JACET Journal 65 (2021) 23–45

Development of a Vocabulary Size Test for Japanese EFL Learners Using the New JACET List of 8,000 Basic Words

HAMADA, Akira
Meikai University

ISO, Tatsuo
Tokyo Denki University

KOJIMA, Masumi
Gifu City Women's College

AIZAWA, Kazumi
Tokyo Denki University

HOSHINO, Yuko
Chiba University

SATO, Kento
Tokyo Denki University

SATO, Ryoko
Reitaku University

CHUJO, Junko
Tokyo Denki University

YAMAUCHI, Yutaka
Soka University

Abstract

This study developed a vocabulary size test for Japanese learners of English as a foreign language (EFL) using the New JACET List of 8,000 Basic Words (VST-NJ8). Data from the vocabulary size tests were interpreted to assess, diagnose, and predict learners' lexical proficiency. Yet, existing tests for Japanese EFL learners involve two critical issues regarding unrepresentative contents and inflated estimation of vocabulary size due to methodological restrictions. For these reasons, a large-scale item bank was created in Study 1 by sampling 800 target words and 2,400 distractors from the New JACET List of 8,000 Basic Words. The data from 2,155 test-takers were analyzed using the three-parameter logistic (3PL) model of the item response theory to obtain the invariant parameters of difficulty, discriminability, and pseudo-guessing probability for every item. In Study 2, an automatic test assembly was conducted in consideration of the number of items to be included, the distribution of parts of speech, and the test information functions estimated by the 3PL model. The item characteristics were argued for validation in terms of the content, substantive, structural, generalizability, and external aspects of construct validity. The administration, scoring, and interpretation methods of the VST-NJ8 are fully provided.

Keywords: test development, vocabulary size, educational word list,
item response theory, automatic test assembly

Acquiring a vast number of words is an essential factor in promoting second/foreign language (L2) learning. Nation (2013) suggests evaluating the achievement levels of vocabulary teaching goals regarding how much vocabulary is learned. A variety of vocabulary size tests (VSTs) have been developed to assess learners' breadth of vocabulary knowledge. For example, Mochizuki's (1998) VST has widely been employed for Japanese learners of English as a foreign language (EFL). The test scores are used to estimate learners' vocabulary size, to predict what types of real-life activities they can perform (Koizumi & Mochizuki, 2011), and to "chart the growth of learners' vocabularies" (Nation & Beglar, 2007, p. 9).

However, the VSTs developed in the relatively distant past entail two critical concerns. First, Mochizuki's VST is not likely to represent the current status of English education in Japan because the word list used for target-word sampling is old. A discrepancy between test contents and target constructs reduces the construct validity of the test (Messick, 1995). Second, VSTs adopt raw scores to estimate the number of words that test-takers know. Since the raw scores are not informative about item difficulty, item discriminability, and effects of using guessing strategies (van der Linden, 2016), the vocabulary size estimated is likely to be inflated (e.g., Gyllstad, Vilkaitė, & Schmitt, 2015). Given the clear importance of a validated assessment of L2 vocabulary size, the present study aimed to develop a new VST for Japanese EFL learners by solving these two issues.

Literature Review

Vocabulary Size in a Second Language

Vocabulary knowledge is one of the most important factors in successful L2 learning. Among the various aspects of vocabulary knowledge, a receptive vocabulary size has been reported to be highly correlated with receptive language skills such as reading and listening (e.g., Alderson, 2005; Hamada, 2011, 2014, 2015; Laufer, 1992; Laufer & Ravenhorst-Kalovski, 2010; Zhang & Zhang, 2020). This suggests that a large vocabulary size is necessary in comprehending messages in target languages (Hamada, 2017). For example, Laufer (1992) revealed strong correlations between two types of VSTs and reading comprehension test scores developed in her research. Vocabulary size was also found to be strongly associated with listening comprehension (Stæhr, 2009). More recently, a research synthesis by Zhang and Zhang (2020) showed that vocabulary knowledge (meaning recognition) had robust correlations with reading comprehension (k = 14, r = .53, 95% CI [.49, .57]) and listening comprehension (k = 22, r = .50, 95% CI [.41, .58]). These findings led to the studies on predicting L2 learners' abilities based on their receptive vocabulary sizes (Schmitt, Nation, & Kremmel, 2020).

The other approach to the importance of vocabulary size is to examine how many words learners should know to comprehend L2 messages. For example, Hu and Nation (2000) and Laufer (1992) suggested that 95%–98% of running words in a given text are necessary to be known in reading comprehension, although the decisive threshold level is still controversial (Schmitt, Jiang, & Grabe, 2011). Similarly, van Zeeland and Schmitt (2013) showed that knowing 98% of words is necessary to accurately understand texts in listening comprehension. According to corpus-based research, text coverage of 95%–98% words indicates that L2 learners should acquire vocabulary ranging from 4,000–5,000 to 8,000-word families (Nation, 2013; Schmitt, 2010). Milton (2010) also stated that 2,000-word families are necessary to achieve an A2 level in the Common European Framework of Reference for Languages, 3,000-word families to reach B1 to B2 levels, and 5,000-word families to achieve C1 to C2 levels. Considering these achievement goals in L2 vocabulary instruction and learning, measuring vocabulary size helps both teachers and learners aim at their future objectives and plan their teaching and learning.

A variety of VSTs have been developed to measure L2 learners' breadth of vocabulary knowledge (see Schmitt, 2010, for review). VSTs have been used to estimate the number of words learners know, diagnose what frequency range of vocabulary learners should acquire, and as placement and proficiency tests (e.g., Alderson, 2005; Nation, 2013; Read, 2000). In the next section, we will overview the major types of VSTs developed so far from the perspectives of their specifications, including the sampling methods of the test items, test formats, and validation procedures.

Development and Validation of Vocabulary Size Tests

The first widely used VST for L2 learners was Nation's (1983) Vocabulary Levels Test (VLT). Using a helpful thumbnail sketch offered by the VLT, two other representative VSTs were developed, including the Eurocentres Vocabulary Size Test (a form of lexical judgment called the Yes/No test; Meara & Buxton, 1987) and the Vocabulary Size Test (Nation & Beglar, 2007). In Japan, the VSTs developed by Aizawa (1998) and modified by Mochizuki (1998) are widely used for Japanese EFL learners. Although the examinees, test items, and test formats differ from each other, they commonly considered the following features in their test development: (a) sampling test items from frequency-based word lists, (b) creating test formats suitable for vocabulary size estimation, and (c) verifying the construct validity of the tests.

Sampling methods of test items. The VST development concerns that it is practically impossible to assess the knowledge of every word to be evaluated (e.g., Gyllstad et al., 2015; Schmitt & Schmitt, 2012). To balance between test practicality and representativeness, previous studies have referred to frequency-based word lists compiled from large-scale corpora. Since the words randomly sampled from each frequency level (i.e., stratified random sampling) can be considered to represent a population of the words listed (Schmitt, 2010), the VSTs are able to estimate the knowledge of the entire vocabulary. For example, Nation (1983) used the frequency information from the General Service List (West, 1953) to sample test items for 2,000, 3,000, 5,000, and 10,000 word-frequency levels. The University Word List (Xue & Nation, 1984) was also adopted to include the words necessary for academic situations at universities. These two lists provide frequency information based on a word family or the base form of a word and its inflected and derived forms made with affixes (Nation, 2013). By contrast, the most widely used VSTs in Japan were developed using lemma-based frequency information from the (revised) Hokkaido University Learning Vocabulary List (Sonoda, 1996; see also Aizawa, 1998; Mochizuki, 1998; Koizumi & Mochizuki, 2011) and the Standard Vocabulary List 12,000 (ALC, 2001; see also Mizumoto, 2006). The underlying rationale is that the knowledge of base forms cannot necessarily guarantee the knowledge of those derived forms because Japanese EFL learners with low proficiency have limited knowledge of affixes (Mochizuki & Aizawa, 2000).

The vocabulary knowledge assessed by the VSTs is reflective of the nature of the word lists used for the sampling of test items (Schmitt & Schmitt, 2012); therefore, both Meara's and Nation's tests were modified as the word frequency information of the source corpora was updated (Nation, 2013). However, the VSTs for Japanese EFL learners have not changed the word lists, which no longer reflect the current status of English education in Japan. The Hokkaido University Learning Vocabulary List was mainly composed of the vocabulary used in 1980s' and 1990s' authorized English textbooks and dictionaries (Sonoda, 1996). The state-of-the-art educational word list for Japanese EFL learners is the New JACET List of 8,000 Basic Words (Committee of Revising the JACET Basic Words, 2016). The inclusion criteria for this list were that the words were necessary for Japanese EFL learners to (a) communicate enough to prevent misunderstanding in everyday life, (b) make presentations and write reports and articles for academic purposes, and (c) be assessed by standardized English proficiency tests. As a result of the reference corpora being compiled based on updated source corpora and educational materials, 947 words were replaced from the previous version (Aizawa et al., 2005) and 35% and 26% of previously listed words were moved to the higher and lower word-frequency levels, respectively. Because these modifications reflect the current use of English in the educational settings of Japan (see Committee of Revising the JACET Basic Words, 2016, for review), it is rational to adopt the New JACET List of 8,000 Basic Words when estimating Japanese EFL learners' vocabulary size.

