Studies in Second Language Acquisition (2023), 1–27
doi:10.1017/S027226312300044X

RESEARCH ARTICLE

Individual differences in L2 listening proficiency revisited: Roles of form, meaning, and use aspects of phonological vocabulary knowledge

Kazuya Saito[1], Takumi Uchihara[2], Kotaro Takizawa[3] and Yui Suzukida[1,4]

[1]University College London, Institute of Education, London, United Kingdom; [2]Tohoku University, Graduate School of International Cultural Studies, Sendai, Japan; [3]Waseda University, School of Education, Tokyo, Japan; [4]Juntendo University, Faculty of Medicine, Tokyo, Japan
Corresponding author: Kazuya Saito; Email: k.saito@ucl.ac.uk

(Received 31 January 2023; Revised 14 July 2023; Accepted 05 August 2023)

Abstract

The present study revisits the differential roles of form, meaning, and use aspects of phonological vocabulary knowledge in L2 listening proficiency. A total of 126 Japanese English-as-a-foreign-language listeners completed the TOEIC Listening test, working memory and auditory processing tests, the Metacognitive Awareness Listening Questionnaire, and several tasks designed to tap into three broad aspects of phonological vocabulary knowledge: (1) the ability to access phonological forms without any orthographic cues (phonologization), (2) the ability to recognize words regardless of the talker (generalization), and (3) the ability to determine the semantic and collocational appropriateness of words in global contexts in a fast and stable manner (automatization). Whereas the perceptual, cognitive, and metacognitive variables made relatively small contributions to L2 listening proficiency (0.4%–21.3%), the vocabulary factors explained a large amount of the variance (77.6%) in the full regression model (R^2 = .507). These large lexical effects uniquely derived from the three different aspects of phonological vocabulary knowledge—automatization (55.3%), phonologization (20.8%), and generalization (1.5%). The findings suggest that successful L2 listening skill acquisition draws on not only various levels of phonological form-meaning mapping (phonologization, generalization) but also the spontaneous and robust retrieval of such vocabulary knowledge in relation to surrounding words (automatization).

Introduction

There is general agreement that second language (L2) listening proficiency is essential for successful communication, as it allows individuals to quickly understand the message of their interlocutor and respond appropriately in social, academic, and business contexts. This skill is increasingly important in foreign language contexts, as L2 learners can now access authentic input through various media outlets (e.g., the

internet, movies, and TV) and benefit from L2 speech learning experiences outside of the classroom. To achieve advanced L2 listening proficiency, learners must develop, access, and use both bottom-up knowledge (e.g., vocabulary, grammar, discourse) and top-down knowledge (e.g., strategies, topic, genre, culture). However, as Vandergrift noted in his 2007 review, "L2 listening remains the least researched of all four language skills" (p. 191). Since then, an increasing number of studies have sought to identify the mechanisms underlying successful L2 listening proficiency by examining the linguistic, perceptual-cognitive, and metacognitive profiles of L2 learners with varying levels of listening proficiency (In'nami et al., 2023; Smith, 2019; Zhang & Zhang, 2020, for meta-analyses). There is a general consensus that individual differences in L2 listening proficiency can be largely explained by learners' vocabulary knowledge and to a lesser extent by a range of perceptual-cognitive and metacognitive factors (e.g., Andringa et al., 2012; Vafaee & Suzuki, 2020; Vandergrift & Baker, 2015, 2018; Stæhr, 2009; Wallace, 2022; Wang & Treffers-Daller, 2017).

There is an emerging paradigm that seeks to reconceptualize vocabulary knowledge relevant to L2 listening (i.e., phonological vocabulary knowledge) as a multifaceted phenomenon (McLean et al., 2015). As stated in Nation's (2013) framework of *spoken* vocabulary knowledge, learners need to understand not only what target words sound like and mean (i.e., form-meaning mapping) but also how they interact with other words in a semantically, collocationally, and grammatically appropriate manner (i.e., use-in-context). In his critical review, Schmitt (2019) noted that although an increasing number of researchers have explored different levels of form-meaning mapping, using both recognition and recall task formats (Zhang & Zhang, 2020), few studies have examined whether, to what degree, and how L2 learners access already-learned vocabulary knowledge in context during real-life L2 speech comprehension (Read, 2020).

Drawing on the usage-based account of L2 comprehension (Ellis, 2006), we propose that such phonological vocabulary comprises three broad psycholinguistic abilities: (1) the ability to access phonological forms without any orthographic cues (phonologization); (2) the ability to recognize words regardless of the talker (generalization); and (3) the ability to encode the semantic and collocational appropriateness of words in global contexts in a fast, stable, and automatic manner (automatization). In line with Nation's model (2013), both phonologization and generalization pertain to the form-meaning mapping aspect of vocabulary knowledge (therefore measured via meaning recognition formats), whereas automatization relates to the use-in-context aspect of vocabulary knowledge (thus measured via lexicosemantic judgments). To validate our proposed framework, we first examine these three distinct aspects of phonological vocabulary knowledge (phonologization, generalization, and automatization) among 126 Japanese students studying English as a foreign language. Then, we investigate the unique contribution of such knowledge to overall L2 listening proficiency, taking into account perceptual-cognitive and metacognitive abilities.

Background

Mechanisms and individual differences in L2 listening

There is ample research examining the linguistic, perceptual-cognitive, and metacognitive factors that contribute to L2 listening outcomes. As for linguistic factors, learners initially segment auditory input into word units (word recognition: Norris & McQueen, 2008). Subsequently, listeners must identify not only the morphological features of each

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

word (morphological processing) and the grammatical structures within a sentence (grammatical parsing; Vafaee & Suzuki, 2022), but they must also grasp the speaker's intended meaning in line with different discursive, social, and cultural contexts (pragmatic processing; Taguchi, 2011). Finally, listeners are required to decipher the meaning of a sentence by employing paralinguistic elements such as tone of voice, facial expressions, and body gestures (Kamiya, 2022).

Listeners' perceptual-cognitive skills enhance their ability to process both linguistic and paralinguistic cues. More precise perceptual abilities significantly aid listeners in encoding the acoustic features of input, thus optimizing phoneme and word recognition (see Saito et al., 2020, for auditory processing). Likewise, enhanced cognitive abilities can facilitate the retention of information while processing incoming speech, thereby aiding in the comprehension of extended discourse (see Linck et al., 2013, for working memory). Listeners who possess higher levels of metacognitive awareness can more consciously, and therefore more effectively, hone their L2 listening skills. Although they may not comprehend every word, they are still capable of grasping the gist of L2 passages by employing a variety of relevant strategies. These strategies encompass preparing for a relevant task using background knowledge, maintaining focus on the task, guessing the meaning of unfamiliar words, avoiding direct translation, and fostering increased confidence (Vandergrift & Goh, 2012).

All the linguistic, perceptual-cognitive, and metacognitive factors are associated with L2 listening, but which among them is then *relatively* more critical in determining successful comprehension? It has been shown that out of all these factors, learners' vocabulary knowledge can explain the greatest amount of variance in global listening test scores (Vandergrift & Baker, 2015, 2018; Stæhr, 2009; Wallace, 2022). According to the results of meta-analyses (Smith, 2019; Zhang & Zhang, 2020), the relationship between L2 vocabulary knowledge and listening proficiency is moderate to strong (r = .50–.70). Other secondary affecting factors (r = .30–.50) include a range of perceptual-cognitive abilities (Linck et al., 2013, for working memory; Vandergrift & Baker, 2018, for auditory processing; Hui & Godfroid, 2021, for processing speed) and metacognition and strategy use (In'nami et al., 2023; Vandergrift & Goh, 2012). There is emerging evidence suggesting that vocabulary knowledge is a primary determinant of L2 listening proficiency for learners at all proficiency levels, whereas secondary variables affect listening proficiency among relatively advanced L2 learners who have already established a sufficient amount of vocabulary knowledge (Milliner & Dimoski, 2021; Wallace, 2022; Yanagawa, 2023). Without a strong linguistic foundation for optimal L2 comprehension (e.g., 3–4K word families), L2 learners have difficulty encoding L2 aural discourse even when they are familiar with the topic and equipped with good strategies (van Zeeland & Schmitt, 2013; see also Schmitt et al., 2017).

In many existing studies, however, L2 learners' vocabulary knowledge has been examined via *written* vocabulary tests (Schmitt et al., 2001, for the Vocabulary Levels Test; Read, 1998, for the Word Associates Test). Scholars are beginning to draw attention to the potential gap between written and aural measures of L2 learners' vocabulary knowledge. That is, many classroom L2 learners can recognize words when they are presented in written formats but demonstrate much difficulty in doing so when they are presented aurally (Cheng & Mathew, 2018; Hamada & Yanagawa, 2023; Milton & Hopkins, 2006). This is because many of them learn new vocabulary items from textbooks and flashcards without much opportunity for listening or speaking practice (Uchihara & Harada, 2018) and phonological training is generally lacking in the classroom (Saito, 2014, for Japanese EFL teachers and learners).

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

Traditional foreign language syllabi focus on the orthographic aspects of words, but learning the phonological aspects of new words is a difficult task that may require a great deal of intensive exposure and explicit phonological training (Saito & Plonsky, 2019), especially when these L2 phonetic features are absent in learners' L1 systems on segmental (Flege & Bohn, 2021) and suprasegmental levels (Trofimovich & Baker, 2006). Given the significant role of learners' lexical knowledge and the effect of input modality (written vs. aural) in relation to L2 listening proficiency, researchers suggest that L2 vocabulary should be taught and assessed in aural rather than written modalities (Uchihara, 2023). Here, we argue that it is important to extend this line of research with a view to developing a more nuanced understanding of the complex relationship between phonological vocabulary and successful L2 comprehension. In the following sections, we revisit how L2 learners acquire the lexical aspects of L2 speech, what characterizes L2 phonological vocabulary knowledge, and how we can assess the multilayered nature of such knowledge.

