Timed vs. Untimed Lexicosemantic Judgement Task for Measuring Automatized Phonological Vocabulary Knowledge

Kazuya Saito[1, 2]
Izumi Hosaka[1]
Yui Suzukida[1, 2]
Kotaro Takizawa[1, 3]
Takumi Uchihara[2]

[1] University College London, UK
[2] Tohoku University, Japan
[3] Waseda University, Tokyo, Japan

Correspondence concerning this article should be addressed to Kazuya Saito, University College London, 20 Bedford Way, London, WC1H0AL, United Kingdom Email: k.saito@ucl.ac.uk

Acknowledgement

This research was supported by a Leverhulme Trust Grant (RPG-2024-391) and a UK-ISPF Grant (1185702223), awarded to Kazuya Saito. We thank the anonymous reviewers and Editor Guilherme Garcia for their insightful comments on an earlier version of the manuscript.

Abstract

Given the critical role of phonological vocabulary in L2 listening proficiency, scholars have increasingly focused on measuring not only the form-meaning mapping aspects of vocabulary knowledge but also the extent to which learners can access this knowledge accurately and promptly in real-time contexts. To assess the *automatized* dimension of vocabulary knowledge, we built upon prior efforts to develop the Lexicosemantic Judgment Task (LJT). Study 1 successfully replicated previous findings, showing that LJT outcomes could be distinguished from measures of declarative vocabulary knowledge (meaning recognition) and that LJT exhibited relatively strong associations with general L2 listening proficiency, even when controlling for other influencing factors. To further refine the construct validity of LJT as a measure of automatized L2 knowledge, Studies 2 and 3 developed and tested a timed version of the LJT. The results of the timed LJT, compared to the untimed version in Study 1, revealed similar links with listening proficiency, suggesting that the effect of time pressure on performance may be relatively minor. The test materials are provided in various formats to facilitate future research and pedagogical applications.

Key words: Listening, Vocabulary, Automatization, Comprehension, Individual Differences

Introduction

Successful second language (L2) comprehension is a fundamental skill in effective L2 speech communication and involves two broader levels of processing: bottom-up and top-down (Vandergrift & Goh, 2012). In bottom-up processing, learners begin by detecting prosodic structures to segment speech into words (Cutler, Dahan, & Van Donselaar, 1997) and use segmental details to distinguish phonologically similar words (e.g., minimal pairs; Munro & Derwing, 2006), which helps them identify a wider range of infrequent and sophisticated words (Wallace, 2007). They also recognize morphological features to interpret grammatical patterns (Vafaee & Suzuki, 2020) and understand a speaker's intention by contextualizing conversational, societal, and cultural cues (Taguchi, 2011) as well as interpreting extralinguistic cues such as vocal tone, facial expressions, and gestures (Kamiya, 2022). Regarding top-down processing, effective listeners are adept at employing various strategies, including predicting and inferring meaning, as well as planning, monitoring, and evaluating their comprehension (Vandergrift & Goh, 2012).

Given that L2 listening comprehension requires the development of various types of knowledge and processing skills, scholars have extensively investigated the factors that contribute to the attainment of L2 listening comprehension (e.g., Vafaee & Suzuki, 2020; Vandergrift & Baker, 2015; Wallace, 2022). This line of research is critical for developing an optimized syllabus for learners seeking to improve their listening skills in an efficient and effective manner. The extant literature has consistently identified phonological vocabulary knowledge as by far the strongest determinant: Learners who achieve higher scores in general L2 listening proficiency tests tend to possess larger phonological vocabularies (r = .5-.7; see Zhang & Zhang, 2020 for a meta-analysis). In his methodological synthesis of L2 vocabulary research, however, Schmitt (2019) rightly pointed out that most extant studies have focused almost exclusively on the form-meaning mapping aspects of vocabulary knowledge, typically measured through meaning recognition and recall tasks.

According to Nation's (2013) oft-cited framework, phonological vocabulary knowledge relevant to L2 listening comprises two aspects of vocabulary knowledge: form-meaning mapping and use-in-context. The first dimension, form-meaning mapping, involves linking the sound of a word to its meaning. The second dimension, use-in-context, refers to accessing form-meaning knowledge in semantically, collocationally, and grammatically appropriate ways relative to surrounding words. The former pertains to recognising target words in isolation (which may be one or two to three words), while the latter highlights the ability to comprehend target words in relation to other words at the broader clause and sentence levels. In essence, L2 listening proficiency develops through a transition from form-meaning mapping to use-in-context aspects of phonological vocabulary knowledge. As Schmitt (2019) highlighted, however, it is surprising that few studies have examined the use-in-context aspects of vocabulary knowledge, even though this dimension better captures the lexical processing required in real-world L2 listening conditions.

To provide further theoretical and methodological advancements on this topic, certain scholars (Saito et al., 2025; Uchihara et al., 2025) have begun conceptualizing the two-step developmental aspects of phonological vocabulary knowledge (form-meaning mapping → use-in-context) within the framework of the declarative, proceduralization, and automatization distinction proposed in skill acquisition theory for instructed L2 learning—a framework traditionally used to model the development of rule-based morphosyntactic knowledge

(DeKeyser, 2017). Notably, this represents the first attempt to introduce the skill acquisition paradigm in the context of L2 vocabulary learning and teaching.

According to the skill acquisition framework, L2 learning is characterized by a transitional progression across three types of knowledge: declarative knowledge ("knowing *what*"), procedural knowledge ("knowing *how*" in controlled contexts), and automatized knowledge ("knowing how" in *real-life* contexts). Given the parallels between these models, Saito et al. (2025) and Uchihara et al. (2025) have taken an initial step towards aligning Nation's (2013) model of spoken vocabulary knowledge—which distinguishes between form-meaning mapping and use-in-context—with the declarative–procedural–automatized continuum outlined in skill acquisition theory (DeKeyser, 2017).

In this interdisciplinary conceptualization, Saito and Uchihara argue that form-meaning mapping corresponds to declarative and procedural knowledge—that is, the explicit association between a word's sound and meaning, typically assessed through meaning recognition tasks. At this stage, learners establish declarative representations of words through explicit, form-focused instruction and practice accessing these representations through controlled activities, such as multiple-choice quizzes and flashcards. This reflects the declarative-procedural distinction, involving representational changes tied to the symbolic aspects of knowledge.

Next, Saito and Uchihara conceptualize use-in-context as aligned with automatized knowledge—that is, the ability to access and integrate lexical knowledge quickly, accurately, and consistently within broader sentential contexts. Early on, learners gradually develop the ability to use vocabulary appropriately in context through repeated practice in sentence-level activities that require conscious attention. With prolonged exposure to authentic L2 aural input (e.g., TV shows, films, and conversations with L1 and L2 speakers), learners increasingly integrate lexical knowledge into surrounding linguistic material as cohesive chunks, enabling more accurate, rapid, and stable processing with less attentional demand (Ellis, 2006). This reflects the nonautomatic–automatic distinction, involving improvements in the quality of the same representations operating at the subsymbolic level of knowledge. Here, "subsymbolic" refers to knowledge that is executed rapidly and implicitly through procedural memory rather than consciously accessed symbolic representations, consistent with how automatization is characterized in skill acquisition theory.

Crucially, the latter stage, automatized knowledge, is viewed as the ultimate goal, representing the level of knowledge typically attained by advanced L2 users in communicative listening (and speaking) contexts. Here, this framework does *not* view the automatization of lexical access and the integration of lexical knowledge into context as entirely separate processes; rather, it considers context-appropriate lexical retrieval a core manifestation of automatized phonological vocabulary knowledge in authentic L2 listening. Following DeKeyser's (2017) view, automatized knowledge can be reflected through examining the automatized processing behaviors of advanced L2 learners. By extension, automatized phonological vocabulary knowledge can be viewed as synonymous with routinised, fluent lexical access in context—on the assumption that enhanced lexical access reflects enhanced lexical knowledge.

In his more recent paper, DeKeyser has begun to revise and extend skill acquisition theory to a broader range of instructed L2 learning areas, with an emphasis on deriving more direct pedagogical implications for classroom practice (see DeKeyser & Suzuki, 2025). More directly relevant to the present study, Suzuki and DeKeyser (in press) have explicitly discussed phonological vocabulary knowledge and L2 listening from a skill acquisition perspective:

> Traditionally, scores from vocabulary size tests are interpreted as demonstrating that words are "known" in a declarative form (Word X means Y). This narrow view, however, fails to capture what vocabulary scholars call "employable knowledge"…From a skill acquisition perspective, this shift from declarative to employable (i.e., automatized) knowledge is crucial: initial declarative knowledge, such as that called upon in form-meaning recognition and recall, must be automatized through repeated contextual use to become truly employable…Achieving automaticity enables accurate and efficient lexical processing needed to support fluent comprehension and production. This is not just about knowing what a word means, but being able to access and integrate that meaning effortlessly during real-time language use.

The question has now become: How can we measure such automatized phonological vocabulary knowledge and processing? As stated in skill acquisition theory for instructed L2 learning (DeKeyser & Suzuki, 2025), dualities play a crucial role in assessing the declarative and automatized dimensions of L2 knowledge. By definition, declarative knowledge is typically measured through *single*-modal tasks that focus on participants' relatively controlled performance when they can fully concentrate on the specific knowledge in question. In the context of phonological vocabulary, declarative knowledge is often assessed via meaning recognition tasks (McLean et al., 2015) or meaning recall tasks (Cheng et al., 2023), which require learners to demonstrate their explicit knowledge of form–meaning connections. In contrast, automatized knowledge is measured using dual-modal tasks that assess participants' ability to access the target linguistic structures while simultaneously using language for meaningful communication. This dual-task condition simulates real-life situations in which learners must apply target knowledge accurately, promptly, and subconsciously while attending to multiple aspects of language at once. In these contexts, learners manage both linguistic and communicative demands, reflecting the automatic, fluent use of language required for successful communication. Despite its significance for research and practice, however, surprisingly little has been known about how to measure the automatized dimension of phonological vocabulary knowledge (Suzuki & Elgort, 2025).[1]

In L2 morphosyntax literature, one of the most commonly used tests for measuring automatized knowledge is the grammaticality judgement test (GJT). Turning to L2 *morphosyntax*, one of the most commonly used tests for measuring automatized knowledge is the grammaticality judgement test (GJT). In GJTs, learners are presented with sentences containing manipulated target morphosyntactic structures and must quickly and intuitively judge whether these sentences are grammatically correct or incorrect. While understanding the overall meaning of the sentences, they need to assess the grammatical accuracy of specific features within them (see Plonsky et al., 2020 for a review).

---

[1] In the present study, we conceptualize meaning recognition/recall tasks as reflecting lexical knowledge that is primarily declarative but may involve initial proceduralization in isolated word recognition. This early procedural stage reflects learners' ability to retrieve form-meaning mappings with increasing efficiency, although processing remains largely controlled. In contrast, the LJT is intended to capture a later point on the continuum, where lexical knowledge is accessed fluently and integrated with other linguistic systems during sentence-level comprehension, spanning collocation, grammar, semantics, and pragmatic appropriateness. Because procedural and automatized knowledge are gradual and difficult to isolate empirically, our working assumption is that LJT performance reflects a more advanced, automatized stage of lexical processing rather than a fully distinct construct from procedural knowledge.