Creating test formats. The formats of VSTs are roughly categorized into either Yes/No or multiple-choice formats. In the Eurocentres VST, test-takers give Yes/No responses to each item to indicate whether it is a real or pseudo word (Meara & Buxton, 1987). The score is calculated with four patterns of Yes/No responses to each item using the stimulus detection theory. However, the limitation of the Yes/No format is that the score does not necessarily reflect the knowledge of word meanings (Read, 2000).

Nation's type of VSTs adopts a multiple-choice format, although the way to present test items differs among the existing tests. In the VLT, test-takers choose three answers from six options per item (Nation, 1983). A similar format was applied by Aizawa (1998) and Mochizuki (1998), in which test-takers choose two answers from six options. In contrast, recent VSTs employed the format of one answer for each question to avoid violating the assumption of local independence among test items (Koizumi & Mochizuki, 2011; Mizumoto, 2006; Nation & Beglar, 2007). This allowed them to analyze item characteristics based on the item response theory (IRT; Beglar, 2010). The critical issue concerning the multiple-choice format is the measurement error caused by a pseudo-guessing strategy, resulting in an inflation of the estimated vocabulary size (Gyllstad et al., 2015; Stewart, 2014). To solve this, Gibson and Stewart (2014) and Tseng (2013) suggested the practical use of IRT to calibrate the expected probability of correct responses based on the invariant parameters of test items.

Regarding the VSTs for Japanese EFL learners, it is also necessary to consider the effects of a first language (L1) used for definitions of item stems on vocabulary size estimation. Mochizuki's types of VSTs require test-takers to choose one of the English word forms that correspond to its Japanese definition (Aizawa, 1998; Koizumi & Mochizuki, 2011; Mizumoto, 2006; Mochizuki, 1998). This is different from Nation's VLT and VST because they ask test-takers to look at a target word form and choose the corresponding English definition from multiple choices (usually easier words than the target word). Elgort (2013) showed that response accuracy to individual test items was approximately 10% higher when the meanings of options were presented in an L1, particularly for low-proficiency learners. Meanwhile, Nation (2013) supported the bilingual version of the VSTs because the L1 definitions avoided the difficult grammar of English definitions, and they were immediately comprehensible to the test-takers.

Verifying the test validity. For the construct validity of the VSTs, early research depended on providing convergent evidence in terms of significant associations between VST scores and other test batteries (e.g., Aizawa, 1998; Mochida & Harrington, 2006; Mochizuki, 1998). Yet, theories in language testing suggested providing multiple sources of evidence for construct validity (Messick, 1995) in the process of making a validity argument (Kane, 2006). For example, Koizumi and Mochizuki (2011) validated the computer-based Mochizuki's VST within Messick's (1995) validation framework (i.e., see details in Overview of This Study). More recently, Schmitt et al. (2020) specified the validity argument for VST development. The validation begins with the domain definition by explaining rationale of the test design and purpose. Then, evidence is presented to ensure the appropriateness of the scoring procedure. A generalizability aspect of the test is further argued in terms of whether the scores are reflective of the test-takers' vocabulary knowledge. For extrapolation, it is important to examine the relationship between the VST scores and other skills that are closely related to the vocabulary knowledge measured in the test.

A statistical method used in recent studies is the latent trait models. Beglar (2010) argues for the validation of Nation's VST using the Rasch model. The results showed that the item difficulty was stratified into seven and more groups, and 10 items in each frequency level should be sufficient for vocabulary size estimation because the items whose difficulty was similar to each other were included in almost all frequency levels. Koizumi and Mochizuki (2011) also applied the Rasch model to the response data of Mochizuki's VST and indicated that the measurement invariance was ensured because there were few items that did not fit the Rasch model, the item difficulty increased as the word frequency became low, and the test scores were highly correlated with test-takers' English proficiency defined by TOEIC® Listening & Reading Test. Other studies employed the IRT as well to verify the item characteristics of their newly developed VSTs (e.g., Tseng, 2013) and vocabulary size estimation (e.g., Gibson & Stewart, 2014).

Overview of This Study

The development and validation of VSTs are continuously necessary to assess L2 learners' vocabulary knowledge and predict their L2 competence. Following Messick's (1995) framework of validity of psychological assessment, the present study conducted two studies to show multiple sources of evidence in construct validity of the new VST for Japanese EFL learners (hereafter, VST-NJ8). Study 1 was designed to validate the content and substantive aspects of construct validity through the creation of a large-scale item bank. As discussed earlier, the New JACET 8,000 Basic Word List was employed to the stratified random sampling in order to increase content relevance and representativeness. The substantive aspect refers to the theoretical mental processes being actually observed in test responses, and accordingly, the present study adopted the L1-L2 matching format to examine learners' receptive knowledge of form-meaning connections (see, e.g., Nation, 2013; Schmitt, 2010; Zhang & Zhang, 2020, for review).

In Study 2, this study examined the following five research questions (RQs) concerning the structural, generalizability, and external aspects of the construct validity.

Structural aspects:

RQ1. Is the distribution of item difficulty wide enough to measure a wide range of test-takers' latent trait and correspond to the word-frequency levels without gaps?

RQ2. Does the test score reflect a single construct to be measured?

Generalizability aspects:

RQ3. Is measurement invariance ensured in terms of item difficulty, item discriminability, and pseudo-guessing probability?

RQ4. Does the item information for each frequency level cover test-takers' latent trait to be measured?

External aspect:

RQ5. Is the estimated vocabulary size correlated with a proficiency test?

These three aspects were investigated using the IRT. Although the vocabulary size estimation has been biased according to sampled test items, test reliability, and learners' guessing strategy (e.g., Gyllstad et al., 2015; Stewart, 2014; Tseng, 2013), the IRT can solve the problems by recognizing each response as the outcome of a distinct probability estimated from separate parameters for the item and test-taker effects (van der Linden, 2016). Particularly, using the three-parameter logistic (3PL) model, we can identify the probability of person j providing a positive answer to dichotomous item i as

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + exp\{-1.702a_i(\theta_j - b_i)\}} \tag{1}$$

where θ_j represents the latent trait of person j, and a_i, b_i, and c_i represent the discriminability, difficulty, and pseudo-guessing probability of item i, respectively (Kato, Yamada, & Kawabata, 2014). By considering the goodness of the 3PL model, the consistencies between the test scores and the internal structure of the test (i.e., the structural aspect) were investigated in terms of (a) the invariability of item difficulty, item discriminability, and pseudo-guessing probability; (b) distribution of item difficulty (spreading widely without gaps); and (c) unidimensionality. The generalizability aspect was further checked by item information functions estimated by the 3PL model. Finally, the external aspect was argued based on the relationships between the estimated vocabulary size and a proficiency test score.

Study 1: Item Bank Construction

Design

The purpose of this item bank construction was to pick out the items with the following features: (a) item difficulty was in line with eight word-frequency levels, (b) item discriminability was high enough to differentiate levels of test-takers' latent trait, and (c) pseudo-guessing probability was lower than chance level (25%). In Study 1, a matrix sampling design was applied to estimate these three parameters for the items sampled from the New JACET 8,000 Basic Word List. In this design, different subsets of test items were assigned to different groups of participants. Every subset included "common" items assigned to two or more subgroups of the participants as well as "unique" items that were only administered to one particular subgroup. The matrix sampling design allowed for obtaining information about the characteristics of large numbers of test items (see Kato et al., 2014, for review). Specifically, the present study included 800 target items (and 2,400 distractors) in the item bank. This sampling ratio was much higher than that of previous VSTs for Japanese EFL learners (Aizawa, 1998; Mizumoto, 2006; Mochizuki, 1998), which can support the inference that the sampled items were the representatives of the target construct (i.e., the content aspect of construct validity).

Participants

Test-takers for the VST-NJ8 included 2,189 undergraduate and graduate students from 16 Japanese universities (29 faculties) located in urban, suburban, and rural areas. Data obtained from 2,155 participants were analyzed (age: *M* = 19.27, *SD* = 2.06), excluding 34 participants due to data recording errors or disagreement on data use. As for participants' gender, 62% were males (*n* = 1,340), 33% were females (*n* = 712), and 5% were nonreported (*n* = 103). Table 1 summarizes their self-reported TOEIC® Listening & Reading Test scores and Eiken grades.