Three-stage model of phonological vocabulary knowledge

A variety of models pertaining to L2 speech comprehension exist, and each model emphasizes certain aspects of relevant linguistic, perceptual-cognitive, and metacognitive processing. Given that the primary focus of the study lies in phonological vocabulary knowledge, we adhere to the usage-based account of L2 speech comprehension as our theoretical foundation. This model specifically highlights the intricate mechanisms underpinning word recognition (Ellis, 2006; Norris & McQueen, 2008).

When exposed to auditory input, listeners use their knowledge of rhythmic and phonotactic structures to segment the speech stream into words and phrases (Cutler et al., 1997). Once listeners hear the initial syllable, a range of competing lexical candidates is triggered (i.e., word activation). As more input is received, this group of candidates narrows (i.e., word competition; Norris & McQueen, 2008). When the incoming input has a high phonological neighborhood density, there is a greater number of competing candidates (Pisoni & Luce, 1987). For instance, the input [leit] activates not only "late" but also "lace" (substitution), "ate" (deletion), and "plate" (addition). Listeners must attend to segmental details to discriminate the target words from phonologically similar words (Werker, 2018). Input with high neighborhood density may activate an even greater number of candidates for L2 listeners due to L1 phonetic interference (Bradlow & Pisoni, 1999). When exposed to the input [leit], for example, Japanese listeners, who tend to neutralize the English [r] and [l] contrast, may activate both L-words (e.g., "late," "lace," "plate") and R-words (e.g., "rate," "race," "pray"; Flege et al., 1996).

Ellis (2006) further posits that a fluent listener functions as "an optimal word processor" (p. 2). In this view, the selection of the most suitable candidates is influenced by three major factors: (a) frequency (the number of times certain words have been encountered in the past), (b) recency (how long ago these words were last accessed), and (c) context (the lexical contexts in which they occur). Listeners are able to better recognize words when they have been encountered with greater frequency (Webb et al., 2023) and on more recent occasions (Nakata, 2015) and when they appear in conjunction with words that are more likely to co-occur with them (Ellis et al., 2008). For example, "jog" is more likely to be chosen as the best lexical candidate over "job" and "jot" when listeners have heard it more frequently and more recently and when it appears together with words such as "sports," "health," and "outside."

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

Given the nature of real-life L2 speech comprehension, what type of phonological vocabulary knowledge is necessary for the attainment of successful L2 listening proficiency? In this paper, we argue that the development of phonological vocabulary knowledge comprises three different stages—phonologization, generalization, and automatization. The three stages concur with the proceduralization and automatization of declarative knowledge stated in the skill acquisition model for instructed second language acquisition (SLA; DeKeyser, 2017; Suzuki, 2023). Under this view, learners initially form their lexical representations as controlled knowledge through explicit vocabulary training in classroom environments. As they gain exposure to more language input opportunities, they continue to refine and enhance these representations (phonologization and generalization) while improving their accessibility to this knowledge under varying processing conditions (automatization). These stages are likely to occur simultaneously and influence each other even though they relate to different aspects of the learning process.

Phonologization

Many theories in L2 speech learning (see, Flege & Bohn, 2021, for a speech learning model) suggest that learners face considerable difficulty in perceiving L2 words. This is because they need to adjust not only to novel prosodic patterns to detect word, sentence, and discoursal units in auditory input but also to segmental details in order to identify a range of phonologically similar words. The level of difficulty could be particularly high when these L2 prosodic and segmental features are completely (or partially) absent in learners' L1 systems and thus need to be established as new phonetic and phonological categories—for example, the identification of four distinct lexical tones among non-tonal learners of Mandarin (Wang et al., 2003), and the encoding of third formant [F3] variation for the English [r] and [l] contrast (Saito, 2013).

Despite the inherent challenges of spoken L2 word recognition, growing evidence suggests that L2 learners often focus on written form-meaning mappings, largely due to the limited amount of communicatively authentic input in many foreign language classrooms (Uchihara & Harada, 2018). Although L2 learners possess explicit knowledge of form-meaning mappings for words, their access to these mappings may sometimes be confined to written modes and not extend to aural ones. Consequently, many low-proficiency and inexperienced L2 learners may struggle to recognize words when they are presented aurally, even though they can read the same words when they are presented in written form; these learners need to develop the ability to use such explicit knowledge across various modalities for successful L2 listening skill acquisition (Du et al., 2022). Indeed, research has found that listening proficiency is more closely correlated with vocabulary knowledge when assessed through aural modalities (Hamada & Yanagawa, 2023; Masrai, 2020). A central aspect of phonological vocabulary knowledge, therefore, is the ability to recognize words when they are presented aurally, without any orthographic cues.

Generalization

Another crucial aspect of phonological vocabulary knowledge involves the ability to recognize words spoken by different individuals. In the context of first language acquisition, children initially rely heavily on familiar voices (e.g., parents) when forming representations of spoken words. By the end of their first year, their understanding becomes more refined and robust, enabling them to process words from both familiar and unfamiliar speakers (Houston & Jusczyk, 2000). The acquisition of

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

accurate yet *generalizable* knowledge that can be applied across different talkers is also a topic extensively discussed in L2 speech learning. Numerous training studies show that although learning might be constrained when exposed to a single speaker's voice, training with multiple speakers can lead to learning that transcends individual speaker variations (Thomson, 2018, for a comprehensive overview of low vs. high variability phonetic training).

In reality, the phonological characteristics of words are not identical among different talkers. Spectral information in speech can differ (e.g., in vowels) due to anatomical differences (e.g., vocal tract length; Adank et al., 2004), and temporal information in speech can also vary (e.g., some speak faster than others; de Jong et al., 2015). When exposed to a multitude of speakers, L2 learners are believed to perform statistical analyses to identify which acoustic parameters reliably predict target sounds and words (e.g., F3 < 2000 Hz for English [r] and F3 > 2400 for English [l]). This allows them to effectively recognize sounds and words by selectively attending to primary acoustic variables while disregarding the surface-level acoustic differences between speakers (Flege & Bohn, 2021).

Automatization

As stressed in the skill acquisition theory of instructed SLA (DeKeyser, 2017; Suzuki, 2023), the ultimate goal for students should be the gradual automatization of controlled knowledge. One critical issue is processing speed and stability, which is the cognitive foundation of fluency (Segalowitz, 2010). When learners repeatedly practice using their learned vocabulary knowledge, their retrieval speed quickly goes up, but after a certain amount of practice, retrieval speed plateaus, after which it remains stable and invariant across different lexical conditions (i.e., coefficients of variance being smaller; Hui & Godfroid, 2021).

Another important issue related to automaticity concerns the retrieval of learned vocabulary knowledge in more spontaneous and communicatively authentic contexts. Surprisingly, the previous literature on the assessment and development of phonological vocabulary has exclusively relied on controlled and decontextualized tasks. For example, listeners are asked to engage in multiple-choice meaning recognition (McLean et al., 2015), meaning recall (Cheng et al., 2023), partial dictation (Cheng & Matthews, 2018) and yes/no form recognition (Milton & Hopkins, 2006). In such tasks, listeners focus solely on the explicit analysis of one to two words in isolation. Although these tasks can assess L2 learners' understanding of what words sound like and signify (i.e., form-meaning mapping), it remains unclear how to measure learners' ability to access their explicit word knowledge accurately and rapidly within sentence context during successful L2 comprehension (i.e., use-in-context; Schmitt, 2019).

The notion of automatization in the skill acquisition theory for instructed SLA (DeKeyser, 2017; Suzuki, 2023) suggests that advanced L2 learners have the capacity to access learned phonological vocabulary not just in single-task conditions but also in real-life language processing. In the latter conditions, learners must consider the context and collocational properties of both the target word and surrounding words while simultaneously processing other aspects of language (e.g., morphology, syntax, and discourse; Ellis et al., 2008).

In the field of applied linguistics, automatized L2 knowledge (i.e., accurate and fluent use of acquired knowledge) is often measured using acceptability judgement tasks (Spinner & Gass, 2019). To measure *automatized* morphosyntactic knowledge, for

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

example, L2 learners hear or read a set of sentences, some of which include a particular morphosyntactic error, and judge whether they are grammatically accurate (grammaticality judgement task [GJT]; Plonsky et al., 2020). It has been shown that GJT results can differ substantially from those of controlled tasks measuring the same morphosyntactic knowledge (e.g., fill-in-the-blank tests; Gutiérrez, 2013). Such automatized explicit knowledge may anchor the development of implicit knowledge (Suzuki & DeKeyser, 2017). In Suzuki and Elgort's (2023) comprehensive review of measurement practices for automatized L2 knowledge, they found surprisingly little attention has been devoted to the development and application of tasks for assessing automaticity in auditory lexical processing (the main focus of the current study).

In the context of L2 spoken word recognition, when L2 phonological vocabulary knowledge is automatized, listeners store it together with strongly collocated words as a chunk (Tavakoli & Uchihara, 2020). We propose that acceptability judgements be used to measure this automatized phonological vocabulary knowledge in global contexts. During such tasks, L2 listeners would be asked to accept target words when they appear in contextually appropriate sentences and reject them when they do not match the context of a sentence. The degree of contextual appropriateness is related to the collocational relationship between target words and surrounding words (Saito, 2020).

To date, very few studies have adopted acceptability judgements as a way to assess L2 vocabulary knowledge. Ellis et al. (2008) asked both L1 and L2 participants to read a series of word strings (e.g., "by the way" vs. "by way the") and judge whether they exist in English. The results revealed that reaction time was determined by different factors for L1 and L2 participants (the strength of collocation association vs. word frequency). Foster et al. (2014) asked highly experienced L2 participants to read narratives where a set of nonnativelike word combinations were embedded. The results showed that more advanced L2 participants (with earlier age of arrival) showed more accurate identification of nonnativelike word selections. Note that the stimuli were presented in written modalities in these studies (Ellis et al; Foster et al).

