Nadasdy, P., Iso, T., Aizawa, K., & Sato, K. (2026). Development and validation of an auditory lexical size test for Japanese learners of English. *Accents Asia, 21*(3), 47-68.

Development and Validation of an Auditory Lexical Size Test for Japanese Learners of English

Paul Nadasdy[1], Tatsuo Iso[2], Kazumi Aizawa[3], Kento Sato[4]
Tokyo Denki University

ABSTRACT

Although several aural vocabulary size tests have been developed, including those by Fountain and Nation (2000) and Milton and Hopkins (2006), relatively few have been developed and validated with Japanese learners of English in mind. One notable exception is McLean, Kramer, and Beglar's (2015) Listening Vocabulary Levels Test (LVLT), which was specifically developed to assess Japanese learners' aural vocabulary knowledge. However, there remains scope for developing tests that combine validated aural assessment formats with lexical content selected specifically for Japanese EFL learners. This study introduces the Auditory Lexical Size Test (ALEXIS), a diagnostic tool designed to measure the aural vocabulary size of Japanese university students. ALEXIS draws on content from Hamada et al.'s Vocabulary Size Test for Japanese EFL learners (VST-NJ8) and adopts key features of McLean, Kramer, and Beglar's (2015) LVLT. Data from 280 students at a technical college in Japan were collected, and a Rasch analysis focusing on item difficulty and unidimensionality was conducted. Cronbach's alpha indicated high internal consistency (α = .948). Additionally, ALEXIS scores correlated positively with overall TOEIC scores (r = .54), providing evidence of criterion-related validity. Although some items showed misfit or overfit, potentially related to abstraction, polysemy, and distractor-related confusion, the findings provide preliminary evidence supporting the use of ALEXIS as a measure of aural lexical knowledge. The initial findings suggest that ALEXIS may provide a reliable means of assessing aural vocabulary knowledge and contribute to research on vocabulary assessment for Japanese learners.

Keywords: auditory vocabulary; vocabulary size testing; Japanese EFL learners; listening vocabulary; language assessment

[1] Paul Nadasdy is an Associate Professor at Tokyo Denki University, Japan. His research interests include English language education, vocabulary assessment, listening, and the relationship between language and media. Correspondence should be sent to: nadasdy@cck.dendai.ac.jp

[2] Tatsuo Iso is a Professor of English Education at Tokyo Denki University, Japan. His research interests include second language vocabulary acquisition, lexical automaticity, and language assessment. Correspondence should be sent to: tiso@mail.dendai.ac.jp

[3] Kazumi Aizawa is a Professor of English Education at Tokyo Denki University, Japan. His research interests include second language vocabulary acquisition, vocabulary assessment, and language testing. Correspondence should be sent to: aizawa@cck.dendai.ac.jp

[4] Kento Sato is a Lecturer in the Department of English Education at Tokyo Denki University, Japan. His research interests include vocabulary acquisition, computer-assisted language learning (CALL), and English language education. Correspondence should be sent to: ksato@mail.dendai.ac.jp

INTRODUCTION

Vocabulary acquisition is essential for second or foreign language learning. A larger vocabulary enables learners to communicate more effectively, making vocabulary development an important goal for language teachers. As Webb and Nation (2017) state, "words are the building blocks of language," and at all stages of education, building vocabulary knowledge is an essential part of the learning process (pp. 5–6). To achieve learning goals, teachers need to assess how much vocabulary their students know. To assess lexical knowledge, vocabulary tests are often used because they help identify gaps in students' knowledge and highlight their communicative needs (Read, 2000, pp. 1–3).

Traditionally, vocabulary testing has focused on the written rather than spoken forms of words (Mizumoto & Shimamoto, 2008). However, with the increased access to digital technologies, the ways in which learners interact with foreign languages have changed. Through videos and other online media, learners now have greater opportunities to encounter spoken language outside the classroom (Dizon, 2021). This increased exposure highlights the importance of assessing whether learners can recognize and understand vocabulary in its spoken form, and therefore the need for greater attention to aural vocabulary assessment. This may be particularly important in Japan, where learners often experience difficulty recognizing in speech words that they know in written form (Mizumoto & Shimamoto, 2008). Despite this, dedicated tests of aural vocabulary knowledge remain relatively uncommon (Mizumoto & Shimamoto, 2008).

Although it has been suggested that a separate aural vocabulary test may not be necessary for predicting listening ability when learners have already taken a written vocabulary test (Katagiri, 2001, as cited in Mizumoto & Shimamoto, 2008, p. 48), studies comparing written and aural vocabulary knowledge have identified differences between the two modalities among Japanese learners. One possible explanation is that Japanese learners are often exposed to written forms before developing sufficiently robust phonological representations, which may subsequently impede recognition of words in speech (Nadasdy, 2026).

To address this, further development of validated aural vocabulary tests tailored to Japanese learners is needed, particularly given learners' increased exposure to spoken English through digital and online media. Existing aural tests tend to have limitations, as some rely on outdated word lists, contain potential confounding factors, or use procedures that may not reflect authentic listening conditions (McLean, Kramer, & Beglar, 2015). Furthermore, most vocabulary size tests focus on written rather than spoken vocabulary, leaving scope for assessments that specifically measure aural lexical knowledge. Greater attention should therefore be given to developing reliable tests that can accurately identify gaps in learners' spoken vocabulary knowledge and support more effective language learning (Read, 2000, p. 26).

LITERATURE REVIEW

Vocabulary Testing

Vocabulary testing has evolved over the past four decades. One of the most widely used tests is Nation's Vocabulary Levels Test (VLT; 1983, 1990), a diagnostic tool which measures

learners' knowledge of both high-frequency and lower-frequency words and covers up to the 10,000-word level. Another well-known test is Meara's Yes/No test (1992). The testing approach differs from others in that it presents a mix of real and pseudo-words and asks students to identify which ones they know. Meara's test is considered a quick and efficient way to assess vocabulary size, though it does not provide results at individual thousand-word levels. In 2007, Nation and Beglar developed the Vocabulary Size Test (VST) to estimate total vocabulary size. The test was designed using Nation's (2006) 14,000-word-family lists, which were revised based on word frequency data from the 10-million-token spoken subsection of the British National Corpus (BNC) (Nation & Beglar, 2007, p. 10). The test is considered reliable, while its multiple-choice format facilitates scoring and administration. Careful distractor selection also helps reduce the effects of guessing, although random guessing cannot be eliminated entirely (McLean, Kramer, & Stewart, 2015, p. 27). Based on more modern corpora than the VLT, McLean, Kramer, and Beglar (2015) created the New Vocabulary Levels Test (NVLT), a written-receptive, multiple-choice test that runs parallel to the Listening Vocabulary Levels Test (LVLT). The authors addressed the limitations of the original VLT, improving reliability by ensuring items discriminated well between different ability levels and by reducing random guessing effects (McLean, Kramer, & Beglar, 2015, pp. 4–6). The main objective of the tests mentioned above is to assess written vocabulary knowledge. Consequently, vocabulary assessment has traditionally placed greater emphasis on written rather than spoken forms, with relatively few validated tests available for assessing aural vocabulary knowledge (McLean, Kramer, & Beglar, 2015).

Aural Lexical Size Testing

Aural vocabulary tests are relatively limited (Milton, 2009; Stæhr, 2009). Three aural tests stand out in the literature. First, Fountain and Nation's (2000) Graded Dictation Test measures recognition of high-frequency spoken words but may conflate listening ability with vocabulary knowledge. Reliance on contextual information may also introduce additional sources of variability in test performance. The second is Milton and Hopkins's (2006) A-Lex, a computer-based yes/no test of receptive vocabulary. Criticisms include that its authenticity may be reduced by unlimited reuse of the same words and that its reliance on response speed may measure processing ability rather than vocabulary size (Buck, 2001; Milton, 2009). Finally, McLean, Kramer, and Beglar's (2015) Listening Vocabulary Levels Test (LVLT). The LVLT tests high-frequency word knowledge using Nation's BNC/COCA word list and Coxhead's Academic Word List (AWL) (Coxhead, 2000). Though the LVLT "provides a comprehensive measure of aural vocabulary knowledge" (McLean, Kramer, & Beglar, 2015, p. 16), it uses general word lists, which may not be entirely suitable for the specific needs of Japanese learners. The limitations of contextual interference, test format, and word list appropriateness indicate that existing tools do not fully address the needs of Japanese learners. This provides a rationale for the development of a targeted and empirically validated aural vocabulary test.