Table 1
Descriptive and Demographic Statistics of Test-Takers' TOEIC® Scores and Eiken Grades

TOEIC® (descriptive statistics)		Eiken (demographic statistics)	
n	1,298	1st	10
M	504.94	Pre-1st	19
95% CI	[497.94, 511.93]	2nd	467
SD	128.50	Pre-2nd	333
Min	100	3rd	315
Mdn	510	4th	30
Max	990	5th	15

Item Bank

Sampling of target items. The item bank was constructed through stratified random sampling of every 100 words from Levels 1–8 of the New JACET 8,000 Basic Word List. Selection criteria of the target items and underlying rationales were as follows:

- Only the content words (i.e., nouns, verbs, adjectives, and adverbs) were included in the same way as the existing VSTs (see Mochizuki, 1998), although three ordinal number words (i.e., *first*, *second*, and *third*) were the candidates as Mochizuki's VST included them.
- The distribution ratio of nouns, verbs, adjectives, and adverbs were the same as per each frequency level of the New JACET 8,000 Basic Word List.
- Because the VST-NJ8 was designed as the lemma-based vocabulary size test, (a) when the base and inflectional forms of target items were sampled simultaneously, only the base forms

were adopted to the item bank; and (b) when the base and derivational forms of target items were sampled simultaneously, both forms were included, but either of the two forms were presented to the identical test-takers.

- When Japanese katakana notation was conventionally used to express the meanings of the target items, particularly corresponding to those English pronunciation, they were excluded from the item bank (see Laufer & McLean, 2016).

Table 2 shows the number of nouns, verbs, adjectives, and adverbs sampled per frequency level. Since the New JACET 8,000 Basic Word List did not contain any semantic information about the entries, the meanings of the target items sampled were provided according to the dominant meanings as listed in the previous version of the JACET List of 8,000 Basic Words (Aizawa et al., 2005). When the target words were not listed, Taishukan's *Unabridged Genius English–Japanese Dictionary* was referred to in the same way. In cases where multiple meanings for the items were registered in the dictionary, the most typically used ones were determined through discussion among the authors.

Table 2
Number of Nouns, Verbs, Adjectives, and Adverbs Sampled per Frequency Level

Parts of speech	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8
Noun	50	60	55	60	60	55	65	60
Verb	20	15	15	15	10	20	5	5
Adjective	20	15	20	20	25	20	25	30
Adverb	10	10	10	5	5	5	5	5

Construction of distractors. Three distractors per target item were randomly sampled from the New JACET 8,000 Basic Word List (i.e., 2,400 words). Note that the distractors were selected from the same frequency level of the target items to avoid salience effects on response and guessing strategies. Moreover, the parts of speech of the test items and distractors were all the same because the VST-NJ8 did not target learners' lexico-grammatical knowledge to differentiate the grammatical nature of parts of speech.

Versions of test booklet. For the matrix sampling of the response data, this study divided 800 target items into 100 and created 10 separate booklets. Of the 100 target items, 60 were unique items that were only assigned to a particular subgroup of test-takers. The other 20 and 20 target items were common items further assigned to the other two subgroups of test-takers. The common items functioned as anchors to measure all the test-takers' latent trait of vocabulary knowledge. Accordingly, they were selected ranging from Levels 2–5 to prevent unreasonably easy and difficult words from being concentrated in a particular booklet. Moreover, the frequency levels in each booklet were manually adjusted to make the perceived difficulty approximately the same among 10 separate booklets. The assignment of each booklet and the presentation order of target items were random per participant, which ensured the stable estimation of the IRT parameters (Kato et al., 2014). Table 3 displays the details of the matrix sampling design.

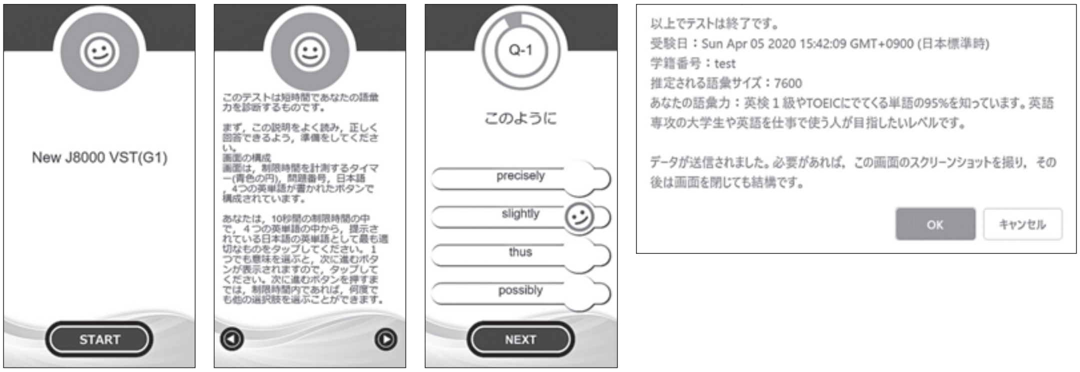
Table 3
The Matrix Sampling Design of Response Data

	Common	Unique			Common
Booklet	Block 1	Block 2	Block 3	Block 4	Block 5
Book 1	Level 2-a	Level 3-a	Level 5-a	Level 7-a	Level 3-b
Book 2	Level 3-b	Level 2-b	Level 6-a	Level 8-a	Level 4-a
Book 3	Level 4-a	Level 1-a	Level 2-c	Level 7-b	Level 5-b
Book 4	Level 5-b	Level 6-b	Level 2-d	Level 8-b	Level 3-c
Book 5	Level 3-c	Level 1-b	Level 5-c	Level 8-c	Level 4-b
Book 6	Level 4-b	Level 7-c	Level 6-c	Level 1-b	Level 5-c
Book 7	Level 5-c	Level 4-c	Level 7-d	Level 1-c	Level 3-d
Book 8	Level 3-d	Level 8-d	Level 2-e	Level 5-d	Level 4-d
Book 9	Level 4-d	Level 6-d	Level 7-e	Level 1-e	Level 3-e
Book 10	Level 3-e	Level 4-e	Level 6-e	Level 8-e	Level 2-a

Note. The target words in every frequency level were divided into five equal blocks.

Procedure

The survey was conducted individually or in regular English classes from September 2019 to December 2019. It involved informed consent, demographic information reporting (university, faculty, age, sex, TOEIC® score, and Eiken grade), five practice questions of the test, and the main test (15–20 minutes long). As Figure 1 shows, the participants took the survey online through an interactive application developed by the authors. In the survey, a test administrator read information on how to use the interactive application and answer the VST-NJ8 to the participants. They were instructed to choose the most appropriate English word that corresponded to the Japanese equivalent within 10 seconds per item. When the test was over, individualized feedback on the test results (i.e., estimated vocabulary size, estimated TOEIC® score and Eiken grade, and recommended vocabulary learning strategies) were given to each participant automatically.



このテストは短時間であなたの語彙力を診断するものです。

まず，この説明をよく読み，正しく回答できるよう，準備をしてください。
画面の構成
画面は，制限時間を計測するタイマー(青色の円)，問題番号，日本語，4つの英単語が書かれたボタンで構成されています。

あなたは，10秒間の制限時間の中で，4つの英単語の中から，提示されている日本語の英単語として最も適切なものをタップしてください。1つでも意味を選ぶと，次に進むボタンが表示されますので，タップしてください。次に進むボタンを押すまでは，制限時間内であれば，何度でも他の選択肢を選ぶことができます。

このように

precisely
slightly
thus
possibly

以上でテストは終了です。
受験日：Sun Apr 05 2020 15:42:09 GMT+0900 (日本標準時)
学籍番号：test
推定される語彙サイズ：7600
あなたの語彙力：英検１級やTOEICにでてくる単語の95%を知っています。英語専攻の大学生や英語を仕事で使う人が目指したいレベルです。

データが送信されました。必要があれば，この画面のスクリーンショットを撮り，その後は画面を閉じても結構です。

Figure 1. The interactive application of the VST-NJ8: The examples of separate booklets, directions, questions, and individualized feedback.

Results and Discussion

Goodness of fit to the 3PL model. The 3PL IRT analysis was conducted for the response data with the irtoys (Partchev, 2017) and mirt (Chalmers, 2012) packages for R. In IRT modeling, it is first necessary to confirm how well the latent trait model fit to both the observed response patterns (i.e., person fit) and the item characteristics (i.e., item fit). First, the goodness of person fit to the 3PL model was examined in terms of the discrepancy between the observed response patterns and the response patterns expected under the 3PL model. The result showed that 47 out of 2,155 (2.18%) participants' response patterns did not fit to the 3PL model. Given that there is inevitably approximately a 5% misfit to the IRT model (Kato et al., 2014), it is highly possible that the 3PL model fit to the observed response data well.