Our prior work (Uchihara et al., 2023) proposed, developed, and tested the Lexicosemantic Judgement Task (LJT) as a way to tap into L2 listeners' automatized, spontaneous, and contextualized phonological vocabulary knowledge. Participants listened to a series of sentences. After each sentence, they were prompted to make an intuitive judgement of whether it sounded semantically appropriate or not. All the sentences were grammatically correct and included only high-frequency words. In half of the sentences, target words were embedded in semantically appropriate contexts ("*He has published many books*"), but in the other half, target words were embedded in semantically inappropriate contexts ("*Mary published her left hand*").

Advanced L2 listeners were expected to attend to the degree of semantic and collocational associations between the target and surrounding words, find the most contextually appropriate combinations, process the sentence as a series of lexical chunks, and grasp the overall meaning in an efficient manner. For instance, advanced L2 listeners can quickly accept the former appropriate sentence as the combination of "publish" and "books" is semantically transparent and frequently used, but they can intuitively reject the latter inappropriate sentence because "publish" does not match with an animated object (semantically incongruent) and is rarely collocated with "hand" (weak collocational associations).

As part of our efforts to validate the LJT as an outcome measure for the automatized use-in-context dimension of L2 phonological vocabulary knowledge, our team's precursor project (Uchihara et al., 2023) examined the characteristics of 114 Japanese EFL

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

learners' L2 phonological vocabulary knowledge via multiple tasks. The findings indicated distinct phonological vocabulary performance when assessed via LJT as opposed to when measured using two tasks typically employed to evaluate the form-meaning aspects of L2 phonological vocabulary knowledge—that is, meaning recognition and meaning recall. 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.

Current study

To date, scholars have emphasized the multilayered nature of phonological vocabulary knowledge and its relation to overall L2 listening proficiency. However, much of the work has concentrated on different levels of form-meaning mapping aspects of phonological vocabulary knowledge (e.g., recognition vs. recall; Zhang & Zhang, 2020), with less attention given to the use-in-context aspects of phonological vocabulary knowledge (see Schmitt, 2019). In our prior work (Uchihara et al., 2023), we established the validity of two distinct phonological vocabulary assessments—multiple choice for declarative knowledge and acceptability judgement for automatized knowledge. In accordance with Nation's (2013) model, multiple choice, termed as meaning recognition, is considered to evaluate the form-meaning mapping aspect of vocabulary knowledge, and the judgement task is assumed to assess the use-in-context aspect of vocabulary knowledge (see Suzuki & DeKeyser, 2017, for the discussion and methodology for assessing automatized explicit knowledge in L2 morphosyntax).

Replicating and extending this line of research, the primary objective of the current study was to test the validity of our proposed model of phonological vocabulary knowledge that can be conceptualized as three different stages of L2 spoken word learning in EFL classrooms—namely, phonologization (the ability to recognize words without orthographic cues), generalization (the ability to recognize words across different speakers), and automatization (the ability to recognize the semantic and collocational associations between words). To achieve this objective, we first developed three phonological vocabulary tasks: the Phonological Multiple-Choice task (PhonMC), the Generalization Multiple-Choice task (GenMC), and the Phonological Lexicosemantic Judgement task (LJT). Next, we examined how different types of phonological vocabulary knowledge correlate with overall L2 listening proficiency test scores, controlling for individual differences in perception, cognition, and metacognition. Last, we sought to compare the role of phonological vocabulary in L2 listening proficiency across three different listener groups (low, mid, and high). Accordingly, we formulated the following research questions and predictions:

RQ1: How are different aspects of phonological vocabulary knowledge associated with L2 listening proficiency?

Based on the results of our prior work (Uchihara et al., 2023), it was predicted that all vocabulary measures would correlate to L2 listening proficiency (McLean et al., 2015) and the predictive power of automatized phonological vocabulary knowledge would be stronger than controlled phonological knowledge (r = .60–.70 for LJT vs. .30–.40 for PhonMC).

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

RQ2: How do different aspects of phonological vocabulary knowledge and other cognitive and metacognitive abilities predict L2 listening proficiency?

As shown in existing studies (Wallace, 2022), L2 listening proficiency (measured via the TOEIC Listening test) is primarily determined by phonological vocabulary knowledge (r = .60–.70) and secondarily by perceptual-cognitive and metacognitive abilities (r = .20–.30). Based on the results of the precursor projects (Uchihara et al. 2023), it was predicted that the link between vocabulary knowledge and L2 listening proficiency would be observed most clearly when the analyses focused on spontaneous tasks (acceptability judgments) rather than controlled tasks (multiple choice).

RQ3: How does the link between vocabulary and listening proficiency vary among low, mid, and high-level Japanese EFL listeners?

Given that the initial stage of L2 listening proficiency development is phonologization, it was predicted that participants' explicit analyses of phonological words (PhonMC for phonologization) would be a significant predictor of low- to mid-level L2 listening proficiency (Du et al., 2022). Given that the automatization of more robust phonological knowledge develops at a later stage, it was predicted that differences between mid- and high-level L2 listeners would be related to the extent to which they could access their phonological vocabulary knowledge regardless of talker (GenMC for generalization; Thomson, 2018) and processing conditions (LJT for automatization; Suzuki & DeKeyser, 2017).

Method

Participants

A total of 126 Japanese learners of English in Japan participated in the current study (75 females, 51 males; M_{age} = 20.5 years; $Range_{age}$ = 18–26 years). As part of a larger investigation to explore the assessment and training of L2 listening and speaking proficiency, an electronic flyer was distributed to various universities in Japan. The flyer explicitly stated that the participants needed to have primarily studied English in classroom settings, without extensive experience abroad (more than 1 month). As knowledge of the first 1–2-K word families is suggested as a minimum requirement for global L2 listening comprehension in everyday situations (Adolphs & Schmitt, 2003), participants who fell below this lexical threshold were excluded. Such listeners would primarily be guessing on the global listening test, which would render their data irrelevant to the purposes of the project. To determine whether participants met this threshold, they first completed a PhonMC test focusing on 10 target words from the first 1,000 word families available in the BNC-COCA corpus in Cobb's Vocab Profilers (https://www.lextutor.ca/). If they did not attain 80% accuracy, they were removed from the study. This was under the assumption that they lacked the fundamental vocabulary knowledge required for basic comprehension of L2 speech. The selection of 80% as a cutoff point aligns with existing practices in numerous vocabulary size tests (e.g., Dang et al., 2022, for 80%; Hu & Nation, 2000, for 80%; Schmitt et al., 2001, for 86.6%). According to the results of the L2 listening proficiency test (measured via the Test of English for International Communication [TOEIC]; for further information, see below), the remaining participants' English proficiency ranged from A2 to C1 on the Common European Framework of Reference for Languages scale. Ethical clearance for

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

this research was secured from the review board at University College London, in accordance with the guidelines for studies involving human participants.

General setup

Due to the pandemic, all the data collection (phonological vocabulary tests, the TOEIC, perception and cognition tests, and metacognition questionnaire) were administered using a set of online tools. Participants were asked to use headphones and a computer with stable Internet access in a quiet room. Several steps were taken to help participants follow the procedure and monitor their performance. Participants were asked to complete the screening vocabulary test first, which allowed us to ensure that there were not any major problems with their technical setup before they proceeded to the main tasks. To facilitate their understanding of test procedures, the participants who participated in the main data collection were asked to read a handout that detailed all the task instructions in Japanese. They were encouraged to ask questions regarding anything they found to be unclear. After they were assigned to a 1-hr time slot, they participated in a Zoom session with a total of 5 to 20 participants. In each session, participants took the TOEIC Listening test with an invigilator's guidance. Their responses were recorded via Google Forms. Once they completed the TOEIC Listening test, they were given a URL link that allowed them to access the three phonological vocabulary tests (LJT, PhonMC, and GenMC in this order) as well as an EFL Experience Questionnaire via the online data collection platform Gorilla (Anwyl-Irvine et al., 2020).

Efforts were made to minimize the potential influence of fatigue on data quality. As part of the screening process, the participants first completed the aptitude tests (5 min for auditory processing, 5 min for working memory). During the main testing session, they initially took the TOEIC test (40 min), followed by a short intermission (5–10 min). Subsequently, they undertook the three vocabulary tasks—LJT for 10–15 min, PhonMC for 5 min, and GenMC for 5 min. The average duration of the pretest session was around 70 min. Notably, this procedure was vetted during our precursor projects (Uchihara et al., 2023), and no participants reported experiencing fatigue.

Listening proficiency test

The TOEIC Listening test was chosen as this type of composite proficiency test taps into L2 learners' ability to process various kinds of realistic spoken language and has previously been used to measure L2 listening proficiency (Hamada & Yanagawa, 2023; McLean et al., 2015; Cheng et al., 2023). Three (out of four) parts of the New Official Workbook (Educational Testing Service, Vol.4)—Question-Response, Conversation, and Monologue—were used for the current study.

- In Question-Response (k = 30), participants listened to 30 single-sentence questions (5–10 words), each accompanied by three response options, and then selected the most appropriate response. Performance on this section provided an assessment of the participants' ability to understand linguistically and semantically simple input.
- In Conversations (k = 30), participants listened to 10 dialogues between a male and a female speaker. For each dialogue (80–100 words), they answered three comprehension questions by selecting the most appropriate response from among four options.

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

This section was used to assess comprehension of interactional speech including frequent turn taking (approximately 20–25 words per turn).

- In Monologues (k = 30), participants listened to 10 business-related monologues spoken by a single person. For each monologue (80–100 words), they answered three comprehension questions by selecting the most appropriate response from four options. This section was used to provide an assessment of the participants' ability to understand linguistically and semantically complex input.