Orthographic versus Phonological Vocabulary Size

As many teachers of English as a foreign language in Japan know, Japanese learners of English tend to struggle with the spoken form of words. One possible explanation is that learners are often exposed to the written form of a word before hearing it, which may result in weaker knowledge of its spoken form (Mizumoto & Shimamoto, 2008; Nadasdy, 2026). Differences between the phonological systems of Japanese and English may also contribute to difficulties with listening comprehension and pronunciation (Thompson, 2001, p. 297). In

parallel testing between orthographic and phonological forms, it has been shown that written forms are better known than spoken forms (Aizawa et al., 2017), with lower-proficiency Japanese learners sometimes failing to retrieve the meaning of a spoken word unless they see its written form (Mizumoto & Shimamoto, 2008, p. 48). Mizumoto and Shimamoto highlight tests conducted by Yamauchi (2005) and Katagiri (2001) to further support this. Yamauchi (2005, as cited in Mizumoto & Shimamoto, 2008, p. 38) used a web-based vocabulary test constructed from Nation's VLT and reported a stronger relationship between vocabulary size and reading ability (r = .49) than between vocabulary size and listening ability (r = .32). Using Mochizuki's vocabulary size test (1998) with high school students, Katagiri (2001) found that listening tests were more challenging than written vocabulary tests and that the written version showed a stronger correlation with TOEIC scores (Mizumoto & Shimamoto, 2008, pp. 38–39). However, Katagiri (2001, as cited in Mizumoto & Shimamoto, 2008, p. 48) suggested that if the aim is to predict learners' listening ability, administering a separate aural vocabulary size test may not be necessary when learners have already taken the written version. Aural vocabulary assessment can, however, provide information about learners' ability to recognize spoken words and help identify vocabulary-related weaknesses that may contribute to difficulties in listening comprehension.

Word Lists and Test Formats

Vocabulary tests depend on two key components: reliable word lists and effective test formats. Word lists that have been commonly used for vocabulary size tests include Thorndike and Lorge's lemma-based list (1944), West's General Service List (GSL) (1953), Coxhead's AWL (2000), the BNC/COCA word lists developed by Nation and Webb (2011), and the New General Service List (NGSL) developed by Browne (2013). In Japan, several word lists have been favored for their relevance to Japanese students. The Hokkaido University English Vocabulary List (HUEVL), created by Sonoda (1996), is widely used, as is the Standard Vocabulary List (SVL 12000), used for TOEIC and TOEFL® preparation, which also aligns with the Eiken test (Chujo & Oghigian, 2009). Another word list favored in Japan is the New JACET list of 8,000 Basic Words (Committee of Revising the JACET Basic Words, 2016). The developers of the JACET word list believe that there should be a specific set of vocabulary targeting Japanese learners, not solely based on frequency, but on pedagogical considerations (Ishikawa, 2019, p. 2). The New JACET List of 8,000 Basic Words was constructed by analyzing word frequency data from the BNC, focusing specifically on catering to the needs of Japanese learners. Comparing the word lists made in Japan to those designed in English-speaking countries, Aizawa (1998) suggested that some vocabulary tests based on the lists used, as in Nation (1990), Meara (1992), Read (1993), and Wesche and Paribakht (1996), may not be completely appropriate for testing Japanese students (pp. 74–75). Because the contents are designed for students starting their education at universities in English-speaking countries, they are not wholly appropriate for Japanese learners of English at a lower proficiency level. Compared to other available word lists, the New JACET List of 8,000 Basic Words targets Japanese learners specifically (Hamada et al., 2021, p. 26). Words were added to the list to help students communicate in everyday conversations, write reports, support their academic endeavors, and align with standardized English proficiency tests (Hamada et al., 2021, p. 26). This word list closely matches current educational contexts in Japan, and results suggest that it provides "an accurate measure of vocabulary difficulty" (Stewart & Stewart, 2016, p. 62). Moreover, "the list can be used for predictive purposes, i.e., choosing appropriate texts for college courses, and also for productive purposes, i.e., writing syllabuses and entrance examination questions" (Stewart & Stewart, 2016, p. 62).

For test formats, multiple-choice questions are widely favored. Compared with some yes/no formats, they may reduce the effects of guessing (Hamada et al., 2021). Multiple-choice items also allow for advanced analyses using Item Response Theory (IRT), improving accuracy and test design (Beglar, 2010). Yes/no and multiple-choice tests have each been criticized for encouraging guessing, such as Meara and Buxton's yes/no test (Mochida & Harrington, 2006, p. 80), or for being too easy, as in Nation and Beglar's VST (Milton 2009, p. 119). Multiple-choice questions might be favorable for testing Japanese students, as they sometimes underestimate their ability to understand a word and might indicate "No" on a test when they do, in fact, have some knowledge of the word (Mochida & Harrington, 2006, as cited in Stubbe, 2014, p. 32; Stubbe, Stewart & Pritchard, 2010, as cited in Stubbe, 2014, p. 33). Furthermore, although yes/no tests can be administered quickly, they usually require many test items to get an insight into students' vocabulary size.

THE CURRENT STUDY

In this study, a new auditory lexical size test (ALEXIS) was developed and evaluated to measure the auditory vocabulary knowledge of Japanese learners of English. ALEXIS was designed as a diagnostic tool to assess how well students can recognize and understand spoken English words. To achieve this, content from established vocabulary size tests was adapted, with particular attention given to the specific needs of Japanese EFL learners. The research questions are as follows:

RQ1: What evidence supports the reliability and validity of the initial version of ALEXIS as a measure of aural vocabulary knowledge?
RQ2: What items demonstrate problematic or weak performance, and what factors may account for their performance?

Test Construction

ALEXIS was developed to assess Japanese learners' aural vocabulary knowledge, adapting content from Hamada et al.'s VST-NJ8 (See Appendix B for a full list of test items and content) and using the question format of McLean et al.'s LVLT.

Content and Format

Content was taken directly from the VST-NJ8, which is based on the New JACET List of 8,000 Basic Words. Minor changes were made to ensure suitability for auditory testing. The final ALEXIS test content is provided in Appendix B. Unlike the VST-NJ8, in which test takers see Japanese words and select the corresponding meaning from four written English options, ALEXIS presents each item in the following format:

1. The target English word is read aloud once in isolation.
2. A short context sentence follows to provide contextual information and, where appropriate, clarify the target word's grammatical function.
3. Four possible written Japanese options are presented, along with a fifth "don't know" option to discourage guessing.

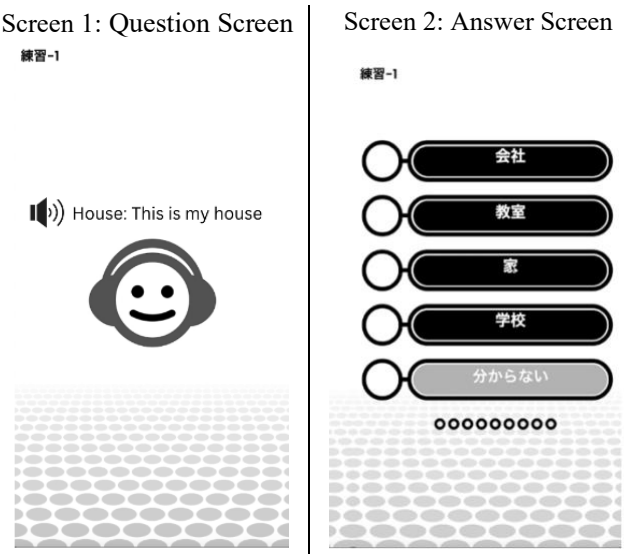
The format was designed to assess auditory vocabulary recognition while reducing the influence of written-form recognition. By presenting the target word aurally with written Japanese options, the test focuses on the learner's ability to recognize the spoken word and identify its meaning. The context sentences were simplified to provide sufficient contextual information while minimizing opportunities to infer the answer from context alone.