Second, the goodness of item fit was examined with respect to the discrepancy between the observed response patterns and the item characteristic curves expected under the 3PL model. Before that, 28 items were excluded from the analysis because their correct response rates were 100% and not useful to evaluate the item characteristics with a maximum likelihood method. Statistical testing showed significant differences between observed and expected probabilities of correct responses in 83 out of 772 (11%) target items. Specifically, two misfit items were found in Level 1, 11 in Level 2, 11 in Level 3, 11 in Level 4, 17 in Level 5, 15 in Level 6, nine in Level 7, and seven in Level 8. These target items were omitted from the item bank to ensure the measurement invariance of the VST-NJ8.

Item difficulty, discriminability, and pseudo-guessing probability. Figure 2 shows the distributions and outliers of item difficulty, discriminability, and pseudo-guessing probability estimated by the 3PL model. First, the average difficulty increased from Levels 1–3. However, the differences in the item difficulty were not obvious between Levels 4–8. The boxplot also indicates the high overlap of the item difficulty between the adjacent frequency levels. To pick out the target items that should represent the difficulty of each frequency level, we tagged the following items as candidates for exclusion: (a) conventional outliers beyond 1.5 inter-quantile ranges per frequency level and (b) items whose difficulty was still much higher/lower than the average difficulty of the adjacent high/low frequency levels after outlier elimination.

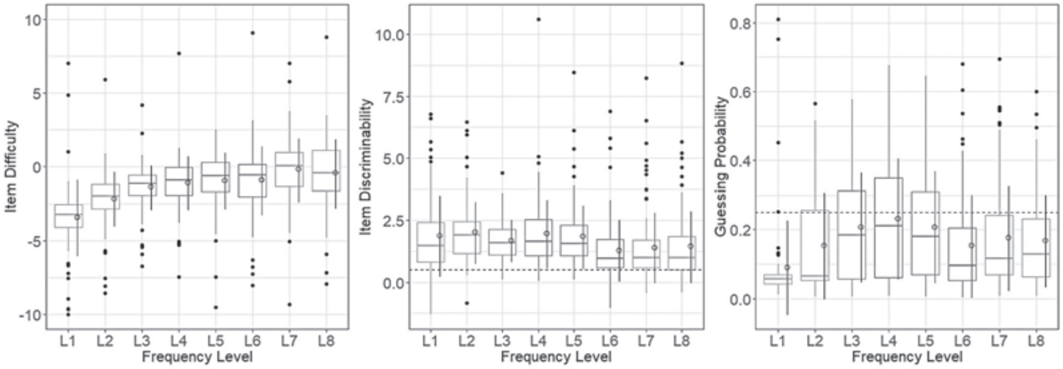


Figure 2. Boxplots for the distributions of item difficulty, item discriminability, and pseudo-guessing probability with bars of means and ±1 standard deviations in the item bank. Black dots represent conventional outliers beyond 1.5 inter-quantile ranges.

The means of item discriminability were higher than 1.0 (a reference point of the parameter to be discussed; Beglar, 2010) in every frequency level (see the dotted line of Figure 2). However, the discriminability of some items was relatively low, as shown in the whiskers of the boxplot. Although the mean probabilities of pseudo-guessing were lower than the chance level (25%) per frequency level (see the dotted line of Figure 2), the boxplot also showed the high probabilities of pseudo-guessing for some items. These defect items reduced the item information related to the measurement accuracy (e.g., van der Linden, 2016). Therefore, this study tagged the target items whose discriminability parameter was below 0.5 and pseudo-guessing probability was higher than the chance level as candidates for exclusion while considering the balance between the sum of item information and the number of items included in the finalized version of the VST-NJ8 (see details in Study 2).

The IRT-based item bank allowed for assembling a set of test items that maximize as much test information (i.e., sum of each item information) as possible. Under the 3PL model, the item information function for item i is defined as

$$I_i(\theta) = \frac{1.702^2 a_i^2 (1 - P_{i\theta})(P_{i\theta} - c_i)^2}{P_{i\theta}(1 - c_i)^2} \tag{2}$$

where $P_{i\theta}$ = 1 / [1 + $exp\{-1.702a_i\ (\theta - b_i)\}$] (Kato et al., 2014). In Study 2, we first assembled the best set of the target items based on formula (2). Then, the structural and generalizability aspects of construct validity of the finalized version of the VST-NJ8 were verified. Finally, the appropriateness of score interpretation was argued from the external aspect or to what extent the test score reflects Japanese EFL learners' vocabulary size.

Study 2: Test Assembly

Design

Automatic test assembly was conducted with the gipkAPI package for R (Gelius-Dietrich, 2020). This method can find the best combinations of target items from the possible candidates within predetermined test specifications. The assembled test set was further evaluated with respect to whether (RQ1) the difficulty of target items was spread widely without gaps and was in line with the predicted order; (RQ2) the test scores reflected a single construct; (RQ3) item difficulty, item discriminability, and pseudo-guessing probability were invariant; (RQ4) the item information per frequency level covered test-takers' latent trait; and (RQ5) the test scores correlated strongly to the proficiency test.

Automatic Test Assembly

To conduct the automatic test assembly, this study determined the test specifications of the VST-NJ8 and underlying rationales as follows:

- Each frequency level included 20 target items (a total of 160 items) to balance between the estimation accuracy of test-takers' latent trait and test practicability. Since the time limit for each item, frequency level, and the entire test was 10 seconds, 200 seconds, and 26 minutes and 40 seconds, respectively, these were not much different from the existing VSTs.
- The distribution ratios of parts of speech were kept as they were in Study 1 (see Table 2).
- Every item was from different word families among each other. Thus, certain word forms

were made not to be a cue to answer other target items.

- The items that could maximize test information ranging from −2 to 2 in the test-takers' latent trait (θ) were selected.

Using 0–1 integer programming, the automatic test assembly was run to find the most suitable combinations of the target items that satisfied these conditions. For more details on the algorithm of the automatic test assembly, see, e.g., Kato et al. (2014). After the automatic test assembly, any target items and distractors whose meanings seemed educationally inappropriate were manually replaced with similar ones with respect to their item characteristics. Supplementary materials displayed the target items, distractors, and three parameters of item characteristics.

Results and Discussion

Structural aspects of the construct validity. Figure 3 and Table 4 show the distributions of item difficulty, item discriminability, and pseudo-guessing probability for the finalized version of the VST-NJ8. In response to RQ1, the result of the automatic test assembly showed that the item difficulty spread widely from −4.0 to 3.0 logits without gaps (see the scatter dots). Compared to the entire items in the item bank (see Figure 2), the average item difficulty monotonically increased, and a one-way analysis of variance with Holm's adjustment showed significant differences between the adjacent frequency levels (ps < .05). The items whose discriminability parameter was below 0.5 logits were not included, and the means ranged from 1.94 to 2.87 logits. Because the discriminability parameter of items is statistically fixed to 1.0 logit in the Rasch model (Kato et al., 2014; van der Linden, 2016), the VST-NJ8 can more accurately differentiate learners' latent trait of vocabulary knowledge than the existing Rasch-based VST (e.g., Beglar, 2010; Koizumi & Mochizuki, 2011; Mizumoto, 2006). Finally, the average pseudo-guessing probabilities were all less than the chance level of 25%. Although those of some items exceeded the chance level in Levels 4, 7, and 8, these were included in the finalized version of the VST-NJ8 in consideration of the overall balance of the test information. These results ensure a part of the structural aspects of construct validity.

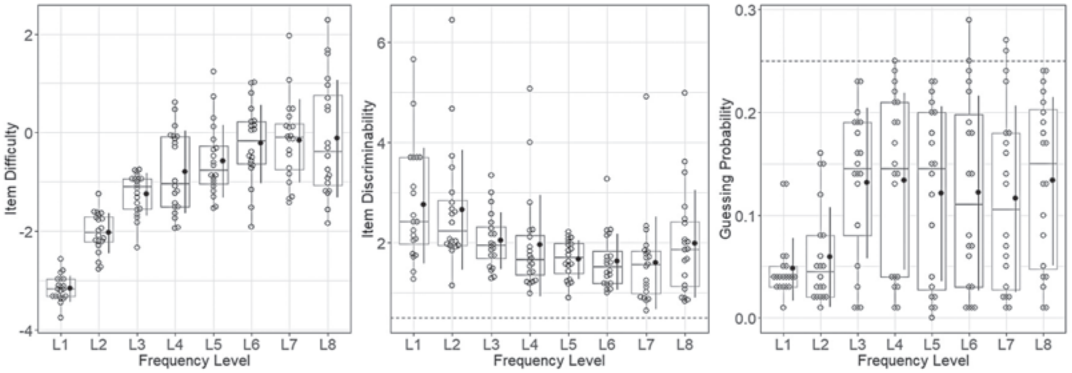


Figure 3. Box-scatter plots for the distributions of item difficulty, item discriminability, and pseudo-guessing probability with bars of means and ±1 standard deviations in the finalized version of the VST-NJ8.