Research has examined the constructs underlying the GJT in relation to those measured by traditional tests such as metalinguistic knowledge tests (Ellis, 2005; Gutiérrez, 2013). Findings generally indicate that grammatical items in the GJT tend to load onto a single factor, which is often interpreted as representing automatized or implicit knowledge. In contrast, ungrammatical items and metalinguistic tests typically load onto a separate factor, generally associated with explicit knowledge. This distinction implies that L2 learners' ability to accept grammatical sentences in the GJT may tap into a construct that differs from what traditional metalinguistic tests measure, potentially serving as a measure of automatized L2 knowledge.

Building on the methodological practices in L2 morphosyntax (i.e., GJT), the recent paradigm has attempted to measure the *automatized* phonological vocabulary knowledge via a Lexicosemantic Judgement Task (LJT) wherein L2 learners judge whether a target word (e.g. "promotion") is used appropriately in a given sentence (e.g., "I work hard for promotion" vs. "I ate a promotion last night"; e.g., Saito et al., 2025; Uchihara et al., 2025). Following the notion of dual-modal tasks, the test is thought to tap into learners' lexical processing in the middle of global listening activities wherein they process not only lexical but also other (e.g., phonological, morphosyntactic, and pragmatic) aspects of language for grasping overall meaning. Unlike recognition tasks, which focus on isolated lexical knowledge, LJTs require learners to integrate lexical information within complex linguistic contexts, engaging multiple linguistic subsystems in real-time processing. This complexity is crucial for demonstrating the level of automaticity needed for fluent and accurate language comprehension.

In the development and validation studies (Saito et al., 2025; Uchihara et al., 2025), Japanese learners of English took a range of vocabulary tests (including LJT) and global L2 listening proficiency tests (TOEIC). Factor analyses reported by Uchihara et al. revealed a two-factor structure, with meaning recognition and meaning recall clustering together as one latent construct, and LJT loading separately as a second factor, suggesting a distinction between declarative and automatized vocabulary knowledge. This in turn suggests that the first two tasks tap into the declarative dimension and the last task taps into the automatized dimension. In terms of the vocabulary-listening link, participants' LJT scores yielded significantly stronger correlations with TOEIC listening scores (r = .6-.7) than their meaning recognition and recall scores did (r = .4-.5). Follow-up research has further shown that LJT (but not meaning-recognition) scores uniquely distinguished advanced EFL learners who were regularly engaged in what typically matters for automatization in classroom L2 learning—not only extensive amounts of EFL education but also a range of extracurricular activities, such as daily conversational use and study abroad (Saito & Uchihara, 2025; Takizawa et al., 2025). Under intensive phonological vocabulary training, L2 learners' improvement more clearly predicted global L2 listening proficiency development when measured via the LJT rather than meaning recognition (Saito et al., 2024); however, such change in the LJT relative to meaning recognition is likely to be slow, gradual, and time-extensive (Saito et al., 2026).

Motivation for Current Study

We propose the LJT as one method to assess the extent to which vocabulary knowledge has been automatized, and consequently, how prepared learners are for real-life L2 listening comprehension. In the context of our target population (i.e., university-level EFL learners), the current project was designed to develop a paper-pencil version of the LJT that teachers and learners can easily utilize in classroom settings. To further enhance both the theoretical and pedagogical potential of the LJT, the study aimed to achieve two main objectives through a series of studies (Studies 1, 2, and 3).

Study 1 aimed to conduct a replication of previous validation studies involving a total of 240 Japanese EFL students (n = 126 in Saito et al., 2025; n = 114 in Uchihara et al., 2025) with a similar population of 134 Japanese EFL learners. Following McManus (2024), the study could be defined as *exact* replication as it drew on "the selected study's entire design, methods, and procedure…without alteration" (p. 1302). In line with skill acquisition theory, which posits a developmental relationship between declarative lexical knowledge, automatized lexical knowledge, and successful L2 listening comprehension, we anticipated that the vocabulary-listening link would be more robust when vocabulary knowledge is measured through the LJT rather than through meaning recognition tasks. Thus, we predicted that the findings of the prior studies (Saito et al.; Uchihara et al.) would be replicated, specifically that the LJT would show stronger correlations with TOEIC listening scores than meaning recognition tasks (MRTs; r = .6-.7 vs. .4-.5).

Next, Studies 2 and 3 sought to further investigate a key methodological aspect of the LJT—the role of time pressure. Notably, in the prior studies, the LJT was administered without any time restrictions, allowing participants as much time as they needed. From a theoretical standpoint, this approach could be considered problematic, as automatized knowledge should be characterized not only by accurate but also rapid and stable access to target structures. In the L2 morphosyntax literature, time pressure has been highlighted as a crucial factor in measuring automatization. Research consistently indicates that timed and untimed GJTs assess distinct constructs, likely tapping into automatized (relatively implicit) knowledge versus explicit and declarative knowledge, respectively (Ellis, 2005; Gutiérrez, 2013). Eye-tracking studies further suggest that the degree of automatization can be gauged by how quickly learners access this knowledge (ideally approaching native-like speeds). Under time pressure, L2 speakers exhibit fewer regressions (i.e., backward eye movements), suggesting differences in automatic versus controlled processing between timed and untimed tasks (Godfroid et al., 2015).

While Study 2 aimed to develop a timed LJT, Study 3 further examined the relationship between the timed LJT scores and TOEIC listening scores among 61 Japanese EFL learners with varied proficiency levels. We predicted that a timed LJT would better capture the construct of automatized phonological vocabulary knowledge, and that it would show even stronger predictive power for L2 listening proficiency than the previous studies using untimed LJTs (i.e., r = .7-.8 vs. .6-.7).

Study 1: Partial Replication

Study 1 aimed to replicate key findings from prior studies (Saito et al., 2025; Uchihara et al., 2025). Specifically, the focus was on examining the triangular correlations between participants' declarative vocabulary knowledge (measured via MRT), automatized vocabulary knowledge (measured via LJT), and general listening proficiency (measured via TOEIC). Additionally, as in Saito et al. (2023), we explored whether the vocabulary-listening link would change when other influencing variables (aptitude and listening strategy use) were taken into consideration. Consequently, Study 1 can be viewed as a partial replication of Saito et al. (2023) and Uchihara et al. (2025). Given that the same methodological procedures were used, the method description has been minimized in this section. For comparability with the original studies (Saito et al.; Uchihara et al.) and other existing work (e.g., Wallace, 2022; Vafaee & Suzuki, 2020), we focused on the linguistic and cognitive correlates of L2 listening proficiency. We refer readers to Saito and Uchihara (2024), which further examined what kinds of learners (in terms of age and length of language learning) could achieve such declarative and automatized L2 phonological vocabulary knowledge and, by extension, more advanced L2 listening proficiency.

General Setup

To align with the original studies, maximize data collection, and account for the challenges posed by the global pandemic, all data collection was conducted remotely with careful guidance from trained research assistants via a video-conferencing tool (Zoom) and an online psychology experiment builder (Gorilla; Anwyl-Irvine et al., 2020). An electronic flyer was distributed to multiple universities across Japan during Summer 2023, resulting in over 150 participants expressing interest via email. These participants first received a Gorilla link, where they were invited to take an aural version of a vocabulary test (meaning recognition of 20 highly frequent words from the first 1,000-word list in the BNC/COCA Corpus; Nation, 2012), a working memory test, and a background questionnaire. This screening process ensured that participants scored at least 80% (16 out of 20 words) on the vocabulary test, indicating sufficient vocabulary knowledge for taking the listening proficiency test, and recalled at least four numbers in the forward digit span task, indicating adequate capacity to handle an online experiment of this nature with focused attention. A total of 134 Japanese EFL learners met the criteria and proceeded to the main study.

After discussing scheduling preferences with the research team, participants chose a convenient time slot and joined a Zoom room (typically together with 5-10 other participants). Guided by a trained research assistant, participants first completed the TOEIC Listening Proficiency Test (measuring general listening proficiency). Next, they were given a unique Gorilla URL and asked to complete the LJT and MRT (measuring automatized and declarative phonological vocabulary knowledge), tests of auditory processing and working memory (for perceptual-cognitive individual differences), and the MALQ (assessing listening strategy use) on their own. Whenever they had questions, they could ask the assistant via Zoom, ensuring a good level of support and smooth data collection. The entire session lasted 80-90 minutes, with brief intermissions.

Participants

The participants (N = 134) comprised 80 females and 54 males (M age = 19.9 years; Range = 18-26 years). The participants' previous EFL learning backgrounds varied widely in terms of age of learning onset (M = 10.6 years, SD = 2.9, Range = 3-14) and the total length of EFL learning prior to the university entrance (M = 1782 hours, SD = 567, Range = 600-5100 hours).

Listening Proficiency Test

Participants' general L2 listening proficiency was measured using the TOEIC test. The test materials were adopted from the Educational Testing Service's the New Official Workbook Volume 4. The test was designed to examine how well learners can understand various types of aural discourse in L2 English and it has widely been used to measure L2 learners' general listening proficiency (Cheng et al., 2022; Hamada & Yanagawa, 2023; McLean et al., 2015). Part 1 (30 items) required participants to select the best response from three options for single-sentence questions. Part 2 involved listening to a conversation between a male and a female speaker and then answering three comprehension questions by choosing the most appropriate response from four options. Part 3 asked participants to listen to a business announcement delivered by a single speaker and respond to three comprehension questions by selecting the best answer from four options. The total scores comprised 90 points.

Vocabulary Tests

Based on the in-house corpus developed from a retired version of the TOEIC test, a total of 80 target words were carefully selected for used in the LJT and MRT. Selection was made

according to word frequency (2K-8K from the BNC/COCA Corpus) and cognate status (loanwords commonly used in Japanese were excluded). As detailed in Saito et al. (2025), certain vocabulary items were prioritized when they featured phonological characteristics known to pose challenges for Japanese EFL learners. Specifically, the selected words were predominantly iambic and contained a greater number of syllables. They also included difficult segmentals—namely the English phonemes [r] and [l]—and featured consonant clusters (Saito, 2014). To capture participants' more spontaneous processing of target words and minimize overly conscious focus on these items, they first completed the LJT before moving on to the MRT. This ordering also aligns directly with prior LJT work (Saito et al., 2025; Uchihara et al., 2025), allowing for valid replication and comparability across datasets. All instructions and choices were presented in Japanese to ensure participants fully understood the tasks (see Figure 1).

For the LJT, a total of 160 unique sentences were constructed, with each sentence containing one target word. To prompt listeners to process the entire context and prevent an over-reliance on the target word's position, the target items were never placed in the sentence-initial position. To minimize the burden on working memory and ensure clear contextual understanding (independent of the target item), the sentences were intentionally kept short (ranging from four to eight words), grammatically accurate, and structurally simple, lacking any subordination. The surrounding vocabulary was carefully controlled: 93% of the non-target words were drawn from the 1K most frequent word families or were proper names. While the remaining 7% originated from the 2K most frequent word families, these specific items were limited to established Japanese loan words. The target words were used in a semantically appropriate manner in half of the total sentences (80 items) and in a semantically inappropriate manner in the remaining 80 sentences.