Participants hear: "House: This is my house."

(1)
a. 会社 (company)
b. 教室 (classroom)
c. 学校 (school)
d. 家 (house)

As shown in Figure 1, the ALEXIS interface consists of two main screens.

FIGURE 1
Sample screen of ALEXIS



Specifications

Microsoft Bing Translator (2023 version) was used to record the audio files. Bing Translator was selected for its clear American English pronunciation. The audio files were edited using Audio Hijack (Rogue Amoeba Software, Inc., version 4.1.1). ALEXIS was developed in JavaScript and designed to run on both personal computers and smartphones. Students completed the test in a classroom setting with access to a stable Wi-Fi connection.

METHODOLOGY

Participants

The study included 280 first-year Japanese university students enrolled at a technical college in Tokyo. Participants' majors included engineering, architecture, robotics, and system design. Participants' TOEIC scores ranged from 130 to 750, with a median of 360. The mean score was 353.81 (95% CI [340.92, 366.71]), with a standard deviation of 108.13.

Test Administration

The beta version of ALEXIS was administered over two weeks during regular English classes. At the start of the test, instructions were displayed in Japanese as a guide, followed by five example questions to familiarize students with the format and procedure.

Test Design

The test consists of eight levels of difficulty and takes approximately 40 minutes to complete. Levels 1–4 were presented without breaks. A one-minute break was provided between Levels 4 and 5 to reduce test fatigue. Each question was presented with a single playback of the target word followed by a non-defining context sentence. Participants were given ten seconds to respond, during which a countdown timer was displayed. A five-second pause was provided between questions to allow time for processing. At the end of the test, participants received individual feedback on their estimated vocabulary size score.

Test Environment

Tests were administered online, with each student using a personal laptop in the classroom. Students were instructed to answer each question to the best of their ability and were encouraged to select the "don't know" option when uncertain, to reduce the likelihood of guessing.

RESULTS

Validity and Reliability

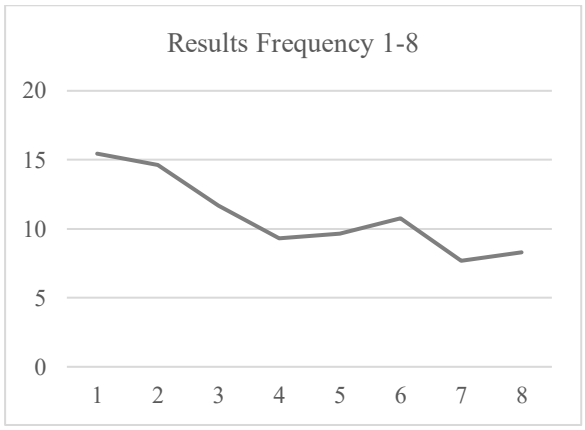
ALEXIS consists of 160 items distributed across eight frequency levels, with 20 items per level. The mean scores for each level are presented in Table 1. Table 1 and Figure 2 show that mean scores generally decrease as word difficulty increases, but this pattern becomes less consistent at Levels 5–8, especially in Level 6, where mean scores increase unexpectedly.

TABLE 1
Vocabulary Size Across Proficiency Levels

Level	M	SD	Min	Max
L1	15.45	3.39	2.00	20.00
L2	14.61	4.24	2.00	20.00
L3	11.66	4.27	2.00	20.00
L4	9.31	4.14	0.00	20.00
L5	9.66	3.43	1.00	19.00
L6	10.78	3.38	1.00	19.00
L7	7.68	2.98	0.00	17.00
L8	8.31	3.21	0.00	17.00
Total	87.46	23.98	21.00	150.00

Note: *M* = Mean; *SD* = Standard Deviation.

FIGURE 2
Results Based on Frequency



ANOVA and Post-Hoc Comparisons

A one-way repeated-measures ANOVA was conducted to determine whether mean scores differed significantly across the eight frequency levels. The results showed a significant main effect of frequency level, $F(7, 1953) = 444.26$, $p < .001$, $\eta^2 = .614$, indicating a large effect of frequency level. Holm-adjusted post hoc comparisons were then conducted to identify significant differences between levels. The difference between Levels 4 and 5 was not statistically significant, although the mean score for Level 5 was slightly higher than that of Level 4. The difference between Levels 5 and 6 was significant, with Level 6 showing an unexpectedly higher mean score. The difference between Levels 7 and 8 was also statistically significant, with Level 8 showing a higher mean score. These deviations from the expected pattern suggest that minor adjustments to some items at the higher frequency levels may be necessary.

Item Fit

Most items demonstrated acceptable fit with the Rasch model and covered a wide range of difficulty. Among the misfit items, two were particularly noteworthy, including one that was

especially evident in the graphical analysis. Item 6 (Level 1) exhibited lower-than-expected performance despite belonging to the easiest frequency level. Item 101 (Level 6) showed poor discrimination in its Item Characteristic Curve (ICC).