Table 4

Descriptive Statistics of the Item Characteristics for the Finalized Version of the VST-NJ8

	Item difficulty			Item discriminability			Guessing probability		
Frequency	*M*	95% CI	*SD*	*M*	95% CI	*SD*	*M*	95% CI	*SD*
Level 1	−2.21	[−2.70, −1.71]	1.07	2.21	[1.44, 2.97]	1.64	0.07	[0.05, 0.10]	0.05
Level 2	−1.51	[−1.71, −1.31]	0.43	2.87	[2.33, 3.42]	1.16	0.08	[0.05, 0.11]	0.07
Level 3	−0.70	[−0.85, −0.56]	0.31	2.01	[1.73, 2.28]	0.59	0.12	[0.08, 0.16]	0.08
Level 4	−0.08	[−0.29, 0.14]	0.45	2.72	[1.69, 3.75]	2.21	0.13	[0.09, 0.17]	0.08
Level 5	0.75	[0.50, 1.00]	0.53	2.18	[1.56, 2.81]	1.34	0.10	[0.06, 0.13]	0.08
Level 6	1.15	[0.84, 1.46]	0.66	1.94	[1.47, 2.40]	0.99	0.08	[0.05, 0.12]	0.08
Level 7	1.50	[1.26, 1.75]	0.53	2.22	[1.42, 3.01]	1.70	0.12	[0.08, 0.17]	0.09
Level 8	2.09	[1.73, 2.45]	0.76	2.21	[1.72, 2.70]	1.06	0.15	[0.10, 0.20]	0.11

Note. Each frequency level contains 20 items.

Regarding the other part of the structural aspect of the VST-NJ8, whether the test score reflected a single construct was examined (RQ2). Because the response patterns for the finalized version of the VST-NJ8 were not directly observed, the validity was inferred from the data on the item bank. Table 5 shows the results of principle component analyses per booklet. The average percentages of variances explained by the first and second principle components were 34.90% and 5.92%, respectively. This result satisfied the 3-to-1 ratio of the first to second eigenvalues for the unidimensionality of the VST-NJ8 (see van der Linden, 2016, for review). Moreover, all the booklets satisfied this criterion because the 3-to-1 ratios of the first to second eigenvalues ranged from 3.07 to 11.11. Therefore, the entire VST-NJ8 and individual booklets were likely to measure a single construct of the test-takers' latent trait.

Table 5

The Percentages of Variance Explained by the First and Second Principle Components

Booklet	% of variance explained by Component 1	% of variance explained by Component 2	Ratios of Component 1 to Component 2
Book 1	44.93	4.73	9.51
Book 2	38.75	5.72	6.77
Book 3	51.54	6.04	8.53
Book 4	45.46	4.09	11.11
Book 5	49.02	6.67	7.35
Book 6	31.71	8.20	3.87
Book 7	24.46	6.66	3.67
Book 8	23.78	5.72	4.16
Book 9	18.65	6.07	3.07
Book 10	20.72	5.27	3.93

Note. The principle component analyses were based on a tetrachoric-correlation matrix.

Generalizability aspects of the construct validity. In Study 1, the measurement invariance of the item bank was validated in terms of the insignificant differences between the observed and expected response patterns (i.e., the goodness of item fit). Because the items selected from the item bank all fit to the 3PL model, this result provides evidence for the measurement invariance of the VST-NJ8 (RQ3). In other words, the item characteristics, such as difficulty, discriminability, and pseudo-guessing probability, can be generalized across different groups of test-takers.

In the IRT models, the test information functions are referred to as test reliability indices, as opposed to any reliability coefficients in the classical testing theory (Kato et al., 2014). Figure 4 visualizes the test information functions per frequency level. In response to RQ4, the items selected for Levels 1–8 were likely to cover the test-takers' latent trait ranging from –2 to 2 logits, as intended in the automatic test assembly. Considering that the range of being able to accurately measure the latent trait differed according to the frequency levels, these test information functions are informative about what frequency levels should be presented to test-takers. Specifically, the higher-frequency words should be used to measure lower-proficiency L2 learners, and vice versa. The frequency levels misfit to the learners' lexical proficiency will cause time waste and measurement errors. If the number of items for the estimation of vocabulary size is reliable enough, some items can be omitted to reduce test redundancy (e.g., Beglar, 2010; Enayat & Babaii, 2018). This is also related to the development of computer adaptive VSTs (e.g., Laufer & Goldstein, 2004; Nation, 2013; Tseng, 2013).

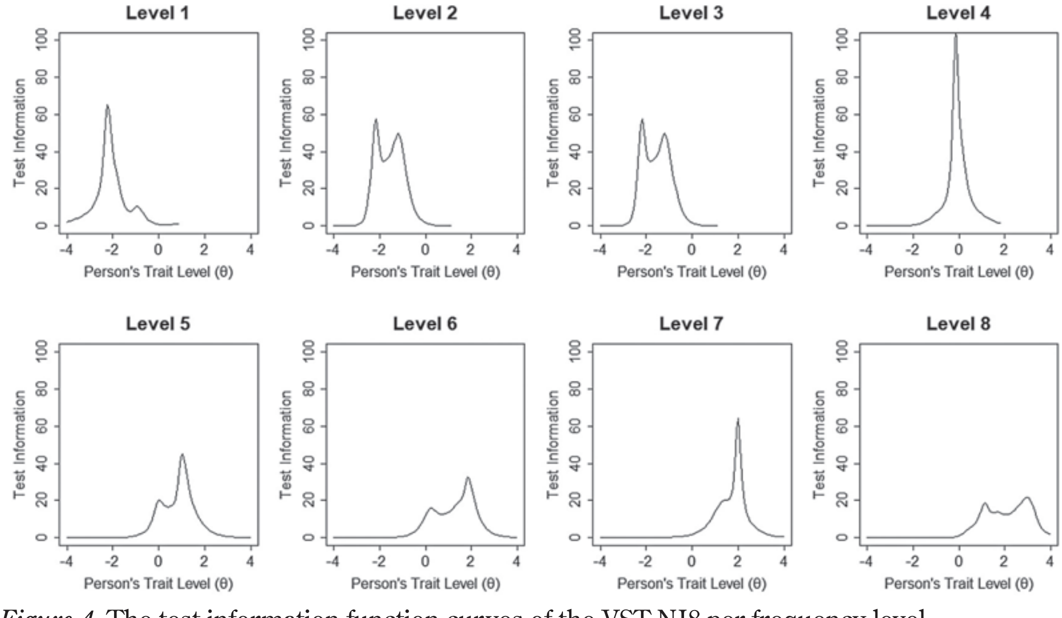


Figure 4. The test information function curves of the VST-NJ8 per frequency level.

External aspect of construct validity. Table 6 shows the correlation between the proficiency test score and the vocabulary size estimated by the finalized version of the VST-NJ8. For calculation of the vocabulary size, formula (1) and three IRT parameters (see Table 4) were used: e.g., when a participant's latent trait (θ) was 0.00, the mean probability of correct responses for Level 4 was calculated as $0.13 + (1 - 0.13)*\mathit{exp}\{2.72*(0.00 + 0.08)\} / [1 + \mathit{exp}\{2.72*(0.00 + 0.08)\}]$. In this case, the average expected probability was approximately 64%, indicating that the participant knew 640 words in Level 4. The result showed a significant correlation between the proficiency test score and vocabulary size (r = .62, 95% CI [.60, .65], p < .001). This strong correlation was consistent with the meta-analysis result of Zhang and Zhang (2020). Together with the content validity of the VST-NJ8, whose contents and formats are similar to existing VSTs (e.g., Koizumi & Mochizuki, 2011; Mizumoto, 2006), the finding of this study indicates that the VST-NJ8 scores can be interpreted as reflecting L2 learners' vocabulary size (RQ5).

Table 6

Descriptive Statistics and the Correlation of the Proficiency Test Score and Vocabulary Size

Variables	*M*	95% CI	*SD*	*r*	95% CI
TOEIC® score	503.46	[496.33, 510.58]	131.03	.62	[.60, .65]
Vocabulary size	4575.81	[4510.36, 4641.26]	1203.81		

Note. n = 1,302.