Analyses

Corpus analyses were conducted using the methodology suggested by Révész and Brunfaut (2013). Results indicated that the three tasks differed in difficulty. The degree of grammatical complexity varied in the following order: Question-Response < Conversations < Monologues. Complexity was reflected in the number of words per sentence (M = 5.7 vs. 12.1 vs. 14.8) and the number of words before main verbs (M = 0.9 vs. 1.2 vs. 3.4). As shown in Supporting Information S1, significant task effects were found in the participants' listening performance on the three tasks—that is, Question-Response (M = 18.4) < Conversations (M = 16.6) < Monologues (M = 14.3). Their total scores (out of 90 points) were used as a measure of L2 listening proficiency. The participants were divided into different proficiency subgroups based on the results of cluster analyses of their scores in each subtest (Question-Response, Conversations, and Monologues).

Phonological vocabulary tests

To examine the generalizability of the controlled and spontaneous phonological vocabulary tests proposed in Uchihara et al. (2023), the same formats were used in the current study: multiple choice (PhonMC, GenMC) and acceptability judgements (LJT). All the test materials are deposited in the open science platform, L2 Speech Tools (Mora-Plaza et al., 2022: http://sla-speech-tools.com/).

Target words

As detailed in our precursor projects (Uchihara et al., 2023), we selected 80 target words with the goal of developing measures of vocabulary knowledge that are relevant to L2 listening proficiency among Japanese learners of English. Typically, scholars select a fixed number of words per major word-frequency list to measure L2 learners' vocabulary knowledge (and its relevance to their reading proficiency), assuming an even distribution of frequent to infrequent words (up to 10–12K) in L2 discourse. However, recent evidence suggests that the majority of aural L2 discourse comprises 6K word families (Mathew, 2018), and the distribution of frequent to infrequent words is uneven (Nation, 2006).

Given our primary focus on assessing participants' phonological vocabulary knowledge, we took into account not only the uneven-distribution nature of word frequency but also other factors that particularly affect Japanese learners' understanding of L2 English speech discourse, such as cognates (Uchihara et al., 2023) and phonological difficulty (Saito, 2014). We believe that this composite approach is more ecologically valid and better represents real-life L2 listening experiences.

To capture the lexical profiles of L2 speech experience, we first created a speech corpus based on scripts from one retired version of the TOEIC Listening test. We chose

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

the test/materials because they cover different types of L2 discourse (e.g., conversations, monologues), and this is the test format we adopted to measure participants' L2 listening proficiency (a separate version was used for this purpose; see below).

In total, 2,731 tokens used in the passages were evaluated based on the following three factors, and the top 80 most phonologically challenging words were selected:

1. **Word frequency:** We placed an emphasis on less frequent words using the BNC/COCA word family lists (Nation, 2012).
2. **Cognate status:** We excluded cognates because they might aid in L2 comprehension (Uchihara et al., 2023).
3. **Phonological difficulty:** We prioritized words with phonological features that pose difficulty for Japanese learners of English. They were iambic words with more syllables, difficult segmentals (English [r] and [l]), and consonant clusters (Saito, 2014).

The 80 target words ranged in frequency; the most frequent words fell within the top 2K word families and the least frequent within the top 8K. It has been suggested that this range (2K–8K) covers 98% of the words used in spoken discourse (e.g., Nation, 2006) and is thus sufficient for advanced L2 comprehension (Van Zeeland & Schmitt, 2013). A relatively higher proportion of these 80 words was high frequency (22 words in 2K, 35 words in 3K), and a lower proportion was mid frequency (13 words in 4K, and 10 words in 5K–8K). This ratio is appropriate given that knowledge of high-frequency vocabulary is a significant predictor of L2 listening test scores, accounting for over 50% of the observed variance in performance or an even higher proportion in the case of EFL listeners without much experience overseas (i.e., the focus of the current study; Matthews, 2018).

Importantly, even though the distribution of word frequency appeared to be uneven, the selection of the target words was also determined by cognate status and phonological difficulty. We posit that assessing learners' knowledge of the 80 most *phonologically* difficult target words in the TOEIC test (representing various types of real-life L2 discourse) can provide a rough but overall estimate of their phonological vocabulary knowledge, which is most directly relevant to global listening abilities.

Procedure

The 80 target words were assessed in three different task formats in the following order: one spontaneous task (LJT) and two controlled tasks (PhonMC, GenMC). Following the methods in L2 speech research (Saito, 2013), participants first took the LJT, where their ability to spontaneously access the target words without much planning was tested. Subsequently, they took the PhonMC test, which assessed the presence (or absence) of their declarative knowledge of the target words, even when they were given ample time. Finally, they took the GenMC test to examine whether the results of their PhonMC test could be replicated when the materials were presented by a different speaker.

Lexicosemantic judgement task

Adapting the acceptability judgements commonly used in L2 grammar research, a lexicosemantic judgement task (LJT) was developed. The task consisted of 160 short sentences spoken by a female native speaker of General American English. Upon hearing each sentence, participants were asked to select one of two options:

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

"semantically appropriate" or "semantically inappropriate." Each sentence featured a single target word. To encourage listeners to pay attention to the entire sentence, target words did not appear in the initial position. To ensure participants understood the context (except for the target word) without too much burden on their working memory, the sentences were kept short (4–8 words), grammatically accurate, and simple without any subordination. Most of the words were chosen from the 1K most frequent word families or were proper names (93%). Although a few words (7%) were from the 2K most frequent word families, these were Japanese loan words. The target words were used in a semantically appropriate way in half of the sentences (80 out of 160) but in a semantically inappropriate way in the other 80 sentences. For example, in the case of the target word "estate," listeners were presented with the semantically appropriate sentence "*My grandfather bought an estate*" and the semantically inappropriate sentence "*My friend's estate was very kind.*" The rest of the sentence was grammatically accurate and lexically comprehensible, so it was only the use of the target word that determined the degree of semantic appropriateness. After the researchers drafted the sentences, their semantical appropriateness was carefully checked, revised, and piloted by three linguistically trained L1 English speakers numerous times until they all came to a consensus.

The 160 sentences were read by a female native speaker of General American (Talker A). The audio stimuli were presented to participants in a randomized order. For each target word, 0.5 points were given if a participant correctly accepted a contextually appropriate sentence or rejected a contextually inappropriate sentence. The total score was out of 80 points.

Phonological multiple choice

Using a format similar to that found in previous studies (e.g., McLean et al., 2015), a phonological multiple-choice task (PhonMC) was developed to tap into participants' ability to explicitly analyze form-meaning mappings of words without any orthographic cues (i.e., phonologization). The 80 target words were read aloud by one female native speaker of American English. After hearing each stimulus, participants were asked to select the correct meaning from four options (one correct answer and three distractors). All four answer options were the same part of speech, and all distractors were selected from a list of words frequently found in TOEIC test materials. As in McLean et al. (2015), all the answers and distracters were translated into Japanese in order to reduce difficulty by avoiding answer options that could be potentially confusing for Japanese learners. Before the study was conducted, three fluent Japanese speakers with extensive EFL teaching experience reviewed and provided feedback on both versions of the test (PhonMC and GenMC). Any issues with the translations of the answers or distractors into Japanese, such as translations that did not fit the context of the passage, were corrected, and any distractors that could be potentially considered correct were revised. The 80 target words were read by Talker A (a female native speaker of General American). The audio stimuli were presented in a randomized order. For each target word, 1 point was given when a participant chose the correct response. The total score was out of 80 points.

Generalization multiple choice

Another version of the PhonMC was developed in which the target words were read by a different male speaker of General American (Talker B). Listeners with robust phonological representations should be able to recognize the words in both talker

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

conditions. Listeners with less robust phonological knowledge may recognize target words produced by one talker but not the other. The total score was out of 80 points.

Analyses

Using one spontaneous measure (LJT) and two controlled measures (PhonMC and GenMC), three stages of L2 phonological vocabulary knowledge were analyzed as follows:

- **Phonologization:** PhonMC scores (1 point × 80 words) were used to determine the degree to which participants were able to recognize the target words without orthographic cues.
- **Generalization:** The Euclidean distance between scores on the PhonMC produced by Talker A (1 point × 80 words) and GenMC produced by Talker B (1 point × 80 words) was calculated. This allowed us to determine the extent to which participants' phonological vocabulary knowledge could be applied across varying speaker conditions. A longer distance indicated that one's phonological vocabulary score was higher for Talker A than Talker B or vice versa, whereas a smaller distance suggested that participants had more robust knowledge.
- **Automatization**: To measure automatization, both accuracy and fluency scores of the spontaneous task (LJT) were used. Accuracy scores were determined by the total number of correct responses (0.5 points × 80 words × 2 contexts [contextually appropriate and inappropriate]). Fluency was operationalized as the coefficient of variation (CV), calculated by dividing the standard deviation of participants' reaction times by their mean reaction times. It has been argued that CV can be used as an index of automaticity when reaction time and standard deviation values are positively correlated (Segalowitz, 2010). In the context of L2 vocabulary learning, when new lexical items are more strongly integrated into the system through repeated encounters, such knowledge can be accessed more quickly with less variability across trials (Elgort & Warren, 2014). L2 listeners' vocabulary processing speed and stability (operationalized via CVs) have been found to correlate with listening proficiency (Hui & Godfroid, 2021). In this study, higher accuracy scores and lower CV scores in the LJT were interpreted as evidence of strongly automatized knowledge, reflecting participants' ability to accept semantically appropriate sentences and reject semantically inappropriate sentences in a faster and more stable manner.

Perceptual-cognitive measures

To determine participants' individual differences in perceptual-cognitive abilities, which have been found to be related to L2 listening proficiency, they completed a digit span task (to measure working memory; Wallace, 2022) and an auditory discrimination task (to measure auditory processing; Vandergrift & Baker, 2015). The digit span task consisted of two parts: the forward span and the backward span. For the forward span, participants were presented with a series of digits (with each digit shown for 500 ms) and then asked to recall them in the same order. For the backward span, they were asked to recall the digits in reverse order. Participants entered their responses in a space provided on their computer screen. The series of digits ranged from three to eleven digits, and participants completed two trials at each length. The highest number of digits in the series that they correctly recalled in both trials served as their score for each

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

span. Their working memory scores were determined by averaging their scores for the forward and backward spans.