FIGURE 3
Item Characteristic Curves from Levels 1 & 6

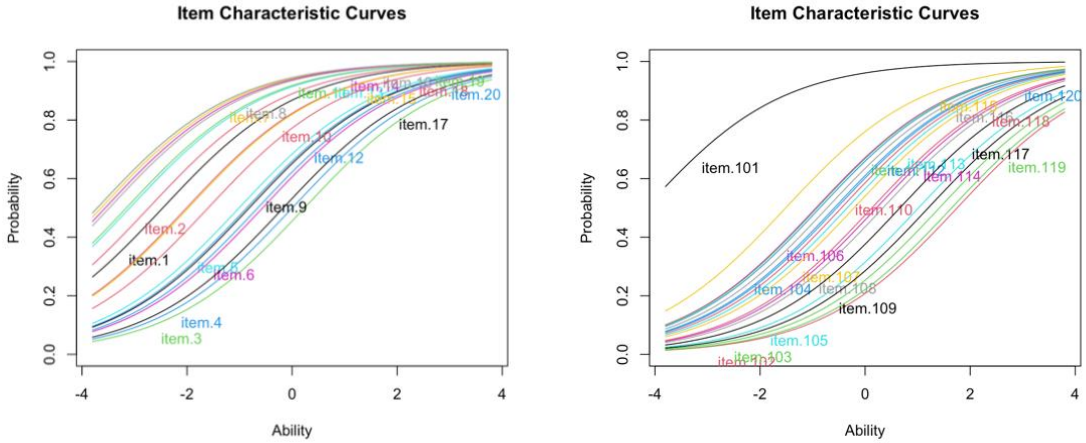


FIGURE 4
Plot of the Person Parameters

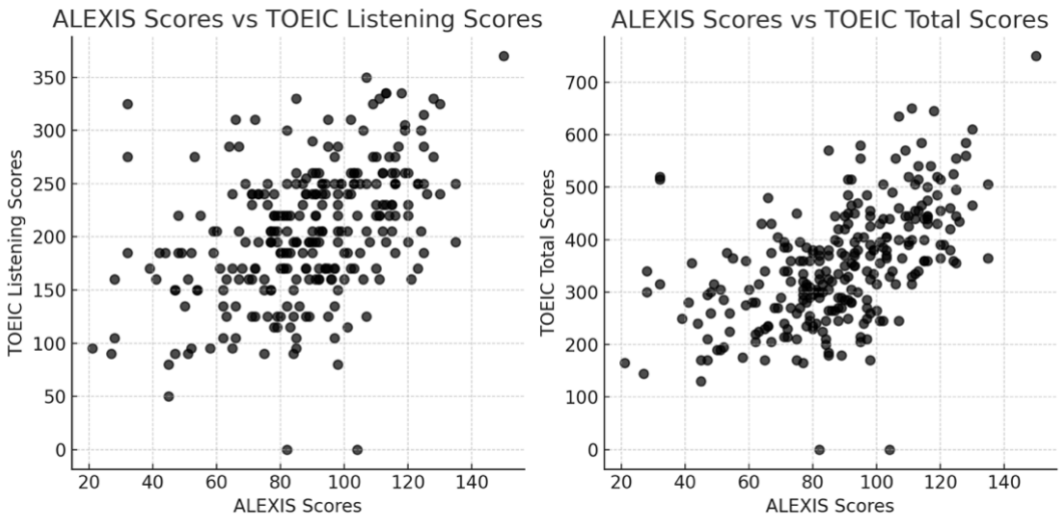
Internal Consistency

Cronbach's alpha was calculated as α = .948, indicating a high level of internal consistency among the test items. This provides evidence of the reliability of ALEXIS scores.

Correlation with TOEIC Scores

To measure criterion-related validity, Pearson correlation coefficients were calculated to assess the relationships between ALEXIS scores and TOEIC scores. ALEXIS scores showed a moderate positive correlation with TOEIC Listening scores (r = .44) and a stronger positive correlation with overall TOEIC scores (r = .54).

FIGURE 5
Correlation Between ALEXIS and TOEIC Listening and Overall Scores



DISCUSSION

Reliability and Validity (RQ1)

Overall, the results provide evidence that ALEXIS has strong internal consistency and is positively associated with established measures of English proficiency. The high internal consistency (Cronbach's α = .948) indicates that the test is reliable. Although a high alpha may sometimes indicate redundancy in item content, no item pairs exceeded .8, suggesting that redundancy is not a major concern. ALEXIS scores showed a moderate positive correlation with TOEIC Listening scores (r = .44) and a stronger positive correlation with overall TOEIC scores (r = .54). These findings, together with the correlation reported for the VST-NJ8 (r = .62), provide additional evidence of the test's validity as a measure of vocabulary knowledge. However, the Rasch analysis identified issues with item performance, particularly for items within Levels 1 and 4 and Levels 5–8. Therefore, although ALEXIS demonstrates promising reliability and validity evidence, further item-level analysis and refinement may be necessary.

Problematic Items (RQ2): Qualitative and Quantitative Analysis

Level 1:

In Level 1, Item 6 ("usually") was a clear misfit, despite being in the easiest level. Responses to Item 6 deviated from the model's expectations, with a chi-square statistic of $\chi^2(279) = 347.02$, p = .003, indicating statistically significant item misfit.

More evidence that Item 6 is misfitting can be seen in both its Outfit MSQ = 1.239 and Infit MSQ = 1.19 (a fit statistic, where values near 1.0 indicate good fit), showing variability in test-taker responses. High t-values (Outfit t = 3.388 / Infit t = 3.639) exceed the ±2 threshold, providing additional evidence of misfit. In addition, the item's low

discrimination (0.119) indicates that it does not distinguish between test takers of varying proficiency levels. This reveals possible weaknesses in the context sentences or distractor choices. Overall, Item 6 can be classified as a statistically significant misfit.

Other items in Level 1 that warrant attention include Item 7 (“area”), Item 11 (“special”), Item 14 (“begin”), and Item 16 (“room”). Although these items did not show statistically significant misfit, they had low MSQ values and negative *t*-values. MSQ values below 1.0 indicate less variation than expected by the model and therefore potential overfit, whereas values above 1.0 indicate greater variation or noise and potential underfit. These items should therefore be considered for further review and possible revision in future iterations of the test.

Level 2 & 3:

There were some minor misfitting items in Levels 2 and 3; however, these were less pronounced than the item-level issues identified at other frequency levels.

Level 4:

The main issue in Level 4 was overfit. Three items stood out in the data: Item 62 (“literally”), Item 74 (“restrict”), and Item 78 (“visible”). Although these items did not show statistically significant misfit, their low MSQ values and negative *t*-values indicated potential overfit. The least problematic, Item 62 (“literally”), had an Infit *t*-value of −3.417 and an Outfit MSQ of 0.853. Item 74 (“restrict”) had an Infit *t*-value of −3.56 and an Outfit MSQ of 0.825. Similarly, Item 78 (“visible”), the most overfitting item, had an Infit *t*-value of −3.996 and an Outfit MSQ of 0.825. These items displayed more predictable response patterns than expected by the model and should therefore be considered for further review and possible revision in future iterations of the test (see Table A.2 in Appendix A for detailed overfit statistics).

Level 5–8:

In Levels 5–8, a range of linguistic and contextual factors may have contributed to unusual item performance, including abstraction, polysemy, cultural factors, vague or insufficient context, and interference from distractors.

Level 5

In Level 5, for example, Item 85 (“rating”) produced $\chi^2(279) = 312.68$, $p = .081$. Although this did not indicate statistically significant misfit, the slightly elevated Outfit MSQ (1.117) indicates greater response variability than expected and may reflect unexpected responses or other sources of noise. Learners may not have been familiar with how the word “rating” is commonly used or may have been confused by its potential overlap with words such as “grade” or “score.” Item 90 (“listing”) also showed unexpected response patterns, with an Outfit MSQ of 1.353 and Outfit $t = 2.90$, providing evidence of item misfit. The low discrimination estimate (0.103) suggests that the item may have limited ability to differentiate among test takers with different proficiency levels. There may have been issues with the distractors (寛容さ [“tolerance”], 礼儀作法 [“courtesy”], 傾斜面 [“inclination”]). Although 24% answered correctly, 32% selected “courtesy” and 17% selected “tolerance,” suggesting that the distractors may have contributed to confusion. Item 92 (“accomplishment”) also showed signs of misfit (Outfit $t = 1.99$; Infit $t = 2.21$), which may relate to abstraction and/or

word length. Item 96 ("prayer") showed statistically significant misfit, $\chi^2(279) = 324.06$, $p = .033$. Its Outfit MSQ of 1.157 indicates greater response variability than expected, while its Infit MSQ of 1.042 was close to the model expectation of 1.00. Its distractors (地球 ["earth"], 窒素 ["nitrogen"], 雄羊 ["ram"]) are semantically distinct, and its non-defining context sentence ("It is a prayer") appears unlikely to have provided substantial clues to the answer. Although 67% of participants answered correctly, the Rasch analysis identified unexpected response patterns among test takers of differing ability levels, which may have resulted from guessing or other sources of response variability. Another significantly misfitting item was Item 99 ("integral"), $\chi^2(279) = 334.88$, $p = .012$, with an Outfit MSQ of 1.196, Infit MSQ of 1.165, Outfit $t = 3.35$, and Infit $t = 4.00$. The low discrimination estimate (0.145) suggests that the item may have limited ability to differentiate among test takers with different proficiency levels. Of the distractors ("offshore," "legislation," and "triple"), "legislation" was the most frequently selected incorrect answer. Test takers may have been able to eliminate the other distractors relatively easily, effectively reducing the item to a choice between "integral" and "legislation." The abstract nature of both terms may have contributed to this response pattern, and additional contextual information (e.g., "This idea is integral") may improve the item in future versions.