Conclusions

In the present study, we developed and validated the VST-NJ8 for the assessment of Japanese EFL learners' vocabulary size. As opposed to existing VSTs, a large-scale item bank for test representativeness was created and applied to the 3PL model for vocabulary size estimation. The results of the automatic test assembly and test validation showed the robustness and credibility of the VST-NJ8 in terms of structural, generalizability, and external aspects of construct validity. The strength of the VST-NJ8 can be described as (a) the items reflect the current vocabulary use in English education in Japan, (b) the IRT parameters can be used to ensure the accuracy of vocabulary size estimation, and (c) the test scores can be used to predict the levels of reading and listening skills. These advantages are further related to the future use of the VST-NJ8 with respect to developments of parallel tests, computer adaptive tests, and diagnosis tests. In the Supplementary Materials, the administration, scoring, and interpretation methods of the VST-NJ8 are presented.

This study, however, is not without limitations. First, the test-takers sampled for the application of the IRT were limited to university students, although their proficiency levels varied significantly. It is necessary to obtain response data from younger learners and further calibrate the item parameters. Second, the present study did not examine the relationships between the VST-NJ8 and other standardized VSTs for extrapolation. Higher correlations between the VST-NJ8 and other standardized VSTs than the correlation between the VST-NJ8 and TOEIC® scores will indicate the convergent and discriminant validity of the VST-NJ8. Although these limitations are beyond the scope of this study, it is important to build big data on the use of the VST-NJ8.

Acknowledgements

This research was supported by the Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research (B) 20H01287. We would like to thank the handling editors and anonymous reviewers of *JACET Journal*. We are also grateful to Kazumichi Enokida, Shusaku Kida, Mitsuhiro Morita, Atsushi Nakagawa, Tatsuya Sakaue, and Yuka Takahashi (Hiroshima University), Shuichi Takaki (Fukushima University), Yuichiro Kobayashi (Nihon University), Chikako Sugita (Chuo University), and Yuka Muraoka (Seigakuin University) for their help in

data collection.

References

Aizawa, K. (1998). Developing a vocabulary size test for Japanese EFL learners. *ARELE: Annual Review of English Language Education in Japan*, *9*, 75–85. doi:10.20581/arele.9.0_75

Aizawa, K., Ishikawa, S., Murata, M., Iso, T., Uemura, T., Kokawa, T., Shimizu, S., Sugimori, N., Haisa, A., & Mochizuki, M. (2005). *The JACET list of 8,000 basic words*. Tokyo, Japan: Kirihara Shoten.

ALC Inc. (2001). *Standard vocabulary list 12,000*. Retrieved from https://www.alc.co.jp/vocgram/article/svl/

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London, England: Continuum.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing*, *27*, 101–118. doi:10.1177/0265532209340194

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. doi:10.18637/jss.v048.i06

Committee of Revising the JACET Basic Words. (Ed.). (2016). *The new JACET list of 8,000 basic words*. Tokyo, Japan: Kirihara Shoten.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing*, *30*, 253–272. doi:10.1177/0265532212459028

Enayat, M. J., & Babaii, E. (2018). Reliable predictors of reduced redundancy test performance: The interaction between lexical bonds and test takers' depth and breadth of vocabulary knowledge. *Language Testing*, *35*, 121–144. doi:10.1177/0265532216683223

Gelius-Dietrich, G. (2020). R interface to C API of GLPK [R package]. Retrieved from https://cran.r-project.org/web/packages/glpkAPI/glpkAPI.pdf

Gibson, A., & Stewart, J. (2014). Estimating learners' vocabulary size under item response theory. *Vocabulary Learning and Instruction*, *3*, 78–84. doi:10.7820/vli.v03.2.gibson.stewart

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *International Journal of Applied Linguistics*, *166*, 278–306. doi:10.1075/itl.166.2.04gyl

Hamada, A. (2011). Lexical inferencing and the acquisition of vocabulary depth: Focusing on strategy use and contextual information. *ARELE: Annual Review of English Language Education in Japan*, *22*, 313–328. doi:10.20581/arele.22.0_313

Hamada, A. (2014). Using latent semantic analysis to promote the effectiveness of contextualized vocabulary learning. *JACET Journal*, *58*, 1–20. doi:10.32234/jacetjournal.58.0_1

Hamada, A. (2015). Linguistic variables determining the difficulty of Eiken reading passages. *JLTA Journal*, *18*, 57–77. doi:10.20622/jltajournal.18.0_57

Hamada, A. (2017). Estimating input quantity for L2 vocabulary acquisition: A preliminary study of statistical language analysis. *JACET Journal*, *61*, 109–129. doi:10.32234/jacetjournal.61.0_109

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *23*, 403–430. Retrieved from https://nflrc.hawaii.edu/rfl/PastIssues/rfl131hsuehchao.pdf

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kato, K., Yamada, T., & Kawabata, K. (2014). *Item response theory with R*. Tokyo, Japan: Ohmsha.

Koizumi, R., & Mochizuki, M. (2011). Development and validation of the PC version of the Mochizuki Vocabulary Size Test. *JACET Journal*, *53*, 35–55. Retrieved from https://ci.nii.ac.jp/naid/110008897624

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 126–132). London, England: Macmillan Academic and Professional.

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

Laufer, B., & McLean, S. (2016). Loanwords and vocabulary size test scores: A case of different estimates for different L1 learners. *Language Assessment Quarterly*, *13*, 202–217. doi:10.1080/15434303.2016.1210611

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*, 15–30. Retrieved from https://nflrc.hawaii.edu/rfl/April2010/articles/laufer.pdf

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*, 142–154. doi:10.1177/026553228700400202

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749. doi:10.1037/0003-066X.50.9.741

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels: A common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, and textbooks across Europe. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 211–232). Amsterdam, the Netherlands: European Second Language Association.

Mizumoto, A. (2006). What do vocabulary size tests measure? Problems involved in developing vocabulary size tests. *The Institute of Statistical Mathematics Cooperative Research Report*, *190*, 71–80. Retrieved from https://www.mizumot.com/files/VocSizeMeasure.pdf

Mochida, A., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing*, *23*, 73–98. doi:10.1191/0265532206lt321oa

Mochizuki, M. (1998). A vocabulary size test for Japanese learners of English. *The IRLT Bulletin*, *12*, 27–53. Retrieved from https://ci.nii.ac.jp/naid/40004654197

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System*, *28*, 291–304. doi:10.1016/S0346-251X(00)00013-0

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

Nation, P. (2013). *Learning vocabulary in another language* (2nd ed.). New York, NY: Cambridge University Press.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13. Retrieved from https://jalt-publications.org/files/pdf/the_language_teacher/07_2007tlt.pdf

Partchev, I. (2017). irtoys: A collection of functions related to Item Response Theory (IRT) [R package]. Retrieved from https://cran.r-project.org/web/packages/irtoys/irtoys.pdf

Read. J. (2000). *Assessing vocabulary*. New York, NY: Cambridge University Press.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Hampshire, England: Palgrave Macmillan.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*, 26–43. doi:10.1111/j.1540-4781.2011.01146.x

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *53*, 109–120. doi:10.1017/S0261444819000326

Schmitt, N., & Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*, 484–503. doi:10.1017/S0261444812000018

Sonoda, K. (1996). *Daigakusei eigo goihyo no tameno kisoteki kenkyu* [Fundamental research on English word list for university students]. Hokkaido, Japan: Hokkaido University Press.

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, *31*, 577–607. doi:10.1017/S0272263109990039

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*, 271–282. doi:10.1080/15434303.2014.922977

Tseng, W.-T. (2013). Validating a pictorial vocabulary size test via the 3PL-IRT model. *Vocabulary Learning and Instruction*, *2*, 64–73. doi:10.7820/vli.v02.1.tseng

van der Linden, W. J. (2016). *Handbook of item response theory (vol. 1): Models*. Boca Raton, FL: Chapman and Hall/CRC.

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

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

Xue, G., & Nation, P. (1984). A university word list. *Language Learning and Communication*, *3*, 215–229.

