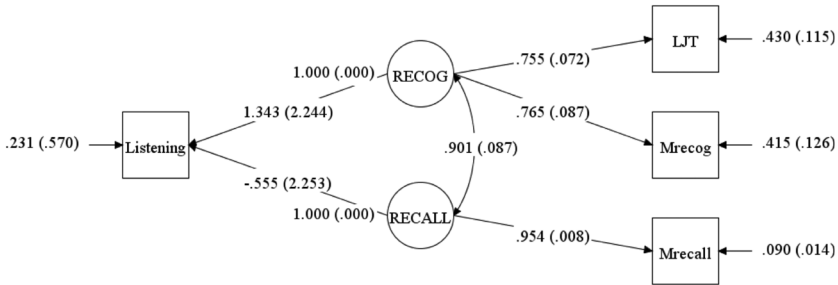


Figure 6 Recall–recognition model. RECOG = recognition; LJT = lexicosemantic judgment task; Mrecog = meaning recognition; Mrecall = meaning recall.

types of phonological vocabulary knowledge can reflect learners’ lexical processing during L2 listening comprehension. We hypothesized that participants’ phonological knowledge of form–meaning mapping (i.e., declarative knowledge) would continue to develop with L2 exposure and practice, and it would eventually become automatized so that learners can retrieve word meanings with efficiency and stability without being disturbed by cognitively demanding conditions (i.e., automatized knowledge). We proposed that such robust lexical knowledge be conceptualized as advanced phonological vocabulary knowledge that is closely tied to employability of word knowledge in real-life L2 listening. The findings of this study have provided initial evidence for the discriminant validity of the two vocabulary constructs (automatized and declarative knowledge) and the important role of automatized phonological vocabulary in successful word use during L2 listening comprehension. In what follows,

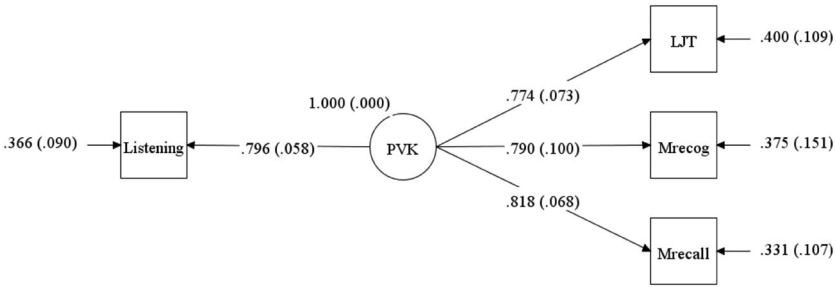


Figure 7 One-factor phonological vocabulary knowledge (PVK) model. LJT = lexico-semantic judgment task; Mrecog = meaning recognition; Mrecall = meaning recall.

we consider the results of this study in answer to each of the two research questions.

In response to RQ 1 (concerning the relationship between phonological vocabulary and L2 listening), the results showed that all vocabulary measures significantly correlated with listening comprehension. The finding that the correlation with L2 listening for meaning recall ($r = .62$) was slightly higher than, but comparable to, that for meaning recognition ($r = .57$) is consistent with findings from previous studies on vocabulary and L2 comprehension skills (Zhang & Zhang, 2022). However, the LJT revealed an even stronger correlation ($r = .71$) and explained by far the most variance in listening comprehension scores (dominance weight = 49.95%) compared with meaning recall (28.95%) and meaning recognition (21.09%). These findings suggest that automatized lexical knowledge (measured with the LJT), which allows learners to retrieve contextually constrained meanings of target vocabulary spontaneously, is what is indeed required by and more closely reflective of L2 speech processing in global and communicatively authentic contexts. To illustrate with the word *estate* as a target item, upon hearing *My grandfather bought an estate* (semantically appropriate), participants needed to judge the plausibility of the sentence quickly by drawing on a range of semantic and collocational knowledge resources (e.g., *estate* denotes a purchasable entity, *buy*[verb] + *estate*[noun] is a frequently occurring combination). Similarly, rejecting a semantically inappropriate sentence (*My friend's estate is kind*) involved judging that a personality-describing adjective such as *kind* does not match with an inanimate object (semantically incongruent) and is rarely collocated with *estate* (weak collocational associations). Such advanced lexicons consist of not only the knowledge of a form–meaning connection for

an individual word (indicated by meaning recognition or recall) but also the use-in-context knowledge of the word, encompassing contextual, grammatical, and collocational information associated with it (Nation, 2022).