Afterwards, participants completed the two subcomponents of the individually adaptive auditory discrimination task (spectral and temporal processing; Saito et al., 2020). In each subtest, they were asked to complete a series of AXB discrimination tasks. They were presented with three sounds and asked to identify which one was different from the others. The stimuli were nonverbal sounds that were identical except for the target acoustic dimension (second formant for spectral processing; amplitude rise time for temporal processing). The tests were designed to determine how small of a difference in sound participants could detect. It is thought that learners with precise auditory processing abilities are able to encode more detailed acoustic characteristics of sounds and thus demonstrate more advanced L2 speech proficiency. Following Kachlicka et al. (2019), participants' spectral and temporal processing scores were standardized and averaged to provide a single auditory processing score per participant.

Metacognitive awareness measures

Another crucial variable that affects L2 listening proficiency is metacognitive awareness, defined as "listener awareness of the cognitive processes involved in comprehension, and the ability to oversee, regulate, and direct these processes" (Vandergrift & Baker, 2015, p. 395). In this view, metacognitive awareness of L2 listening comprises five components: (a) problem-solving (guessing what has not been understood), (b) planning and evaluation (using strategies to prepare for L2 listening tasks), (c) translation (avoiding direct translation), (d) person knowledge (perception of difficulty and self-efficacy in L2 listening), and (e) directed attention (concentrating and staying on task). The literature has shown that those with greater metacognitive awareness tend to demonstrate more advanced L2 listening proficiency (In'nami et al., 2023). To measure participants' metacognitive awareness of L2 listening, they completed the Metacognitive Awareness Listening Questionnaire (MALQ; Vandergrift et al., 2006). Participants responded to a total of 21 statements on a 6-point scale (1 = strongly disagree, 6 = strongly agree). To avoid confusion, the original statements in the MALQ were translated into Japanese. Following Vandergrift et al., raw scores were averaged for each of the five components.

Results

Relationships between the LJT, PhonMC, and GenMC

All the vocabulary tasks demonstrated adequate internal consistency: LJT, $\alpha = .93$, 95% CI [.92, .95]; PhonMC, $\alpha = .95$, 95% CI [.94, .96]; and GenMC, $\alpha = .98$, 95% CI [.97, .99]. As summarized in **Supporting Information S1**, the descriptive statistics showed that participants' LJT, PhonMC, and GenMC scores were comparable to a normal distribution ($D = .062$, .044, and .052, $p > .05$). As for temporal measures, participants' reaction time and standard deviation values in the LJT were positively correlated, $r = .682$, $p < .001$, 95% CI [.511, .853]. Thus, CV could be used to measure the speed and stability of lexical retrieval (Segalowitz, 2010). As participants' CV scores in the LJT task demonstrated significant deviation from a normal distribution ($D = .380$, $p < .05$), raw values were transformed via a log10 function. The resulting scores did not significantly deviate from normal distribution ($D = .105$ $p = .113$) and were used for the rest of the analyses.

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

The results of Pearson correlation analyses showed that the PhonMC scores were moderately correlated with accuracy scores on the phonological lexicosemantic judgement task, LJT: $r = .50$, $p < .001$, 95% *CI* [.36, .62]. A paired-sample *t* test found that PhonMC scores were significantly higher than LJT scores, with large effects, $t = 12.787$, $p < .001$, $d = 7.54$. Despite some degree of overlap in the constructs of these tasks, it is possible that they tap into two distinct modalities of L2 knowledge, with the PhonMC measuring more controlled processing abilities and the LJT measuring more spontaneous processing abilities; the ability of L2 learners to encode contextual and collocational associations between words may differ from their ability to recognize these words presented in isolation. The two different versions of the multiple-choice task (PhonMC for Talker A, GenMC for Talker B) demonstrated strong correlations, $r = .92$, $p < .001$, 95% *CI* [.89, .94]. Participants' performance did not significantly differ between the two talkers ($t = -0.060$, $p = .952$, $d < .01$).

Measurements of phonologization (PhonMC), generalization (Euclidian distance between Talker A [PhonMC] and Taker B [GenMC]), and automatization (accuracy and CV on the LJT) were compared via Pearson correlation analyses (alpha was set at .016; Bonferroni corrected). Phonologization scores were not significantly associated with generalization scores, $r = -.147$, $p = .103$, 95% CI [−.31, .03]. As shown above, phonologization was significantly associated with LJT-accuracy ($r = .50$); interestingly, the link between generalization and LJT-accuracy also reached statistical significance, $r = -.301$, $p < .001$, 95% CI [−.45, −.13]. Neither phonologization nor generalization was significantly correlated with LJT-CV ($p > .238$). The results suggest that those who can better recognize words regardless of talker conditions (for whom there is a smaller distance between Talkers A and B) tend to demonstrate more automatized vocabulary knowledge.

Finally, a confirmatory factor analysis was performed to test our assumption that the four different vocabulary scores (LJT-accuracy, LJT-CV, PhonMC, GenMC) can be relatively independent (without much overlap) and they uniquely contribute to a single latent construct of phonological vocabulary knowledge. The four vocabulary scores were used as observed variables. The model showed a good overall fit with the data ($\chi^2 = 1.635$, $df = 2$, $p = .442$; comparative fit index = 1.000; Tucker–Lewis index = 1.038; root mean square error of approximation = .000; standardized root mean square residual = .029). As visually depicted in Figure 1, the four vocabulary measures could be considered separate measures but all of them tap into participants' phonological vocabulary

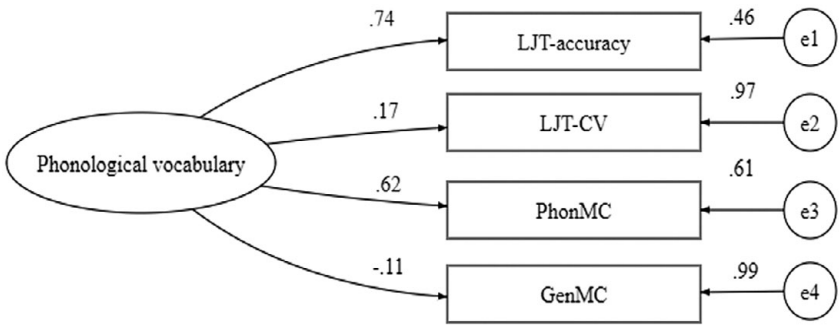

Figure 1. Final model of phonological vocabulary knowledge relative to four different vocabulary test scores. All values were standardized.

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

knowledge. The results of the factor analyses echoed the above-mentioned weak-to-medium correlations among the four different vocabulary scores.[1]

Roles of phonological vocabulary in L2 listening proficiency

The next objective of statistical analyses was to examine the relationship between phonological vocabulary knowledge and L2 listening proficiency (measured via the TOEIC). Participants' TOEIC performance showed high reliability, α = .92, 95% CI [.91, .94]. Participants' general listening proficiency scores varied widely: M = 49.4, SD = 9.8, 95% CI [47.6, 51.1], and their scores did not show significant deviation from a normal distribution (D = .257, p = .085). As summarized in Figure 2, Pearson correlation analyses were performed to examine how different aspects of phonological vocabulary knowledge (phonologization, generalization, and automatization) were related to listening proficiency (alpha was set at .010; Bonferroni corrected). Although participants' controlled vocabulary knowledge (PhonMC) was moderately associated with their TOEIC scores (r = .43), their spontaneous phonological knowledge (LJT-accuracy) showed medium-to-strong correlations with TOEIC scores (r = .66). The relationship between LJT-CV and TOEIC scores was marginal (r = −.15, p = .095).

Next, we examined whether the predictive power of phonological vocabulary knowledge varied according to participants' listening proficiency levels—low, mid, and high. A total of 126 participants were divided into three proficiency groups (for descriptive statistics, see **Supporting Information S2**). Their scores on the three subcomponents of the TOEIC (i.e., 30 points in Question-Response, 30 points in Conversations, and 30 points in Monologues) were submitted to cluster analysis using the k-means method with the number of clusters defined at three. This resulted in the following three groups: low (n = 27), mid (n = 61), and high (n = 38). The values for the 95% CIs showed that total TOEIC scores differed substantially between low, M = 35.6, SD = 5.7, 95% CI [33.3, 37.9]; mid, M = 47.9, SD = 3.6, 95% CI [47.0, 48.8]; and, high, M = 61.0, SD = 6.0, 95% CI [59.0, 63.0], groups. To detect which lexical factors—phonologization (PhonMC), generalization (Euclidian distance between PhonMC and GenMC), and/or automatization (LJT-accuracy and -CV)—distinguished between the two group contrasts (low vs. mid, mid vs. high), a set of nonparametric Mann–Whitney U tests were performed (Bonferroni corrected; the alpha level was set at .025). As summarized in Table 1, participants' LJT scores significantly differed across all proficiency levels (p < .01), phonologization scores were predictive of placement in the low- versus mid-proficiency group (p < .001), and generalization scores were predictive of placement in the mid- versus high-proficiency group (p = .001). The findings suggest that automatization plays an important role in every stage of L2 listening development, phonologization is developed at a relatively early stage, and generalization is developed at a later stage.

[1]The correlation and factor analyses strongly support our initial hypothesis about the relative independence among the four measures (LJT-accuracy, LJT-CV, PhonMC, GenMC). Although there may be models that provide a better fit to the data, we are reluctant to conduct further model comparisons via another round of confirmatory factor analyses. This is primarily because such analyses not only deviate from the core focus of our study but could also invite Type 1 and Type 2 errors due to the limited size of our dataset. Importantly, our study consisted of only 126 participants, a number barely surpassing the "minimum" threshold (n = 100) as proposed by Kline (2005). For a more robust confirmatory factor analysis, an ideal sample size would be around 200 (e.g., Wallace, 2022, for a sample size of n = 226).