These test stimuli were recorded by a female native speaker of General American English and were presented in a randomized order. After listening to each sentence, participants were asked to decide whether it was semantically appropriate or inappropriate. To avoid participants' too much conscious processing of a target sentence, and to better reflect real-life L2 listening comprehension, they were asked to make the judgment as quickly as possible once the sentence had finished (see Figure 1). To help participants familiarize themselves with the task procedure, they first started with four example questions (two appropriate sentences and two inappropriate sentences). Only after we confirmed their clear understanding of the task, did they proceed to the main sentences. 0.5 points were given if a learner could correctly either accept a semantically appropriate sentence or reject a semantically inappropriate sentence. The total possible score for the LJT was 80 (80 target words × 2 semantic contexts [appropriate, inappropriate]).

FIGURE 1 Screenshot of the Lexicosemantic Judgement Task (in Japanese). The on-screen instruction in Japanese asks: “Is the sentence you just heard semantically appropriate or inappropriate?” After hearing the sentence, participants are asked to judge its semantic appropriateness by selecting the left square (“appropriate”) or the right square (“inappropriate”).

In the MRT, participants’ ability to identify target words in isolation, regardless of talker variability, was assessed. A total of 160 recordings were used, consisting of the 80 target words produced by both the female L1 speaker and a male L1 speaker of General American English, presented in a randomized order. Upon hearing a target word, participants were asked to choose the most suitable Japanese translation from four options (see Figure 2). They first familiarized themselves with a total of four practice questions, followed by the main task session (160 trials). 0.5 points were given if a learner could recognize a target word produced by one of the two talkers. The total possible score for the MRT was 80 (80 target words × 2 talkers).

FIGURE 2 Screenshot of the Meaning Recognition Task (in Japanese). The on-screen instruction in Japanese asks: “Which word did you hear?” After hearing the word (e.g., “estate”), participants are asked to select the most suitable Japanese translation from four options (e.g., “estate,” “souvenir,” “appendix,” and “customs”).

Aptitude Tests

To control for the effects of participants’ perceptual-cognitive individual differences on the vocabulary-listening link, we focused on two domain-general abilities: (a) auditory processing (how precisely participants can encode acoustic details of sounds) and (b) working memory (how well they can maintain and elaborate on perceived information).

For auditory processing, participants completed two subsets of widely used adaptive discrimination tests (Saito et al., 2025). These tests were designed to measure the smallest difference participants could detect in a specific acoustic dimension—formant frequency and amplitude rise time. Smaller threshold values (out of 100) indicated more precise auditory processing. The test-retest reliability of the test was reported to be relatively medium to high (r = .5-.6; Saito & Tierney, 2025). Each participant’s raw scores were standardized for formant and amplitude rise time, and the average was taken to create a composite auditory processing score.

For working memory, we followed Olsthoorn et al. (2014) to assess the phonological loop (responsible for storing information) and the central executive (responsible for processing information for cognitive tasks) using forward and backward digit span tasks, respectively. In the forward span task, participants were asked to memorize a series of digits and then recall them in the original order; in the backward span task, they were asked to recall the digits in reverse order.

Participants entered their responses using a keyboard. Raw scores from both tasks were combined to create a composite working memory score. The test showed high reliability (Cronbach's α = .91).

Listening Strategy Use

To isolate the pure effects of vocabulary knowledge on listening proficiency, it is essential to control for participants' listening strategy use. Research indicates that L2 learners who effectively utilize strategies are metacognitively aware of five key components involved in successful listening processing: (a) problem-solving (inferring information that was not fully understood), (b) planning and evaluation (employing strategies to prepare for L2 listening tasks and assessing performance), (c) translation (minimizing reliance on direct translation), (d) person knowledge (perceptions of task difficulty and self-efficacy in L2 listening), and (e) directed attention (maintaining focus and staying on task). To assess participants' metacognitive awareness of these five components, we administered the Metacognitive Awareness Listening Questionnaire (MALQ; Vandergrift & Goh, 2012). Participants responded to 21 statements on a 6-point Likert scale (1 = *strongly disagree*, 6 = *strongly agree*). To ensure clarity, the original MALQ statements were translated into Japanese. The reliability of the MALQ scores was high (Cronbach's α = .90). Following Vandergrift et al., raw scores were averaged for each of the five components, providing a comprehensive profile of each participant's listening strategy use.

Results

Both vocabulary tasks demonstrated satisfactory internal consistency: LJT (α = .91) and MRT (α = .93). As shown in Supplementary Information-S1, the vocabulary scores did not significantly deviate from a normal distribution ($p > .05$). Grubb's test did not find statistically significant outliers in any instances ($z = 3.48$ as a threshold). The results of Pearson correlation analyses revealed significant but moderate associations between LJT and MRT ($r = .51$, 95% *CI* [.37, .62]). A paired-samples t-test indicated that participants scored significantly lower on the LJT compared to the MRT, with a large effect size ($t = 17.686$, $p < .001$, $d = 1.54$). These findings suggest that while the LJT and MRT scores overlap, they tap into different constructs of phonological vocabulary (possibly declarative vs. automatized) with distinct levels of difficulty (LJT > MRT).

Regarding the vocabulary-listening link, another set of Pearson correlation analyses showed that participants' general listening proficiency (TOEIC) scores were more strongly correlated with LJT ($r = .64$ [.53, .73]) than with MRT ($r = .45$ [.31, .58]). This indicates that LJT, which likely measures more automatized knowledge, has stronger predictive power for L2 listening proficiency compared to MRT, which is more aligned with declarative knowledge. For a visual summary, see Figure 3.

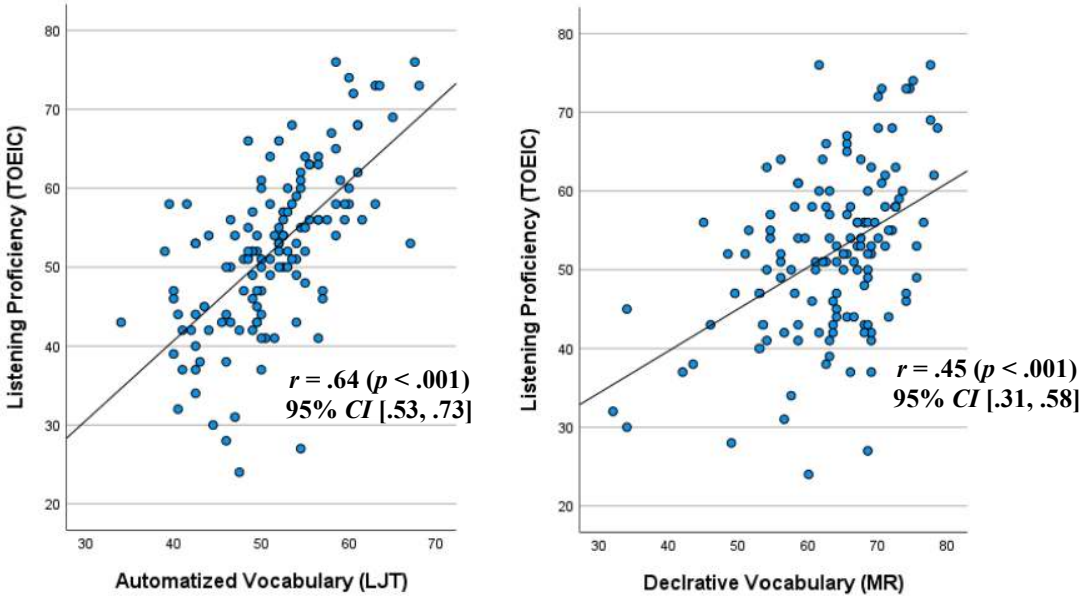

FIGURE 3 Correlations between L2 listening proficiency (y-axis) and phonological vocabulary knowledge (x-axis). The Lexicosemantic Judgement Task (left panel) was used to assess automatized vocabulary knowledge, while the Meaning Recognition Task (right panel) measured declarative vocabulary knowledge. Participants' TOEIC listening scores showed a stronger correlation with LJT scores (r = .64, p < .001) than with MRT scores (r = .45, p < .001). The correlation coefficients (r = .64 vs. .45) were significantly different (Z =3.33, p < .001).

To examine the relationship between aptitude, strategy use, vocabulary knowledge, and listening proficiency, a general multiple regression analysis was conducted with participants' TOEIC listening scores as the dependent variable and nine predictors—LJT, MRT, auditory processing, working memory, and the five components of the MALQ. To address multicollinearity concerns, variance inflation factors (VIF) were checked, and no predictor variables exceeded a VIF of 2 (Range = 1.08 to 1.59). The composite model significantly explained 66.6% of the variance in participants' listening proficiency scores (F = 9.474, p < .001). As summarized in Table 1, the significant and marginally significant predictors included LJT (p < .001), personal knowledge (p = .045), and MRT (p = .052). To determine the relative importance of vocabulary and strategy use within the model (R^2 = .666), a dominance analysis was performed (Mizumoto, 2022). Dominance analysis is a statistical technique that evaluates the relative contribution of each predictor variable by examining its additional explanatory power across all possible combinations of predictors in the model. This approach provides a more nuanced estimate of predictor importance than standardised regression coefficients, particularly when predictors are correlated, and helps clarify which variables play a more central role in explaining the outcome. The analysis revealed that listening proficiency was primarily explained by vocabulary factors, accounting for a total of 70.6% of the variance (with LJT contributing 51.8% and MRT contributing 18.8%). Strategy use, particularly person knowledge, explained 18.2% of the variance. The roles of aptitude factors were minor (1.3% for auditory processing; 0.2% for working memory).[2]

[2] Given that some scholars have argued that the stability of participants' access to lexical items constitutes a key component of automaticity (Segalowitz, 2010), participants' coefficient of variation (CV) was measured for the LJT,

TABLE 1

Summary of Multiple Regression of Listening Proficiency Relative to Lexical, Perceptual, Cognitive, and Metacognitive Predictors

	B	*SE*	95% *CI (B)*		*β*	*t*	*p*	*Relative weight*	
			Upper	*Lower*				*Raw weight*	*Rescaled weight*
Intercept	-0.840	10.610	−36.642	3.222		-0.079	.937		
LJT	0.372	0.073	.513	1.214	.452	5.111	<.001*	.230	51.8%
MRT	0.097	0.055	−.016	.446	.150	1.76	.081†	.083	18.8%
Auditory processing[a]	0.098	0.517	−1.940	2.635	.014	0.189	.850	.005	1.3%
Working memory[b]	-0.012	0.425	−1.022	2.125	-.002	-0.028	.978	.001	0.2%
Problem−solving[c]	0.079	1.294	−3.604	2.223	.005	0.061	.951	.005	1.2%
Planning and evaluation[c]	-0.515	1.095	−3.020	2.777	-.043	-0.47	.639	.007	1.5%
Translation[c]	0.386	0.883	.309	4.561	.040	0.438	.662	.024	5.5%
Person knowledge[c]	2.601	1.134	−.989	2.448	.208	2.294	.024*	.081	18.2%
Directed attention[c]	-0.914	1.395	−2.173	1.750	-.051	-0.655	.514	.005	1.1%

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

Discussion

Given the developmental sequences outlined in the skill acquisition theory (declarative → automatized) and the emerging methodological paradigm for phonological vocabulary knowledge (MRT for declarative knowledge; LJT for automatized knowledge), we initially predicted that (a) LJT and MRT scores represent overlapping but separable constructs of vocabulary knowledge, and (b) the link between vocabulary knowledge and listening proficiency would be more clearly observed in LJT than in MRT. As summarized in Table 2, Study 1 (n = 134) successfully replicated several key findings from the original studies by Saito et al. (2025) and Uchihara (2025), thus confirming our predictions.