Level 6

In Level 6, as noted in the visual analysis of the ICC (Figure 3), Item 101 ("gorgeous") appeared to be easier than expected for its assigned level. The low Outfit MSQ (0.777) and Infit MSQ (0.905), together with the negative *t*-values (Outfit $t = -0.553$; Infit $t = -0.324$), provided evidence of overfit, suggesting that responses were more predictable than expected. The loanword "ゴージャス" may have made the item more familiar to Japanese test takers than intended. The item should therefore be considered for revision or removal in future iterations of the test.

Several other items in Level 6 also showed problematic response patterns. Item 102 ("hence") showed statistically significant misfit, $\chi^2(279) = 321.98$, $p = .039$, although the Infit and Outfit *t*-statistics (1.211 and 1.341, respectively) were below the ±2.00 threshold, suggesting that the magnitude of the deviation was relatively modest. Its low discrimination estimate suggests that the item may have limited ability to differentiate among test takers with different proficiency levels. Item 109 ("liberate") also showed statistically significant misfit. Its Outfit *t* of 1.815 and Infit *t* of 1.495 were below the ±2.00 threshold, while its low discrimination estimate (0.198) suggests limited differentiation among test takers. The relatively low frequency of "liberate" compared with the more common "free" may have contributed to its difficulty.

Item 114 ("discrepancy") also showed statistically significant misfit, $\chi^2(279) = 326.05$, $p = .028$. Its elevated Outfit MSQ (1.164) and Infit MSQ (1.104), together with Outfit $t = 2.863$ and Infit $t = 2.585$, provided further evidence of unexpected response variation. The abstract nature and relatively low frequency of the word may have contributed to this response pattern. Finally, Item 119 ("verification") showed substantial evidence of misfit. Its elevated Outfit MSQ (1.257) and Infit MSQ (1.15) indicated greater response variability than expected, while its low discrimination estimate (0.1) suggests that it may have limited ability to differentiate among test takers with different proficiency levels. One possible source of difficulty is the context sentence "Please show some verification," which may not provide sufficient contextual support and should be considered for revision in future versions of the test.

Level 7

In Level 7, three items stood out to differing degrees: Item 122 ("warrant"), Item 123 ("freight"), and Item 127 ("motorist"). Item 122 showed statistically significant misfit, $\chi^2(279) = 320.18$, $p = .045$, with an Outfit MSQ of 1.144, Infit MSQ of 1.105, Outfit $t = 2.324$, and Infit $t = 2.299$. Its relatively specialized legal meaning may have contributed to its difficulty for this sample. Item 123 ("freight") also showed statistically significant misfit, $\chi^2(279) = 344.18$, $p = .005$. Its Outfit MSQ of 1.229 and Infit MSQ of 1.077 indicated greater response variability than expected. Its low discrimination estimate (0.145) also suggests that the item may have had limited ability to differentiate among test takers with different proficiency levels.

Item 127 ("motorist") showed particularly strong evidence of misfit, $\chi^2(279) = 480.41$, $p < .001$. Its elevated Outfit MSQ (1.716) and Outfit t (4.056) indicated substantial unexpected response variation, while its low discrimination estimate (0.06) suggests limited differentiation among test takers. Comparison with the orthographic VST-NJ8 provides additional insight, as "motorist" and "cardiac" (see Appendix A.1) were also difficult in the written test. "Motorist" is less frequent in everyday English than the more common word "driver," which may have contributed to its difficulty. Similarly, the relatively specialized meaning of "cardiac" may have contributed to its difficulty for this particular sample.

Level 8

Finally, in Level 8, several items showed evidence of misfit. The three most notable were Item 148 ("anthem"), Item 150 ("perennial"), and Item 154 ("decree"). Item 148 showed statistically significant misfit, with $\chi^2(279) = 340.202$, $p = .007$, an Outfit MSQ of 1.215, and Outfit $t = 2.837$, indicating greater response variability than expected. Item 150 ("perennial") showed statistically significant misfit, $\chi^2(279) = 372.772$, $p < .001$, with an Outfit MSQ of 1.331. Its low discrimination estimate (0.052) suggests that the item may have limited ability to differentiate among test takers with different proficiency levels. Finally, Item 154 ("decree") showed particularly strong evidence of misfit, $\chi^2(279) = 442.871$, $p < .001$, with an Outfit MSQ of 1.582. Its very low discrimination estimate (0.021) similarly suggests limited differentiation among test takers. These findings identify several items in the higher-frequency levels that warrant further review. Modifying distractors and context sentences and reducing potential ambiguity may help improve these items in future iterations of ALEXIS. For further details of the item-fit statistics, including overfitting and misfitting items, see Tables A.1–A.3 in Appendix A.

CONCLUSION

In this study, an auditory lexical size test of English was developed, and its validity and reliability were evaluated. ALEXIS was constructed using vocabulary content from Hamada et al.'s VST-NJ8 (2021) and the format of McLean et al.'s LVLT (2015) and was designed particularly for Japanese learners of English. The results provide promising evidence of reliability and validity. ALEXIS demonstrated high internal consistency and was positively associated with external measures of English proficiency, including TOEIC Listening and overall scores. The results also indicate that the test can differentiate learners across different levels of vocabulary proficiency.

Although ALEXIS showed promising evidence of validity and reliability overall, several limitations were identified. In particular, several items showed misfit or overfit at specific frequency levels. These patterns may be partly attributable to the direct adoption of items from the VST-NJ8, some of which may be less suitable for assessing aural vocabulary knowledge. Future versions of the test will therefore incorporate revisions to distractors and context sentences, while individual vocabulary items will be modified or removed where necessary. Further validation will also require data from a larger and more diverse sample of participants. With continued refinement and data collection, ALEXIS has the potential to become a practical and reliable tool for measuring aural vocabulary knowledge, a skill that is increasingly relevant in contemporary language-learning environments.

APPENDIX A

TABLE A.1
Item Fit Statistics

Level	Item	Chisq	df	*p*-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t	Discrim	Dif	Context Sentence	Stem
1	**Item 6**	347.022	279	0.003	1.239	1.19	3.388	3.639	0.119	-0.75	I usually study alone.	usually
	Item 7	225.174	279	0.992	0.804	0.883	-0.636	-0.565	0.349	-3.86	I love this area.	area
	Item 11	217.187	279	0.998	0.776	0.849	-0.722	-0.729	0.413	-3.94	He is special.	special
	Item 14	210.638	279	0.999	0.752	0.883	-1.927	-1.238	0.471	-3.78	Let's begin.	begin
	Item 16	194.066	279	1	0.693	0.828	-1.063	-0.848	0.419	-3.94	This is the room.	room
4	**Item 62**	238.973	279	0.96	0.853	0.856	-2.26	-3.417	0.49	0.67	Don't take it literally.	literally
	Item 74	230.997	279	0.984	0.825	0.866	-3.376	-3.56	0.507	0.04	This will restrict its use.	restrict
	Item 78	230.934	279	0.984	0.825	0.847	-3.385	-3.996	0.53	-0.15	It is not visible.	visible
5	**Item 85**	312.683	279	0.081	1.117	1.109	2.054	2.585	0.221	-0.22	Please give a rating.	rating
	Item 90	378.721	279	0.0	1.353	1.115	2.896	1.595	0.103	1.63	This listing is new.	listing
	Item 92	313.254	279	0.077	1.119	1.09	1.992	2.207	0.223	0.30	It is an accomplishment.	accomplishment
	Item 96	324.059	279	0.033	1.157	1.042	1.895	0.727	0.297	-1.13	It is a prayer.	prayer
	Item 99	334.88	279	0.012	1.196	1.165	3.35	4.001	0.145	0.10	This idea is integral.	integral
6	**Item 101**	217.517	279	0.997	0.777	0.905	-0.553	-0.324	0.309	-4.42	It is gorgeous.	gorgeous
	Item 102	321.976	279	0.039	1.15	1.085	1.341	1.211	0.165	1.60	I won, hence my mood.	hence
	Item 108	348.787	279	0.003	1.246	1.186	4.087	4.484	0.108	0.17	I have the inclination to leave.	inclination
	Item 109	322.929	279	0.036	1.153	1.079	1.815	1.495	0.198	1.04	Please liberate them.	liberate
	Item 114	326.046	279	0.028	1.164	1.104	2.863	2.585	0.212	0.04	There is a discrepancy.	discrepancy
	Item 119	352.083	279	0.002	1.257	1.15	2.331	2.206	0.1	1.49	Please show some verification.	verification
7	**Item 122**	320.183	279	0.045	1.144	1.105	2.324	2.299	0.218	-0.50	This is a warrant.	warrant
	Item 123	344.18	279	0.005	1.229	1.077	1.73	0.955	0.145	1.88	It is in the freight.	freight
	Item 127	480.407	279	0.0	1.716	1.093	4.056	0.985	0.06	2.21	He is a motorist.	motorist
	Item 148	340.202	279	0.007	1.215	1.08	2.837	1.705	0.204	0.78	This is an anthem.	anthem
	Item 150	372.772	279	0.0	1.331	1.156	2.279	1.759	0.052	2.01	It is perennial.	perennial
	Item 154	442.871	279	0.0	1.582	1.142	3.352	1.44	0.021	2.24	This law is by decree.	decree