In response to RQ 2 (concerning the relative contribution of automatized and declarative phonological vocabulary to word usage during real-life L2 listening comprehension), we first demonstrated that automatized lexical knowledge, measured by the LJT, is related to yet empirically separable from declarative lexical knowledge, measured by the two declarative measures of form–meaning knowledge (meaning recognition and meaning recall). More importantly, the automatized-knowledge factor significantly and more accurately predicted L2 listening test scores ($b = 0.566$, $p = .036$), whereas the declarative-knowledge factor failed to be a significant predictor ($b = 0.287$, $p = .291$). This does not mean that declarative knowledge is irrelevant to L2 listening, given a relatively large factor correlation ($r = .740$) and the presence of an indirect effect of declarative knowledge on listening comprehension ($b = 0.419$, 95% CI [0.220, 0.802]). Using declarative form–meaning knowledge does help learners to comprehend L2 speech to some degree (Matthews et al., 2024; McLean et al., 2015), yet successful listening performance is largely dependent on the extent to which learners can use automatized lexical knowledge. Whereas controlled and explicit analyses of the meanings of individual words may act as an indirect contributor to their employability in L2 listening comprehension, advanced listeners should develop the ability to encode and evaluate the degree of semantic and collocational association between target vocabulary and surrounding words, retrieve the contextually appropriate word meaning spontaneously, and use it for speech comprehension.

This study, originally driven by the skill acquisition theory, offers some important insights into L2 phonological vocabulary acquisition together with consideration of the widely accepted frameworks of L2 vocabulary acquisition (Laufer & Goldstein, 2004; Nation, 2022; Schmitt, 2019). Based on our working model of L2 phonological vocabulary acquisition, L2 learners first encode novel forms of words when they are encountered in speech (form acquisition; Bordag et al., 2021) and deliberately connect their aural forms to L1 meanings through explicit training (form–meaning mapping; Jiang, 2000). In this regard, increase in phonological vocabulary *size* could be equated to increase in the number of words for which declarative knowledge of form–meaning connections is developed. Given that language processing is driven by comprehension of meanings (Ellis, 2006), learners start to employ declarative knowledge for comprehending utterances through retrieving lexical meanings while optimizing the efficiency of the retrieval process (proceduralization). Retrieval opti-

mization involves increasing the precision of phonological representations of words (Saito et al., 2023) and the robustness of form–meaning links (Hui & Godfroid, 2021). It also entails encoding the contextual, grammatical, and collocational associations of target vocabulary with other L2 words (Ellis et al., 2008). In other words, in light of Nation’s (2022, pp. 86–92) framework of word knowledge and development, enhancing and refining lexical knowledge in strength (precision and robustness of form and meaning knowledge), depth (encoding of use-in-context knowledge: collocations, grammatical functions, and constraints on use), and organization (integration of form, meaning, and use-in-context knowledge) can be considered to increase word-processing efficiency (automatization) and the employment of word knowledge in communicative and authentic comprehension.

Limitations and Future Directions

This research provides practical implications for vocabulary teaching and learning. Measuring automatized phonological vocabulary knowledge with a judgment task may serve as a useful tool to evaluate the employability of learners’ word knowledge in L2 listening comprehension. The additional information gained from the LJT, besides that from existing controlled measures (multiple-choice and translation tasks), may help teachers to diagnose learners’ phonological vocabulary knowledge comprehensively and adjust their instructional approach effectively (e.g., changing instructional focus to fluency development and meaning-focused input). Although the LJT may have a pedagogical value, we suggest that more empirical evidence needs to be collected to fully support the validity of the test. Future research should conduct item-level analyses to systematically examine test reliability and validity (e.g., dichotomous Rasch model; McLean et al., 2015) as well as the generalizability of the current findings to another population of learners with different L1 backgrounds and L2 proficiency levels. For instance, given the taxing nature of the task, it is possible that the LJT may not be suitable for lower proficiency learners as they might rely on random guessing and test-taking strategies.