First, the relationship between LJT and the traditional vocabulary test format measuring form-meaning mapping (i.e., MRT) was moderately strong (r = .51 [.36, .62]). This suggests that LJT and MRT tests tap into somewhat overlapping but essentially different constructs, potentially reflecting declarative and automatized knowledge, respectively. Second, LJT demonstrated stronger correlations with general listening proficiency (TOEIC; r = .64 [.53, .73]) compared to MRT (r = .45 [.31, .58]). Regarding the relative importance of LJT and MRT scores in predicting listening proficiency, with other variables (aptitude, listening strategy use) controlled for, LJT emerged as the primary predictor (51.8%), followed by MRT (18.8%).

TABLE 2 Comparisons of Vocabulary and Listening Proficiency Between the Current Study and Precursor Studies (Saito et al.; Uchihara et al.)

	Saito et al. (2025) n = 126	Uchihara et al. (2025) n = 114	Current Study n = 134
Key correlations			
LJT – MRT	r = .51 [.37, .62]	r = .58 [.44, .69]	r = .50 [.36, .62]
LJT – TOEIC	r = .66 [.54, .75]	r = .71 [.61, .79]	r = .64 [.53, .73]
MRT – TOEIC	r = .43 [.27, .56]	r = .57 [.43, .68]	r = .45 [.31, .58]
Relative importance			
LJT	49.9%		51.8%
MRT	20.8%		18.8%
Aptitude	1.1%		1.5%
Strategy use	21.3%		28.1%

Study 2: Development of Timed LJT

As a task intended to measure the automatized dimension of phonological vocabulary, the LJT successfully distinguished itself from MRT and demonstrated stronger predictive power for general L2 listening proficiency. For the test format to fully capture participants' prompt and stable access to knowledge—another key component of automatization, Study 2 aimed to develop a timed LJT. To achieve this, we first determined an appropriate time limit for a specific group of L2 learners—college-level Japanese EFL learners—by examining the processing times of 10 L1 speakers (600-700 ms) as an ideal benchmark and 10 advanced Japanese EFL learners (1600-2000 ms) as a realistic benchmark.

The development of the timed GJT has seen surprising variation in how scholars determine optimal time limits for L2 learners to judge sentence grammaticality. Some studies have piloted their materials with native speakers and then added an extra 20% relative to native speaker performance for each stimulus (e.g., Godfroid et al., 2015; Ellis, 2005), while others have applied a fixed 10-second limit for all stimuli. Using a native speaker baseline, Godfroid

and Kim (2021) adopted a different approach, calculating the time limit based on the median length of each audio-recorded sentence and adding 50% of that length as a buffer for each stimulus.

In Study 2, we aimed to establish an optimal time limit for the LJT for this specific group of L2 learners (college-level Japanese EFL learners), building on various approaches from prior GJT studies. It is noteworthy that all of the aforementioned GJT studies relied on native speakers as a baseline. In many EFL classrooms, both learners and educators strive for nativelike proficiency, and most teaching methods focus on achieving this goal. However, an increasing number of scholars have questioned the feasibility of this approach, arguing that attaining nativelike proficiency is an ideal that very few learners will achieve in their lifetime (Abrahamsson & Hyltenstam, 2009; cf. Saito & Hanzawa, 2016 for the ultimate attainment of Japanese L2 learners of English in classroom settings without any experience abroad). These scholars propose that learners should aim for a more realistic goal: proficiency levels akin to those of advanced, highly functional L2 users, even if these levels differ significantly from those of native speakers (Levis, 2018; Ortega, 2013).

With this perspective in mind, we decided to use not only native speakers but also advanced L2 users as benchmarks for the participants in this study—college-level Japanese EFL learners. This approach offers a more attainable goal for the learners in this context, one that aligns with the realistic objectives of L2 proficiency development.

Participants

For the native speaker baseline, a total of 10 L1 users of British English were recruited from the Greater London area (3 males, 7 females). All participants were university students, with a mean age of 23.5 years (Range = 21-25 years). Although they had some knowledge of foreign languages (e.g., French, Spanish), none had reported any immersion or residence abroad experience. They grew up in families where both parents were native speakers of English. For the advanced Japanese EFL learners, we revisited the original dataset of 126 Japanese EFL students in Saito et al. (2023). Ten participants (top 9%) were identified as relatively advanced based on their TOEIC test scores, which exceeded 80% (> 64 out of 90 points). According to the Common European Framework of Reference for Languages aligned with TOEIC listening scores, their proficiency levels ranged from vantage Independent (B2) to Proficient (C1). This score benchmark (i.e., > 80%) is what many university-level Japanese EFL students are recommended to aim for according to TOEIC guidelines. These advanced participants (5 males, 6 females) were university students in Tokyo, Japan, with no prior immersion or residence abroad at the time of the project (M_{age} = 20.3 years; Range = 18-22). A total of 21 participants (10 native speakers, 11 advanced Japanese EFL learners) in Study 2 was comparable to similar methodological studies (e.g., n = 20 natives in Godfroid et al., 2015).

Procedure

As in Study 1, the L1 participants completed the untimed version of the LJT using the online psychology experimental platform, Gorilla (Anwyl-Irvine et al., 2020). A trained research assistant was available remotely to provide technical assistance if needed. The Japanese participants were given the same instructions (see Figure 1). The English participants received the instructions in English. After listening to each target sentence, where a target word could be used either appropriately or inappropriately in context, they were asked to decide whether the sentence they had heard made semantic sense. To ensure they listened to the full sentence, participants were only allowed to click on the screen after the sentence had fully played. We made this methodological decision to prevent participants from guessing the answer without

fully processing the content of the stimulus. To avoid overemphasis on linguistic form and to better reflect real-life L2 comprehension processes, participants were encouraged to make these judgments as quickly as possible. However, the task was *untimed* in the sense that there was no fixed time limit for making each sentence judgment. Reaction time (RT) was recorded to measure how quickly participants selected a response (semantically appropriate vs. inappropriate) immediately after finishing listening to each sentence. Participants were only allowed to click on the screen after the sentence had fully played. The LJT performance of the 11 advanced Japanese EFL students (previously reported in Saito et al. [2023]) was also used to calculate RT. Following Hui and Godfroid's (2020) the procedures for using RT in L2 listening research, RTs were only included for correct responses (i.e., correctly accepting semantically appropriate sentences and rejecting semantically inappropriate sentences). To prevent the influence of excessively slow outlier responses on the overall dataset, responses exceeding 6000 ms were excluded. This threshold of 6000 ms was determined as 1.5 standard deviations beyond the mean RT among the 126 Japanese EFL participants in Saito et al. (2023).

Results

Given the insights that participants' responses to correct acceptance and rejection seem to represent two different behaviors in L2 morphosyntax (Ellis, 2005) and vocabulary literature (Uchihara et al., forthcoming), the cleaned data were analyzed separately not only for types of participants (Group: L1 vs. L2) but also for types of sentences (Stimulus: semantically appropriate vs. inappropriate). Descriptive statistics for both accuracy and RT are summarized in Supplementary Information-S2. The L1 participants demonstrated satisfactory reliability scores for their LJT performance ($\alpha = .90$). According to the Kolmogorov-Smirnov test, the data did not significantly deviate from normal distribution ($p > .05$).

For accuracy, participants' scores were submitted to a two-way ANOVA (Group × Stimulus), finding significant main effects of Group, $F = 114.394$, $p < .001$, $\eta^2 = .864$. This indicates that the accuracy of L1 participants was consistently higher (> 90%) than that of L2 participants (70-80%), with a large effect size, regardless of stimulus type (appropriate vs. inappropriate sentences).

For RT, a two-way ANOVA detected significant main effects not only of Group, $F = 33.952$, $p < .001$, $\eta^2 = .654$, and Stimulus, $F = 25.933$, $p < .001$, $\eta^2 = .590$, but also of interaction effects, $F = 13.949$, $p < .001$, $\eta^2 = .437$. According to post-hoc multiple comparisons, L1 participants' RT (600-700ms) did not significantly differ between their judgments of appropriate and inappropriate sentences ($F = .922$, $p = .350$, $\eta^2 = .049$). However, L2 participants' RT was significantly slower when rejecting inappropriate stimuli (1500-1600ms) compared to accepting appropriate stimuli (1800-2000ms; $p < .05$, $\eta^2 = .684$).

Development of Timed LJT

The findings revealed significant differences in LJT performance between L1 and L2 participants. L1 participants consistently demonstrated high accuracy (> 90%), quick response times (600-700 ms), and stable performance regardless of the type of stimulus. In contrast, L2 participants exhibited more variability in both accuracy (70-80%) and response times, with notably slower processing when dealing with inappropriate sentences compared to appropriate ones.

In prior L2 morphosyntax studies (e.g., Godfroid et al., 2015), time limits for tests were typically set based on L1 baseline data. However, the current project focused on developing a tool specifically for university-level Japanese EFL learners. While achieving native-like proficiency in English may be ideal, it is often extremely challenging and may not be necessary

for adult L2 learners in particular (Abrahamsson & Hyltenstam, 2009; Levis, 2018). Therefore, we prioritized establishing a realistic time limit that reflects the abilities of advanced Japanese EFL learners, who serve as a more attainable benchmark for the broader target population.

To set these time limits, we analyzed the 95% CI data from the advanced L2 participants. The results indicated that their response times ranged from 1513 ms to 1639 ms for accepting appropriate sentences and from 1848 ms to 1966 ms for rejecting inappropriate sentences. Based on these findings, we developed a new timed version of the LJT, setting the time limits at 1600ms for appropriate sentences and 2000 ms for inappropriate sentences.

These separate time limits were adopted because Study 2 revealed clear differences in response behaviour between accepting appropriate sentences and rejecting inappropriate ones. While L1 speakers processed both sentence types at similar speeds (600–700 ms), advanced L2 users showed markedly slower processing when rejecting inappropriate sentences than when accepting appropriate ones (1800–2000 ms vs. 1500–1600 ms). This asymmetry aligns with previous findings in GJT work (Ellis, 2005) and in LJT validation research (Uchihara et al., forthcoming), which indicate that detecting semantic misfit requires deeper lexical-context integration and may draw more heavily on automatized processing. For this reason, a uniform time-limit risked disproportionate penalty against inappropriate-sentence trials. The differentiated limits were therefore empirically motivated rather than arbitrary, designed to reflect realistic processing load for advanced L2 users while still imposing measurable time pressure.