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

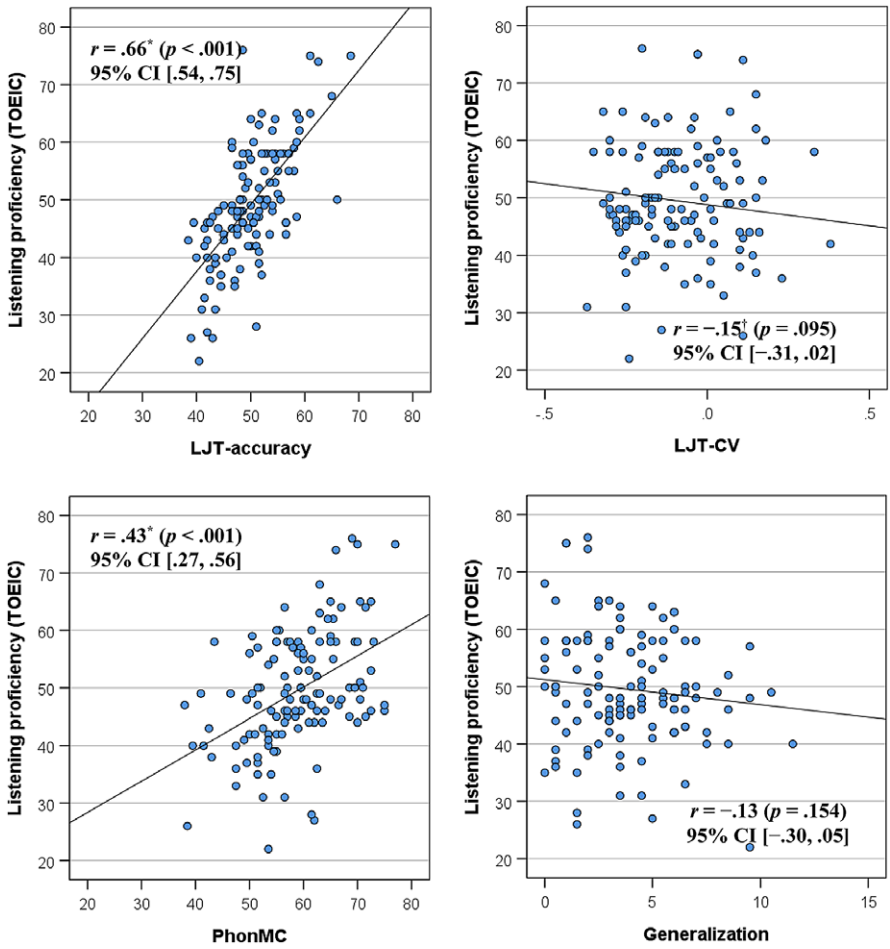

Figure 2. Correlations between L2 listening proficiency (*y*-axis) and phonological vocabulary knowledge (*x*-axis). The lexicosemantic judgement task (LJT-accuracy, LJT-CV) was used to measure automatization, the PhonMC was used to measure phonologization, and the GenMC was used to measure generalization. $^{*}p < .05$; $^{\dagger}p < .10$.

The final objective of the analyses was to investigate the influence of phonological vocabulary knowledge on L2 listening proficiency while controlling for other perceptual, cognitive, and metacognitive factors. To this end, a multiple regression analysis was conducted using participants' TOEIC scores as the dependent variable and various other factors as predictor variables, including different aspects of phonological vocabulary knowledge, auditory processing, working memory, and five dimensions of metacognitive abilities (problem-solving, planning and evaluation, translation, person knowledge, and directed attention). To avoid issues of multicollinearity, we confirmed that no predictor variables exceeded a variance inflation factor of 2 (Range = 1.11 to 1.80). A compromise power analysis via G*Power (Faul et al., 2009) revealed that the model of 11 predictors with 126 participants possessed sufficiently strong power (.994) to achieve a relatively large effect size (R^2 = .69; Wallace, 2022). The entire model was

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

Table 1. Summary of phonologization, generalization, and automatization scores as per different levels of L2 listening proficiency

	Low vs. Mid (Mann−Whitney *U*)			Mid vs. High (Mann−Whitney *U*)		
	z	*p*	*d*	*z*	*p*	*d*
Phonologization						
PhonMC (80 points)	−4.601	< .001*	1.13	−1.746	.081†	0.35
Generalization						
Euclidian distance (PhonMC vs. GenMC)	−.599	.549	0.16	−2.849	.004*	0.63
Automatization						
LJT−accuracy (80 points)	−4.242	< .001*	1.14	−3.382	< .001*	0.80
LJT−CV	−.982	.326	0.37	−.664	.507	0.13

Note. *$p < .025$; †$p < .10$.

found to significantly explain 50.7% of the variance in participants' listening test scores, $F(11, 115) = 8.692$, $p < .001$. As summarized in Table 2, significant and marginally significant predictors were, in descending order, LJT-accuracy ($\beta = .503$, $p < .001$), person knowledge ($\beta = .201$, $p = .025$), LJT-CV ($\beta = -.195$, $p = .014$), and PhonMC ($\beta = .185$, $p = .067$). To investigate the relative importance of the lexical, perceptual, cognitive, and metacognitive predictors in the full model ($R^2 = .507$), dominance analysis was conducted (Mizumoto, 2023). The results showed that 77.6% of the variance in the regression model for L2 listening proficiency ($R^2 = .507$) could be explained by three lexical phenomena—automatization (55.3% [LJT-accuracy, LJT-CV]), phonologization (20.8%), and generalization (1.5%)—and a lesser amount by metacognitive strategy use (21.3%) and perceptual-cognitive abilities (1.2%).

Discussion and future directions

Nation (2013) posited that knowing a spoken word encompasses understanding what words sound like and what they signify (i.e., form-meaning mapping) as well as how they should be used on sentence levels (use-in-context). Building on Nation's framework of spoken vocabulary knowledge and guided by psycholinguistic views on L2 comprehension (Ellis, 2006), we propose the three-stage model of phonological vocabulary knowledge. In the context of adult L2 speech learning, such knowledge comprises the ability to recognize phonological forms without any orthographic cues (phonologization), the ability to retrieve such lexical knowledge across different talkers (generalization), and the ability to access the semantic and collocational aspects of words in a fast and stable manner (automatization).

This three-stage model represents an extension of Nation's (2013) framework of vocabulary knowledge, suggesting that phonologization and generalization are connected to the form-meaning aspect of vocabulary knowledge, whereas automatization corresponds to the use-in-context aspect. Our argument is twofold: (a) phonological form-meaning mapping can be conceived on different levels—specifically, perceiving the phonological form of target words regardless of speaker variations—and (b) use-in-context can be operationalized as learners' rapid, efficient, stable, and effortless retrieval of target words in relation to surrounding words as part of automatized lexical chunks.

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

Table 2. Summary of multiple regression of listening proficiency relative to lexical, perceptual, cognitive, and metacognitive predictors

			95% CI (B)					Relative weight	
	B	*SE*	Upper	Lower	*β*	*t*	*p*	Raw weight	Rescaled weight
Intercept	−16.710	10.037	−36.642	3.222		−1.665	.099		
LJT−accuracy	.864	.176	.513	1.214	.503	4.898	<.001*	.252	49.9%
LJT−CV	−9.337	3.735	−16.755	−1.919	−.195	−2.500	.014*	.027	5.4%
PhonMC	.215	.116	−.016	.446	.185	1.850	.067†	.105	20.8%
Generalization	.244	.261	−.274	.761	.075	.935	.352	.007	1.5%
Auditory processing[a]	.348	1.152	−1.940	2.635	.025	.302	.763	.002	0.5%
Working memory[b]	.552	.792	−1.022	2.125	.053	.696	.488	.003	0.6%
Problem−solving[c]	−.121	1.459	−3.020	2.777	−.008	−.083	.934	.002	0.4%
Planning and evaluation[c]	−.212	.988	−2.173	1.750	−.020	−.214	.831	.004	0.8%
Translation[c]	.730	.865	−.989	2.448	.073	.843	.401	.028	5.6%
Person knowledge[c]	2.435	1.071	.309	4.561	.201	2.274	.025*	.071	14.1%
Directed attention[c]	−.690	1.467	−3.604	2.223	−.040	−.471	.639	.002	0.4%

Note. *$p < .025$; †$p < .10$.
[a]for combined scores of spectral and temporal processing.
[b]for combined scores of forward and digit span.
[c]from Metacognitive Listening Awareness Questionnaire.

The current study reexamined the differential roles of phonologization, generalization, and automatization facets of phonological vocabulary knowledge in L2 listening proficiency (measured via the TOEIC) for 126 Japanese EFL learners, taking into account their proficiency level (low, mid, or high) and other factors related to perception (auditory processing), cognition (working memory), and metacognition. Following the methodological discussions and validations of phonological vocabulary knowledge assessment from previous studies (McLean et al., 2015; Uchihara et al., 2023), we employed two controlled measures for phonologization and generalization (multiple choice [PhonMC, GenMC]) and one spontaneous measure for automatization (LJT) in this study. Overall, results yielded three specific findings.

First, we successfully replicated the results of a prior project (Uchihara et al., 2023), suggesting that L2 listening proficiency may be more strongly associated with spontaneous, contextualized, and automatized phonological vocabulary knowledge ($r = .66$ for LJT-accuracy) than controlled and declarative phonological vocabulary knowledge ($r = .43$ for PhonMC). Second, participants' automatized knowledge (LJT-accuracy) served as the primary determinant of L2 listening comprehension at all proficiency levels, phonologization (PhonMC) was a significant predictor of low-to-mid L2 listening proficiency, and generalization (Euclidian distance between PhonMC and GenMC) was a significant predictor of mid-to-high L2 listening proficiency. Third, the amount of variance explained by phonological vocabulary remained strong (77.6%) even after participants' perceptual and cognitive abilities (1.2%) and metacognitive awareness (21.3%) were factored into the full regression model ($R^2 = .507$).