TABLE A.2
Overfits (Too Predictable Performance)

Level	Item	Chisq	*p*-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t	Discrim	Stem
4	**Item 62**	238.973	0.96	0.853	0.856	-2.26	-3.417	0.49	literally
	Item 74	230.997	0.984	0.825	0.866	-3.376	-3.56	0.507	restrict
	Item 78	230.934	0.984	0.825	0.847	-3.385	-3.996	0.53	visible

TABLE A.3
Misfits

Level	Item	Chisq	*p*-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t	Discrim	Stem
1	Item 6	347.022	0.003	1.239	1.19	3.388	3.639	0.119	usually

Level	Item	Chisq	*p*-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t	Discrim	Stem
5	Item 85	312.683	0.081	1.117	1.109	2.054	2.585	0.221	rating
	Item 90	378.721	0.0	1.353	1.115	2.896	1.595	0.103	listing
	Item 92	313.254	0.077	1.119	1.09	1.992	2.207	0.223	accomplishment
	Item 96	324.059	0.033	1.157	1.042	1.895	0.727	0.297	prayer
	Item 98	318.182	0.053	1.136	1.038	2.134	0.924	0.253	erect
	Item 99	334.88	0.012	1.196	1.165	3.35	4.001	0.145	integral
6	Item 101	217.517	0.997	0.777	0.905	-0.553	-0.324	0.309	gorgeous
	Item 102	321.976	0.039	1.15	1.085	1.341	1.211	0.165	hence
	Item 108	348.787	0.003	1.246	1.186	4.087	4.484	0.108	inclination
	Item 109	322.929	0.036	1.153	1.079	1.815	1.495	0.198	liberate
	Item 114	326.046	0.028	1.164	1.104	2.863	2.585	0.212	discrepancy
	Item 119	352.083	0.002	1.257	1.15	2.331	2.206	0.1	verification
7	Item 122	320.183	0.045	1.144	1.105	2.324	2.299	0.218	warrant
	Item 123	344.18	0.005	1.229	1.077	1.73	0.955	0.145	freight
	Item 124	315.817	0.064	1.128	1.12	2.266	2.929	0.202	assertion
	Item 127	480.407	0.0	1.716	1.093	4.056	0.985	0.06	motorist
	Item 129	352.219	0.002	1.258	1.153	3.018	2.891	0.113	raider
	Item 131	332.663	0.015	1.188	1.138	2.117	2.465	0.136	poise
	Item 133	510.716	0.0	1.824	0.985	2.992	-0.053	0.109	cardiac
	Item 135	345.156	0.004	1.233	1.029	2.211	0.485	0.209	profoundly
	Item 136	383.12	0.0	1.368	1.196	3.267	2.882	0.029	constituency
	Item 140	392.333	0.0	1.401	1.091	1.92	0.746	0.067	attic
8	Item 145	334.86	0.012	1.196	1.082	1.417	0.965	0.137	womb
	Item 148	340.202	0.007	1.215	1.08	2.837	1.705	0.204	anthem
	Item 150	372.772	0.0	1.331	1.156	2.279	1.759	0.052	perennial
	Item 154	442.871	0.0	1.582	1.142	3.352	1.44	0.021	decree
	Item 156	319.51	0.048	1.141	1.093	1.658	1.735	0.18	omission
	Item 159	350.541	0.002	1.252	1.063	1.581	0.667	0.132	chore

APPENDIX B

Supplementary Materials: Final Content from ALEXIS.

No.	Item stem	Gram.	Cor. Ans	D1	D2	D3	Dif.	Disc.
				Frequency Level 1				
1	department	N	売り場	少年	会議	町	-2.68	0.33
2	wish	V	願う	戻る	取る	感じる	-2.95	0.50
3	let	V	許可する	試す	買う	費やす	0.06	0.35
4	round	ADJ	丸い	用意のできた	単純な	簡単な	-0.18	0.29
5	whole	ADJ	全体の	重要な	その他の	大きい	-1.20	0.36
6	usually	ADV	普通は	決して…でない	常に	ついに	-0.75	0.12
7	area	N	地域	様式	芸術	記録	-3.86	0.35
8	bank	N	銀行	話題	年齢	世紀	-3.71	0.48
9	performance	N	実行	**事業**	心	大統領	-1.01	0.35
10	year	N	年	状態	力	父親	-2.22	0.37
11	special	ADJ	特別な	西洋の	重要な	古い	-3.94	0.41
12	then	ADV	それから	単独で	他に	遠く	-1.06	0.33
13	pressure	N	圧力	野原	**援助**	質問	-3.32	0.43
14	begin	V	始める	立つ	する	欲しい	-3.78	0.40
15	past	ADJ	過去の	主要な	それぞれの	赤い	-2.18	0.47
16	room	N	部屋	娘	増加	正面	-3.94	0.42
17	course	N	講座	一団	音楽	順番	-0.35	0.35
18	hall	N	会館	大きさ	場所	一員	-1.80	0.29
19	break	N	休憩	心臓	時間	寝台	-3.38	0.34
20	hold	V	持つ	呼ぶ	覚えている	持ってくる	-0.84	0.36
				Frequency Level 2				
1	suppose	V	...だと思う	横たえる	成功する	得る	-0.48	0.52
2	perhaps	ADV	たぶん	かなり	同じように	効果的に	-0.13	0.40
3	bill	N	請求書	社会	場面	巻き尺	-0.68	0.43
4	range	N	範囲	多量	器具	表面	-1.44	0.41
5	general	ADJ	一般的な	最近の	空の	違法な	-1.74	0.37
6	region	N	地域	銀	内容	顧客	-0.64	0.43
7	structure	N	構造	隣人	不足	申し込み	-1.63	0.45
8	control	N	制御	利益	胃袋	供給	-2.90	0.43
9	type	N	型	議論	糸	気体	-3.63	0.42
10	patient	N	患者	技術	利用者	補助	-1.66	0.44
11	pain	N	痛み	ウェブサイト	訴え	解決策	-2.85	0.53
12	consider	V	みなす	乗り越える	保持する	承認する	-0.37	0.38
13	available	ADJ	利用可能な	中央の	日常的な	巨大な	-1.23	0.44
14	recently	ADV	最近	どこかで	新たに	重く	-1.49	0.39
15	issue	N	問題	型	装置	賞	-1.36	0.51
16	term	N	期間	子供	教壇	比較	-1.28	0.47
17	economy	N	経済	印象	義務	十字架	-3.50	0.37
18	political	ADJ	政治的な	完全な	先輩の	とがった	-0.68	0.50
19	process	N	過程	砂糖	たくらみ	経路	-2.59	0.28
20	steal	V	盗む	創造する	押す	破壊する	-2.99	0.39