To ensure the practicality of these time limits, five experienced Japanese EFL teachers were invited to take the test. They unanimously agreed that the test, while challenging, was feasible and realistic for university-level Japanese EFL learners with varied proficiency levels. This assumption was based on the teachers' extensive experience working with university-level Japanese EFL students, who typically vary greatly in proficiency (A1 to C1) but generally possess knowledge of approximately 3,000-4,000 word families—covering all the words included in the LJT—thanks to around 8–10 years of EFL education in Japan (McLean et al., 2014). Additionally, we tested an ideal or native-like time limit of 750 ms for both appropriate and inappropriate sentences, based on the L1 RT data (CI range: 653 ms to 765 ms). However, the teachers agreed that this version of the LJT was too fast and unsuitable for the target population, affirming that the originally established time limits (1600 ms and 2000 ms) were more appropriate for the intended learners.

Study 3: Comparisons of Timed vs. Untimed LJT

Given that the untimed LJT has consistently shown medium-to-strong correlations with general L2 listening proficiency (r = .66 [.54, .75] in Saito et al., 2025; r = .71 [.61, .79] in Uchihara et al., 2025; r = .64 [.53, .73] in Study 1), Study 3 aimed to investigate whether and to what extent the use of a timed LJT would affect the vocabulary-listening relationship. Specifically, we examined the performance of 61 Japanese EFL learners on the timed LJT and correlated their scores with TOEIC listening scores to determine whether the strength of this relationship would be comparable to or greater than that observed in previous studies using the untimed LJT.

Participants

Initially, 76 university-level Japanese EFL learners were recruited via an electronic flyer. Using the same data collection procedure and screening process as in Study 1, interested participants first took brief MRT (focusing on 1K words), working memory tests (forward and backward span), and completed a background questionnaire. Of these, 61 participants met the

eligibility criteria and proceeded to take the TOEIC test followed by the timed LJT. The final group of participants (32 males, 29 females) varied in age (M = 21.5 years, Range = 18-22 years), the amount of EFL instruction they had received prior to university (M = 1662.30 hours, Range = 600-3150 hours), and the age at which they began learning EFL (M = 10.21 years, Range = 3-15 years). The characteristics of this participant group are highly comparable to those in prior studies (Saito et al., 2025; Uchihara et al., 2025; Study 1), making it an appropriate population for examining the effects of the timed LJT on the vocabulary-listening relationship.

Vocabulary and Listening Proficiency Tests

The timed LJT developed in Study 2 was used to measure participants' automatized phonological vocabulary knowledge (out of 80 points). Unlike the untimed LJT, where participants were simply asked to choose their answer (semantically appropriate vs. inappropriate) as quickly as possible, participants in the timed LJT were explicitly informed (a) that each trial had a time limit to assess their spontaneous processing of L2 sentences, and (b) that they would automatically be moved to the next trial if they did not complete the judgment within the given time limit. To familiarize themselves with the task procedure, participants practiced with four example questions before proceeding to the main session, which consisted of 160 trials. The same TOEIC test used in Study 1 was used to measure their general listening proficiency (90 points).

Results

As indicated in Supplementary Information-S3, the Kolmogorov-Smirnov test revealed that participants' LJT and TOEIC listening scores did not significantly deviate from a normal distribution ($p > .05$). Grubb's test did not identify statistically significant outliers in the dataset (z =3.20 as a threshold). As shown in Figure 4, the Pearson correlation analysis demonstrated a moderate-to-strong correlation between participants' scores on the timed LJT and TOEIC (r = .73 [.59, .83]). Given the significant overlap in the 95% confidence intervals, these correlation coefficients appear comparable to those reported in previous studies (r = .66 [.54, .75] in Saito et al., 2025; r = .71 [.61, .79] in Uchihara et al., 2025; r = .64 [.53, .73] in Study 1). Overall, these findings suggest that the presence of time pressure did not noticeably enhance the strength of the vocabulary-listening relationship.

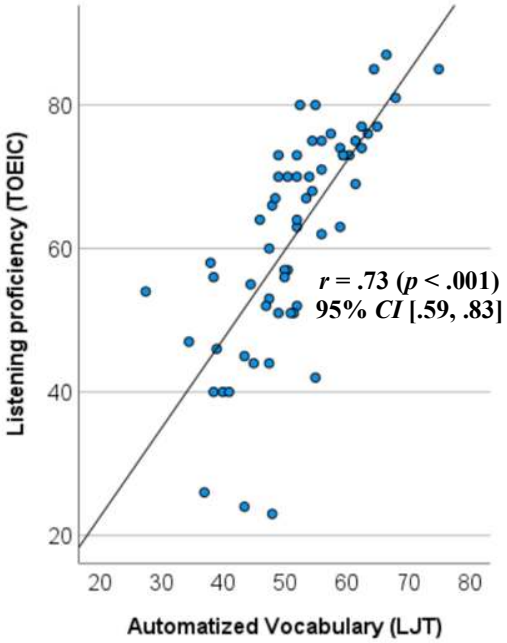

FIGURE 4 Correlation Between L2 Listening Proficiency (y-axis) and Phonological Vocabulary Knowledge (timed LJT scores; x-axis). Participants' TOEIC listening scores were strongly correlated with LJT scores (r = .73, p < .001). The difference in the correlation coefficients between Studies 1 and 3 (r = .64 vs. .73) failed to reach statistical significance (z = 1.081, p = .140).

As shown in Supplementary Information-S3, participants experienced a certain number of timeouts (M = 12.3, SD = 11.4), meaning that 7.6% of their responses were recorded as zero because they were unable to make a judgment within the fixed time limit. In prior GJT studies, timeout scores were typically classified as "incorrect responses." However, such instances could also reflect that some participants failed to respond due to generally slow lexical decoding abilities—a crucial individual difference that can affect L2 acceptability judgments (Maie & Godfroid, 2021). Taking this possibility into account, we conducted a follow-up analysis using transformed LJT scores, where the number of timeouts was statistically controlled:

$$\textbf{Transformed LJT scores} = \frac{\textbf{No. of correct responses}}{\textbf{Total instances (160 points)} - \textbf{No. of timeouts}}$$

The correlation between the transformed LJT scores and TOEIC listening scores increased (r = .80 [.68, .87]; see Figure 5). Additionally, the results of the timed LJT significantly differed from those of the untimed LJT (Fisher's z-transformation; z = 2.06, p = .039).

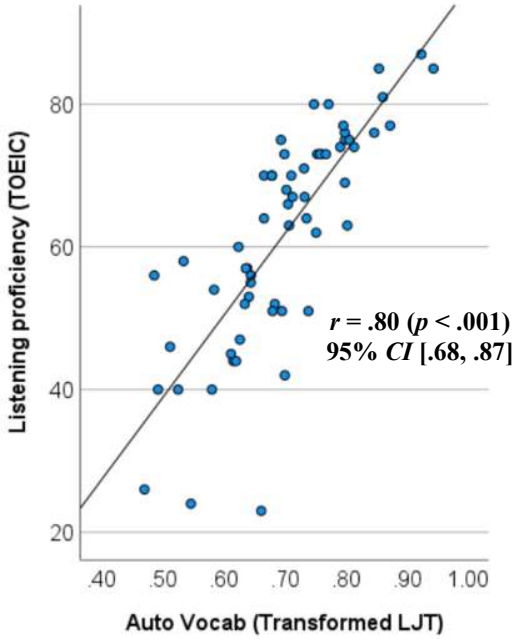

FIGURE 5 Correlation Between L2 Listening Proficiency (y-axis) and Transformed Phonological Vocabulary Knowledge (timed LJT scores with the number of timeouts statistically controlled for; x-axis). The listening-vocabulary link was strong (r = .80, p < .001) relative to Study 1 (r = .64). The difference in correlation coefficient is statistically significant (z = 2.06, p = .039).

General Discussion

Overall, the current investigations yielded three key findings. First, replicating the results of previous studies (Saito et al., 2025; Uchihara et al., 2025), the untimed Lexicosemantic Judgement Task was found to produce medium-to-strong correlations with general listening proficiency scores (r = .6-.7), while Meaning Recognition Task exhibited small-to-medium correlations (r = .4-.5). Second, the response behaviors of L1 participants and advanced Japanese EFL participants (CEFR levels B2 and C1) differed significantly. The former group responded consistently faster (600-700 ms). The latter group not only responded more slowly overall but also displayed sensitivity to stimulus type (1600 ms for accepting semantically appropriate sentences, 2000 ms for rejecting semantically incorrect sentences). Third, the strength of the vocabulary-listening link did not significantly change whether the LJT was untimed or untimed (based on the response times of advanced Japanese EFL learners).

These findings have two overarching implications. On the one hand, the LJT may be an effective measure of the automatized dimension of phonological vocabulary, which is closely linked to general listening proficiency, in contrast to the declarative dimension of phonological vocabulary, which is typically assessed through MRT and recall tasks. On the other hand, the role of time pressure in this task format remains unclear and subject to more rigorous research investigations.

Although exploratory, several possible reasons may explain why time pressure did not enhance the predictive power of the LJT for general L2 listening proficiency. In the context of L2 morphosyntax, while many studies have supported the role of time pressure in GJTs, it is also important to note that some have reported potentially null effects. Crucially, studies supporting the role of time pressure have predominantly used written GJTs. In Plonsky et al.'s (2020)

methodological synthesis, they found that modality (aural vs. written) and time pressure (timed vs. untimed) effects may have been confounded in the existing literature. Among their sample, they identified only 22 GJTs that were both timed and aural, with only a very small subset contributing to modality or timing comparisons. To our knowledge, Bialystok (1979) was such exception, reporting no differences in performance between auditory GJTs with varying levels of time pressure, suggesting that the auditory modality itself may inherently encourage participants to draw on implicit or automatized knowledge rather than declarative knowledge. As Plonsky et al. (2020) pointed out, "by definition, aural language happens at a fixed pace for all participants, so is, in a sense, 'timed,' and so, generally with these comparisons between aural and written, 'modality' is inevitably conflated with timing, unless one times both modalities to keep timing constant across them" (p. 598).

More recently, Maie and Godfroid (2021) found that time pressure could negatively affect both automatic and controlled processing, implying that the influence of time pressure may be more closely related to processing speed than to the degree of automatization. Our follow-up analysis supported Maie and Godfroid's (2021) argument in particular, hinting at the possibility that (a) certain participants with slower lexical decoding abilities may have failed to provide correct responses regardless of the degree of their automatized vocabulary knowledge, and (b) the timed LJT showed stronger predictive power for listening proficiency (r = .80) compared to the untimed LJT (r = .63) when the number of timeouts was statistically controlled. However, given the exploratory nature of this follow-up analysis, we are hesitant to draw any firm conclusions about the relative effectiveness of timed versus untimed LJT.

It is worth noting that response speed was not recorded in Study 3, which prevents direct comparison of RT distributions between the timed and untimed LJT conditions. However, the presence of timeouts in Study 3 (about 7.6% of trials) suggests that the time limits constrained response behavior relative to untimed performance. Because the LJT requires semantic evaluation, some participants may have prioritised accuracy over speed, which may have reduced the sensitivity of timing to automatized knowledge. Future work should incorporate RT recording within the timed format and vary instructional emphasis (speed-prioritized vs. accuracy-prioritized) to better isolate how timing interacts with automatized lexical processing.