The results of the current study have a range of implications for our understanding of the mechanisms underlying the attainment of successful L2 listening skills. Our study aligns with other research that has found that although listeners' top-down skills (i.e., metacognitive strategy use) are weakly associated with L2 listening proficiency ($r = .20$–$.30$; In'nami et al., 2023) and can explain approximately a quarter of the variance (21.3%), vocabulary factors account for the largest amount of variance ($r = .60$–$.07$; Smith, 2019; Zhang & Zhang, 2020).

Furthermore, the detailed analyses of phonological vocabulary revealed that these large lexical effects (dominance weight = 77.6%) uniquely derived from three different abilities—automatization (55.3%), phonologization (20.8%), and generalization (1.5%). Thus far, most studies have focused on the phonologization aspect of vocabulary knowledge, typically assessed through controlled tasks such as phonological multiple choice tests (McLean et al., 2015). The phonological vocabulary knowledge evaluated here aligns with what is referred to as "form-meaning mappings" in prevalent vocabulary knowledge models (i.e., understanding the sound of the word and its associated meaning; Nation, 2013). However, few studies have investigated the generalizability of spoken word recognition (i.e., the ability to perceive a word irrespective of talker variations; Thomson, 2018) and the use aspect of phonological vocabulary knowledge (how the word occurs in patterns, what words it aligns with, and where/when/how it is used; Schmitt, 2019).

Drawing on the usage-based account of L2 comprehension (e.g., Ellis, 2006), we adopted both GenMC and LJT. The former mirrors the controlled measure of PhonMC but with a different speaker, whereas the latter was designed to assess participants' spontaneous ability to select (and reject) words in a manner that is semantically and collocationally compatible with the surrounding context. Crucially, our study implies that the contribution of phonologization to L2 listening proficiency (20.8%) can be differentiated from that of generalization (1.5%) and is substantially less than that of automatization (55.3%).

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

These results provide empirical backing for our proposed three-stage model of phonological vocabulary knowledge and assessment: (a) conventional controlled measures (PhonMC) explore the initial stage of form-meaning mapping (i.e., whether learners can recognize the meaning of a word without orthographic cues; McLean et al., 2015), (b) the same measure with a different speaker (GenMC) is necessary to capture L2 learners' process of generalizing phonological vocabulary knowledge (i.e., whether learners' perception of the word can resist the influence of differing speaker voices; Thomson, 2018), and (c) spontaneous measures (LJT) may provide a more accurate representation of L2 learners' vocabulary knowledge that is directly relevant to real-world L2 listening experiences (i.e., the extent to which learners can contextually use the word as a part of automatized lexical chunks; Ellis, 2006).

Finally, an examination of the link between three different aspects of phonological vocabulary knowledge (phonologization, generalization, automatization) and different levels of L2 listening proficiency (low, mid, high) provides suggestive patterns about how L2 phonological vocabulary develops over time and how it should be taught. As the initial stage of L2 listening proficiency development (low → mid) is characterized by phonologization, learners at this stage should be guided to attend to both orthographic and phonological forms of words (McLean et al., 2015). Suprasegmental instruction can be effective for segmentation on word and sentence levels, whereas segmental instruction can facilitate the phoneme-level refinement of words (Kissling, 2018). As generalization relates to a later stage of L2 listening proficiency development (mid → high), learners at this stage should be encouraged to attain more robust phonological representations via exposure to the phonological forms of words as they are produced by multiple talkers (Thomson, 2018, for high-variability phonetic training) and under various conditions (Leong et al., 2018, for noise-based training).

The most critical implication of our analyses is that automatization may be central to every phase of L2 listening proficiency development (low → mid → high). In line with the skill acquisition account of instructed SLA (DeKeyser, 2017; Suzuki, 2023), more attention should be given to the teaching of automatization because this is the aspect of vocabulary knowledge that is most essential for L2 listening proficiency. For successful L2 comprehension, listeners need to be able not only to explicitly analyze phonological form-meaning mappings (phonologization) but also to access such acquired knowledge in more global and real-life contexts across different talker conditions (generalization) and processing abilities (automatization).

However, surprisingly little is known about how to facilitate the acquisition of automatized and spontaneous phonological vocabulary knowledge in classroom contexts. To become proficient listeners, students must not only learn the phonological form and meaning of new words but also attend to their semantic and collocational associations with surrounding words and increase their processing speed and stability. Though limited in the literature, there are some suggestions on how to facilitate this process, such as explicit instruction on multiword items rather than single words (Pellicer-Sánchez, 2019), enhancing awareness of collocation and advanced L2 proficiency (Tavakoli & Uchihara, 2020, for fluency; Saito, 2020, for comprehensibility and appropriateness), and simultaneous focus on form and meaning (Ellis et al., 2019, for task-based language teaching).

With an eye toward future studies on the relationship between phonological vocabulary and L2 listening, there is a range of topics worthy of further investigation. First, the findings of the current study were limited to Japanese EFL learners without any experience abroad. To examine the generalizability of the findings, they should be replicated with other groups of L2 learners with different levels of immersion experience (Trofimovich & Baker, 2006), L1 backgrounds (e.g., Indo vs. non-Indo-European

https://doi.org/10.1017/S027226312300044X Published online by Cambridge University Press

languages; Saito et al., 2019), L1–L2 distance (Jaekel et al., 2023), L2 proficiency (Grüter et al., 2023), and aptitude profiles (Linck et al., 2013). Second, many scholars are increasingly conceptualizing L2 listening proficiency as the ability to understand not only L1 speakers but also L2 speakers. Future research could further explore L2 phonological knowledge by employing both L1 and L2 listeners' voices (Thir, 2023). Third, although we took a first step toward using the LJT as a measure of automatized phonological vocabulary knowledge (Uchihara et al., 2023), the construct validity of the task needs to be further studied by examining its relation to other measures of implicit lexical knowledge (Elgort & Warren, 2014, for form and semantic priming in lexical decision tasks; for a synthesis, Suzuki & Elgort, 2023) and aptitude for proceduralization and automatization (Suzuki & DeKeyser, 2017, for procedural memory). Last, all the acquisitional and pedagogical suggestions in the current study were based on cross-sectional data. Future studies should conduct longitudinal investigations of the differential roles of phonologization, generalization, and automatization in L2 listening proficiency. It would be interesting to examine how the provision of different types of practice—for example, multiple choice for controlled knowledge, high-variability input for generalization, and timed lexicosemantic judgements for spontaneous and automatized knowledge—could help L2 learners attain different levels of L2 listening proficiency (low to mid vs. mid to high).

For the purposes of future replication and extension research, and as classroom assessments of students' L2 listening vocabulary knowledge, the three phonological vocabulary tasks (PhonMC, GenMC, and LJT) can be accessed on the open science platform for both researchers and teachers, L2 Speech Tools (Mora-Plaza et al., 2022: http://sla-speech-tools.com/).

Supplementary materials. The supplementary material for this article can be found at http://doi.org/10.1017/S027226312300044X.

Acknowledgments. We extend our gratitude to Eguchi Masaki, Satsuki Kurokawa, Noriaki Mikajiri, Noriko Nakanishi, Nobuhiro Kamiya, Konstantinos Macmillan, and Magdalena Kachlicka for their invaluable assistance in data collection and analysis. Our appreciation also extends to Yuichi Suzuki, three anonymous SSLA reviewers, and Handling Editor, Luke Plonsky, for their insightful comments. This project was made possible through funding by the Leverhulme Trust (RPG-2019-039) and the Spencer Foundation (202100074).

Competing interest. The authors declare no conflicting interests,

References

Adank, P., Smits, R., & Van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, *116*, 3099–3107. https://doi.org/10.1121/1.1795335

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*, 425–438. https://doi.org/10.1093/applin/24.4.425

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language Learning*, *62*, 49–78. https://doi.org/10.1111/j.1467-9922.2012.00706.x

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*, 388–407. https://doi.org/10.3758/s13428-019-01237-x

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*, 2074–2085. https://doi.org/10.1121/1.427952

Cheng, J., & Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing*, *35*, 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*, 213–241. https://doi.org/10.1002/tesq.3137

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, *40*, 141–201. https://doi.org/10.1177/002383099704000203

Dang, T. N. Y., Webb, S., & Coxhead, A. (2022). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research*, *26*, 617–641. https://doi.org/10.1177/1362168820911

de Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*, 223–243. https://doi.org/10.1017/S0142716413000210

DeKeyser, R. M. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of second language acquisition* (pp. 15–32). Routledge.

Du, G., Hasim, Z., & Chew, F. P. (2022). Contribution of English aural vocabulary size levels to L2 listening comprehension. *International Review of Applied Linguistics in Language Teaching*, *60*, 937–956. https://doi.org/10.1515/iral-2020-0004

Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning*, *64*, 365–414. https://doi.org/10.1111/lang.12052

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied linguistics*, *27*, 1–24. https://doi.org/10.1093/applin/ami038

Ellis, N. C., Simpson-Vlach, R. I. T. A., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, *42*, 375–396.

Ellis, R., Skehan, P., Li, S., Shintani, N., & Lambert, C. (2019). *Task-based language teaching: Theory and practice*. Cambridge University Press.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.

Flege, J., & Bohn, O-S. (2021). The revised speech learning model. In R. Wayland (Ed.), Second language speech learning: Theoretical and empirical progress (pp. 3–83). Cambridge University Press.