It is also important to note that this study tested vocabulary items sampled from the listening materials for the purpose of examining the direct impact of word knowledge. Thus, the extent to which the LJT serves as a proxy of general listening proficiency needs to be further explored. Saito et al. (2023) is the only exception that examined the predictive power of the LJT for general listening proficiency ($r = .66$), although the effect size was slightly smaller than that reported in this study ($r = .71$). Relatedly, our selection of target vocabulary did not consider the relevance of target vocabulary to listening success

(i.e., whether and to what degree knowledge of L2 words would be required for answering comprehension questions). Taking the direct approach while considering the relevance of target lexical items, future studies may clarify the direct impact of lexical knowledge in relation to the knowledge required to demonstrate successful comprehension (i.e., knowing relevant words vs. less relevant words).

We should also acknowledge that although the LJT was intended to tap how learners use multiple aspects of word knowledge (e.g., collocational probability) to achieve improved efficiency in retrieval of contextually appropriate meanings of single-word items, it did not primarily assess knowledge of multiword items as a target construct. To measure automatization of multiword items, future research, as well as assessing declarative knowledge (e.g., recall of meanings of phrasal verbs; Matthews et al., 2024), may additionally evaluate how fluently and accurately learners can judge the appropriateness and inappropriateness of the use of multiword items in sentential contexts (e.g., *He figured out the problem* vs. **He figured out the banana*). Lastly, all the acquisitional and pedagogical suggestions drawn from this study have been based on cross-sectional data. Longitudinal investigations tracking the development of automatized and declarative phonological vocabulary knowledge and examining the time-varying relationships between them will further inform our understanding of L2 phonological vocabulary development.

Conclusion

The current study has been the first endeavor to examine multiple aspects of phonological vocabulary knowledge in relation to L2 listening comprehension. Following previous research using acceptability judgments to measure automatized grammatical knowledge and building on the framework of skill acquisition theory, we developed a LJT with the goal of measuring phonological vocabulary knowledge closely relevant to successful listening performance. Our findings confirmed that the LJT assessed a construct of lexical knowledge more closely related to actual usage of L2 words in listening comprehension compared to controlled measures of declarative form–meaning knowledge (meaning recognition and meaning recall). Our data also suggested that automatized and declarative knowledge are related yet independent constructs, each of which contributed to L2 listening performance in a different way. These findings offer new insights into the developmental sequence of L2 phonological vocabulary acquisition (Schmitt, 2017, 2019): no knowledge → declarative knowledge (meaning recognition → meaning recall) → automatized

knowledge (use-in-context aspects) → appropriate use (processing L2 words in speech comprehension).

Final revised version accepted 19 June 2024

Open Research Badges



This article has earned Open Data and Open Materials badges for making publicly available the digitally-shareable data and the components of the research methods needed to reproduce the reported procedure and results. All data and materials that the authors have used and have the right to share are available at <https://osf.io/jgudx/>. All proprietary materials have been precisely identified in the manuscript.

References

- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, 52(1), 388–407. <https://doi.org/10.3758/s13428-019-01237-x>
- Bordag, D., Gor, K., & Opitz, A. (2021). Ontogenesis model of the L2 lexical representation. *Bilingualism: Language and Cognition*. Advance online publication. <https://doi.org/10.1017/S1366728921000250>
- Cheng, J., & Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing*, 35(1), 3–25. <https://doi.org/10.1177/0265532216676851>
- Cheng, J., Matthews, J., Lange, K., & McLean, S. (2023). Aural single-word and aural phrasal verb knowledge and their relationships to L2 listening comprehension. *TESOL Quarterly*, 57(1), 213–241. <https://doi.org/10.1002/tesq.3137>
- Cutler, A., & Clifton, C. (1999). Comprehending spoken language: A blueprint of the listener. In C. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 123–166). Oxford University Press.
- DeKeyser, R. M. (2020). Skill acquisition theory. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition: An introduction* (3rd ed., pp. 83–104). Routledge.
- Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, 61(2), 367–413. <https://doi.org/10.1111/j.1467-9922.2010.00613.x>
- Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, 27(1), 1–24. <https://doi.org/10.1093/applin/ami038>
- Ellis, N. C. (2022). Fuzzy representations. *Bilingualism: Language and Cognition*, 25(2), 210–211. <https://doi.org/10.1017/S1366728921000638>
- Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL.