Alternatively, the findings may suggest that the set time pressure needs to be re-evaluated in conjunction with other groups of L2 learners with varying proficiency levels to determine an optimal reaction time. In the current study, we focused on advanced college-level Japanese EFL students (CEFR levels B2 and C1), given their realistic learning goals. However, the lack of statistical differences between the timed and untimed LJT suggests that the time limits used in this study (1600-2000 ms) may have been too lenient. Additionally, participants' sense of time pressure may have been similar across both conditions, regardless of the presence of a formal time limit. Even in the untimed LJT, participants were encouraged to respond as quickly as possible.

It is important to note that during the development of the timed LJT, experienced Japanese EFL teachers piloted the task and concurred that the L1 baseline (600-700 ms) was too fast, particularly for the group of participants in this study (beginner-to-intermediate Japanese university-level students). This suggests that, for this specific population, the LJT itself may be sufficiently demanding, regardless of the presence of time pressure. In future research, it would be valuable to replicate this study with Japanese learners of English who have higher L2 proficiency, such as those with substantial immersion and residential experience in English-speaking environments. These learners may better represent the ultimate goals of EFL

participants and provide a more appropriate level of time pressure, which could better capture the automatization of phonological vocabulary.

Finally, the findings of the study may have been influenced by a range of confounding variables that were not intentionally controlled for. For example, the null effects of time-related variables on participants' LJT performance in this study may be attributable to the possibility that some participants prioritised accuracy over speed (i.e., speed-accuracy trade-offs). Future research could consider participants' individual preferences for speed versus accuracy, which might be particularly evident when task cognitive load is relatively high and task familiarity relatively low. However, the existence of such trade-offs in relation to task complexity remains a matter of debate (e.g., Robinson, 2001 vs. Skehan, 1998).

Sharing Test Materials & Pedagogical Implications

As a measure of L2 learners' phonological vocabulary knowledge, the LJT materials are shared through various platforms. For researchers planning to use the same materials as those in the current study, the materials are available in both Japanese and English as open materials on Gorilla (https://app.gorilla.sc/openmaterials/654106). The tests are untimed, allowing participants to complete them at their own pace. Researchers can integrate the LJT into online experiments via Gorilla or conduct face-to-face experiments, provided participants have access to a computer and the internet (see Figure 6).

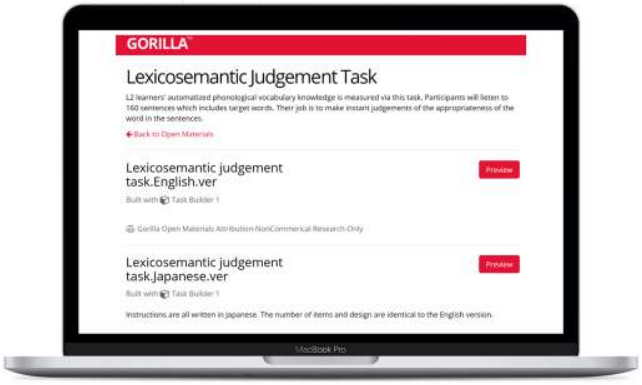

FIGURE 6 Screenshot of the Lexicosemantic Judgement Task on the Gorilla platform (https://app.gorilla.sc/openmaterials/654106).

For practitioners (teachers and students seeking to measure vocabulary knowledge relevant to real-life listening comprehension), a pencil-and-paper version of the LJT has been developed, allowing it to be administered to multiple participants simultaneously. In this format, the test is timed, with a limit of 1600 ms for appropriate stimuli and 2000 ms for inappropriate stimuli. Participants are informed at the beginning of the test that they will have limited time for each sentence and are instructed to leave an answer blank if they miss a sentence (see Figure 7). The test materials are deposited in L2 Speech Tools (Mora-Plaza et al., 2021: https://sla-speech-tools.com/) and found in Supplementary Information-S4.

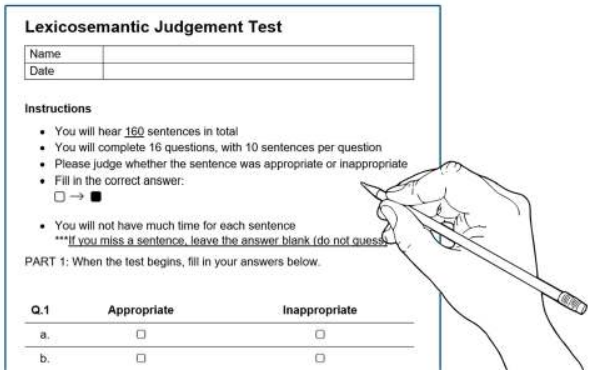

FIGURE 7 Screenshot of a pencil-and-paper version of the Lexicosemantic Judgement Task available in L2 Speech Tools and Supplementary Information-S4

There are several suggested methods for using the LJT in pedagogical contexts. One valuable application is using the LJT to raise learners' awareness. It would be insightful for participants to complete both the LJT and the MRT and compare their performance and perceptions. According to the skill acquisition theory for instructed L2 learning (DeKeyser & Suzuki, 2025; Suzuki & DeKeyser, in press), while the initial stage of learning involves linking a word's sound to its meaning (i.e., form-meaning mapping), this knowledge must be proceduralized and ultimately automatized through repeated practice. See examples of how declarative and automatized vocabulary knowledge differentially develop during brief vocabulary training (Saito et al., 2026) and after extensive amounts of classroom L2 learning (Saito & Uchihara, 2025).

To this end, learners should be encouraged to notice the gap between words they can merely recognize or recall in isolation and words they can analyze for contextual appropriateness when embedded in sentences. This noticing could be a critical first step toward the automatization of form-meaning mapping for partially acquired words. Enhanced awareness of what constitutes automatized (vs. declarative) phonological vocabulary knowledge may prompt learners to attend to vocabulary at the sentence level, paying more attention to the collocational, grammatical, and semantic relationships between target words and surrounding words. This process can significantly improve bottom-up processing in L2 listening comprehension (Hui & Godfroid, 2018), which in turn promotes the development of L2 speaking abilities (Takizawa et al., 2025).

The LJT can also be used as a training tool to develop automatized phonological vocabulary. While practitioners and researchers have predominantly focused on the form-meaning mapping aspects of vocabulary knowledge through flashcards and recall activities, there has been little discussion on how to help automatize this knowledge so it can be directly applied in real-life L2 listening comprehension. If learners practice target words through LJT activities, it could greatly enhance their ability to transfer control over isolated words to broader contextual use (for the differential effects of lexicosemantic judgment vs. meaning recognition *training*, see Saito et al., 2024).

Finally, since LJT scores are a strong predictor of successful L2 comprehension, they can serve as an approximate and indirect measure of listening proficiency. This approach follows the same conceptual and methodological rationale as the use of nonword judgment tasks to approximate L2 proficiency (i.e., LexTALE: Lemhöfer & Broersma, 2012). In the current study,

we compared the LJT and TOEIC listening scores of 195 university-level Japanese EFL students, enabling us to provide CEFR-based proficiency levels (see Supplementary Information S5). Thus, LJT scores can offer learners valuable insights into how much vocabulary knowledge they have acquired and what they still need to master in order to reach their target listening proficiency levels (Saito et al., under review).

References

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language learning*, *59*(2), 249-306. https://doi.org/10.1111/j.1467-9922.2009.00507.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*(1), 388-407. https://doi.org/10.3758/s13428-019-01237-x

Bialystok, E. (1979). Explicit and implicit judgments of L2 grammaticality. *Language Learning*, *29*(1), 81–103. https://doi.org/10.1111/j.1467-1770.1979.tb01053.x

Cheng, J., Matthews, J., Lange, K., & McLean, S. (2023). Aural single-word and aural phrasal verb knowledge and their relationships to L2 listening comprehension. *TESOL Quarterly*, *57*(1), 213-241. https://doi.org/10.1002/tesq.3137

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and speech*, *40*(2), 141-201. https://doi.org/10.1177/002383099704000203

DeKeyser, R. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 15-32). Routledge. https://doi.org/10.4324/9781315676968-2

DeKeyser, R. M., & Suzuki, Y. (2025). Skill acquisition theory. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition: An introduction* (4th ed., pp. 157-182). Routledge.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*(2), 141-172. http://dx.doi.org/10.1017/S0272263105050096

Godfroid, A., & Kim, K. M. (2021). The contributions of implicit-statistical learning aptitude to implicit second-language knowledge. *Studies in Second Language Acquisition*, *43*(3), 606-634. https://doi.org/10.1017/S0272263121000085

Godfroid, A., Loewen, S., Jung, S., Park, J. H., Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge: Evidence from eye-movement patterns. *Studies in Second Language Acquisition*, *37*(2), 269-297. http://doi.org/10.1017/S0272263114000850

Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition*, *35*(3), 423-449. https://doi.org/10.1017/S0272263113000041

Hamada, Y., & Yanagawa, K. (2024). Aural vocabulary, orthographic vocabulary, and listening comprehension. *International Review of Applied Linguistics in Language Teaching*, *62*(2), 953-975. https://doi.org/10.1515/iral-2022-0100

Kamiya, N. (2022). The limited effects of visual and audio modalities on second language listening comprehension. *Language Teaching Research, 29*(4), 1688-1714. https://doi.org/10.1177/13621688221096213

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*(2), 325-343. https://doi.org/10.3758/s13428-011-0146-0

Levis, J. M. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press. https://doi.org/10.1017/9781108241564

Maie, R., & Godfroid, A. (2022). Controlled and automatic processing in the acceptability judgment task: an eye-tracking study. *Language Learning*, *72*(1), 158-197. https://doi.org/10.1111/lang.12474

McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' English vocabulary sizes using the vocabulary size test. *Vocabulary Learning and Instruction*, *3*(2), 47-55. https://doi.org/10.7820/vli.v03.2.mclean.et.al

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*(6), 741–760. https://doi.org/10.1177/1362168814567889

McManus, K. (2024). Replication studies in second language acquisition research: Definitions, issues, resources, and future directions: Introduction to the special issue. *Studies in Second Language Acquisition*; 46, 1299-1319. https://doi.org/10.1017/S0272263124000652

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, *34*(4), 520-531. https://doi.org/10.1016/j.system.2006.09.004

Nation, I. S. P. (2012). *The BNC/COCA word family lists.* Victoria University of Wellington. https://www.wgtn.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P. (2013). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Olsthoorn, N. M., Andringa, S., & Hulstijn, J. H. (2014). Visual and auditory digit-span performance in native and non-native speakers. *International Journal of Bilingualism*, *18*(6), 663-673. https://doi.org/10.1177/1367006912466314

Ortega, L. (2013). SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn. *Language learning*, *63*, 1-24. https://doi.org/10.1111/j.1467-9922.2012.00735.x

Plonsky, L., Marsden, E., Crowther, D., Gass, S. M., & Spinner, P. (2020). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research*, *36*(4), 583-621. https://doi.org/10.1177/0267658319828413

Saito, K. (2014). Experienced teachers' perspectives on priorities for improved intelligible pronunciation: The case of Japanese learners of English. *International Journal of Applied Linguistics*, *24*(2), 250-277. https://doi.org/10.1111/ijal.12026

Saito, K. (under review). Validation of the lexicosemantic judgment task as a rapid measure of L2 listening proficiency in EFL classrooms.