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

Supplementary Materials

The Finalized Version of the VST-NJ8

No.	Item stem	POS	Correct answer	Distractor 1	Distractor 2	Distractor 3	Dffclt	Dscrmn	Gussng
				Frequency Level 1					
1	売り場	N	department	boy	meeting	town	–2.38	1.98	0.13
2	願う	V	wish	return	take	feel	–1.85	4.15	0.06
3	許可する	V	let	try	buy	spend	–1.18	1.65	0.02
4	丸い	Adj.	round	ready	simple	easy	–2.40	2.18	0.05
5	全体の	Adj.	whole	important	other	large	–0.95	3.83	0.25
6	普通は	Adv.	usually	never	always	finally	–3.07	0.86	0.06
7	地域	N	area	style	art	record	–2.47	3.33	0.02
8	銀行	N	bank	subject	age	century	–3.54	0.87	0.06
9	実行	N	performance	project	mind	president	1.00	1.40	0.15
10	1年	N	year	state	power	father	–3.99	0.92	0.07
11	特別の	Adj.	special	western	key	old	–2.25	6.79	0.02
12	それから	Adv.	then	alone	else	far	–2.18	0.86	0.05
13	押すこと	N	pressure	field	support	question	–2.04	1.35	0.10
14	始める	V	begin	stand	do	want	–2.97	1.53	0.05
15	過去の	Adj.	past	major	each	red	–2.12	5.34	0.02
16	部屋	N	room	daughter	increase	front	–2.87	1.30	0.06
17	進路	N	course	group	music	order	–2.28	1.62	0.09
18	玄関	N	hall	size	place	member	–1.70	1.38	0.08
19	小休止	N	break	heart	hour	bed	–1.64	1.40	0.07
20	持つ	V	hold	call	remember	bring	–3.21	1.47	0.08
				Frequency Level 2					
1	…だと思う	V	suppose	lay	succeed	gain	–1.19	2.13	0.16
2	たぶん	Adv.	perhaps	quite	similarly	effectively	–1.13	1.62	0.04
3	請求書	N	bill	society	scene	measure	–1.10	2.47	0.00
4	範囲	N	range	plenty	instrument	surface	–1.76	2.03	0.03
5	一般的な	Adj.	general	recent	empty	illegal	–1.73	2.03	0.05
6	地域	N	region	silver	content	customer	–1.15	2.78	0.22
7	構造	N	structure	neighborhood	lack	application	–1.25	3.50	0.15
8	制御	N	control	benefit	stomach	supply	–1.94	2.80	0.02
9	型	N	type	discussion	string	gas	–2.19	6.45	0.02
10	患者	N	patient	technique	user	aid	–1.67	1.97	0.05
11	苦痛	N	pain	website	appeal	solution	–1.21	3.02	0.25
12	よく考える	V	consider	overcome	preserve	approve	–0.92	2.40	0.09
13	利用可能な	Adj.	available	central	everyday	huge	–1.62	3.73	0.08
14	最近	Adv.	recently	somewhere	newly	heavily	–1.05	2.14	0.01
15	問題	N	issue	pattern	device	award	–1.21	4.18	0.11
16	期間	N	term	kid	platform	contrast	–1.06	2.44	0.15
17	経済	N	economy	impression	duty	cross	–2.09	2.58	0.01
18	政治の	Adj.	political	complete	senior	sharp	–1.77	2.67	0.02
19	過程	N	process	sugar	trick	channel	–1.92	1.84	0.06
20	盗む	V	steal	create	push	destroy	–2.29	4.67	0.06

The Finalized Version of the VST-NJ8 (Continued)

No.	Item stem	POS	Correct answer	Distractor 1	Distractor 2	Distractor 3	Dffclt	Dscrmn	Gussng
Frequency Level 3									
1	種類	N	sort	complaint	client	youth	–0.95	1.31	0.01
2	評議会	N	council	soil	producer	strategy	–0.76	1.75	0.14
3	分析	N	analysis	stream	presence	monitor	–0.57	1.89	0.16
4	反対	N	opposition	failure	arrival	awareness	–0.94	1.96	0.19
5	追う	V	pursue	calculate	negotiate	deserve	–0.80	2.18	0.23
6	重要な	Adj.	significant	immediate	multiple	flexible	–0.48	1.69	0.00
7	明らかに	Adv.	obviously	fairly	somewhat	additionally	–1.13	1.93	0.19
8	革	N	leather	participant	reputation	crack	–0.73	1.29	0.01
9	比較	N	comparison	quarter	expertise	introduction	–1.07	2.17	0.23
10	出会い	N	encounter	stock	tension	register	–0.66	2.40	0.15
11	おおよそ	Adv.	approximately	precisely	possibly	slightly	–0.14	2.03	0.21
12	もの（物質）	N	stuff	statistic	contract	characteristic	–0.13	1.78	0.06
13	求める	V	seek	launch	deny	abandon	–0.23	2.29	0.16
14	学問の	Adj.	academic	unlikely	extraordinary	mass	–0.31	3.58	0.22
15	視力	N	vision	preparation	boss	administration	–0.83	1.86	0.03
16	欠点	N	fault	reduction	existence	description	–0.95	1.49	0.05
17	身元	N	identity	editor	wealth	ancestor	–1.00	1.75	0.13
18	提出する	V	submit	convince	enable	satisfy	–0.55	1.99	0.09
19	批判的な	Adj.	critical	raw	brief	unnecessary	–0.86	3.34	0.14
20	国内の	Adj.	domestic	permanent	constant	enthusiastic	–0.95	1.41	0.01
Frequency Level 4									
1	信念	N	faith	disposal	density	staple	–0.12	1.19	0.03
2	文字通りに	Adv.	literally	publicly	presumably	formally	–0.18	1.57	0.04
3	権利を与える	V	entitle	assess	cultivate	seize	0.13	4.00	0.21
4	先の	Adj.	prior	snowy	causal	stressful	–0.98	2.26	0.15
5	大量の	Adj.	massive	harsh	primitive	gradual	–0.22	2.25	0.19
6	角（度）	N	angle	herb	stove	manual	–0.93	1.41	0.01
7	歳入	N	revenue	ambition	coral	ballet	0.01	4.80	0.11
8	金融	N	finance	tablet	farmland	exploration	–0.23	1.91	0.21
9	知覚する	V	perceive	qualify	frustrate	renew	–0.27	2.90	0.28
10	法人の	Adj.	corporate	occasional	noticeable	unavoidable	0.82	1.36	0.01
11	統合	N	integration	defeat	artifact	wallet	–0.17	10.62	0.13
12	計画	N	scheme	ant	pamphlet	merit	0.09	1.75	0.22
13	収容設備	N	accommodation	miracle	grace	conservation	–0.08	2.11	0.17
14	制限する	V	restrict	depict	dedicate	abolish	–0.29	5.07	0.25
15	所有（物）	N	possession	chemist	personnel	chamber	–0.06	1.21	0.13
16	不安	N	anxiety	habitat	dialect	convention	0.50	1.64	0.04
17	横顔	N	profile	refund	bloom	penalty	0.61	1.24	0.14
18	目に見える	Adj.	visible	vivid	pleasant	stylish	–0.51	1.83	0.01
19	不動産	N	estate	summit	mud	neglect	–0.10	3.76	0.12
20	政治制度	N	regime	voyage	investigator	punch	0.47	1.53	0.14

The Finalized Version of the VST-NJ8 (Continued)