Flege, J. E., Takagi, N., & Mann, V. (1996). Lexical familiarity and English-language experience affect Japanese adults' perception of/ɹ/and/l. *The Journal of the Acoustical Society of America*, *99*, 1161–1173. https://doi.org/10.1121/1.414884

Foster, P., Bolibaugh, C., & Kotula, A. (2014). Knowledge of nativelike selections in a L2: The influence of exposure, memory, age of onset, and motivation in foreign language and immersion settings. *Studies in Second Language Acquisition*, *36*, 101–132. https://doi.org/10.1017/S0272263113000624

Grüter, T., Kim, J., Nishizawa, H., Wang, J., Alzahrani, R., Chang, Y. T., Nguyen, H., Nuesser, M., Onba, A., Ross, S., & Yusa, M. (2023). Language proficiency modulates listeners' selective attention to a talker's mouth: A conceptual replication of Birulés et al. (2020). *Studies in Second Language Acquisition.* Advance online publication. https://doi.org/10.1017/S0272263123000086

Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition*, *35*, 423–449. https://doi.org/10.1017/S0272263113000041

Hamada, Y., & Yanagawa, K. (2023). Aural vocabulary, orthographic vocabulary, and listening comprehension. *International Review of Applied Linguistics in Language Teaching.* https://doi.org/10.1515/iral-2022-0100

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 1570–1582. https://doi.org/10.1037/0096-1523.26.5.1570

Hui, B., & Godfroid, A. (2021). Testing the role of processing speed and automaticity in second language listening. *Applied Psycholinguistics*, *42*, 1089–1115. https://doi.org/10.1017/S0142716420000193

In'nami, Y., Cheung, M. W.-L., Koizumi, R., & Wallace, M. P. (2023). Examining second language listening and metacognitive awareness: A meta-analytic structural equation modeling approach. *Language Learning*, *73*, 759–798. https://doi.org/10.1111/lang.12548

Jaekel, N., Ritter, M., & Jaekel, J. (2023). Associations of students' linguistic distance to the language of instruction and classroom composition with English reading and listening skills. *Studies in Second Language Acquisition.* Advance online publication. https://doi.org/10.1017/S0272263123000268

Kachlicka, M., Saito, K., & Tierney, A. (2019). Successful second language learning is tied to robust domain-general auditory processing and stable neural representation of sound. *Brain and Language*, *192*, 15–24. https://doi.org/10.1016/j.bandl.2019.02.004

Kamiya, N. (2022). The limited effects of visual and audio modalities on second language listening comprehension. *Language Teaching Research.* Advance online publication. https://doi.org/10.1177/13621688221096213

Kissling, E. M. (2018). Pronunciation instruction can improve L2 learners' bottom-up processing for listening. *The Modern Language Journal*, *102*, 653–675. https://doi.org/10.1111/modl.12512

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). Guilford.

Leong, C. X. R., Price, J. M., Pitchford, N. J., & van Heuven, W. J. (2018). High variability phonetic training in adaptive adverse conditions is rapid, effective, and sustained. *PloS One*, *13*, Article e0204888.

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., Smith, B. K., Bunting, M. F., & Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning*, *63*, 530–566. https://doi.org/10.1111/lang.12011

Masrai, A. (2020). Exploring the impact of individual differences in aural vocabulary knowledge, written vocabulary knowledge and working memory capacity on explaining L2 learners' listening comprehension. *Applied Linguistics Review*, *11*, 423–447. https://doi.org/10.1515/applirev-2018-0106

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

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*, 741–760. https://doi.org/10.1177/1362168814567889

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*, 127–147. https://doi.org/10.3138/cmlr.63.1.127

Milliner, B., & Dimoski, B. (2021). The effects of a metacognitive intervention on lower-proficiency EFL learners' listening comprehension and listening self-efficacy. *Language Teaching Research.* Advance online publication. https://doi.org/10.1177/13621688211004646

Mizumoto, A. (2023). Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. *Language Learning*, *73*, 161–196. https://doi.org/10.1111/lang.12518

Mora-Plaza, I., Saito, K., Suzukida, Y., Dewaele, J. M., & Tierney, A. (2022). Tools for second language speech research and teaching. http://sla-speech-tools.com/

Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, *37*, 677–711. https://doi.org/10.1017/S0272263114000825

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*, 59–82. https://doi.org/10.3138/cmlr.63.1.59

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*. Cambridge University Press.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*, 357–395. https://doi.org/10.1037/0033-295X.115.2.357

Pellicer-Sánchez, A. (2019). Learning single words vs. multiword items. In *The Routledge handbook of vocabulary studies* (pp. 158–173). Routledge.

Pisoni, D. B., & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, *25*(1–2), 21–52. https://doi.org/10.1016/0010-0277(87)90003-5

Plonsky, L., Marsden, E., Crowther, D., Gass, S., & Spinner, P. (2020). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research*, *36*, 583–621. https://doi.org/10.1177/0267658319828413

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnen (Ed.), *Validation in language assessment* (pp. 41–60). Erlbaum.

Read, J. (2020). Key issues in measuring vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 545–560). Routledge.

Révész, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, *35*, 31–65. https://doi.org/10.1017/S0272263112000678

Saito, K. (2013). Age effects on late bilingualism: The production development of/ɹ/by high-proficiency Japanese learners of English. *Journal of Memory and Language*, *69*, 546–562. https://doi.org/10.1016/j.jml.2013.07.003

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*, 250–277. https://doi.org/10.1111/ijal.12026

Saito, K. (2020). Multi-or single-word units? The role of collocation use in comprehensible and contextually appropriate second language speech. *Language Learning*, *70*, 548–588. https://doi.org/10.1111/lang.12387

Saito, K., Kachlicka, M., Sun, H., & Tierney, A. (2020). Domain-general auditory processing, age, experience, and post-pubertal L2 speech learning: A behavioral and neurophysiological investigation. *Journal of Memory and Language*, *115*, Article 104168. https://doi.org/10.1016/j.jml.2020.104168

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*, 652–708. https://doi.org/10.1111/lang.12345

Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (2019). How do second language listeners perceive the comprehensibility of foreign-accented speech? Roles of first language profiles, second language proficiency, age, experience, familiarity, and metacognition. *Studies in Second Language Acquisition*, *41*, 1133–1149. https://doi.org/10.1017/S0272263119000226

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*, 261–274. https://doi.org/10.1017/S0261444819000053

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching*, *50*, 212–226. https://doi.org/10.1017/S0261444815000075

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language testing*, *18*, 55–88. https://doi.org/10.1177/026553220101800104

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

Smith, G. (2019, March 9). *The relationship between L2 vocabulary knowledge and listening comprehension ability: A meta-analysis* [Paper presentation]. 2019 Conference of the American Association for Applied Linguistics, Atlanta, GA, United States.

Spinner, P., & Gass, S. M. (2019). *Using judgments 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*, 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. https://doi.org/10.4324/9781003414643

Suzuki, Y., & DeKeyser, R. (2017). The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning*, *67*, 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* (pp. 206–234). Routledge.

Taguchi, N. (2011). Teaching pragmatics: Trends and issues. *Annual Review of Applied Linguistics*, *31*, 289–310.

Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning*, *70*, 506–547. https://doi.org/10.1111/lang.12384

Thir, V. (2023). Co-text, context, and listening proficiency as crucial variables in intelligibility among nonnative users of English. *Studies in Second Language Acquisition*. Advance online publication. https://doi.org/10.1017/S0272263123000207

Thomson, R. I. (2018). High variability [pronunciation] training (HVPT): A proven technique about which every language teacher and learner ought to know. *Journal of Second Language Pronunciation*, *4*, 208–231. https://doi.org/10.1075/jslp.17038.tho

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, *28*, 1–30. https://doi.org/10.1017/S0272263106060013

Uchihara, T. (2023). How does the test modality of weekly quizzes influence learning the spoken forms of second language vocabulary? *TESOL Quarterly*, *57*, 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*, 564–587. https://doi.org/10.1002/tesq.453

Uchihara, T., Saito, K., Kurokawa, S., Takizawa, K., & Suzukida, Y. (2023). *Declarative and automatized phonological vocabulary: Recognition, recall, lexicosemantic judgement, and employability of words in L2 Listening* [Manuscript submitted for publication]. Tohoku University, Graduate School of International Cultural Studies.

Vafaee, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition*, *42*, 383–410. https://doi.org/10.1017/S0272263119000676

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*, 457–479. https://doi.org/10.1093/applin/ams074

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, *40*, 191–210. https://doi.org/10.1017/S0261444807004338

Vandergrift, L., & Goh, C. C. (2012). *Teaching and learning second language listening: Metacognition in action*. Routledge.

Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, *65*, 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*, 79–100. https://doi.org/10.3138/cmlr.3906

Vandergrift, L., Goh, C. C., Mareschal, C. J., & Tafaghodtari, M. H. (2006). The metacognitive awareness listening questionnaire: Development and validation. *Language Learning*, *56*, 431–462. https://doi.org/10.1111/j.1467-9922.2006.00373.x

Wallace, M. P. (2022). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*, *72*, 5–44. https://doi.org/10.1111/lang.12424

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, *113*, 1033–1043. https://doi.org/10.1121/1.1531176

Wang, Y., & Treffers-Daller, J. (2017). Explaining listening comprehension among L2 learners of English: The contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System*, *65*, 139–150. https://doi.org/10.1016/j.system.2016.12.013

Webb, S., Uchihara, T., & Yanagisawa, A. (2023). How effective is second language incidental vocabulary learning? A meta-analysis. *Language Teaching*, *56*, 161–180. https://doi.org/10.1017/S0261444822000507

Werker, J. F. (2018). Perceptual beginnings to language acquisition. *Applied Psycholinguistics*, *39*, 703–728.

Yanagawa, K. (2023). The role of bottom-up strategy instruction and proficiency level in L2 listening test performance: an intervention study. *Language Awareness.* Advance online publication. https://doi.org/10.1080/09658416.2022.2161557

Zhang, S., & Zhang, X. (2020). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, 1362168820913998. https://doi.org/10.1177/1362168820913998

Cite this article: 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,* 1–27. https://doi.org/10.1017/S027226312300044X