Saito, K., & Uchihara, T. (2025). Experiential, perceptual, and cognitive individual differences in the development of declarative and automatized phonological vocabulary knowledge. *Bilingualism: Language and Cognition*, *28*, 427-443. https://doi.org/10.1017/S1366728924000609

Saito, K., Uchihara, T., Takizawa, K., & Suzukida, Y. (2024). Declarative and automatized phonological vocabulary knowledge in L2 listening proficiency: A training study. *Applied Psycholinguistics*, *45*, 1187-1218. https://doi.org/10.1017/S0142716424000468

Saito, K., Uchihara, T., Takizawa, K., & Suzukida, Y. (2025). Individual differences in L2 listening proficiency revisited: Roles of form, meaning, and use aspects of phonological vocabulary knowledge. *Studies in Second Language Acquisition*, *47*(1), 26-52. https://doi.org/10.1017/S027226312300044X

Saito, K., Fan, X., Pellicer-Sanchez, A., & Uchihara et al. (2026). Beyond form-meaning: Investigating the potential and limits of captioned video in building declarative and automatized vocabulary knowledge. *Language Teaching Research*.

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*(2), 261-274. https://doi.org/10.1017/S0261444819000053

Suzuki, Y., & DeKeyser, R. M. (in press). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), The Routledge handbook of instructed second language acquisition (2nd ed.). Routledge.

Suzuki, Y., & Elgort, I. (2023). Measuring automaticity in a second language: A methodological synthesis of experimental tasks over three decades (1990-2021). In Y. Suzuki (Ed.), *Practice and automatization in second language research: Perspectives from skill acquisition theory and cognitive psychology* (pp. 206-234). New York, NY: Routledge.

Taguchi, N. (2011). Teaching pragmatics: Trends and issues. *Annual review of applied linguistics*, *31*, 289-310. http://doi.org/10.1017/S0267190511000018

Takizawa, K., Saito, K., Suzukida, S., Kurokawa, S., & Uchihara, T. (in press). Automatized knowledge to automaticity in speech: Examining the contribution of automatized phonological vocabulary knowledge to L2 utterance fluency. *Applied Linguistics*. https://doi.org/10.1093/applin/amaf042

Uchihara, T., Saito, K., Kurokawa, S., Takizawa, K., & Suzukida, Y. (2025). Declarative and automatized phonological vocabulary knowledge: Recognition, recall, lexicosemantic judgement, and listening-focused employability of L2 words. *Language Learning*, 75(2), 458-492. https://doi.org/10.1111/lang.12668

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*(2), 383-410. http://doi.org/10.1017/S0272263119000676

Vandergrift, L., & Goh, C. (2012). *Teaching and learning second language listening: Metacognition in action.* Routledge. https://doi.org/10.4324/9780429287749

Wallace, M. P. (2022). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*, *72*(1), 5-44. https://doi.org/10.1111/lang.12424

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

Supplementary information-S1: Descriptive Statistics of Listening, Vocabulary, Perceptual, Cognitive and Metacognitive Profiles

	M	*SD*	95% *CI*		Kolmogorov-Smirnov	
			Lower	Upper	*D*	*p*
A. Listening proficiency						
TOEIC Total (90 points)	52.18	8.35	50.35	54.01	.063	.621
B. Vocabulary						
Lexicosemantic judgement (80 points)	49.6	5.0	48.8	50.4	.051	.852
Meaning recognition (80 points)	62.3	7.7	61.3	63.2	.104	.101
C. Perceptual cognitive profiles						
Working memory (forward; letters)	6.6	1.3	6.3	6.8	.123	.157
Working memory (backward; letters)	7.6	0.6	7.5	7.7	.113	.178
Auditory processing (formants)	45.0	16.2	42.8	47.1	.147	.085
Auditory processing (amplitude rise time)	25.5	18.0	23.5	27.5	.149	.081
D. Metacognitive profiles						
Problem-solving (6 points)	3.5	1.5	3.4	3.6	.125	.152
Planning & evaluation (6 points)	2.9	0.7	2.8	3.0	.132	.137
Mental translation (6 points)	2.3	0.8	2.2	2.4	.138	.112
Person knowledge (6 points)	3.2	1.0	3.1	3.3	.117	.168
Directed attention (6 points)	3.0	0.9	3.0	3.1	.134	.128

Supplementary information-S2: Descriptive Statics of Accuracy and Reaction Time Data Among L1 vs. L2 Participants

		M	*SD*	95% *CI*		Kolmogorov-Smirnov	
				Lower	Upper	*Z*	*p*
A. Accuracy							
L1 participants	Appropriate	77.7	1.4	76.6	78.7	.220	.640
(n = 10)	Inappropriate	76.8	1.3	75.8	77.7	.160	.924
L2 participants	Appropriate	64.8	6.5	60.1	69.4	.178	.852
(n = 10)	Inappropriate	61.0	8.2	55.0	66.9	.276	.359
B. Reaction time							
L1 participants	Appropriate	684 ms	637	653	716	.184	.829
(n = 10)	Inappropriate	735 ms	593	706	765	.240	.534
L2 participants	Appropriate	1576 ms	1279	1513	1639	.224	.620
(n = 10)	Inappropriate	1907 ms	1201	1848	1966	.199	.751

Supplementary information-S3: Descriptive Statistics of Listening and Vocabulary Profiles

	M	*SD*	95% *CI*		Kolmogorov-Smirnov	
			Lower	Upper	*D*	*p*
A. Listening proficiency						
TOEIC Total (90 points)	61.8	15.1	57.9	65.8	.116	.351
B. Vocabulary						
Lexicosemantic judgement (80 points)	51.6	8.9	49.3	53.9	.072	.880
No. of timeouts	12.3	11.4	9.4	15.3	.157	.085

Supplementary information-S4: Paper-and-Pencil Version of the Timed Lexicosemantic Task

Lexicosemantic Judgement Test

Name	
Date	

Test manual [here]
Audio materials [here]

Instructions

- You will hear 160 sentences in total
- You will complete 16 questions, with 10 sentences per question
- Please judge whether the sentence was appropriate or inappropriate
- Fill in the correct answer:
  ▢ → ■
- You will not have much time for each sentence
  ***If you miss a sentence, leave the answer blank (do not guess)

First, you will hear 4 practice questions.

PRACTICE

Q.	Appropriate	Inappropriate
a.	■	▢
b.	■	▢
c.	▢	■
d.	▢	■

PART 1: When the test begins, fill in your answers below.

Q.1	Appropriate	Inappropriate
a.	☐	☐
b.	☐	☐
c.	☐	☐
d.	☐	☐
e.	☐	☐
f.	☐	☐
g.	☐	☐
h.	☐	☐
i.	☐	☐
j.	☐	☐

Q.2	Appropriate	Inappropriate
a.	☐	☐
b.	☐	☐
c.	☐	☐
d.	☐	☐
e.	☐	☐
f.	☐	☐
g.	☐	☐
h.	☐	☐

i.	☐	☐
j.	☐	☐

Q.3	Appropriate	Inappropriate
a.	☐	☐
b.	☐	☐
c.	☐	☐
d.	☐	☐
e.	☐	☐
f.	☐	☐
g.	☐	☐
h.	☐	☐
i.	☐	☐
j.	☐	☐

Q.4	Appropriate	Inappropriate
a.	☐	☐
b.	☐	☐
c.	☐	☐
d.	☐	☐
e.	☐	☐
f.	☐	☐
g.	☐	☐

h.	☐	☐
i.	☐	☐
j.	☐	☐

Q.5	Appropriate	Inappropriate
a.	☐	☐
b.	☐	☐
c.	☐	☐
d.	☐	☐
e.	☐	☐
f.	☐	☐
g.	☐	☐
h.	☐	☐
i.	☐	☐
j.	☐	☐

Q.6	Appropriate	Inappropriate
a.	☐	☐
b.	☐	☐
c.	☐	☐
d.	☐	☐
e.	☐	☐

f.	☐	☐
g.	☐	☐
h.	☐	☐
i.	☐	☐
j.	☐	☐

Q.7	Appropriate	Inappropriate
a.	☐	☐
b.	☐	☐
c.	☐	☐
d.	☐	☐
e.	☐	☐
f.	☐	☐
g.	☐	☐
h.	☐	☐
i.	☐	☐
j.	☐	☐

Q.8	Appropriate	Inappropriate
a.	☐	☐
b.	☐	☐
c.	☐	☐

d.	☐	☐
e.	☐	☐
f.	☐	☐
g.	☐	☐
h.	☐	☐
i.	☐	☐
j.	☐	☐

-BREAK-

PART 2:

Q.9	**Appropriate**	**Inappropriate**
a.	☐	☐
b.	☐	☐
c.	☐	☐
d.	☐	☐
e.	☐	☐
f.	☐	☐
g.	☐	☐
h.	☐	☐
i.	☐	☐
j.	☐	☐

Q.10	Appropriate	Inappropriate
a.	☐	☐
b.	☐	☐
c.	☐	☐
d.	☐	☐
e.	☐	☐
f.	☐	☐
g.	☐	☐
h.	☐	☐
i.	☐	☐
j.	☐	☐

Q.11	Appropriate	Inappropriate
a.	☐	☐
b.	☐	☐
c.	☐	☐
d.	☐	☐
e.	☐	☐
f.	☐	☐
g.	☐	☐
h.	☐	☐

i.	☐	☐
j.	☐	☐

Q.12	Appropriate	Inappropriate
a.	☐	☐
b.	☐	☐
c.	☐	☐
d.	☐	☐
e.	☐	☐
f.	☐	☐
g.	☐	☐
h.	☐	☐
i.	☐	☐
j.	☐	☐

Q.13	Appropriate	Inappropriate
a.	☐	☐
b.	☐	☐
c.	☐	☐
d.	☐	☐
e.	☐	☐
f.	☐	☐

g.	☐	☐
h.	☐	☐
i.	☐	☐
j.	☐	☐

Q.14	Appropriate	Inappropriate
a.	☐	☐
b.	☐	☐
c.	☐	☐
d.	☐	☐
e.	☐	☐
f.	☐	☐
g.	☐	☐
h.	☐	☐
i.	☐	☐
j.	☐	☐

Q.15	Appropriate	Inappropriate
a.	☐	☐
b.	☐	☐
c.	☐	☐
d.	☐	☐
e.	☐	☐

f.	☐	☐
g.	☐	☐
h.	☐	☐
i.	☐	☐
j.	☐	☐

Q.16	**Appropriate**	**Inappropriate**
a.	☐	☐
b.	☐	☐
c.	☐	☐
d.	☐	☐
e.	☐	☐
f.	☐	☐
g.	☐	☐
h.	☐	☐
i.	☐	☐
j.	☐	☐

-END OF TEST-

Lexicosemantic Judgement Test Answer Key

■ = Correct, ☐ = Incorrect

Question	Sub-Question	Appropriate	Inappropriate	Sentence
1	a	☐	■	Dogs usually expand water.
	b	■	☐	You deserve a holiday.
	c	☐	■	Japan is considerably larger than China.
	d	☐	■	People like to get on an assembly train.
	e	■	☐	I want to work for a prestigious company.
	f	■	☐	My aunt gave me an oak cabinet.
	g	■	☐	We have substantial experience here.
	h	☐	■	This greenhouse produces cats.
	i	■	☐	The deadline is the end of this month.
	j	☐	■	I finished the highway very early.