No.	Item stem	POS	Correct answer	Distractor 1	Distractor 2	Distractor 3	Dffclt	Dscrmn	Gussng
				Frequency Level 5					
1	安定	N	stability	astronomer	relaxation	dilemma	0.91	1.57	0.03
2	選ぶ	V	opt	orient	deteriorate	execute	0.14	0.85	0.01
3	（～する）資格のある	Adj.	eligible	partial	optimistic	operational	1.09	2.52	0.17
4	潜在的に	Adv.	potentially	scientifically	remarkably	adequately	0.96	6.13	0.15
5	格付け	N	rating	homeland	discomfort	logo	1.18	1.26	0.01
6	敵意のある	Adj.	hostile	advisory	mandatory	magnetic	0.12	2.08	0.00
7	残酷な	Adj.	cruel	serial	exotic	unrelated	1.34	1.96	0.12
8	身をさらすこと	N	exposure	supplement	hydrogen	detention	0.60	2.21	0.18
9	降下	N	descent	frontier	peer	diesel	−0.05	4.27	0.17
10	一覧表への記入	N	listing	tolerance	protocol	ramp	0.67	1.00	0.01
11	協力	N	collaboration	ambassador	creativity	lawsuit	0.09	2.02	0.09
12	業績	N	accomplishment	constraint	resignation	inventory	−0.32	1.98	0.20
13	使用料	N	toll	sickness	pharmacy	patch	1.12	1.74	0.15
14	枠組み	N	framework	penny	revision	rejection	1.50	1.43	0.01
15	墓	N	tomb	inclusion	separation	surplus	1.53	1.63	0.04
16	祈り	N	prayer	globe	nitrogen	ram	0.87	1.76	0.15
17	一致	N	correspondence	worship	transplant	combat	0.47	2.10	0.23
18	直立した	V	erect	tremble	deprive	renovate	0.73	0.90	0.02
19	完全な	Adj.	integral	offshore	legislative	triple	1.15	4.66	0.19
20	抽象的な	Adj.	abstract	inherent	naïve	archaeological	0.87	1.58	0.02
				Frequency Level 6					
1	とても天気の良い	Adj.	gorgeous	lightweight	inferior	promotional	1.75	1.03	0.11
2	それゆえ	Adv.	hence	urgently	ecologically	nicely	0.14	3.27	0.14
3	概要	N	overview	spacecraft	alteration	mound	0.39	1.17	0.01
4	祖先	N	ancestry	coordination	liability	mathematician	1.85	5.41	0.22
5	子孫	N	descendant	capitalism	acquaintance	ankle	1.48	2.19	0.22
6	実験	N	experimentation	sunglass	courtesy	pathogen	0.10	1.95	0.01
7	原子	N	atom	fungus	rust	fist	1.92	2.14	0.01
8	傾向	N	inclination	projector	resin	infancy	1.15	1.19	0.01
9	自由にする	V	liberate	encompass	alleviate	circulate	1.57	2.24	0.14
10	誠実な	Adj.	sincere	parliamentary	linear	inward	0.23	2.26	0.24
11	市民の	Adj.	civilian	oval	simultaneous	damp	2.10	2.46	0.02
12	統計の	Adj.	statistical	radioactive	unpopular	jumbo	1.52	1.33	0.03
13	居住	N	residency	commander	aerial	lieutenant	0.63	1.60	0.01
14	不一致	N	discrepancy	incorporation	patron	breadth	1.26	1.77	0.02
15	加入する	V	subscribe	withhold	differentiate	murmur	1.05	1.50	0.18
16	抑制する	V	restrain	initiate	comprehend	intrigue	1.51	1.72	0.07
17	従う	V	conform	recite	evacuate	compel	1.70	1.50	0.05
18	嘆き	N	grief	rifle	miner	naturalist	1.47	1.19	0.07
19	確認	N	verification	stagnation	deduction	herd	0.06	1.36	0.03
20	分割	N	partition	motto	pickle	novelty	1.16	1.44	0.08

The Finalized Version of the VST-NJ8 (Continued)

No.	Item stem	POS	Correct answer	Distractor 1	Distractor 2	Distractor 3	Dffclt	Dscrmn	Gussng
Frequency Level 7									
1	誓い	N	oath	repertoire	payroll	ensemble	1.25	2.61	0.15
2	保証書	N	warrant	unison	aerospace	gospel	1.08	1.18	0.14
3	貨物	N	freight	ignorance	mucosa	mainland	2.49	1.58	0.13
4	主張	N	assertion	inquest	biopsy	disclosure	1.13	1.53	0.18
5	固定観念	N	stereotype	precedent	module	dismay	0.72	1.82	0.17
6	補足の	Adj.	complementary	imminent	tricky	eternal	1.33	1.95	0.23
7	運転する人	N	motorist	secrecy	curator	analogy	0.36	1.01	0.02
8	ちらりと見えること	N	glimpse	betrayal	disco	banner	1.96	8.24	0.25
9	侵入者	N	raider	monopoly	skipper	cinnamon	1.11	0.91	0.02
10	外交	N	diplomacy	injunction	farmhouse	embargo	1.12	1.70	0.18
11	バランスをとる	V	poise	chuck	knit	shiver	2.07	1.64	0.01
12	装飾用の	Adj.	decorative	aging	hopeless	marital	1.42	1.85	0.03
13	心臓の	Adj.	cardiac	creamy	apt	dire	2.13	1.24	0.06
14	量の	Adj.	quantitative	relentless	median	comparative	1.31	1.71	0.01
15	深く	Adv.	profoundly	loosely	indirectly	etcetera	2.02	3.36	0.26
16	選挙区	N	constituency	unrest	majesty	baron	1.76	1.43	0.07
17	格言	N	maxim	brigade	accountant	probation	1.97	4.92	0.07
18	軽蔑	N	contempt	communism	stimulation	referendum	1.89	2.45	0.20
19	頑固な	Adj.	stubborn	pragmatic	bureaucratic	molecular	1.45	2.26	0.27
20	屋根裏部屋	N	attic	referral	embargo	goodwill	1.49	0.92	0.01
Frequency Level 8									
1	構文	N	syntax	brewery	hooker	hawk	2.31	1.07	0.01
2	起訴する	V	prosecute	grate	lurk	rinse	2.07	1.97	0.26
3	普遍的に	Adv.	universally	inadvertently	boldly	immensely	1.04	4.64	0.40
4	匿名であること	N	anonymity	broom	snooker	saloon	2.23	1.15	0.01
5	子宮	N	womb	envy	vibration	cooker	2.92	0.86	0.04
6	正気の	Adj.	sane	penal	pristine	pancreatic	2.96	3.40	0.17
7	自殺の	Adj.	suicidal	longitudinal	disruptive	endoscopic	0.87	1.87	0.05
8	聖歌	N	anthem	tempo	elegance	slaughter	2.03	0.90	0.03
9	説得	N	persuasion	persona	variability	heartbeat	1.75	2.39	0.22
10	長期間続く	Adj.	perennial	homogeneous	psychic	fluorescent	2.52	2.02	0.20
11	結果として起こる	Adj.	consequent	doctoral	stray	recurrent	1.48	2.06	0.23
12	没頭	N	preoccupation	monsieur	starvation	gypsy	1.06	3.64	0.26
13	恐怖	N	dread	relegation	deviation	midfielder	1.57	3.16	0.26
14	法令	N	decree	affinity	devastation	parasite	2.67	2.72	0.13
15	贅沢な	Adj.	luxurious	erratic	cunning	indicative	0.57	2.48	0.08
16	省略	N	omission	gem	resonance	retaliation	2.70	0.89	0.04
17	移り変わり	N	flux	coating	whereabouts	jeopardy	3.09	3.62	0.24
18	説教	N	sermon	apprehension	brilliance	pilgrimage	2.65	1.68	0.10
19	雑用	N	chore	wig	fracture	foreground	2.46	2.00	0.18
20	階層性の	Adj.	hierarchical	aristocratic	martial	proprietary	2.80	1.67	0.05

Note. POS = Parts of speech. Dffclt = item difficulty, Dscrmn = item discriminability, Gussng = pseudo-guessing probability.

Terms of Use and Guidelines for the VST-NJ8 Administration, Scoring, and Interpretation

Test name: Vocabulary Size Test Based on the New JACET List of 8,000 Basic Words
大学英語教育学会基本語リスト版語彙サイズテスト (VST-NJ8)

Measurement: The VST-NJ8 measures test-takers' vocabulary size. The vocabulary size in this test indicates how many words listed in the New JACET List of 8,000 Basic Words test-takers know. The 3PL IRT model was applied to the test, ensuring measurement invariance in terms of item difficulty, item discriminability, and pseudo-guessing probability. The item information functions do not differ according to different groups of learners, either.

Population: The main test-takers of the VST-NJ8 are Japanese EFL learners. Although the test has not yet been examined for the younger than university students, it can be administered to different groups of learners. As the test format is Japanese–English matching, test-takers must be proficient in Japanese (this is particularly critical when using lower-frequency words in the test).

Administration: The VST-NJ8 can be administered individually and to groups of test-takers via the paper version of the VST-NJ8 (http://j-varg.sakura.ne.jp/vst-nj8/index.html). The computer-based version of the VST-NJ8 will also be available for delivery and scoring. We set the time limit for response to be 200 seconds (3 minute and 20 seconds) per frequency level. However, enough time can be given to complete the test because it does not intend to measure vocabulary fluency. After explaining the purpose of this test, administrators must instruct students to look at a given Japanese word and then select the most appropriate English translation from four options.

Scoring: A conventional way to calculate vocabulary size is described as

$$\mathit{Vocabulary\ size} = \frac{\mathit{The\ number\ of\ correct\ responses}}{\mathit{The\ number\ of\ items}} \times \mathit{Frequency\ levels}$$

e.g., when Levels 2–6 (a total of 20 items × 5 levels) are administered and the number of correct responses is 50, the vocabulary size is calculated as 50 / 100 × 5 + 1,000 = 3,500, where the scores of the omitted levels that are assumed too easy for test-takers can be considered full. The IRT-based scoring is also available for R users (download R scripts from http://j-varg.sakura.ne.jp/vst-nj8/index.html).

Use licensing: The terms and conditions of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) are applied to the VST-NJ8. The use of the VST-NJ8 is free, without permission, as long as this work is cited in your publications and relevant reports.

Corresponding author: Hamada Akira (JACET Vocabulary Acquisition Research Group)
a.hamada.0218@gmail.com (https://hamada-lab.jp/)