Frequency Level 3

1	sort	N	種類	苦情	得意先	青少年	-0.64	0.38
2	council	N	評議会	土	生産者	戦略	0.65	0.35
3	analysis	N	分析	小川	存在	表示装置	-0.24	0.43
4	opposition	N	反対	失敗	到着	気づき	-0.44	0.36
5	pursue	V	追う	計算する	交渉する	値する	1.08	0.46
6	significant	ADJ	重要な	即時の	複数の	柔軟な	-0.05	0.42
7	obviously	ADV	明らかに	かなり	多少の	さらに	-0.80	0.40
8	leather	N	革	参加者	評判	ひび割れ	-0.71	0.38
9	comparison	N	比較	4 分の 1	専門知識	導入	-0.20	0.48
10	encounter	N	出会い	在庫	緊張	登録	-0.75	0.44
11	approximately	ADV	おおよそ	正確に	おそらく	わずかに	0.74	0.18
12	stuff	N	もの	統計	契約	特徴	-0.03	0.41
13	seek	V	求める	打ち上げる	拒否する	放棄する	-0.46	0.52
14	academic	ADJ	学問の	ありえない	並外れた	多数の	-2.48	0.28
15	vision	N	視力	準備	上司	管理	-1.36	0.40
16	fault	N	欠点	減少	存在	記述	-1.86	0.46
17	identity	N	身元	編集者	富	先祖	-1.63	0.30
18	submit	V	提出する	説得する	可能にする	満足させる	-0.89	0.36
19	critical	ADJ	批判的な	生の	簡潔な	不要な	-0.48	0.38
20	domestic	ADJ	国内の	恒久的な	不変の	熱心な	-0.44	0.38

Frequency Level 4

1	faith	N	信念	処分	密度	重要産物	-1.71	0.33
2	literally	ADV	文字通りに	おおやけに	おそらく	正式に	0.67	0.49
3	entitle	V	権利を与える	評価する	栽培する	取り押さえる	0.43	0.32
4	prior	ADJ	先の	雪の	原因となる	緊張の多い	1.41	0.33
5	massive	ADJ	巨大な	厳しい	原始的な	段階的な	0.78	0.39
6	angle	N	角度	香草	こんろ	説明書	-0.37	0.34
7	revenue	N	収益	野望	さんご	踊りの一種	-0.18	0.32
8	finance	N	金融	錠剤	農地	探査	-0.73	0.37
9	perceive	V	知覚する	資格を得る	挫折させる	更新する	1.06	0.46
10	corporate	ADJ	法人の	時折の	目立つ	避けられない	-0.22	0.31
11	integration	N	統合	敗北	人工物	財布	-0.09	0.44
12	scheme	N	計画	蟻	小冊子	長所	0.17	0.33
13	accommodation	N	宿泊施設	奇跡	優雅さ	保全	0.69	0.33
14	restrict	V	制限する	描く	専念する	廃止する	0.04	0.51
15	possession	N	所有	化学者	社員	部屋	-0.53	0.31
16	anxiety	N	不安	生息地	方言	慣習	1.01	0.33
17	profile	N	横顔	払い戻し	花	罰	1.28	0.34
18	visible	ADJ	目に見える	鮮明な	楽しい	格好いい	-0.15	0.53
19	estate	N	不動産	山頂	泥	放置	0.54	0.27
20	regime	N	政治制度	航海	調査員	一撃	0.04	0.31

Frequency Level 5

1	stability	N	安定	天文学者	くつろぎ	苦しい選択	0.45	0.33
2	opt	V	選ぶ	方向を定める	劣化する	実行する	2.66	0.30
3	eligible	ADJ	資格のある	部分的な	楽観的な	運用可能な	1.71	0.17
4	potentially	ADV	潜在的に	科学的に	驚くほど	十分に	-1.46	0.39
5	rating	N	格付け	母国	不快感	ロゴマーク	-0.22	0.22
6	hostile	ADJ	敵意のある	助言の	強制的な	磁性体の	0.00	0.45
7	cruel	ADJ	残酷な	連続の	異国風の	無関係の	-0.53	0.45
8	exposure	N	露出	補足	水素	拘留	0.56	0.36
9	descent	N	降下	辺境	同僚	ディーゼル	0.02	0.46
10	listing	N	表の項目	寛容さ	礼儀作法	傾斜面	1.63	0.10
11	collaboration	N	協力	大使	創造性	訴訟	-3.26	0.29
12	accomplishment	N	業績	制約	辞職	在庫	0.30	0.22
13	toll	N	使用料	病気	薬局	布片	0.99	0.19
14	framework	N	枠組み	英国の貨幣	改訂	拒絶	-0.82	0.44
15	tomb	N	墓	包含	分離	余剰	0.74	0.24
16	prayer	N	祈り	地球	窒素	雄羊	-1.13	0.30
17	correspondence	N	一致	崇拝	移植	戦闘	0.62	0.43
18	erect	V	建立する	震える	奪う	改装する	0.47	0.25
19	integral	ADJ	完全な	沖合の	立法の	三重の	0.10	0.15
20	abstract	ADJ	抽象的な	生まれ持った	うぶな	考古学の	-1.13	0.44

Frequency Level 6

1	gorgeous	ADJ	豪華な	軽い	粗悪な	宣伝の	-4.42	0.31
2	hence	ADV	それゆえ	緊急に	生態学的に	上手に	1.60	0.17
3	overview	N	概要	宇宙船	変更	高台	1.26	0.22
4	ancestry	N	祖先	調整	責任	数学者	-1.06	0.43
5	descendent	N	子孫	資本主義	知人	足首	0.85	0.27
6	experimentation	N	実験	黒眼鏡	礼儀	病原体	-1.13	0.39
7	atom	N	原子	菌類	さび	こぶし	-0.39	0.27
8	inclination	N	傾向	投影機	樹脂	幼児期	0.17	0.11
9	liberate	V	解放する	包含する	緩和する	循環させる	1.04	0.20
10	sincere	ADJ	誠実な	議会の	線の	内向きな	-0.64	0.34
11	civilian	ADJ	市民の	楕円形の	同時の	湿った	-1.11	0.47
12	statistical	ADJ	統計の	放射性のある	不人気な	大型の	-0.77	0.45
13	residency	N	居住	指揮官	空中の	警部補	-0.53	0.38
14	discrepancy	N	不一致	合併	後援者	幅	0.04	0.21
15	subscribe	V	加入する	保留する	差別化する	つぶやく	-1.71	0.27
16	restrain	V	抑制する	開始する	理解する	興味をそそる	-0.91	0.31
17	conform	V	従う	暗唱する	退避させる	強要する	0.49	0.32
18	grief	N	嘆き	銃	鉱夫	自然主義者	-0.05	0.41
19	verification	N	確認	停滞	演繹	群れ	1.49	0.10
20	partition	N	分割	座右の銘	酢漬け	新規制	-0.73	0.28

Frequency Level 7

1	oath	N	誓い	演目	社員名簿	合奏	-0.31	0.28
2	warrant	N	保証書	一致	航空宇宙学	福音	-0.50	0.22
3	freight	N	貨物	無知	粘膜	本土	1.88	0.15
4	assertion	N	主張	審問	生検	開示	-0.09	0.20
5	stereotype	N	固定観念	前例	組み立て部品	狼狽	-1.25	0.45
6	complementary	ADJ	補足の	差し迫った	扱いにくい	永遠の	0.90	0.24
7	motorist	N	運転者	秘密の状態	館長	たとえ	2.21	0.06
8	glimpse	N	ちらりと見えること	裏切り	ディスコ	横断幕	0.17	0.35
9	raider	N	侵入者	専売	船長	シナモン	0.99	0.11
10	diplomacy	N	外交	差止命令	農家	通商停止	-0.13	0.26
11	poise	V	均衡を保つ	捨てる	編む	震える	1.11	0.14
12	decorative	ADJ	装飾用の	老化した	絶望的な	戦争の	-0.09	0.41
13	cardiac	ADJ	心臓の	なめらかな	…しがちな	悲惨な	3.19	0.11
14	quantitative	ADJ	量的な	容赦ない	中央の	比較の	0.67	0.35
15	profoundly	ADV	深く	だらしなく	間接的に	などなど	1.41	0.21
16	constituency	N	選挙区	不安	威厳	男爵	1.46	0.03
17	maxim	N	格言	旅団	会計士	保護観察	0.58	0.24
18	contempt	N	軽蔑	共産主義	刺激	国民投票	0.85	0.34
19	stubborn	ADJ	頑固な	実利的な	官僚的な	分子の	0.08	0.33
20	attic	N	屋根裏部屋	推薦	禁輸	親善	2.79	0.07