Question	Subquestion	Appropriate	Inappropriate	Sentence
2	a	☐	■	Our division is so bright tonight.
	b	☐	■	You have to deserve social manner.
	c	☐	■	The car was parked in a substantial position.
	d	☐	■	We recommend cabinets for health problems.
	e	☐	■	He has a frequent tongue.
	f	■	☐	There are diners in a restaurant at night.
	g	■	☐	His words of praise embarrassed me.
	h	■	☐	Smoke is coming out of the stove.
	i	■	☐	We record the names of the applicants.

	j	☐	■	She registered towards the office.

Question	Subquestion	Appropriate	Inappropriate	Sentence
3	a	■	☐	He is one of the best accountants.
	b	■	☐	He has published many books.
	c	☐	■	In contemporary Japan, there are many samurais.
	d	☐	■	I went to a resume last night.
	e	■	☐	I attended the conference last year.
	f	■	☐	Our headquarters are near Tokyo Station.
	g	☐	■	Tom swims in a fare.
	h	☐	■	The figures conduct up to 594.
	i	☐	■	I ride expenses to go to school.
	j	■	☐	The company provides good customer support.

Question	Subquestion	Appropriate	Inappropriate	Sentence
4	a	■	☐	He opened the packet slowly.
	b	■	☐	Please use extension 123 to call me.
	c	☐	■	I write a lot of appliances every day.
	d	■	☐	We made an excursion to the mountains yesterday.
	e	■	☐	This is just a temporary change.
	f	■	☐	I bought many antiques in the shop.
	g	■	☐	You need to pay bus fares.
	h	■	☐	We heard the launch of an attack.
	i	☐	■	He transfers an exam in a minute.

	j	■	☐	Brian is writing his autobiography.

Question	Subquestion	Appropriate	Inappropriate	Sentence
5	a	☐	■	Mothers arise from babies.
	b	☐	■	She looked directly at the prestigious sun.
	c	☐	■	He gave me a small bottle of screenplay.
	d	☐	■	It is such a current earth.
	e	☐	■	We drank a lot of competitors last night.
	f	■	☐	This hotel offers a complimentary breakfast each morning.
	g	■	☐	Our students must read this screenplay.
	h	☐	■	I booked a door with keys.
	i	☐	■	Plants confirm water.
	j	☐	■	I always drink executives.

Question	Subquestion	Appropriate	Inappropriate	Sentence
6	a	☐	■	I really distributed my computer.
	b	☐	■	The colleague is an excellent drink.
	c	■	☐	The doctor talked to a patient.
	d	■	☐	He arrived at our destination so late.
	e	☐	■	Eating requires a lot of computers.
	f	■	☐	The executive gets paid well.
	g	■	☐	She likes contemporary art.
	h	☐	■	He will be regional to see you tomorrow.

	i	☐	■	He boiled the conference yesterday.
	j	☐	■	The phone rings at a solution.

Question	Subquestion	Appropriate	Inappropriate	Sentence
7	a	☐	■	She brushed her long, red facility.
	b	☐	■	I cooked extension yesterday.
	c	■	☐	That is a risky investment.
	d	☐	■	Tom has appreciated trees three times.
	e	■	☐	This frame is strong enough for the assembly.
	f	☐	■	My mother sometimes acquires bad weather.
	g	■	☐	We should improve economic cooperation.
	h	☐	■	I forgot to bring a reception.
	i	☐	■	My friend’s estate was very kind.
	j	☐	■	The launch was on sale.

Question	Subquestion	Appropriate	Inappropriate	Sentence
8	a	■	☐	The Internet expands our contact with people.
	b	☐	■	You must not equip any attitude.
	c	☐	■	The amount of money is medicinal.
	d	☐	■	Please give your autobiography at the bank.
	e	■	☐	Throw it in the bin.
	f	☐	■	The applicant is used for commercial purposes.
	g	■	☐	I chose the furniture for his house.

	h	☐	■	Please open patients as soon as possible.
	i	☐	■	A healthy person eats antiques every morning.
	j	☐	■	My investment is sometimes angry.

Question	Subquestion	Appropriate	Inappropriate	Sentence
9	a	☐	■	My praise caught her dress.
	b	☐	■	I put headquarters on a table.
	c	■	☐	I received a warm reception last night.
	d	■	☐	I paid all expenses.
	e	☐	■	Many photocopiers grow on the ground.
	f	☐	■	The sun is located on the earth.
	g	■	☐	He needs to register his car.
	h	■	☐	This room is equipped with air conditioning.
	i	■	☐	My mother is a culinary teacher.
	j	■	☐	I was transferred to a private company.

Question	Subquestion	Appropriate	Inappropriate	Sentence
10	a	☐	■	Mary published her left hand.
	b	■	☐	I really appreciate his support.
	c	■	☐	My house is located in Tokyo.
	d	☐	■	The teenagers bought some supervisors at the store.
	e	☐	■	The children read the avenue.
	f	☐	■	People usually learn math at culinary schools.

	g	■	☐	I work as a supervisor in a shop.
	h	■	☐	He was hired by a famous company.
	i	■	☐	Please send your resume with your photo.
	j	■	☐	My grandfather bought an estate.

Question	Subquestion	Appropriate	Inappropriate	Sentence
11	a	■	☐	Please feel free to inquire about anything.
	b	■	☐	Driving a bus requires a special license.
	c	☐	■	I ate a promotion last night.
	d	☐	■	The woman gave us a challenging cup.
	e	■	☐	I conduct various tours.
	f	☐	■	She mixed water with a stove.
	g	☐	■	I read some refreshments every day.
	h	■	☐	Changing a corporate culture is difficult.
	i	■	☐	We need to attend a committee meeting.
	j	■	☐	I'd like to get off at fifth avenue.

Question	Subquestion	Appropriate	Inappropriate	Sentence
12	a	■	☐	The sun is considerably bigger than the earth.
	b	■	☐	I work in the system development division.
	c	■	☐	I booked a tour.
	d	☐	■	We should water accountants every day.
	e	■	☐	A big problem arose in the company.

	f	■	☐	They produced a brochure on healthy eating.
	g	☐	■	The cat hired on a street.
	h	☐	■	I work hard with my own cooperation.
	i	☐	■	I inquired a bag on the floor.
	j	☐	■	The tentative environment looks beautiful.

Question	Subquestion	Appropriate	Inappropriate	Sentence
13	a	☐	■	The brochure prepared meals for everyone.
	b	☐	■	Please open the file on your diner.
	c	■	☐	Please confirm the details.
	d	■	☐	He works at a regional organization.
	e	☐	■	The man sells financial rice.
	f	■	☐	The facility was finally completed yesterday.
	g	■	☐	We grow many rare plants in the greenhouse.
	h	☐	■	The packet will begin to talk soon.
	i	■	☐	Milk is distributed to the local shops.
	j	☐	■	Their destination is getting more nervous than ours.

Question	Subquestion	Appropriate	Inappropriate	Sentence
14	a	■	☐	Ted is a frequent guest.
	b	☐	■	My mother often cooks an excursion.
	c	☐	■	John catered English for his mother.
	d	■	☐	All these dishes are catered for the company.

	e	■	☐	Only one competitor can be selected this time.
	f	■	☐	You can find microwaves in the appliance section
	g	■	☐	You can drive a car on the highway.
	h	■	☐	They received paper for the photocopier.
	i	☐	■	I found a correction on the lake.
	j	■	☐	We have to find a better solution.

Question	Subquestion	Appropriate	Inappropriate	Sentence
15	a	☐	■	I wrote his presence last night.
	b	■	☐	We reached a very tentative agreement.
	c	☐	■	I don’t know the corporate sky.
	d	☐	■	He asked a bin about the test tomorrow.
	e	■	☐	The company got some financial support.
	f	☐	■	I washed my deadlines carefully.
	g	■	☐	The current situation is different from the past.
	h	■	☐	My colleague works for the company.
	i	■	☐	He acquired his license last year.
	j	☐	■	Trains provide beautiful planets.

Question	Subquestion	Appropriate	Inappropriate	Sentence
16	a	☐	■	Furniture is not angry every day.
	b	■	☐	You need to make some corrections.
	c	■	☐	It was a challenging job for me.

	d	■	☐	I drink tea for medicinal purposes.
	e	☐	■	Our dog has had four committees.
	f	☐	■	He is my temporary cousin.
	g	■	☐	I went to a supermarket for some refreshments.
	h	☐	■	There are complimentary ghosts in this hotel.
	i	■	☐	His presence makes me feel stronger.
	j	■	☐	My sister got a promotion in her company.

Supplementary information-S5: Conversion Matrix of the Timed Lexicosemantic Task

CEFR Levels	LJT scores (160 points)		CEFR descriptors
	Lower	Upper	
C1	**> 135** (84.4%)		Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can understand standard spoken language, live or broadcast, on both familiar and unfamiliar topics.
B2_upper	**> 128.6** (80.4%)		Can understand extended speech even when it is not clearly structured and can understand television programs and films without too much effort.
B2_lower	**> 121.0** (75.6%)		Can understand extended speech and lectures and follow complex lines of argument provided the topic is reasonably familiar.
B1_upper	**> 110.3** (68.9%)		Can understand the main points of clear standard input and some specific information in everyday situations.
B1_lower	**> 103.2** (64.5%)		Can understand the main points of clear standard input on familiar matters encountered in work, school, or leisure
A2	**< 96.8** (60.5%)		Can understand simple, routine phrases and sentences related to personal information and immediate needs.

Using the Lexicosemantic Judgement Task (LJT) scores, we can categorize listeners' proficiency levels according to the CEFR system: A for Basic Users, B for Independent Users, and C for Proficient Users. To create the conversion matrix, we tested a total of 487 Japanese EFL learners, with varied proficiency and experience, using the TOEIC test and various vocabulary tasks (Saito & Uchihara, forthcoming). We then categorized these 487 learners into their respective CEFR proficiency levels using the ETS's conversion table (https://www.ets.org/pdfs/toeic/toeic-mapping-cefr-reference.pdf). Through 95% CI analyses, we determined the LJT score ranges corresponding to each CEFR proficiency level:

- If students take the test, and his score is less than 97 out of 160 points, they could be considered "A2"
- If students take the test, and his score is more than 103 out of 160 points, they could be considered "B1_lower"
- If students take the test, and his score is more than 110 out of 160 points, they could be considered "B1_upper"
- If students take the test, and his score is more than 121 out of 160 points, they could be considered "B2_lower"
- If students take the test, and his score is more than 128 out of 160 points, they could be considered "B2_upper"
- If students take the test, and his score is more than 135 out of 160 points, they could be considered "C1"