- TESOL Quarterly*, 42(3), 375–396.
<https://doi.org/10.1002/j.1545-7249.2008.tb00137.x>
- Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, 39(3), 399–423. <https://doi.org/10.2307/3588487>
- Field, J. (2009). *Listening in the language classroom*. Cambridge University Press.
- Field, J. (2013). Cognitive validity. In A. Geranpaye & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 77–151). Cambridge University Press.
- Field, J. (2019). Second language listening: Current ideas, current issues. In J. W. Schwieter & A. Benati (Eds.), *The Cambridge handbook of language learning* (pp. 283–319). Cambridge University Press.
- Fitzpatrick, T., & Izura, C. (2011). Word association in L1 and L2: An exploratory study of response types, response times, and interlingual mediation. *Studies in Second Language Acquisition*, 33(3), 373–398.
<https://doi.org/10.1017/S0272263111000027>
- Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage.
- González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, 41(4), 481–505. <https://doi.org/10.1093/applin/amy057>
- Gutiérrez, X. (2013). The construct validity of grammaticality judgement tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition*, 35(3), 423–449. <https://doi.org/10.1017/S0272263113000041>
- Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL – International Journal of Applied Linguistics*, 166(2), 278–306.
<https://doi.org/10.1075/itl.166.2.04gyl>
- Hamada, Y., & Yanagawa, K. (2023). Aural vocabulary, orthographic vocabulary, and listening comprehension. *International Review of Applied Linguistics in Language Teaching*. Advance online publication.
<https://doi.org/10.1515/iral-2022-0100>
- Heiberger, R. M., & Holland, B. (2004). *Statistical analysis and data display: An intermediate course with examples in S-PLUS, R, and SAS*. Springer.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
<https://doi.org/10.1080/10705519909540118>
- Hui, B., & Godfroid, A. (2021). Testing the role of processing speed and automaticity in second language listening. *Applied Psycholinguistics*, 42(5), 1089–1115.
<https://doi.org/10.1017/S0142716420000193>
- In'nami, Y., Koizumi, R., Jeon, E. H., & Arai, Y. (2022). L2 listening and its correlates. In E. H. Jeon & Y. In'nami (Eds.), *Understanding L2 proficiency: Theoretical and meta-analytic investigations* (pp. 235–283). John Benjamins.

- Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics*, 21(1), 47–77. <https://doi.org/10.1093/applin/21.1.47>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2–3), 259–284. <https://doi.org/10.1080/01638539809545028>
- Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, 54(3), 399–436. <https://doi.org/10.1111/j.0023-8333.2004.00260.x>
- Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, 44(2), 325–343. <https://doi.org/10.3758/s13428-011-0146-0>
- Luchman, J. (2023). *domir: Tools to support relative importance analysis* (R package; Version 1.0.1) [Computer software]. <https://CRAN.R-project.org/package=domir/>
- MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, 58, 593–614. <https://doi.org/10.1146/annurev.psych.58.110405.085542>
- Matthews, J. (2018). Vocabulary for listening: Emerging evidence for high and mid-frequency vocabulary knowledge. *System*, 72, 23–36. <https://doi.org/10.1016/j.system.2017.10.005>
- Matthews, J., Masrai, A., Lange, K., McLean, S., Alghamdi, E. A., Kim, Y. A., Shinhara, Y., & Tada, S. (2024). Exploring links between aural lexical knowledge and L2 listening in Arabic and Japanese speakers: A close replication of Cheng, Matthews, Lange and McLean (2022). *TESOL Quarterly*, 58(1), 63–90. <https://doi.org/10.1002/tesq.3212>
- McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, 19(6), 741–760. <https://doi.org/10.1177/1362168814567889>
- Mecartty, F. H. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning*, 11(2), 323–348.
- Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *The Canadian Modern Language Review*, 63(1), 127–147. <https://doi.org/10.3138/cmlr.63.1.127>
- Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, M. Torreblanca-López, & M. López-Jiménez (Eds.), *Further insights into non-native vocabulary teaching and learning* (pp. 83–97). Multilingual Matters.