Frequency Level 8

1	syntax	N	構文	醸造所	売春婦	鷹（タカ）	-0.53	0.23
2	prosecute	V	起訴する	すりつぶす	うろつく	すすぐ	-0.09	0.31
3	universally	ADV	普遍的に	うっかりと	大胆に	ものすごく	-1.18	0.27
4	anonymity	N	匿名性	ほうき	ビリヤードの一種	大広間	0.25	0.23
5	womb	N	子宮	ねたみ	振動	調理器	2.01	0.14
6	sane	ADJ	正気の	刑事上の	新品同様の	膵臓（すいぞう）の	-0.64	0.27
7	suicidal	ADJ	自殺の	長期にわたる	破壊的な	内視鏡の	0.06	0.49
8	anthem	N	賛歌	速さ	優雅さ	大量殺人	0.78	0.20
9	persuasion	N	説得	表向きの人格	変わりやすさ	鼓動	0.27	0.32
10	perennial	ADJ	絶え間ない	同質な	超自然的な	蛍光性の	2.01	0.05
11	consequent	ADJ	結果として起こる	博士の	道に迷った	回帰性の	-0.64	0.30
12	preoccupation	N	没頭	男性の敬称	飢え	放浪癖のある人	0.62	0.22
13	dread	N	恐怖	降格	逸脱	ミッドフィルダー	0.97	0.23
14	decree	N	法令	親和性	破滅	寄生虫	2.24	0.02
15	luxurious	ADJ	贅沢な	不規則な	ずるい	暗示して	-1.06	0.41
16	omission	N	省略	宝石	共鳴	報復	1.06	0.18
17	flux	N	移り変わり	塗装	所在	危険	1.46	0.17
18	sermon	N	説教	懸念	輝き	巡礼	0.38	0.32
19	chore	N	雑用	かつら	破片	前景	2.28	0.13
20	hierarchal	ADJ	階層性の	貴族の	武闘派の	専売の	0.94	0.31

Note: Gram = grammar. Dif = item difficulty. Dis = item discriminability.

ACKNOWLEDGEMENTS

This research was supported by JSPS KAKENHI Grant Number 25K04324

REFERENCES

Aizawa, K. (1998). Developing a vocabulary size test for Japanese EFL learners. *Annual Review of English Language Education in Japan*, *9*, 75–82.

Aizawa, K., Iso, T., & Nadasdy, P. (2017). Developing a vocabulary size test measuring two aspects of receptive vocabulary knowledge: Visual versus aural. In K. Borthwick, L. Bradley, & S. Thouësny (Eds.), *CALL in a climate of change: Adapting to turbulent global conditions – short papers from EUROCALL 2017* (pp. 1–6). Research-publishing.net. https://doi.org/10.14705/rpnet.2017.eurocall2017.679

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing, 27*(1), 101–118.

Browne, C. (2013). The new general service list: Celebrating 60 years of vocabulary learning. *The Language Teacher, 37*(4), 13–16.

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Chujo, K., & Oghigian, K. (2009). How many words do you need to know to understand TOEIC, TOEFL & EIKEN? An examination of text coverage and high frequency vocabulary. *Journal of Asia TEFL*, *6*(2) 121–148**.**

Committee of Revising the JACET Basic Words. (Ed.). (2016). *The new JACET list of 8000 basic words*. Kirihara Shoten.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

Dizon, G. (2021). Subscription video streaming for informal foreign language learning: Japanese EFL students' practices and perceptions. *TESOL Journal, 12*(1), e566. https://doi.org/10.1002/tesj.566

Fountain, R.L., & Nation, I.S.P. (2000). A vocabulary-based graded dictation test. *RELC Journal*, *31*(2), 29–44. https://doi.org/10.1177/003368820003100202

Hamada, A., Iso, T., Kojima, M., Aizawa, K., Hoshino, Y., Sato, K., Sato, R., Chujo, J., & Yamauchi, Y. (2021). Development of a vocabulary size test for Japanese EFL learners using the New JACET List of 8,000 Basic Words. *JACET Journal, 65*, 23–45. https://doi.org/10.32234/jacetjournal.65.0_23

Ishikawa, S. (2019). A reconsideration of the construct of "a vocabulary for Japanese learners of English": A critical comparison of the JACET wordlists and new general service lists. *Vocabulary Learning and Instruction, 8*(1), 1–7. https://doi.org/10.7820/vli.v08.1.ishikawa

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

McLean, S., Kramer, B., & Stewart, J. (2015). An empirical examination of the effect of guessing on vocabulary size test scores. *Vocabulary Learning and Instruction, 4*(1), 26–35. https://doi.org/10.7820/vli.v04.1.mclean.et.al

Meara, P. (1992). *EFL vocabulary tests*. University College Swansea.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Matters.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *The Canadian Modern Language Review, 63*, 127–147.

Mizumoto, A., & Shimamoto, T. (2008). A comparison of aural and written vocabulary size of Japanese EFL university learners. *Language Education & Technology, 45,* 35–51.

Mochida, A., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing, 23*(1), 73–98.

Mochizuki, M. (1998). Nihonjin eigo gakushusha no tameno goi saizu tesuto [A vocabulary size test for Japanese learners of English]. *Institute for Research in Language Teaching Bulletin, 12*, 27–53.

Nadasdy, P. (2026). Investigating the impact of written-form learning on aural vocabulary recognition in Japanese EFL learners. *Hawaii International Conference on Education 2026 Conference Proceedings*, 218–235.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*(1), 12–25.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Heinle and Heinle.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

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

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10*(3), 355–371.

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Sonoda, K. (1996). *Daigakusei-yō Eigo goihyō no tame no kisoteki kenkyū [A basic study for creating an English vocabulary list for university students].* Institute of Language and Culture Studies, Hokkaido University.

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition, 31*(4), 577–607.

Stewart, J., & Stewart, W. (2016). Using the JACET 8000 word list to evaluate L2 vocabulary and reading difficulty. *Journal of Chikushi Jogakuen University, 11*, 57–66.

Stubbe, R. (2014). The effects of pseudowords in yes/no vocabulary tests for low-level learners. *Language Testing, 31*(1), 15–35.

Stubbe, R., Stewart, J., & Pritchard, T. (2010). Examining the effects of pseudowords in yes/no vocabulary tests for low level learners. *Kyushu Sangyo University Language Education and Research Center Journal, 5*, 5–23.

Thompson, I. (2001). *Japanese speakers*. In M. Swan & B. Smith (Eds.), *Learner English: A teacher's guide to interference and other problems* (2nd ed., pp. 296–309). Cambridge University Press.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. Teachers College, Columbia University.

Webb, S., & Nation, P. (2017). How vocabulary is learned. Oxford University Press.

West, M. (1953). *A general service list of English words.* Longman, Green.

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *The Canadian Modern Language Review, 53*(1), 13–40.

Yamauchi, H. (2005). Web-based vocabulary test for Japanese learners. *CALL-EJ Online, 7*(2), 1–16.