- Mizumoto, A. (2022). Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. *Language Learning*. Advance online publication. <https://doi.org/10.1111/lang.12518>
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Author.
- Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly*, 20(2), 233–253. <https://doi.org/10.2307/747758>
- Nakata, T., Tamura, Y., & Aubrey, S. (2020). Examining the validity of the LexTALE test for Japanese college students. *The Journal of Asia TEFL*, 17(2), 335–348. <https://doi.org/10.18823/asiatefl.2020.17.2.2.335>
- Nation, I. S. P. (2012). *The BNC/COCA word family lists*. Victoria University of Wellington /Te Herenga Waka, Paul Nation's resources. <https://www.wgtn.ac.nz/lals/about/staff/paul-nation>
- Nation, I. S. P. (2022). *Learning vocabulary in another language* (3rd ed.). Cambridge University Press.
- Plonsky, L., Marsden, E., Crowther, D., Gass, S. M., & Spinner, P. (2020). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research*, 36(4), 583–621. <https://doi.org/10.1177/0267658319828413>
- R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.0) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rönkkö, M., & Cho, E. (2022). An updated guideline for assessing discriminant validity. *Organizational Research Methods*, 25(1), 6–14. <https://doi.org/10.1177/1094428120968614>
- Saito, K. (2014). Experienced teachers' perspectives on priorities for improved intelligible pronunciation: The case of Japanese learners of English. *International Journal of Applied Linguistics*, 24(2), 250–277. <https://doi.org/10.1111/ijal.12026>
- Saito, K., Uchihara, T., Takizawa, K., & Suzukida, Y. (2023). Individual differences in L2 listening proficiency revisited: Roles of form, meaning, and use aspects of phonological vocabulary knowledge. *Studies in Second Language Acquisition*. Advance online publication. <https://doi.org/10.1017/S027226312300044X>
- Schmitt, N. (2017, August 30–September 2). *After 25 years of researching vocabulary: A personal reflection on where vocabulary research needs to go next* [Plenary talk]. European Second Language Association (EuroSLA), Reading, UK.
- Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, 52(2), 261–274. <https://doi.org/10.1017/S0261444819000053>
- Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: Acquisition of collocations under different input conditions. *Language Learning*, 63(1), 121–159. <https://doi.org/10.1111/j.1467-9922.2012.00730.x>

- Spinner, P., & Gass, S. (2019). *Using judgements in second language acquisition research*. Routledge.
- Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, 31(4), 577–607. <https://doi.org/10.1017/S0272263109990039>
- Suzuki, Y. (Ed.) (2023). *Practice and automatization in second language research: Perspectives from skill acquisition theory and cognitive psychology*. Routledge.
- Suzuki, Y., & DeKeyser, R. M. (2017). The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning*, 67(4), 747–790. <https://doi.org/10.1111/lang.12241>
- Suzuki, Y., & Elgort, I. (2023). Measuring automaticity in second language comprehension. In Y. Suzuki (Ed.), *Practice and automatization in second language research: Perspectives from skill acquisition theory and cognitive psychology* (pp. 206–234). Routledge.
- Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning*, 70(2), 506–547. <https://doi.org/10.1111/lang.12384>
- Uchihara, T. (2023). How does the test modality of weekly quizzes influence learning the spoken forms of second language vocabulary? *TESOL Quarterly*, 57(2), 595–617. <https://doi.org/10.1002/tesq.3176>
- Uchihara, T., & Harada, T. (2018). Roles of vocabulary knowledge for success in English-medium instruction: Self-perceptions and academic outcomes of Japanese undergraduates. *TESOL Quarterly*, 52(3), 564–587. <https://doi.org/10.1002/tesq.453>
- Vafaei, P., & Kachinske, I. (2019). The inadequate use of confirmatory factor analysis in second language acquisition validation studies. *Studies in Applied Linguistics & TESOL*, 19(2), 1–18. <https://doi.org/10.7916/d8-p2xt-e666>
- Vafaei, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition*, 42(2), 383–410. <https://doi.org/10.1017/S0272263119000676>
- Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, 65(2), 390–416. <https://doi.org/10.1111/lang.12105>
- Vandergrift, L., & Baker, S. C. (2018). Learner variables important for success in L2 listening comprehension in French immersion classrooms. *The Canadian Modern Language Review*, 74(1), 79–100. <https://doi.org/10.3138/cmlr.3906>
- Wallace, M. P. (2022). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*, 72(1), 5–44. <https://doi.org/10.1111/lang.12424>
- Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, 28(1), 46–65. <https://doi.org/10.1093/applin/aml048>

Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, 26(4), 696–725. <https://doi.org/10.1177/1362168820913998>

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Accessible Summary

Appendix S1. Target Items and L1 Japanese Options for Meaning-Recognition Test.

Appendix S2. Prompt Sentences for Lexicosemantic Judgment Task.

Appendix S3. Examples of Vocabulary Test Formats for the Three Vocabulary Tests.

Appendix S4. Descriptive Statistics of Vocabulary and Listening Measures.

Appendix S5. Preliminary Analyses Exploring the Influence of Item-Level Characteristics on Performance of the Lexicosemantic Judgment Task.

Appendix S6. Summary of Mixed-Effects Logistic Regression Analysis.